DATA MINING PER IL MARKETING

Marco Riani

IL MODELLO DI REGRESSIONE LOGISTICA
Introduzione e inferenza

Materiale didattico: Cerioli-Laurini, Il Modello di Regressione Logistica, UNI.NOVA, Parma, 2013

Previsione di una variabile dicotomica

- **La variabile da prevedere** (Y = var. dipendente del modello) è dicotomica: **presenza/assenza di una caratteristica**:
  - Compra / non compra un prodotto o categoria di prodotti
  - Appartiene / non appartiene a un certo profilo o segmento di clientela
  - Aderisce / non aderisce a una campagna promozionale
  - E' solvente / è insolvente
  - …
- **Le variabili esplicative** X_1, X_2, … X_{k-1} forniscono informazioni su fattori ritenuti rilevanti nella previsione di Y. Nel Trade marketing, spesso tali variabili sono tratte dal **database aziendale**:
  - Spesa per prodotti/categorie correlate
  - Comportamento di acquisto precedente
  - Informazioni sul comportamento complessivo di acquisto (spesa tot., scontrino medio, numero di visite in pdv, tipologia di pdv frequentata …)
- **Le variabili esplicative** X_1, X_2, … X_{k-1} possono essere sia **quantitative che qualitative**
- Se disponibili, si possono usare anche **informazioni esterne**:
  - Reddito
  - Età, sesso e caratteristiche socio-demografiche

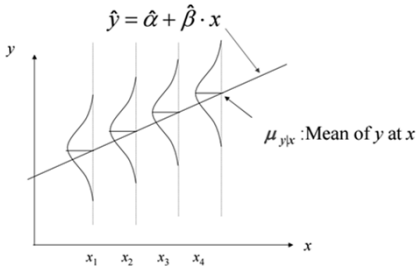
- **Nella regressione lineare**: **Y = variabile dipendente quantitativa** (con distribuzione normale)

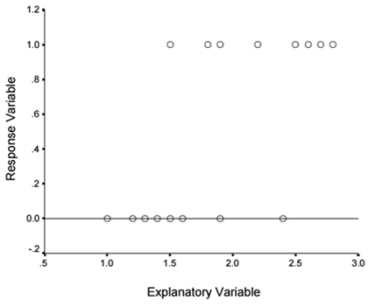
$$E(y|x_i) = \mu_{y|x_i} = \beta_0 + \beta_1 x_i$$

- La combinazione lineare delle variabili esplicative descrive quindi il **valore atteso** di y_i
- **Nel problema in esame**: **Y = variabile dipendente dicotomica** (che rappresentiamo con una variabile aleatoria di **Bernoulli**)

y_i	Probabilità
0	**$1 - \pi_i$**
1	**π_i**
Tot.	**1**

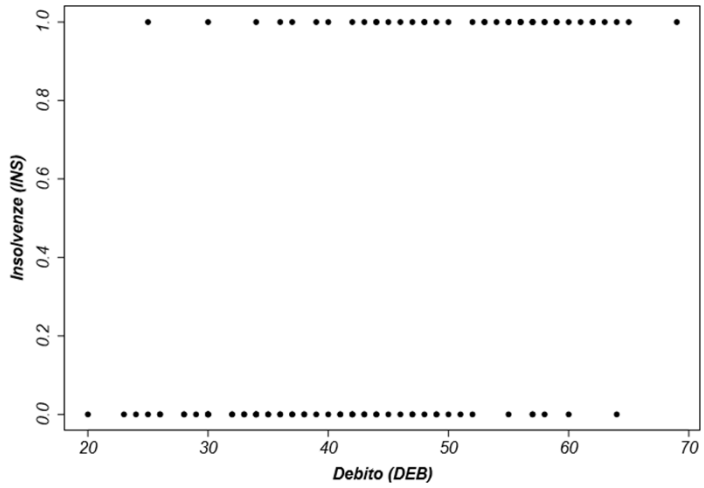
$E(y_i|x_i)=\pi_i$: probabilità di "successo" per l'unità i



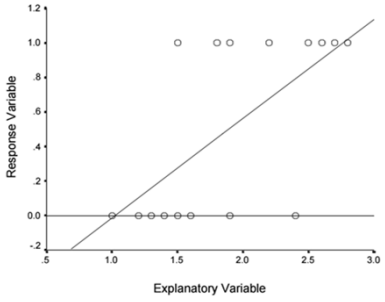


- Nella regressione logistica vogliamo avere un modello che spieghi $\mu_{y|x} = \pi(x)$ ossia la probabilità di osservare l'evento dato x

Esempio 1: Insolvenze in funzione del debito



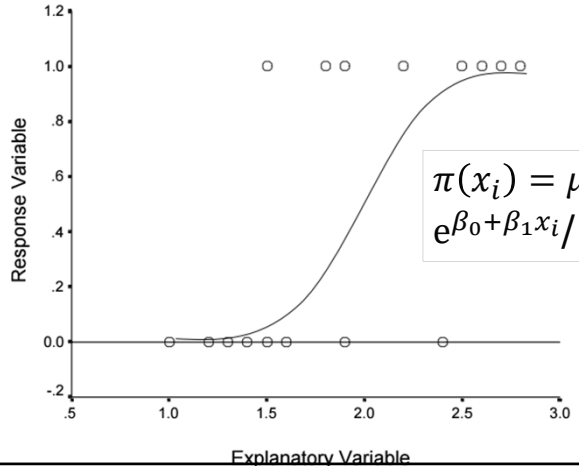
E' possibile modellare la probabilità del successo tramite una retta?



- Problemi
- errori non normali
- varianza non costante
- vincolo $0 \leq \pi(x) = \mu_{y|x} \leq 1$

Come si modellizza la probabilità di successo

- Attraverso una funzione logistica



$$\pi(x_i) = \mu_{y|x_i} = e^{\beta_0+\beta_1 x_i}/\left(1 + e^{\beta_0+\beta_1 x_i}\right)$$

20

La v.c. logistica Λ

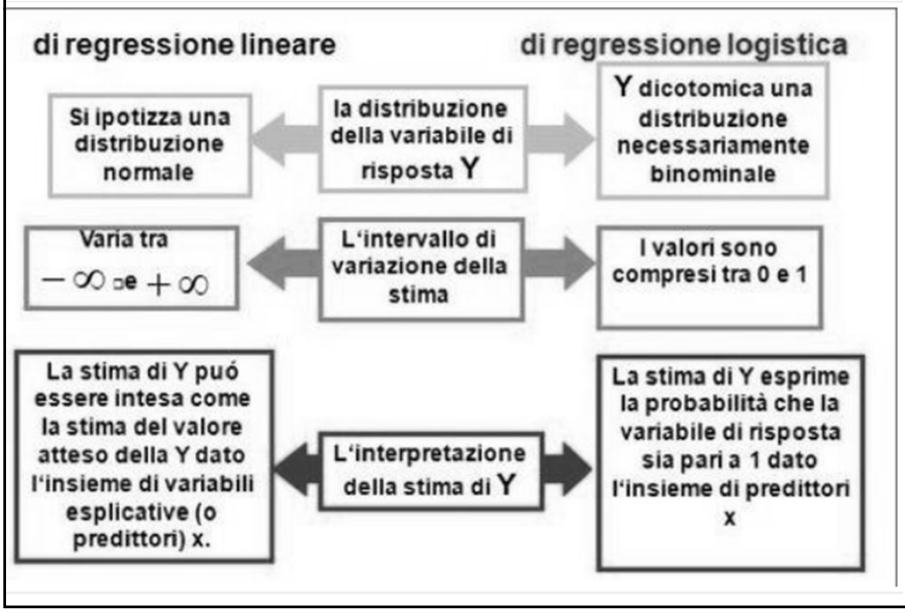
- Funzione di ripartizione (λ ϵ R)

$$\Pr(\Lambda \leq \lambda) = 1 - \frac{1}{1+\exp(\lambda)}.$$

$$\Pr(\Lambda \leq \lambda) = \frac{\exp(\lambda)}{1+\exp(\lambda)} = \frac{1}{1+\exp(-\lambda)},$$

- Es Excel file logistica.xlsx

Differenza regr lineare e regr logistica



Consideriamo per semplicitià il modello con una sola variabile indipendente ($\lambda = \beta_0 + \beta_1 x$)

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}},$$

- $\beta_0 + \beta_1$ sono liberi di variare in $(-\infty, +\infty)$

Probabilità di successo e logit

$$\pi(x_i) = P(Y_i = 1 \mid x_i) = \exp(\beta_0 + \beta_1 x_i)/(1+\exp(\beta_0 + \beta_1 x_i))$$

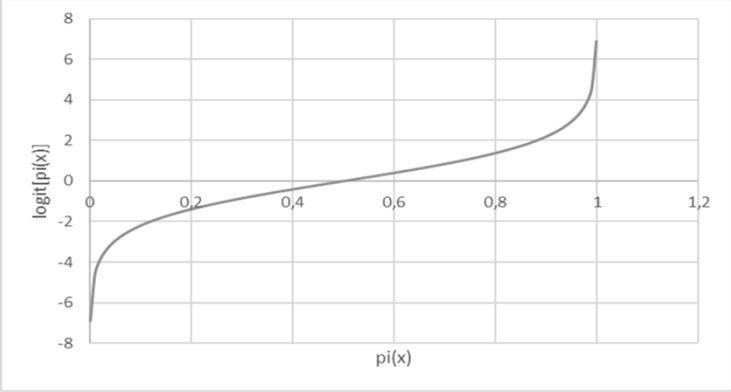
$$\frac{\pi(x_i)}{1-\pi(x_i)} = \exp(\beta_0 + \beta_1 x_i)$$

- **Modello di regressione per logit[$\pi(x_i)$]**

$$\text{logit}[\pi(x_i)] = \log \frac{\pi(x_i)}{1-\pi(x_i)} = \beta_0 + \beta_1 x_i$$

Relazione tra $\pi(x)$ e logit[$\pi(x)$]

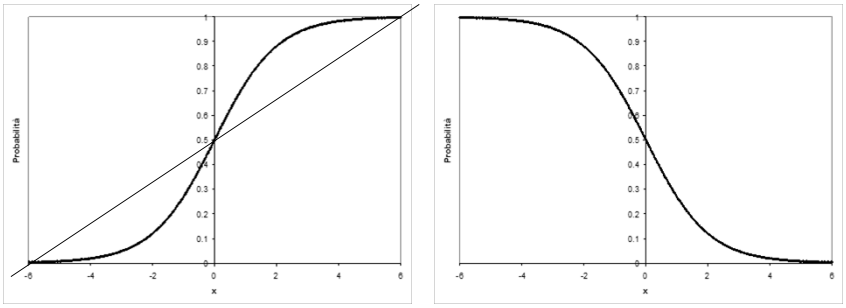
La funzione logit trasforma il range della variabile dipendente (π_i) dall'intervallo [0; 1] all'intero insieme dei numeri reali (-∞;+∞)



Caratteristiche di $\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$,

- La pendenza della curva è data da $\beta_1\pi(x)[1-\pi(x)]$ ➔ andamento crescente quando $\beta_1 > 0$
- L'effetto della variabile esplicativa X sulla probabilità $\pi(x)$ non è costante
- L'effetto è massimo quando $\pi(x) = 0.5$ (punto di ascissa x=-β_0/β_1) $\Rightarrow$ **implicazioni di marketing**
- Tale effetto è simmetrico rispetto a $\pi(x)$=0.5

Funzione logistica (β_0=0; β_1 = $\pm$1)



