Regressione Binomiale

L. Egidi

Autunno 2021

Università di Trieste
Corso di laurea magistrale in Scienze Statistiche ed Attuariali

Regressione binomiale

Modelli per risposta binomiale

Inferenza

Regressione binomiale

Variabili risposta dicotomica

Molto spesso si osservano dati su un insieme di unità in cui la variabile risposta riguarda l'essersi o meno verificato un particolare evento. Per citare alcuni esempi dell'ambito attuariale/finanziario:

- un cliente decide se acquistare una polizza
- un cliente decide se cambiare compagnia alla scadenza di una polizza (*churn*)
- un cliente decide di continuare a lavorare o andare in pensione
- un'azienda diventa insolvente
- un individuo è ancora vivo dopo 5 anni dalla stipula di una polizza

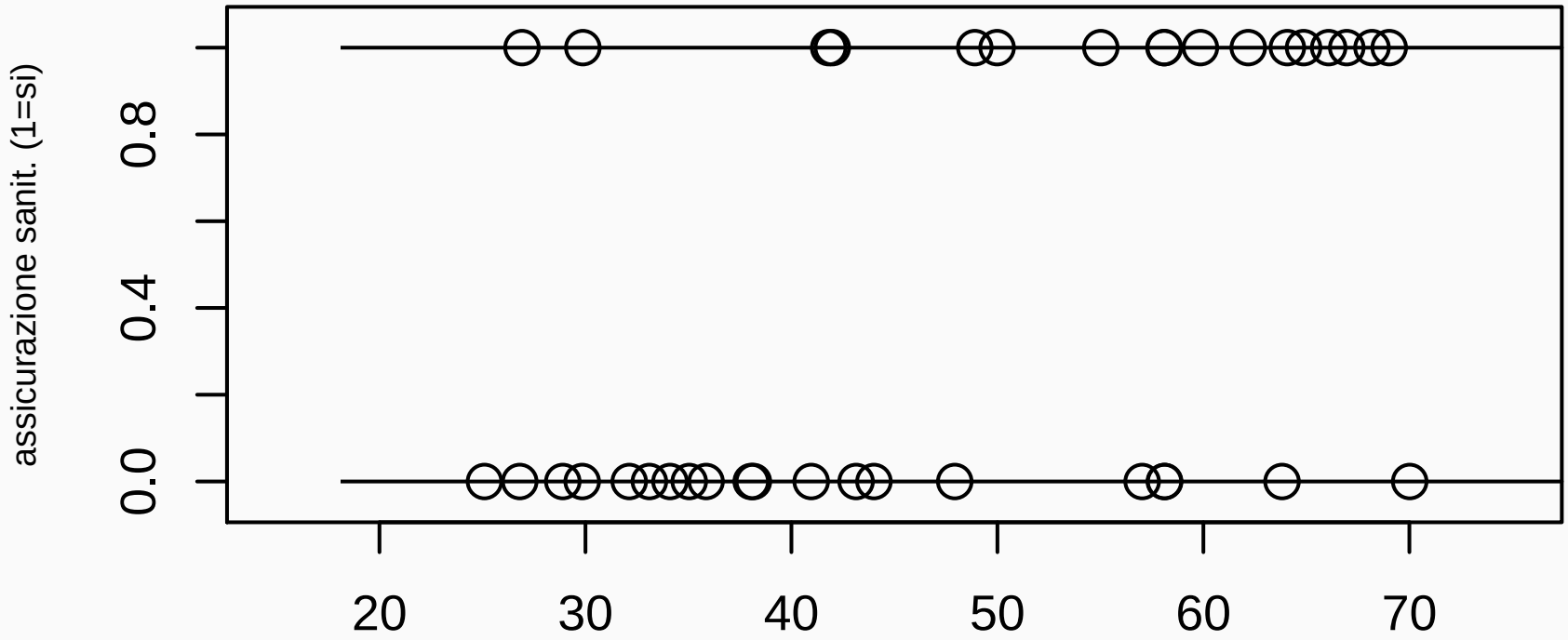
Un modello per variabile dipendente binaria

- Il set informativo di cui si suppone di disporre per l'unità i-esima è costituito da
  - una variabile risposta y_i
  - un insieme di $p-1$ covariate $x_i = (x_{i1}, x_{i2}, \ldots, x_{i(p-1)})$
- Come per il caso di risposta quantitativa siamo interessati a costruire un modello statistico che ci permetta di (prevedere) la media della variabile y_i utilizzando le informazioni sulle covariate x_i. Quindi:
  - $y_i \sim Be(\pi_i)$ per cui $E(y_i) = \pi_i$
  - si definisce il **predittore lineare**

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

  - si ipotizza che $E(y_i) = \pi_i = r(\eta_i) = r(\boldsymbol{x}_i^T \boldsymbol{\beta})$
- La funzione $r(\cdot)$ è detta **funzione risposta** e va scelta in modo opportuno fra quelle continue e invertibili.

Un primo esempio: un'assicurazione sanitaria



Si assuma di avere osservato 37 unità e per ciascuna di esse conosciamo l'età e se ha o meno sottoscritto una polizza sanitaria. Il grafico sembra suggerire che nel campione vi sia maggiore propensione a sottoscrivere una polizza sanitaria da parte dei soggetti più anziani.

1. $y_i \sim \text{Be}(r(\eta_i))$ ove $\eta_i = \boldsymbol{x}_i^T \beta$.
2. Una funzione ragionevole deve essere tale che $r(\cdot) : \mathbb{R} \to [0, 1]$.