- **Funzione non lineare** tra la probabilità π(x) e x $\Rightarrow$ una retta crescerebbe invece indefinitamente (v. grafico)
- **Pendenza della curva**: β_1 **π(x) [1-π(x)]** $\Rightarrow$ β_1>0 andamento crescente; β_1<0 andamento decrescente
- **L'effetto sulla probabilità** di una variazione unitaria di x non è costante: **è max quando π(x)=0.5** (punto di ascissa x=-β_0/β_1) $\Rightarrow$ **implicazioni di marketing**
- Tale effetto è simmetrico rispetto a π(x)=0.5

Logit

- Se **Y dicotomica** (Y=1 oppure Y=0):

$$\text{logit}[P(Y=1)] = \log\frac{P(Y=1)}{1-P(Y=1)} = \log\frac{P(Y=1)}{P(Y=0)} = \log\frac{\pi}{1-\pi}$$

- **Logit = logaritmo della "quota relativa"** (odds):

Odds = $\pi/(1-\pi)$ = P(Y=1)/P(Y=0)

Richiamo tabelle di contingenza (v. Zani-Cerioli, cap. 4)

- Y, X dicotomiche; Y = dipendente {0, 1}; X = esplicativa {A, B};

X\Y	0	1	Tot
A	π_{A0}	π_{A1}	π_{A+}
B	π_{B0}	π_{B1}	π_{B+}
Tot.	π_{+0}	π_{+1}	**1**

$P(Y=1|X=A) = \pi_{A1}/\pi_{A+}$

$P(Y=1|X=B) = \pi_{B1}/\pi_{B+}$

$P(Y=1) = \pi_{+1}$

- **Odds per X=A** = P(Y=1|X=A)/P(Y=0|X=A) = π_{A1}/π_{A0}
- **Odds per X=B** = P(Y=1|X=B)/P(Y=0|X=B) = π_{B1}/π_{B0}
- **Odds Ratio** = OR = Odds(A)/Odds(B) = $(\pi_{A1}\pi_{B0})/(\pi_{A0}\pi_{B1})$
- La definizione di Odds vale per qualunque X (anche non dicotomica):

  $Odds(x_i) = P(Y=1|X=x_i)/P(Y=0|X=x_i) = \pi(x_i)/[1-\pi(x_i)]$

  Quindi: **$logit[\pi(x_i)] = log[Odds(x_i)]$**

REGRESSIONE LOGISTICA MULTIPLA

Punteggio (score) per l'unità i

- **k-1 variabili esplicative**. Per ogni unità i (i=1, …, n):

$$x_i' = (1, x_{i1}, x_{i2}, ..., x_{i,k-1}) \Rightarrow \text{riga } i \text{ della matrice X}$$

$$\text{parte sistematica del modello: } \beta_0 + \beta_1 x_{i1} + ... + \beta_{k-1} x_{i,k-1}$$

- La combinazione lineare fornisce un **punteggio** (score) per l'unità i: **analogia con la regressione lineare**
- Nelle applicazioni di marketing, di solito l' obiettivo è però differente rispetto alla regressione: il punteggio è utilizzato per **prevedere la classe** a cui appartiene l'unità i (previsione di Y_i)
- Per definire un modello dobbiamo specificare il ? a sin. dell'=

$$? = \beta_0 + \beta_1 x_{i1} + ... + \beta_{k-1} x_{i,k-1}$$

- Nella regressione lineare, che cosa c'è a sinistra dell'=?

- **Nella regressione lineare**: **Y = variabile dipendente quantitativa** (con distribuzione normale)

$$E(y \mid X) = \mu_{y|X} = X\beta$$

- La combinazione lineare delle variabili esplicative descrive quindi il **valore atteso** di y_i
- **Nel problema in esame**: **Y = variabile dipendente dicotomica** (che rappresentiamo con una variabile aleatoria di **Bernoulli**)

y_i	Probabilità
0	$1 - \pi_i$
1	π_i
Tot.	1

$E(y_i)=\pi_i$: probabilità di "successo" per l'unità i

Modello di regressione per Y dicotomica

$$E(Y_i \mid x_i) = \pi_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{k-1} x_{i,k-1}$$

- I parametri possono essere stimati con il metodo dei **minimi quadrati** (v. regressione multipla). Però:
  - **Sono violate alcune ipotesi del modello sulla variabile dipendente Y (Quali?)**
  - **Non è detto che la stima di π_i sia compresa in [0; 1]**
- **Soluzione**: Si trasforma il range della variabile dipendente (π_i) dall'intervallo [0; 1] all'intero insieme dei numeri reali $(-\infty;+\infty) \Rightarrow$ **logit**

Logit

- Logit in presenza di **k-1 var. esplicative X_1, X_2, …:**

$$\pi(x_i) = P(Y_i = 1 \mid X_1 = x_{i1}, X_2 = x_{i2}, \ldots, X_{k-1} = x_{i,k-1})$$

$$\text{logit}[\pi(x_i)] = \log \frac{\pi(x_i)}{1-\pi(x_i)}$$

- **La probabilità di successo $\pi(x_i)$ dipende dai valori assunti dalle variabili esplicative (non stocastiche) X_1, X_2, …**
- **Modello di regressione per logit[$\pi(x_i)$]**

$$\log \frac{\pi(x_i)}{1-\pi(x_i)} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{k-1} x_{i,k-1} = x_i^T \beta$$

Modello di regressione logistica

$$\log \frac{\pi(x_i)}{1-\pi(x_i)} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{k-1} x_{i,k-1} = x_i^T \beta$$

- Il modello è **lineare nei parametri**: lo score per l'unità i è una combinazione lineare dei valori osservati x_{i1} … $x_{i,k-1}$
- Il modello non è però lineare in $\pi(x_i)$: non si può più utilizzare il metodo dei minimi quadrati
- Metodo alternativo di stima: **massima verosimiglianza** (maximum likelihood)
- Non esiste una formula esplicita per le stime dei parametri del modello: **algoritmo di stima iterativo** che risolve un sistema di equazioni non lineari
- Le variabili esplicative possono essere qualitative

COME SI STIMANO I PARAMETRI?

Il metodo della massima verosimiglianza

Esempio: dati generati da una distribuzione normale $\mathrm{p}(x|\mu;\sigma^2)$

- Qual è la probabilità che la v.c. X assuma il valore y

$$p(x=y\mid\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\cdot\exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]$$

Date due realizzazioni indipendenti della stessa v.c. Gaussiana $\mathrm{p}(x|\mu;\sigma^2)$ qual è la probabilità di osservare y1 nella prima realizzazione e y2 nella seconda realizzazione

$$p(y1\,|\,\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\cdot\exp\left[-\frac{1}{2}\left(\frac{y1-\mu}{\sigma}\right)^2\right]$$

$$p(y2\,|\,\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\cdot\exp\left[-\frac{1}{2}\left(\frac{y2-\mu}{\sigma}\right)^2\right]$$

$$p(y1,y2\,|\,\mu,\sigma)=p(y1\,|\,\mu,\sigma)\cdot p(y2\,|\,\mu,\sigma)=$$

$$=\frac{1}{\sqrt{2\pi}\sigma}\cdot\exp\left[-\frac{1}{2}\left(\frac{y1-\mu}{\sigma}\right)^2\right]\cdot\frac{1}{\sqrt{2\pi}\sigma}\cdot\exp\left[-\frac{1}{2}\left(\frac{y2-\mu}{\sigma}\right)^2\right]$$

Funzione di verosimiglianza

- Siano date **N variabili casuali <u>indipendenti</u>**… Quale è la **probabilità di misurare il vettore [y1, …, yN]**?

$$p(y1,y2,...,yN)=p(y1)\cdot p(y2)\cdot...p(yN)=L(y1,y2,...,yN)$$

- Questa è la FUNZIONE DI VEROSIMIGLIANZA (Likelihood, L).

Verosimiglianza e probabilità

- In gergo colloquiale spesso "verosimiglianza" è usato come sinonimo di "probabilità", ma in campo statistico vi è una distinzione tecnica precisa.
- Questo esempio chiarisce la differenza tra i due concetti: una persona potrebbe chiedere "Se lanciassi una moneta non truccata 100 volte, qual è la probabilità che esca testa tutte le volte?" oppure "Dato che ho lanciato una moneta 100 volte ed è uscita testa 100 volte, qual è la verosimiglianza che la moneta sia truccata?".

$f(y|\theta)$ e L$(\theta|y)$

- $f(y|\theta)$= (densità di) probabilità di osservare y dato θ
- L$(\theta|y)$ = funzione di verosimiglianza, funzione del parametro θ dato il risultato y (campione osservato), indica quanto verosimilmente è il valore di un parametro è corretto rispetto al risultato osservato = funzione di probabilità condizionata

Inferenza classica e Bayesiana

- Per il teorema di Bayes
- P(B|A)=P(A|B) P(B)/P(A)
- Di conseguenza:
- L$(\theta|y) = f(y|\theta)\, \mathrm{g}(\theta)/f(y)$
- Statistica classica: si concentra solo su $f(y|\theta)$
- Statistica Bayesiana: considera θ come una v.c. e specifica una distribuzione a priori per θ

Notazione

- Nel prosieguo invece di scrivere la funzione di verosimiglianza come

  L$(\theta|y) = f(y|\theta)$

  a volte scriveremo

  L$(\theta;\, y) = f(y;\theta) = f(y_1, \dots, y_n;\theta)$
- Se le osservazioni sono indipendenti
- L$(\theta;\, y) = \prod_{t=1}^{n} f(y_t;\theta)$

Se i dati in esame sono stati generati da f(y|θ) qual è il valore di θ più probabile?

- Siano θ_1e $\theta_2 \in \Theta$
- Mediante il rapporto tra $L(\theta_1; y)$ e $L(\theta_2; y)$ il sostegno empirico che θ^1 riceve da y viene confrontato con quello ricevuto da θ^2
- Se $L(\theta_1; y)/ L(\theta_2; y)$ θ_1è più credibile.

Rappresentazione funzione di verosimiglianza

$$L(\boldsymbol{\theta};\mathbf{y}) = L(\boldsymbol{\theta};\mathbf{y}\,|\,\mathbf{x}) = f(y_1,\ldots,y_n;\boldsymbol{\theta}\,|\,\mathbf{x}_1,\ldots,\mathbf{x}_n) = \prod_{t=1}^{n} f(y_t;\boldsymbol{\theta}\,|\,\mathbf{x}_t) = \prod_{t=1}^{n} f_t(y_t;\boldsymbol{\theta})$$

- Stima di massima verosimiglianza

Definizione – Sia $L(\boldsymbol{\theta};\mathbf{y})$ la funzione di verosimiglianza di un modello relativa al processo delle osservazioni $\{y_t\}_{t=1,\ldots,n}$. Dicesi stima **ML (M**aximum **L**ikelihood o di **Massima Verosimiglianza)** per il parametro $\boldsymbol{\theta}$, il punto di massimo $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y})$ (spesso funzione anche di $\mathbf{x}_1,\ldots,\mathbf{x}_n$) della funzione $L(\boldsymbol{\theta};\mathbf{y})$, **se esiste ed è unico**.

Log verosimiglianza

- La stima di massima verosimiglianza $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y})$ (se esiste) è anche punto di massimo della **funzione obiettivo**

$$Q_n(\boldsymbol{\theta}) \left(= Q_n(\boldsymbol{\theta}; \mathbf{y}) = \frac{1}{n} \log L(\boldsymbol{\theta}; \mathbf{y}) \right) = \frac{1}{n} \sum_{t=1}^{n} l_t(\boldsymbol{\theta}; y_t),$$

avendo posto $l_t(\boldsymbol{\theta}; y_t) = \log f_t(y_t; \boldsymbol{\theta})$.

Proprietà stimatori max verosimiglianza

- Consistenza;
- Asintotica normalità;
- Asintotica efficienza, nella classe degli stimatori asintoticamente corretti e consistenti;
- Invarianza rispetto a trasformazioni invertibili.

Teoria della verosimiglianza

- Il vettore delle derivate prime della funzione di verosimiglianza è chiamato «score function» di Fisher

$$\mathbf{u}(\boldsymbol{\theta}) = \frac{\partial \log \mathrm{L}(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}}.$$

- (meno) il valore atteso delle derivate seconde della funzione di verosimiglianza (matrice Hessiana) è chiamato matrice di informazione

$$\mathbf{I}(\boldsymbol{\theta}) = -\mathrm{E}[\frac{\partial^2 \log \mathrm{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}].$$

Matrice di var cov dello stimatore MLE $\hat{\theta}$ di θ

$$\hat{\boldsymbol{\theta}} \sim N_p(\boldsymbol{\theta}, \mathbf{I}^{-1}(\boldsymbol{\theta}))$$

Lo stimatore MLE ha una distribuzione normale multivariata con media uguale a quella del vero parametro e matrice di covarianza uguale all'inversa della matrice di informazione

Es. stimatore di max verosimiglianza nel contesto della regressione lineare multipla

- Costruiamo lo stimatore di massima verosimiglianza del vettore dei parametri θ

$$\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)'$$

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \quad \text{e} \quad \varepsilon_i \sim n.i.d.(0, \sigma^2).$$

Funzione obiettivo

$$L(y;\beta,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\exp[-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)]$$

$$logL(y;X,\sigma^2) = -\frac{n}{2}\log 2\,\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)$$

$$\frac{\partial \log L}{\partial \beta} = -\frac{1}{2\sigma^2}\frac{\partial}{\partial \beta}(y'\mathrm{y} - 2\mathrm{y}'\mathrm{X}\beta + \beta' X' X\beta)$$

$$\frac{\partial \log L}{\partial \beta} = (X'y - X'X\beta)/\sigma^2$$

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)'(y - X\beta)$$

Stimatori MLE di β e σ^2

- Uguagliando a zero la prima equazione

$$\frac{\partial \log L}{\partial \beta} = \frac{1}{\sigma^2}\,(X'y - X'X\beta)$$

$$\hat{\beta}_{MLE} = (X'X)^{-1}X'y$$

- Questo implica che lo stima di massima verosimiglianza di β è esattamente uguale a quello ottenuto con la tecnica dei minimi quadrati ➔ (Anche in presenza di una non corretta specificazione (errori non normali) gli stimatori hanno buone proprietà)

Stimatori MLE di β e σ^2

- Uguagliando a zero la seconda equazione

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)'(y - X\beta)$$

$$\hat{\sigma}^2{}_{MLE} = \frac{\left(y - X\hat{\beta}\right)'\left(y - X\hat{\beta}\right)}{n} = \frac{e'e}{n}$$

- Questo implica che lo stima di massima verosimiglianza di σ^2 non è uguale a quello ottenuto con la tecnica dei minimi quadrati

Partendo dallo score $u(\theta)$ calcoliamo ora l'Hessiano $H(\theta)$

$$\mathrm{u} = \begin{pmatrix} \frac{\partial \log L}{\partial \beta} = \frac{1}{\sigma^2}\,(X'y - X'X\beta) \\ \frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)'(y - X\beta) \end{pmatrix} \qquad H = \begin{pmatrix} \frac{\partial^2 \log L}{\partial \beta \partial \beta'} & \frac{\partial^2 \log L}{\partial \beta \partial \sigma^2} \\ \frac{\partial^2 \log L}{\partial \sigma^2 \partial \beta} & \frac{\partial^2 \log L}{\partial (\sigma^2)^2} \end{pmatrix}$$

$$H = \begin{pmatrix} -\frac{X'X}{\sigma^2} & -\frac{1}{\sigma^4}\,(X'y - X'X\beta) \\ -\frac{1}{\sigma^4}\,(y'X \quad -\beta'X'X) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(y - X\beta)'(y - X\beta) \end{pmatrix}$$

Partendo da $H(\theta)$ calcoliamo la matrice di informazione $I = -E(H(\theta))$

$$H = \begin{pmatrix} -\frac{X'X}{\sigma^2} & -\frac{1}{\sigma^4}\,(X'y - X'X\beta) \\ -\frac{1}{\sigma^4}\,(y'X \quad -\beta'X'X) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(y - X\beta)'(y - X\beta) \end{pmatrix}$$

- Tenendo presente che $\mathrm{E(y)} = \mathrm{X}\beta$, $\mathrm{E}(y - X\beta)'(y - X\beta) = n\,\sigma^2$

$$I = \begin{pmatrix} \frac{X'X}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

Matrice di var cov dello stimatore MLE $\hat{\theta}$ di θ è uguale all'inversa della matrice di informazione

$$I = \begin{pmatrix} \frac{X'X}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

$$I^{-1} = \begin{pmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix} = \begin{pmatrix} var(\hat{\beta}_{MLE}) & cov(\hat{\beta}_{MLE}, \sigma^2_{MLE}) \\ cov(\sigma^2_{MLE}, \widehat{\beta}'_{MLE}) & var(\sigma^2_{MLE}) \end{pmatrix}$$

Stimatore di max verosimiglianza per una popolazione normale

Siano $X_1, X_2, \ldots, X_n$ variabili aleatorie normali e indipendenti, con media μ e deviazione standard σ, entrambe incognite. La densità congiunta, e quindi la likelihood è data da

$$f(x_1, x_2, \ldots, x_n|\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\}$$
$$= \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sigma^n} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right\} \qquad (7.2.6)$$

La log-likelihood corrispondente è data da

$$\log f(x_1, x_2, \ldots, x_n|\mu, \sigma) = -\frac{n}{2} \log(2\pi) - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

Per trovare le stime $\widehat{\mu}$ e $\widehat{\sigma}$ che contemporaneamente massimizzino la log-likelihood, occorre porre uguali a zero le due derivate parziali, e mettere a sistema le due equazioni trovate.

$$\frac{\partial}{\partial \mu} \log f(x_1, x_2, \ldots, x_n|\mu, \sigma) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)$$

$$\frac{\partial}{\partial \sigma} \log f(x_1, x_2, \ldots, x_n|\mu, \sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2$$

Stimatori di max verosimiglianza di μ e σ

da cui il sistema

$$\begin{cases} \frac{1}{\widehat{\sigma}^2}\sum_{i=1}^{n}(x_i-\widehat{\mu})=0 \\ -\frac{n}{\widehat{\sigma}}+\frac{1}{\widehat{\sigma}^3}\sum_{i=1}^{n}(x_i-\widehat{\mu})^2=0 \end{cases}$$

la cui risoluzione ci porta alle seguenti formule per le stime,

$$\widehat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i$$
$$\widehat{\sigma}=\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i-\widehat{\mu})^2\right\}^{1/2}$$

Quindi, gli stimatori di massima verosimiglianza di μ e σ sono dati rispettivamente da

$$\overline{X} \qquad \text{e} \qquad \sqrt{\frac{1}{n}\sum_{i}(X_i-\overline{X})^2} \qquad (7.2.7)$$

Esercizio

Sia data la variabile casuale unidimensionale X con funzione di densità di probabilità data da:

$$f(x;\vartheta)=\begin{cases} \frac{1}{18\vartheta^4}\cdot e^{-\frac{\sqrt{x}}{3\vartheta^2}} & \text{se } x\geq 0 \\ 0 & \text{altrimenti} \end{cases}$$

con $\vartheta > 0$.

1. Trovare uno stimatore di $\vartheta > 0$ con il metodo di massima verosimiglianza.

1) Costruiamo la funzione di massima verosimiglianza

$$L(x_1,\cdots,x_n,\vartheta)=\frac{1}{18\vartheta^4}e^{-\frac{\sqrt{x_1}}{3\vartheta^2}}\cdot\frac{1}{18\vartheta^4}e^{-\frac{\sqrt{x_2}}{3\vartheta^2}}\cdot\ldots\cdot\frac{1}{18\vartheta^4}e^{-\frac{\sqrt{x_n}}{3\vartheta^2}}=$$
$$=\frac{1}{(18\vartheta^4)^n}e^{-\frac{(\sqrt{x_1}+\sqrt{x_2}+\ldots+\sqrt{x_n})}{3\vartheta^2}}$$

con $x_i \geq 0$, $i = 1,\ldots,n$.

Stimatore di max verosimiglianza

Cerchiamo $\widehat{\vartheta}$ che massimizza $\ln L$ tra le soluzioni di $\frac{d}{d\vartheta}\ln L = 0$. Calcoliamo

$$\ln L = -n\ln 18\vartheta^4 + \ln e^{-\frac{(\sqrt{x_1}+\sqrt{x_2}+\ldots+\sqrt{x_n})}{3\vartheta^2}}$$

$$= -n\ln 18 - 4n\ln\vartheta - \frac{\sqrt{x_1}+\sqrt{x_2}+\ldots+\sqrt{x_n}}{3\vartheta^2}$$

si ha

$$\frac{d}{d\vartheta}\ln L = -\frac{4n}{\vartheta} + \frac{2}{3\vartheta^3}\left(\sqrt{x_1}+\sqrt{x_2}+\ldots+\sqrt{x_n}\right) = 0$$

Essendo $\vartheta > 0$ risulta

$$\widehat{\vartheta} = \sqrt{\frac{\sqrt{x_1}+\sqrt{x_2}+\ldots+\sqrt{x_n}}{6n}}$$

Verifichiamo che sia un massimo

$$\frac{d^2}{d\vartheta^2}\ln L = +\frac{4n}{\vartheta^2} - \frac{2}{\vartheta^4}\left(\sqrt{x_1}+\sqrt{x_2}+\ldots+\sqrt{x_n}\right)$$

$$\left.\frac{d^2}{d\vartheta^2}\ln L\right|_{\vartheta=\widehat{\vartheta}} = \frac{2}{\widehat{\vartheta}^2}\left(2n - \frac{\widehat{\vartheta}^2\cdot 6n}{\widehat{\vartheta}^2}\right) = -\frac{8n}{\widehat{\vartheta}^2} < 0 \Longrightarrow \widehat{\vartheta} \text{ PUNTO DI MAX}$$

Quindi

$$\widehat{\Theta} = \sqrt{\frac{\sum_{i=1}^{n}\sqrt{X_i}}{6n}}$$

è uno stimatore di ϑ di massima verosimiglianza.

LA FUNZIONE DI VEROSIMIGLIANZA NEL CASO DELLA V.C. BERNOULLIANA

Variabile di Bernoulli (bernoulliana)

Esperimento aleatorio con 2 possibili esiti:

Insuccesso: 0

Successo: 1

Y	p(y)
0	$1-\pi$
1	π

$VAR(Y) = E(x-E(Y))^2$

$VAR(Y) = E(Y^2)-[E(Y)]^2 =$

$= \pi - \pi^2 = \pi(1-\pi)$

$= E(Y-\pi)^2 = (0-\pi)^2\ (1-\pi)-(1-\pi)^2\ \pi = \pi(1-\pi)$

50

Funzione di densità della v.c. Bernoulliana

- $f(y, \pi) = \pi^y (1-\pi)^{1-y}$

- Funzione di verosimiglianza =
- $\prod_{i=1}^{n} f(y_i, \pi) = \prod_{i=1}^{n} \pi^{y_i} (1-\pi)^{1-y_i}$

- Log verosimiglianza =
- $\sum_{i=1}^{n} [y_i \log \pi + (1-y_i) \log(1-\pi)]$

Stimatore di massima verosimiglianza di π

- Prendendo la derivata prima della log verosimiglianza ed uguagliando il risultato a 0 otteniamo che
- $\hat{\pi} = \frac{\sum_{i=1}^{n} y_i}{n}$

Nel caso della regressione logistica con una sola variabile esplicativa abbiamo che

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}},$$

in termini di *logit*

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x$$

- Ricordando l'ipotesi di indipendenza delle v.c. $Y_1, Y_2, \ldots, Y_n$ si può scrivere

$$\mathrm{L}(\beta_0, \beta_1 | x) = \prod_{i=1}^{n} f(y_i, \pi(x_i)) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}$$

Funzione di log verosimiglianza

Ricaviamo la funzione di log-verosimiglianza:

$$l(\beta_0,\beta_1) = \sum_{i=1}^{n}\{y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)]\} =$$

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

$$= \sum_{i=1}^{n}\left\{y_i(\beta_0+\beta_1 x_i) + \ln\left(1-\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)\right\} =$$

$$= \sum_{i=1}^{n}\left\{y_i(\beta_0+\beta_1 x_i) + \ln\left[\frac{1}{1+e^{\beta_0+\beta_1 x_i}}\right]\right\} =$$

$$= \sum_{i=1}^{n}\left\{y_i(\beta_0+\beta_1 x_i) - \ln\left[1+e^{\beta_0+\beta_1 x_i}\right]\right\}.$$

Regressione logistica multivariata

- La funzione di (log) verosimiglianza (una sola variabile esplicativa)

$$= \sum_{i=1}^{n} \left\{ y_i (\beta_0 + \beta_1 x_i) - \ln\left[1 + e^{\beta_0 + \beta_1 x_i}\right] \right\}.$$

- diventa nel caso di un vettore di variabili esplicative

$$L(\beta) = \sum_{i=1}^{N} \left[y_i \beta^T x_i - \log(1 + e^{\beta^T x_i}) \right] .$$

Sessione al computer

- File Regr_logistica_Macro_e_Ris.xlsm

- Implementazione funzione di verosimiglianza e massimizzazione tramite il risolutore. Verificare i problemi numerici

- Calcolando le derivate parziali rispetto ai parametri β_0 e β_1 e ponendole uguali a 0 si ricava il sistema delle equazioni di verosimiglianza, la cui soluzione restituisce le stime di max verosimiglianza che indichiamo con b_0 e b_1

$$\begin{cases} \sum_{i=1}^{n}\left(y_i - \frac{1}{1+e^{b_0+b_1x_i}} e^{b_0+b_1x_i} \right) = 0 \\ \sum_{i=1}^{n}\left(y_i x_i - \frac{1}{1+e^{b_0+b_1x_i}} e^{b_0+b_1x_i} x_i \right) = 0 \end{cases}$$

- Eq non lineari nelle incognite b_0 e b_1

Vincoli nella regressione semplice

$$\begin{cases} \sum_{i=1}^{n} e_i = 0 \\ \sum_{i=1}^{n} x_i e_i = 0 \end{cases} \qquad \begin{cases} \sum_{i=1}^{n} (y_i - a - bx_i) = 0 \\ \sum_{i=1}^{n} (y_i - a - bx_i) x_i = 0 \end{cases}$$

2 equazioni e 2 incognite (a e b)

Vincoli nella regressione semplice logistica

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \quad \text{e} \quad \sum_{i=1}^{n} x_i [y_i - \pi(x_i)] = 0.$$

- Prima condizione: la frequenza osservata di casi in cui la risposta è 1 coincide con la somma totale delle probabilità stimate

IL METODO DI NEWTON (RAPHSON) PER L'OTTIMIZZAZIONE NUMERICA

Ob. minimizzare una funzione con una sola incognita $f(\beta)$

- Se facciamo un'espansione in serie di Taylor vicino al minimo β^*

$$f(\beta) \approx f(\beta^*) + \frac{1}{2}(\beta - \beta^*)^2 \left.\frac{d^2 f}{d\beta^2}\right|_{\beta=\beta^*}$$

- Idea: $f(\beta)$ vicino al minimo è (approssimativamente) una funzione quadratica)
- Idea: Al posto di $f(\beta)$ andiamo a considerare un'approssimazione quadratica

Partiamo da un valore β_0

$$f(\beta) \approx f(\beta^{(0)}) + (\beta - \beta^{(0)}) \left.\frac{df}{dw}\right|_{\beta=\beta^{(0)}} + \frac{1}{2}\left(\beta - \beta^{(0)}\right)^2 \left.\frac{d^2 f}{dw^2}\right|_{\beta=\beta^{(0)}}$$

- Minimizziamo il lato destro: prendiamo la derivata prima rispetto a β e valutiamola in un punto β_1 e poniamola uguale a 0

$$\begin{aligned} 0 &= f'(\beta^{(0)}) + \frac{1}{2} f''(\beta^{(0)}) 2(\beta^{(1)} - \beta^{(0)}) \\ \beta^{(1)} &= \beta^{(0)} - \frac{f'(\beta^{(0)})}{f''(\beta^{(0)})} \end{aligned}$$

Procedura iterativa

$$0 = f'(\beta^{(0)}) + \frac{1}{2} f''(\beta^{(0)}) 2(\beta^{(1)} - \beta^{(0)})$$

$$\beta^{(1)} = \beta^{(0)} - \frac{f'(\beta^{(0)})}{f''(\beta^{(0)})}$$

- Esempio: file excel newton_raphson_esempio.xlsx

Caso multivariato $f(\beta_1 \ldots, \beta_p)$

$$\beta^{(n+1)} = \beta^{(n)} - H^{-1}(\beta^{(n)}) \nabla f(\beta^{(n)})$$

- ∇f = gradiente di f = vettore delle derivate parziali $[\frac{\partial f}{\beta_1}, \ldots, \frac{\partial f}{\beta_p}]^T$
- H = Hessiano di f = matrice delle derivate seconde parziali $H_{ij} = \frac{\partial^2 f}{\partial \beta_i \beta_j}$ $H = \frac{\partial^2 f}{\partial \beta \ \beta^T}$

Regressione logistica multivariata

- La funzione di (log) verosimiglianza (una sola variabile esplicativa)

$$= \sum_{i=1}^{n} \left\{ y_i (\beta_0 + \beta_1 x_i) - \ln\left[1 + e^{\beta_0 + \beta_1 x_i}\right] \right\}.$$

- diventa nel caso di un vettore di variabili esplicative

$$L(\beta) = \sum_{i=1}^{N} \left[y_i \beta^T x_i - \log(1 + e^{\beta^T x_i}) \right] .$$

Gradiente ed Hessiano

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{N} x_i (y_i - p(x_i; \beta))$$

$$\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^T} = - \sum_{i=1}^{N} x_i x_i^T p(x_i; \beta)(1 - p(x_i; \beta)) .$$

- Osservazione: $p(x_i; \beta) = \pi(x_i)$

Gradiente ed Hessiano

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{N} x_i (y_i - p(x_i; \beta))$$

$$\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^T} = -\sum_{i=1}^{N} x_i x_i^T p(x_i; \beta)(1 - p(x_i; \beta)) .$$

- In forma matriciale (data W =matrice di dimensione NxN l'iesimo elemento sulla diag principale è $p(x_i;\beta)$(1-$p(x_i;\beta)$ possiamo scrivere

$$\begin{aligned} \frac{\partial L(\beta)}{\partial \beta} &= \mathbf{X}^T(\mathbf{y} - \mathbf{p}) \\ \frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^T} &= -\mathbf{X}^T \mathbf{W} \mathbf{X} . \end{aligned}$$

Newton step

$$\beta^{(n+1)} = \beta^{(n)} - H^{-1}(\beta^{(n)}) \nabla f(\beta^{(n)})$$

$$\begin{aligned} \beta^{new} &= \beta^{old} + (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{y} - \mathbf{p}) \\ &= (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} (\mathbf{X} \beta^{old} + \mathbf{W}^{-1} (\mathbf{y} - \mathbf{p})) \\ &= (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{z} , \end{aligned}$$

$$\mathbf{z} \triangleq \mathbf{X} \beta^{old} + \mathbf{W}^{-1} (\mathbf{y} - \mathbf{p})$$

- Di conseguenza ad ogni passo si effettua una regressione GLS sulla variabile z

Newton step

$$\begin{aligned}\beta^{new} &= \beta^{old} + (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y}-\mathbf{p}) \\ &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(\mathbf{X}\beta^{old} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p})) \\ &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}\,,\end{aligned}$$

- W matrice diagonale con i-esimo elemento sulla diagonale principale pari $p(x_i;\beta^{old})$(1-$p(x_i;\beta^{old})$)
- $\mathbf{p}$= vettore Nx1 contenente le probabilità stimate in con i-esimo elemento $p(x_i;\beta^{old})$

Pseudo codice per ottenere le stime di max verosimiglianza

- Si inizializza β (ad es. β_0=0)
- Si trova **p** con i-esimo elemento
- $p(x_i;\beta) = \exp\beta^T x_i/(1+\exp\beta^T x_i)$
- Si costruisce la matrice W (con i-esimo elemento sulla diagonale $\mathrm{w_i} = p(x_i;\beta)$(1-($x_i;\beta$)
- Si costruisce z =$X\beta + W^{-1}(y-\mathbf{p})$
- Si trova una nuova stima di $\beta = \left(\mathrm{X^T WX}\right)^{-1}\mathrm{X^T Wz}$
- Si va avanti ….

Trucchi computazionali

- Nel calcolo di $\beta = \left(\mathrm{X^T W X}\right)^{-1}\mathrm{X^T W z}$ è possibile evitare di costruire la matrice diagonale di dimensione NxN W , in quanto si può moltiplicare la riga i della matrice X e l'elemento i del vettore z per $w_i^{0.5}$. In altri termini: se definiamo $X_w = W^{0.5}X$ e $y_w = W^{0.5}y$, la nuova stima di β può essere ottenuta come
- $\beta = \left(\mathrm{X_w^T X_w}\right)^{-1}\mathrm{X_w^T z_w}$

Regola di step

- Si va avanti nelle iterazioni fino a quando la differenza tra β_{new} e β_{old} è piccola
- Ad esempio, ci si ferma quando

$$(\beta_{new}-\beta_{old})^T(\beta_{new}-\beta_{old})/(\beta_{new}^T\beta_{new}) < 1e-15$$

Sessione al computer

- Regrlogistica_Iter.xlsm
- Studiare la macro VBA

Interpretazione dei parametri nell'esempio dell'insolvenza

- x= debito espresso in migliaia di Euro

Variabili nell'equazione

		B	S.E.	Wald	gl	Sign.	Exp(B)
Fase 1[a]	Debito	,111	,024	21,254	1	,000	1,117
	Costante	-5,309	1,134	21,935	1	,000	,005

a. Variabili inserite nella fase 1: Debito.

- -5.309 = stima del logit in assenza di debito. La prob di insolvenza quando debito=0 è

$$\hat{\pi}(0) = \frac{e^{-5.30945}}{1 + e^{-5.30945}} = 0.00492 \approx 0.5\%.$$

Interpretazione dei parametri nell'esempio dell'insolvenza

- Interpretazione di β_1
- $logit\ [\pi(x+1)] = \beta_0 + \beta_1(x+1)$
- $logit\ [\pi(x)] = \beta_0 + \beta_1 x$

$$\beta_1 = \text{logit}[\pi(x+1)] - \text{logit}[\pi(x)] = \log\left[\frac{\pi(x+1)}{1-\pi(x+1)}\right] - \log\left[\frac{\pi(x)}{1-\pi(x)}\right].$$

- $\hat{\beta}_1$ = variazione stimata nel logit in corrispondenza di un incremento unitario della variabile indipendente

Interpretazione dei parametri nell'esempio dell'insolvenza

- x= debito espresso in migliaia di Euro

Variabili nell'equazione

		B	S.E.	Wald	gl	Sign.	Exp(B)
Fase 1[a]	Debito	,111	,024	21,254	1	,000	1,117
	Costante	-5,309	1,134	21,935	1	,000	,005

a. Variabili inserite nella fase 1: Debito.

- 0.111 = variazione nel logit di insolvenza in corrispondenza di un incremento di 1000 Euro di debito

$$\beta_1 = \text{logit}[\pi(x+1)] - \text{logit}[\pi(x)] = \log\left[\frac{\pi(x+1)}{1-\pi(x+1)}\right] - \log\left[\frac{\pi(x)}{1-\pi(x)}\right].$$

Interpretazione dei parametri nell'esempio dell'insolvenza

$$\beta_1 = \text{logit}[\pi(x+1)] - \text{logit}[\pi(x)] = \log\left[\frac{\pi(x+1)}{1-\pi(x+1)}\right] - \log\left[\frac{\pi(x)}{1-\pi(x)}\right].$$

$$\beta_1 = \log\left[\frac{\frac{\pi(x+1)}{1-\pi(x+1)}}{\frac{\pi(x)}{1-\pi(x)}}\right].$$

- Interpretazione di $\exp(\beta_1)$
- Stima della variazione dell'odds ratio (OR) in corrispondenza di un incremento unitario della variabile indipendente

Interpretazione dei parametri nell'esempio dell'insolvenza

- x= debito espresso in migliaia di Euro

Variabili nell'equazione

		B	S.E.	Wald	gl	Sign.	Exp(B)
Fase 1[a]	Debito	,111	,024	21,254	1	,000	1,117
	Costante	-5,309	1,134	21,935	1	,000	,005

a. Variabili inserite nella fase 1: Debito.

- 1,117= ad un incremento di 1000 Euro di debito l'odds ratio di insolvenza aumenta dell'11.7%

$$\beta_1 = \log\left[\frac{\frac{\pi(x+1)}{1-\pi(x+1)}}{\frac{\pi(x)}{1-\pi(x)}}\right].$$

Osservazione: OR esprime un effetto moltiplicativo non additivio

- $logit\,[\pi(x+c)] = \beta_0 + \beta_1(x+c)$
- $logit\,[\pi(x)] = \beta_0 + \beta_1 x$
- $log\,\left[\frac{\pi(x+c)}{1-\pi(x+c)}\right] = \beta_0 + \beta_1(x+c)$
- $log\,\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x$
- $\left[\frac{\pi(x+c)}{1-\pi(x+c)}\right] / \left[\frac{\pi(x)}{1-\pi(x)}\right] = \exp(\beta_1 c)$
- $\exp(\beta_1 c)$ = variazione nell'odds ratio aumentando la variabile indipendente di c unità

Interpretazione dei parametri nell'esempio dell'insolvenza

- x= debito espresso in migliaia di Euro

Variabili nell'equazione

		B	S.E.	Wald	gl	Sign.	Exp(B)
Fase 1[a]	Debito	,111	,024	21,254	1	,000	1,117
	Costante	-5,309	1,134	21,935	1	,000	,005

a. Variabili inserite nella fase 1: Debito.

- Ad un incremento di 10000 Euro del debito corrisponde un incremento dell'odds ratio di insolvenza di circa tre volte ($\exp(10 \times 0.111) \approx 3.03$)
- Ad un incremento di 20000 Euro del debito corrisponde un incremento dell'odds ratio di insolvenza di poco superiore alle nove volte ($\exp(20 \times 0.111) \approx 9.21$)

Interpretazione dei parametri

- Di quanto aumenta la probabilità stimata dell'evento ad un incremento marginale di x.
- Matematicamente: dobbiamo calcolare
- $\frac{d\pi(x)}{dx} = \frac{d[\exp(\beta_0+\beta_1 x)/(1+\exp(\beta_0+\beta_1 x)]}{dx}$
- $\frac{d\pi(x)}{dx} = \beta_1\, exp(\beta_0 + \beta_1 x)/[\,1 + exp(\beta_0 + \beta_1 x)]^2$
- Dipende dal valore assunto da x (è proporzionale a β_1). Solitamente si valuta nel punto $\overline{x}$

Interpretazione dei parametri nell'esempio dell'insolvenza

- x= debito espresso in migliaia di Euro

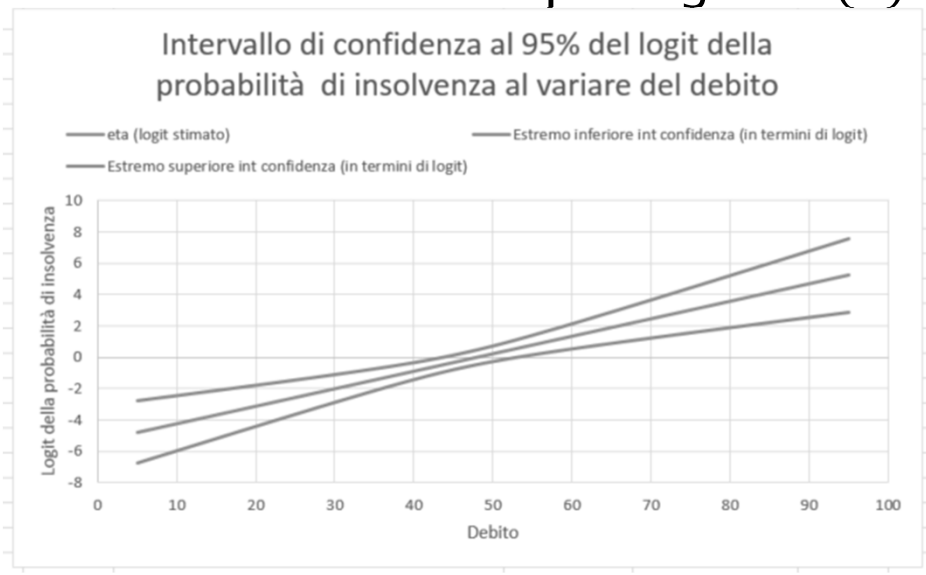
Variabili nell'equazione

		B	S.E.	Wald	gl	Sign.	Exp(B)
Fase 1[a]	Debito	,111	,024	21,254	1	,000	1,117
	Costante	-5,309	1,134	21,935	1	,000	,005

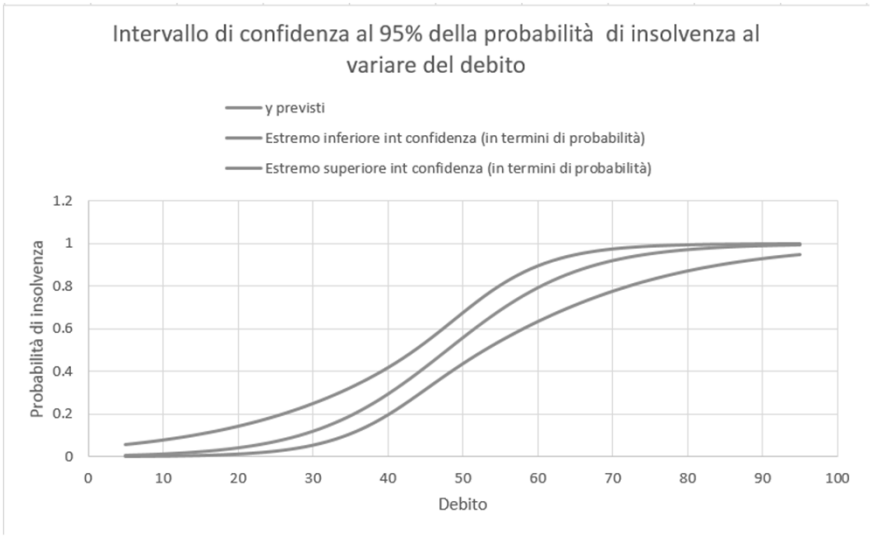
a. Variabili inserite nella fase 1: Debito.

- $\overline{debito}$=44.38
- $\beta_1\, exp\left(exp(\beta_0 + \beta_1\overline{debito})\right)/[\,1 + exp(\beta_0 + \beta_1\overline{debito})]^2$=0.027
- Interpretazione: se il valore del debito è pari a 44380 € circa, ad un incremento marginale del debito corrisponde un incremento della probabilità di insolvenza di 0.027

Intervallo di confidenza per logit e $\pi(x)$



Intervallo di confidenza per logit e $\pi(x)$



Intervallo di confidenza per logit

- $\widehat{logit(x)} = \hat{\beta}_0 + \hat{\beta}_1 x$

- Nel caso multivariato
- $\widehat{logit(x)} = \hat{\beta}^T\, x$
- Per trovare l'intervallo di confidenza del logit occorre trovare ($\mathrm{var}(\widehat{logit(x)})$)

$\Pr(\widehat{logit(x)} - 1.96\ \sqrt{\mathrm{var}(\widehat{logit(x)})}$<logit(x)<$\widehat{logit(x)} + 1.96\ \sqrt{\mathrm{var}(\widehat{logit(x)})}) = 0.95$

Intervallo di confidenza per $\pi(x)$

- Dato l'intervallo di confidenza del logit, si passa poi all'intervallo di confidenza di $\pi(x)$ tenendo presente che
- $\pi(x) = [\exp\ logit(x)]/[1+\exp\ logit(x)]$

Sessione al computer

- File Regr_logistica_logitINT

Regressione logistica - Esempio

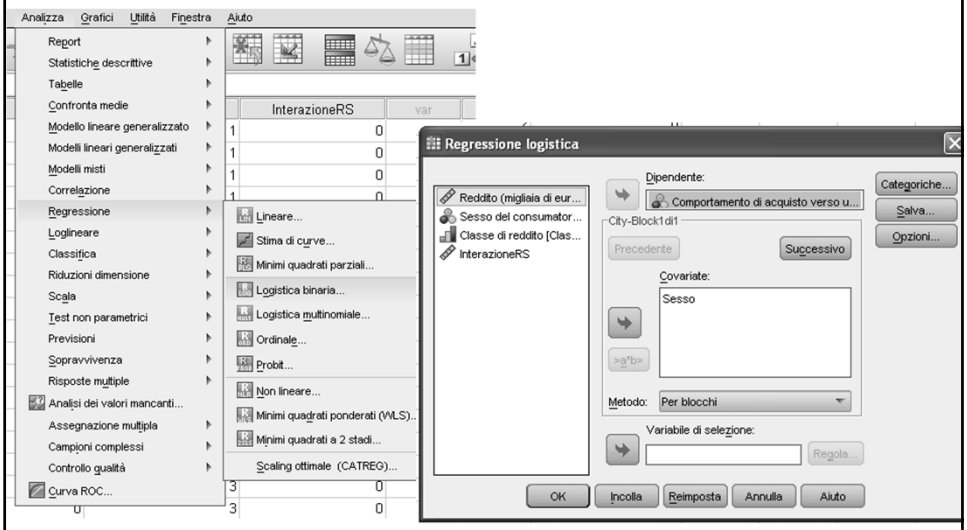
- n = 40 clienti
- Obiettivo (semplificato): prevedere il comportamento di acquisto su un prodotto di largo consumo in base a reddito (in migliaia di Euro supposto noto) e sesso del consumatore
- Prime 10 righe della matrice dei dati:

	Acquisto	Reddito	Sesso	Classe_reddito
1	0	37	0	1
2	0	39	0	1
3	0	39	0	1
4	0	42	0	1
5	0	47	0	2
6	0	48	0	2
7	1	48	0	2
8	0	52	0	2
9	0	53	0	2
10	0	55	0	3

Analisi preliminari

Distribuzione di X=reddito (valori anomali, forma di distribuzione …)
Associazione tra Y e le X

In SPSS



Esempio: Modello 1 (regr. log. semplice)

- Comportamento di acquisto (Y) in funzione di X_1 = sesso del consumatore (variabile dummy):

$$\log \frac{\pi(x_i)}{1-\pi(x_i)} = \beta_0 + \beta_1 x_{i1}$$

- **Codifica X_1** (arbitraria): $x_{i1} = 0$ (F); $x_{i1} = 1$ (M)
- Output SPSS modello logistico:

Variabili nell'equazione

		B	E.S.	Wald	df	Sig.	Exp(B)	95% CI per EXP(B)	
								Inferiore	Superiore
Passo 1[a]	Sesso	1.695	.690	6.030	1	.014	5.444	1.408	21.054
	Costante	-.847	.488	3.015	1	.082	.429		

a. Variabili immesse al passo 1: Sesso.

- Quale **interpretazione** dei parametri?
- **Significatività** dei risultati (test e intervalli di confidenza)

Modello 1 – interpretazione parametri

- **Una sola variabile esplicativa X_1 dicotomica** (dummy)
- Se **$x_{i1} = 0$** (consumatore femmina):

$$\log \frac{\pi(x_i)}{1-\pi(x_i)} = \beta_0$$

$\Rightarrow$ exp(β_0) è la quota relativa (odds) $\pi(x_i)/[1-\pi(x_i)]$ per il gruppo delle femmine

- Se **$x_{i1} = 1$** (consumatore maschio):

$$\log \frac{\pi(x_i)}{1-\pi(x_i)} = \beta_0 + \beta_1$$

$\Rightarrow$ exp($\beta_0+\beta_1$) è la quota relativa (odds) per il gruppo dei maschi

- Quindi β_1 è **la differenza tra il logit per i maschi (X=1) e il logit per le femmine (X=0)**.
- Proprietà di log:

$$\beta_1 = \log \frac{\left[\pi(x_i)/1-\pi(x_i)\right]_M}{\left[\pi(x_i)/1-\pi(x_i)\right]_F} = \log\left[\text{ODDS RATIO}\right]$$

- Pertanto: **exp(β_1) = Odds Ratio**

Modello 1 – aspetti inferenziali

- Output:

Variabili nell'equazione

		B	E.S.	Wald	df	Sig.	Exp(B)	95% CI per EXP(B)	
								Inferiore	Superiore
Passo 1[a]	Sesso	1.695	.690	6.030	1	.014	5.444	1.408	21.054
	Costante	-.847	.488	3.015	1	.082	.429		

a. Variabili immesse al passo 1: Sesso.

- **E.S.**: **errore standard** (asintotico) = $\sqrt{\text{var}(\hat{\beta}_j)}$
- **Wald** = **statistica t^2** per la verifica di H_0: $\beta_j=0$ contro H_1: $\beta_j\neq 0$

$$W = t^2 = \frac{\hat{\beta}_j^2}{\text{var}(\hat{\beta}_j)}$$

- Se H_0 è vera, **in grandi campioni** il test Wald si distribuisce come **Chi-quadrato** con df=1: $[N(0, 1)]^2$

$$W = t^2 \sim \chi_1^2$$

- **P-value:**

$$P(\chi_1^2 > W_{obs} \mid H_0 \text{ vera})$$

- Intervalli di confidenza (asintotici): **interpretazione** (ZC§6.2)

Esempio: Modello 2 (per esercizio)

- Comportamento di acquisto (Y) in funzione del reddito del consumatore (3 classi): $$\log \frac{\pi(x_i)}{1-\pi(x_i)} = \beta_0 + \beta_1 x_{i1}$$
- **Codifica X_1** (parzialmente arbitraria):
  - x_{i1} = 1 se Reddito < 45
  - x_{i1} = 2 se Reddito compreso tra 45 e 54
  - x_{i1} = 3 se Reddito > 54
- Output modello logistico:

Variabili nell'equazione

		B	E.S.	Wald	df	Sig.	Exp(B)	95% CI per EXP(B)	
								Inferiore	Superiore
Passo 1[a]	Classe_reddito	.437	.388	1.270	1	.260	1.548	.724	3.310
	Costante	-.896	.858	1.090	1	.296	.408		

a. Variabili immesse al passo 1: Classe_reddito.

- **Interpretazione dei parametri e inferenza**
- **Effetto della scelta delle classi e della loro quantificazione?**

Esempio: Modello 2bis

- Comportamento di acquisto (Y) in funzione del reddito del consumatore (**variabile quantitativa**):

$$\log \frac{\pi(x_i)}{1-\pi(x_i)} = \beta_0 + \beta_1 x_{i1}$$

- Output modello logistico:

Variabili nell'equazione

		B	E.S.	Wald	df	Sig.	Exp(B)	95% CI per EXP(B)	
								Inferiore	Superiore
Passo 1[a]	Reddito	.054	.036	2.276	1	.131	1.055	.984	1.132
	Costante	-2.734	1.838	2.214	1	.137	.065		

a. Variabili immesse al passo 1: Reddito.

- **Interpretazione dei parametri e inferenza**
- **Effetto del reddito a parità di sesso?**

Esempio: Modello 3 (regr. log. multipla)

- Comportamento di acquisto (Y) in funzione del sesso (X_1) e del reddito (X_2) del consumatore:

$$\log \frac{\pi(x_i)}{1-\pi(x_i)} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

- **X_1 è dummy**:
  - **Se x_{i1} = 1 (M)** $\log \frac{\pi(x_i)}{1-\pi(x_i)} = \beta_0 + \beta_1 + \beta_2 x_{i2}$
  - **Se x_{i1} = 0 (F)** $\log \frac{\pi(x_i)}{1-\pi(x_i)} = \beta_0 + \beta_2 x_{i2}$
- Adattare un modello logistico tra Y, X_1 (dummy) e X_2 (reddito) equivale ad adattare **due modelli logistici** diversi tra Y e X_2: un modello per M e un altro per F.
- **Tali modelli differiscono per l'intercetta; la pendenza è invece la stessa (β_2)**

Esempio: Modello 3 (X_2 = reddito quant.)

$$\log \frac{\pi(x_i)}{1-\pi(x_i)} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

- Output modello logistico:

Variabili nell'equazione

		B	E.S.	Wald	df	Sig.	Exp(B)	95% CI per EXP(B)	
								Inferiore	Superiore
Passo 1[a]	Reddito	.158	.062	6.576	1	.010	1.171	1.038	1.322
	Sesso	3.490	1.199	8.469	1	.004	32.780	3.125	343.834
	Costante	-9.843	3.676	7.171	1	.007	.000		

a. Variabili immesse al passo 1: Reddito, Sesso.

- Interpretazione dei parametri: **coefficienti netti (parziali)**
- **L'effetto del reddito (a parità di sesso) ora è triplicato ed è significativo**

Intepretazione parametri

Commento di 3.49= a parità di reddito annuo del consumatore la stima del logit di acquisto è 3.49 volte più alta nel gruppo dei maschi

Commento di 32.78= controllando l'effetto del reddito, l'odds di acquisto dei maschi è quasi 33 volte superiore a all'odds di acquisto tra le femmine

Commento di 1.171=a parità di sesso del consumatore, ad un aumento di 1000 Euro del reddito annuo corrisponde una variazione positiva nell'odds di acquisto del 17% circa

Modello 3: Stima della probabilità di acquisto

- Ad esempio: **consumatore F** di reddito 40:

$$\log\frac{\hat{\pi}(x_i)}{1-\hat{\pi}(x_i)} = -9.843 + 3.49\cdot 0 + 0.158\cdot 40 = -3.523$$
$$\text{ODDS} = \exp(-3.523) = 0.0295$$
$$\hat{\pi}(x_i) = \frac{\exp(-3.523)}{1+\exp(-3.523)} = \frac{1}{1+\exp(3.523)} = 0.0287$$

- Se invece:**consumatore M** di reddito 40:

$$\log\frac{\hat{\pi}(x_i)}{1-\hat{\pi}(x_i)} = -9.843 + 3.49\cdot 1 + 0.158\cdot 40 = -0.033$$
$$\text{ODDS} = \exp(-0.033) = 0.9675$$
$$\hat{\pi}(x_i) = \frac{\exp(-0.033)}{1+\exp(-0.033)} = \frac{1}{1+\exp(0.033)} = 0.4918$$

- Nel **Modello 2bis** (non distingue tra M e F):

$$\log\frac{\hat{\pi}(x_i)}{1-\hat{\pi}(x_i)} = -2.734 + 0.054\cdot 40 = -0.574$$
$$\text{ODDS} = \exp(-0.574) = 0.5633$$
$$\hat{\pi}(x_i) = \frac{\exp(-0.574)}{1+\exp(-0.574)} = \frac{1}{1+\exp(0.574)} = 0.3603$$

- **E' possibile considerare un coeff. del Reddito diverso per M e F?**

Modello 4: interazione tra le variabili esplicative

- Comportamento di acquisto in funzione del sesso (X_1), del reddito quantitativo (X_2) e di un fattore di **interazione** ($X_1 X_2$):

$$\log \frac{\pi(x_i)}{1-\pi(x_i)} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}$$

- Ora i modelli M e F **differiscono sia per l'intercetta sia per la pendenza**

$$\text{F:} \quad \log \frac{\pi(x_i)}{1-\pi(x_i)} = \beta_0 + \beta_2 x_{i2}$$

$$\text{M:} \quad \log \frac{\pi(x_i)}{1-\pi(x_i)} = (\beta_0 + \beta_1) + \beta_2 x_{i2} + \beta_3 x_{i2} = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) x_{i2}$$

- Output modello logistico:

Variabili nell'equazione

		B	E.S.	Wald	df	Sig.	Exp(B)	95% CI per EXP(B)	
								Inferiore	Superiore
Passo 1[a]	Reddito	.197	.094	4.439	1	.035	1.218	1.014	1.463
	Sesso	7.705	6.760	1.299	1	.254	2218.792	.004	1.259E9
	InterazioneRS	-.082	.126	.423	1	.515	.921	.720	1.179
	Costante	-12.146	5.582	4.736	1	.030	.000		

a. Variabili immesse al passo 1: Reddito, Sesso, InterazioneRS.

- In questo caso **l'interazione non è utile**
- **E' opportuno inserire con parsimonia le interazioni nel modello**

Bontà di adattamento - indici

- L'indice R^2 non è più utilizzabile (perché?) $\Rightarrow$ generalizzazioni in funzione della verosimiglianza
- **Cox & Snell R^2**

Riepilogo del modello

Passo	-2 log verosimiglianza	R-quadrato di Cox e Snell	R-quadrato di Nagelkerke
1	38.917[a]	.339	.451

a. La stima è stata interrotta all'iterazione numero 5 perché le stime dei parametri sono cambiate di meno del .001.

- **Approcci alternativi**:
  - Test sulla bontà di adattamento
  - Misura della capacità predittiva

Bontà di adattamento - indici

- L'indice R^2 non è più utilizzabile (perché?) $\Rightarrow$ generalizzazioni in funzione della verosimiglianza
- **Cox & Snell R^2**

$$R^2_{CS} = 1 - \left\{ \frac{L(0)}{L(\hat{\beta})} \right\}^{2/n}$$

- L(0) = verosimiglianza nel modello che contiene solo l'intercetta
- $L(\hat{\beta})$ = verosimiglianza nel modello che contiene tutte le variabili esplicative prese in esame.
- Idea: Tanto più è piccolo il rapporto $L(0)/L(\hat{\beta})$, tanto più il modello in esame fornisce un adattamento migliore rispetto a quello con solo l'intercetta.
- $L(\hat{\beta})$ è il prodotto di n probabilità, prendendo la radice n-esima si ottiene una stima della verosimiglianza per ciascun valore di y

Cox & Snell R^2

$$R^2_{CS} = 1 - \left\{ \frac{L(0)}{L(\hat{\beta})} \right\}^{2/n}$$

- IL massimo valore dell'indice non è 1
- Se il modello corrente prevede perfettamente y la sua likelihood sarà 1

$$\mathrm{L}\left(\hat{\beta} \middle| x \right) = \prod_{i=1}^{n} f(y_i, \hat{\pi}\,(x_i)) = \prod_{i=1}^{n} \hat{\pi}(x_i)^{\,y_i} (1 - \hat{\pi}(x_i))^{1-y_i}$$

- Di conseguenza $\max(R^2_{CS}) = 1 - L(0)^{\frac{2}{n}} < 1$

Cox & Snell R^2 normalizzato (Nagelkerke)

- $\mathrm{R}_{\mathrm{N}}^2 = \frac{1-\left\{\frac{L(0)}{L(\hat{\beta})}\right\}^{2/n}}{1-L(0)^{2/n}}$

- L'indice vale 1 quando c'è un adattamento perfetto
- L'indice vale 0 quando le variabili esplicative non aumentano in alcun modo la bontà di adattamento del modello ossia quando $L(0) = L(\hat{\beta})$

Quanto vale L(0) = funzione di log verosimiglianza nel modello ridotto?

$$L(\beta) = \sum_{i=1}^{N} \left[y_i \beta^T x_i - \log(1 + e^{\beta^T x_i}) \right] .$$

- Dal vincolo: $\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0$
- Si ottiene che
- $\sum_{i=1}^{n} (y_i - \exp(\hat{\beta}_0)/(1 + \exp(\hat{\beta}_0)) = 0$
- $\hat{\beta}_0 = \log(\overline{y}/(1-\overline{y})$. Nel nostro modello $\overline{y}$=0.5 ➔ $\hat{\beta}_0$=0
- L(0)=$-n \log 2$

Sessione al computer

- File
- Regrlogistica Inf(out).xlsm

Confronto tra modelli

- Nella regressione multipla (con Y quantitativa) il confronto tra modelli avviene con il test F $\Rightarrow$ verifica l'ipotesi:

H_0: β_1=0 ... β_q=0 (q parametri del modello sono = 0)

- Confrontiamo un modello più semplice (M1) con uno più complesso (M2): M1 è ottenuto ponendo **alcuni parametri di M2 = 0** (le X corrispondenti non hanno effetto su Y): **modelli "annidati"**(nested)
- Il test F (per piccoli campioni) richiede però che Y abbia **distribuzione Gaussiana**
- Nella regressione logistica si costruisce un test analogo **per grandi campioni** attraverso la funzione di verosimiglianza $\Rightarrow$ il valore di tale funzione cresce aumentando il numero di parametri nel vettore β (cioè passando da M1 a M2)

Confronto tra modelli - 2

- **L = funzione di verosimiglianza** (Likelihood)
- Il confronto tra valori di **l=log(L)** dà luogo al **Test del rapporto di verosimiglianza**

$$G^2 = -2(l_1 - l_2) = 2(l_2 - l_1)$$

dove l_2**=log(L)** calcolato sul modello **M2** (più complesso) e l_1**=log(L)** calcolato sul modello **M1** (più semplice) $\Rightarrow$ Pertanto: $l_2 \geq l_1$ (entrambi i valori sono < 0) e $G^2 \geq 0$

- **Se H_0 è vera (M1 è il vero modello) e il campione è grande: $G^2 \sim \chi^2$ con q gradi di libertà**
- Nota: non c'è contrasto con il test F della regressione perché in grandi campioni F e G^2 sono equivalenti

Confronto tra modelli – Esempio

- M2 = modello con reddito (quantitativo) e sesso; M1 = modello solo con intercetta (nessuna esplicativa)

$$\log \frac{\pi(x_i)}{1-\pi(x_i)} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

- H_0: $\beta_1=0$, $\beta_2=0$
- Output SPSS (test omnibus sui coefficienti)

Test omnibus dei coefficienti del modello

		Chi-quadrato	df	Sig.
Passo 1	Passo	16.535	2	.000
	Blocco	16.535	2	.000
	Modello	16.535	2	.000

- **$G^2 = 16.535 \Rightarrow$ distribuzione χ^2 con 2 gradi di libertà**
- G^2=16.535 è **fortemente significativo** (**p-value approx = 0**) $\Rightarrow$ forte evidenza che **almeno un coefficiente** tra β_1 e β_2 è ≠ 0 (analogia con il test F dell'ANOVA)

Test di Wald per $\beta_1=0 \ldots \beta_p=0$

- $\hat{\beta} \sim N(\beta, I^{-1})$
- I =matrice di informazione
- Se è vera l'ipotesi $\beta=0$
- $\hat{\beta}^T \ \mathrm{I} \ \hat{\beta} \sim \mathrm{X}_p^2$

Bontà di adattamento - test

- Il test G^2 per il confronto tra modelli può essere usato anche per valutare la bontà di adattamento $\Rightarrow$ si pone:

  M1 = modello stimato

  M2 = modello "saturo": i valori di Y stimati dal modello coincidono con quelli osservati $\Rightarrow$ modello con adattamento perfetto

- **Se G^2 è elevato (significativo) l'adattamento è scadente: le stime fornite dal modello differiscono significativamente dalle osservazioni**

Distribuzione del test $\mathbf{G^2}$

- $\mathbf{G^2 = 2(l_2 - l_1) \sim X^2_{v-p}}$
- **v=numero di parametri del modello saturo (n)**
- **p=numero di parametri stimati del modello corrente**
- $\mathbf{l_2}$ **= verosimiglianza del modello saturo (uguale a 1)**
- $\mathbf{l_1}$**=verosimiglianza del modello corrente**
- **➔** $\mathbf{G^2 = -2\, l_1}$ **= (scaled) DEVIANCE, nella regressione corrisponde alla somma dei quadrati dei residui /**σ^2

Test $\mathbf{G^2}$ nella regressione multipla

- Log likelihood
- $logL(y; X, \sigma^2) = -\frac{n}{2}\log 2\,\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)$
- Log lik modello saturo ($\mathbf{l_2}$)=
- $-\frac{n}{2}\log 2\,\pi - \frac{n}{2}\log \hat{\sigma}^2$
- Log lik modello corrente ($\mathbf{l_1}$)=
- $-\frac{n}{2}\log 2\,\pi - \frac{n}{2}\log \hat{\sigma}^2 - \frac{1}{2\sigma^2}(y - X\hat{\beta})'(y - X\hat{\beta})$
- $\mathbf{2(l_2 - l_1)} = \frac{1}{\sigma^2}(y - X\hat{\beta})'(y - X\hat{\beta})$

Output SPSS

Riepilogo del modello

Passo	-2 log verosimiglianza	R-quadrato di Cox e Snell	R-quadrato di Nagelkerke
1	38.917[a]	.339	.451

a. La stima è stata interrotta all'iterazione numero 5 perché le stime dei parametri sono cambiate di meno del .001.

Bontà di adattamento - test

- L'approx. χ^2 della devianza richiede però che le frequenze con cui osserviamo i valori delle X siano grandi (ad es. > 5): **problemi con variabili esplicative quantitative** (ad es. il reddito)

- Una variante di questo approccio quando si hanno variabili quantitative è il test di **Hosmer-Lemeshow**

Test di Hosmer-Lemeshow

- Si raggruppano le osservazioni in 10 classi approx. della stessa numerosità (decili della distribuzione delle probabilità stimate).
- Test chi-quadrato:

$$HL = \sum_{g=1}^{10} \frac{(s_g - \hat{s}_g)^2}{n_g \hat{\pi}_g (1-\hat{\pi}_g)}$$

$\hat{s}_g$ = frequenza stimata nel gruppo g

$\hat{\pi}_g$ = probabilità (media) stimata nel gruppo g

Test di Hosmer-Lemeshow

Passo	Chi-quadrato	df	Sig.
1	5.229	8	.733

Tabella di contingenza per il test Hosmer-Lemeshow

		Comportamento di acquisto verso un prodotto = No		Comportamento di acquisto verso un prodotto = Si		
		Osservato	Attesa	Osservato	Attesa	Totale
Passo 1	1	4	3.894	0	.106	4
	2	3	3.564	1	.436	4
	3	3	3.028	1	.972	4
	4	3	2.612	1	1.388	4
	5	3	2.069	1	1.931	4
	6	1	1.999	4	3.001	5
	7	2	1.216	2	2.784	4
	8	0	.933	4	3.067	4
	9	1	.572	3	3.428	4
	10	0	.114	3	2.886	3

- Distribuzione (approssimata) χ^2 con 10 – 2 gradi di libertà
- Regola empirica: $n_g \geq 5$
- **Se HL è piccolo (non significativo) l'adattamento è accettabile**: v. esempio

Sessione al computer

- File
- regrlogisticatestHL(out).xlsx
- y previsti+epsilon_i per evitare valori ripetuti =L3+0.0000000001*RIF.RIGA(L3)
- La funzione SCARTO

Selezione del modello

- Il confronto tra modelli (ad es. tramite G^2) può essere usato per scegliere il modello "migliore"
- Criteri automatici (v. regressione multipla):
  - **Backward**: si parte dal modello completo e si rimuove ad ogni passo il parametro con il p-value più elevato; la procedura si arresta quando tutti i parametri hanno p-value < soglia
  - **Forward**: si parte dal modello con solo intercetta e si aggiunge ad ogni passo il parametro con il p-value più piccolo; la procedura si arresta quando tutti i parametri non ancora inclusi hanno p-value > soglia
  - **Stepwise**: alterna passi F e B
- I criteri automatici sono utili con molte variabili. Però:
  - B parte da un modello complesso: le stime possono essere poco precise
  - F esamina le variabili una alla volta: procedura sequenziale in cui ogni passo dipende fortemente da quelli precedenti
  - **Il modello finale dipende dalla sequenza dei passi** e ha gli **stessi inconvenienti visti nella regressione**

Usi del modello

- Il modello logistico è nato con **finalità conoscitive**: ad esempio, per comprendere le relazioni tra fattori di rischio e insorgenza di una malattia (generalizza la regressione)
- Nel marketing esso è utilizzato soprattutto a **fini predittivi**:
  - Individuazione dei **fattori rilevanti** nel comportamento di acquisto: **profilazione dei consumatori** (caratteristiche distintive di acquirenti / non acquirenti)
  - Stima della **probabilità individuale** di acquisto
  - **Classificazione dei consumatori $\Rightarrow$ le classi sono le modalità di Y**
- Come avviene la classificazione?

Classificazione con il modello logistico

- **Stima di P(Y=1|X=x_i)** con il modello scelto:

$$\hat{\pi}(x_i) = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_{k-1} x_{i,k-1}\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_{k-1} x_{i,k-1}\right)}$$

x_i = profilo del cliente i

oppure (se le X sono qualitatitive)

x_i = cella i della tabella di contingenza delle esplicative ⇒ profilo dei clienti di tale cella

- Tale stima misura la **propensione individuale** verso il comportamento descritto da Y=1
- **Regola di classificazione** (2 classi):

$$\text{Se } \hat{\pi}(x_i) > 1 - \hat{\pi}(x_i) \Rightarrow \hat{Y}(x_i) = 1$$

$$\text{Se } \hat{\pi}(x_i) \leq 1 - \hat{\pi}(x_i) \Rightarrow \hat{Y}(x_i) = 0$$

Errori della regola di classificazione

- Probabilità di avere un **falso positivo:**

$$\Pr(\hat{Y}(x_i) = 1 \mid Y(x_i) = 0)$$

- Probabilità di avere un **falso negativo:**

$$\Pr(\hat{Y}(x_i) = 0 \mid Y(x_i) = 1)$$

- Analogia con errori I e II specie nella verifica di ipotesi
- **Specificità** della regola di classificazione:

$$\Pr(\hat{Y}(x_i) = 0 \mid Y(x_i) = 0) = 1 - \Pr(\hat{Y}(x_i) = 1 \mid Y(x_i) = 0)$$

- **Sensitività** della regola di classificazione:

$$\Pr(\hat{Y}(x_i) = 1 \mid Y(x_i) = 1) = 1 - \Pr(\hat{Y}(x_i) = 0 \mid Y(x_i) = 1)$$

- **Tabella di errata classificazione**

Tabella di errata classificazione

- Le n unità possono essere classificate nella **classe prevista (P)** dal modello e nella **classe effettiva (E)**, che è quella osservata: **matrice di confusione**

E\P	0	1	Tot.
0	a	b	a+b
1	c	d	c+d
Tot.	a+c	b+d	n

a + b (numero di unità per cui Y=0) e c + d (numero di unità per cui Y=1) **non dipendono dal modello** ⇒ sono le vere caratteristiche dei dati

a + c (numero di unità per cui Y previsto è = 0) e b + d (numero di unità per cui Y previsto è = 1) **dipendono invece dal modello**

- **Hit rate** = frequenza relativa di unità correttamente classificate: (a+d)/n
- **Specificità**: a/(a+b) (prevedo 0 dato che si verifica 0)
- **Sensitività**: d/(c+d) (prevedo 1 dato che si verifica 1)
- **Falsi positivi**: b/(a+b)
- **Falsi negativi**: c/(c+d)

Tabella di errata classificazione – esempio: acquisto = f(reddito, sesso)

Tabella Classificazione[a]

Osservato			Previsto		
			Comportamento di acquisto verso un prodotto		
			No	Sì	Percentuale corretta
Passo 1	Comportamento di acquisto verso un prodotto	No	15	5	75.0
		Sì	4	16	80.0
	Percentuale globale				77.5

a. Il valore di riferimento è .500

- **Hit rate** = 31/40 = 77.5%
- **Specificità** = 15/20 = 75%
- **Sensitività** = 16/20 = 80%
- **Falsi positivi** = 5/20 = 25%
- **Falsi negativi** = 4/20 = 20%
- **Confronto con il caso** ⇒ se prevediamo Y a caso: la probabilità che la stima di Y sia 1 è costante rispetto a X (nell'esempio è 0.5). In tale caso, avremmo quindi a = b = c = d = 20×0.5 = 10 ⇒ hit rate: 50%

Tabella di errata classificazione – problemi 1

- **Hit rate dà lo stesso peso a entrambi i tipi di errore**: spesso non è ragionevole (costi $\neq$; risente delle frequenze marginali nel campione)
- Si può scegliere una **soglia diversa da 0.5** (in funzione del costo di ciascun errore): ad esempio soglia=0.25 per ridurre i falsi negativi

Tabella Classificazione[a]

Osservato			Previsto		
			Comportamento di acquisto verso un prodotto		
			No	Sì	Percentuale corretta
Passo 1	Comportamento di acquisto verso un prodotto	No	9	11	45.0
		Sì	1	19	95.0
	Percentuale globale				70.0

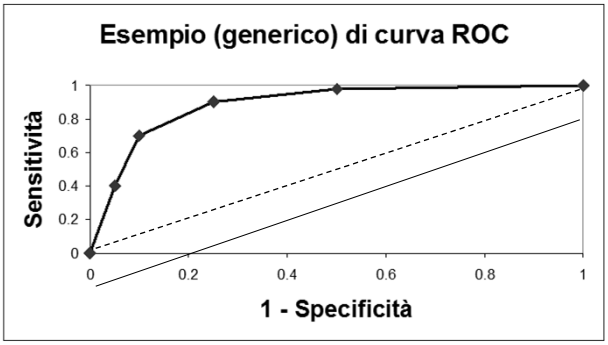
a. Il valore di riferimento è .250

- Aumentano però i falsi positivi $\Rightarrow$ curva ROC (Receiver Operating Characteristic)
- Talvolta il campione è **fortemente sbilanciato** (Y=1 è un evento raro) $\Rightarrow$ sono necessarie correzioni

- **Curva ROC**: grafico di **Sensitività vs. (1 – Specificità)**

cioè di $\Pr(\hat{Y}(x_i)=1 \mid Y(x_i)=1)$ vs. $\Pr(\hat{Y}(x_i)=1 \mid Y(x_i)=0)$

al variare di una caratteristica della regola di classificazione (**soglia per la classificazione**, variabili esplicative …)



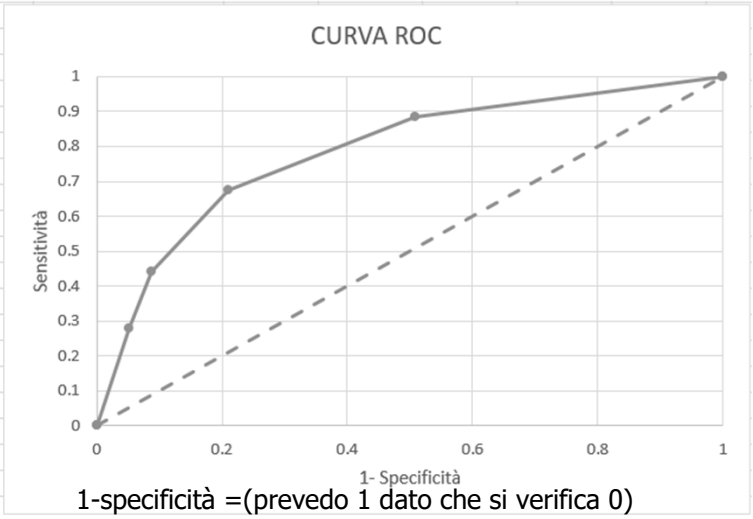
I punti corrispondono a differenti **soglie** per la prob. di classificazione nella classe 1:

1 $\Rightarrow$ $\hat{Y}(x_i) \equiv 0$
0.75
0.60
0.50
0.25
0 $\Rightarrow$ $\hat{Y}(x_i) \equiv 1$

- Ad es. nel primo punto, **tutti i valori previsti di Y sono 0**: non ci sono **Falsi Positivi**, la Sensitività è 0 e la Specificità è 1
- La diagonale rappresenta il **confronto con il caso** $\Rightarrow$ le regole di classificazione migliori sono quelle per cui la curva ROC passa vicino all'angolo in alto a sinistra

Curva ROC nell'esempio del debito

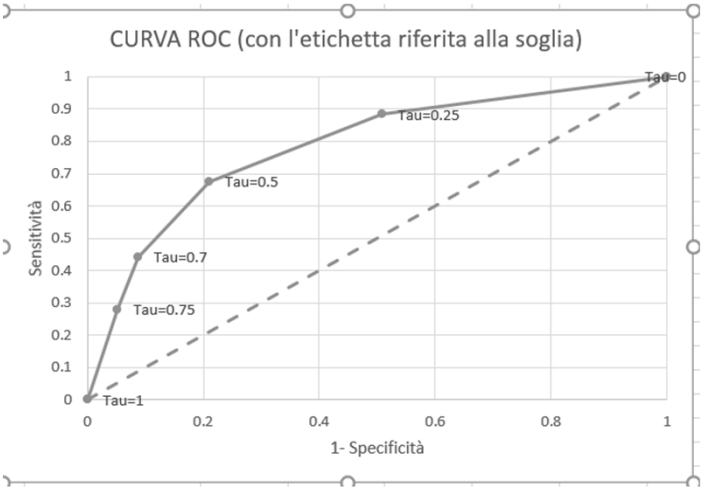
Sensitività =(prevedo 1 dato che si verifica 1)



1-specificità =(prevedo 1 dato che si verifica 0)

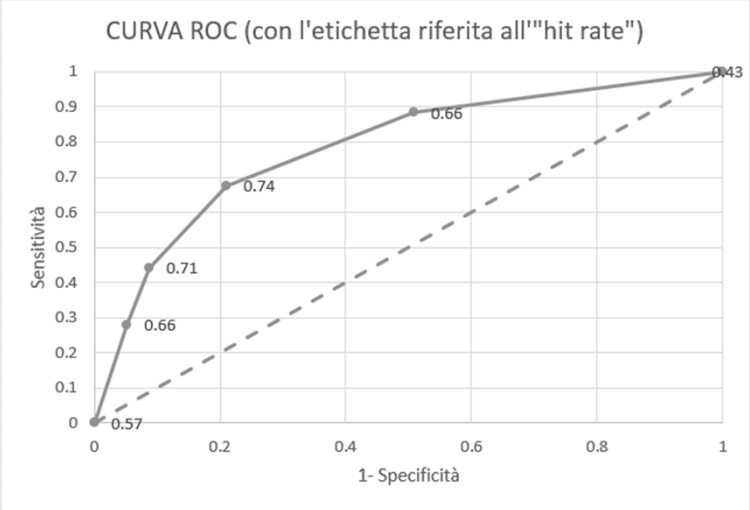
Curva ROC nell'esempio del debito

Sensitività =(prevedo 1 dato che si verifica 1)



1-specificità =(prevedo 1 dato che si verifica 0)

Curva ROC nell'esempio del debito



Sessione al computer

- File
- regrlogistica_ROC(out).xlsx

Tabella di errata classificazione – problemi 2

- In pratica, interessa **la probabilità di errore su una nuova osservazione/unità: errore di generalizzazione**
- La probabilità di errore calcolata dalla tabella di errata classificazione **è una stima** della probabilità di errore di generalizzazione: **si tratta di una stima accurata?**
- Errore sui dati osservati: **errore di apprendimento**
- In realtà la stima dell'errore di apprendimento è **troppo ottimistica** se riferita all'errore di generalizzazione $\Rightarrow$ **perché?**
- **Overfitting**: ottimo adattamento ai dati osservati ma pessima capacità previsiva di nuove osservazioni $\Rightarrow$ **cause ed esempi di overfitting**

- **Possibili soluzioni**: **V. alberi di classificazione**