Possibili scelte per le funzioni $r()$ e $g()$

È sempre possibile esprimere la relazione sopra invertendo la funzione $r()$. Si definisce quindi la **funzione legame** $g(\cdot) = r^{-1}(\cdot)$ e si ha equivalentemente

$$g(E(y_i)) = g(\pi_i) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

Possiamo ottenere modelli alternativi scegliendo fra funzioni $g(\cdot)$ (oppure $r(\cdot)$) alternative. Le due opzioni più note e impiegate sono quelle che conducono ai cosidetti modelli **logit** e **probit**.

- *Modello logit*: si usa la funzione risposta

  $$\pi_i = r(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}} = \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}} \text{ o } g(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

- *Modello probit*: si usa la funzione risposta

  $$\pi_i = \Phi(\eta_i) = \Phi(\boldsymbol{x}_i^T \boldsymbol{\beta}),$$

  ove $\Phi(\cdot)$ è la funzione di ripartizione della Gaussiana standard

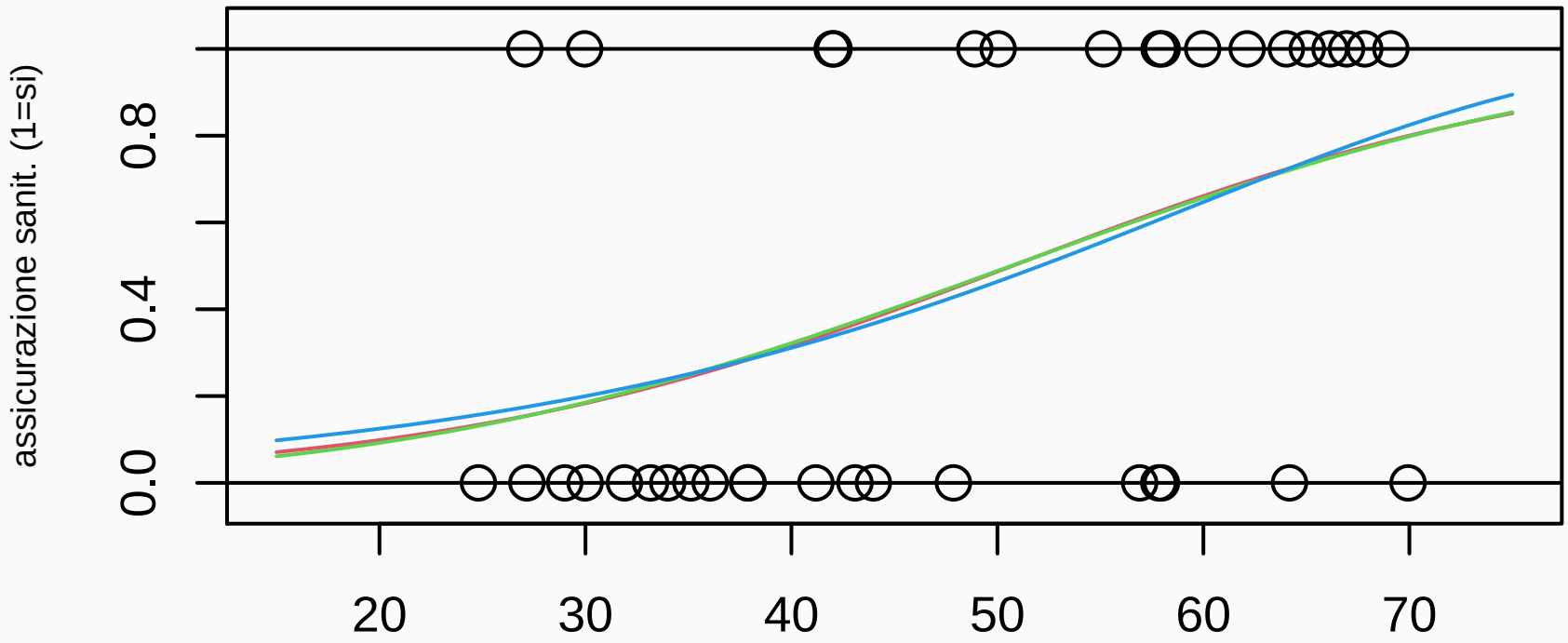
Altre funzioni risposta (e legame)

- Tuttavia altre funzioni possono essere utilizzate, quale ad esempio il **modello log-log complementare**, per il quale si definisce la funzione risposta

$$r(\eta_i) = 1 - \exp(-\exp(\eta_i)) \quad \text{oppure} \quad g(\pi_i) = \log(-\log(1-\pi))$$

- Ogni funzione che mappi la retta reale $\mathbb{R}$ in [0,1] e che sia continua e invertibile potrebbe essere una buona scelta. Pertanto potrebbe essere scelta qualsiasi funzione di ripartizione di una variabile aleatoria continua e con supporto l'asse reale.
- In effetti la funzione risposta logit corrisponde alla ripartizione di una v.c. logistica, quella probit a una v.c. Gaussiana e quella log-log complementare a quella di una v.c valore estremo di I tipo (Gumbel).
- Nulla osta quindi a utilizzare funzioni di ripartizione di altre v.c. continue (quella della t di student con g gradi di libertà ad esempio).

Ancora sull'esempio



Nel grafico vedete le diverse forme delle funzioni $r(\cdot)$ in corrispondenza di un modello logit (rosso), probit (verde), c-log-log (blu).

Si noti che la funzione risposta c-log-log non corrisponde alla funzione di ripartizione di una variabile simmetrica.

Perchè usare il legame logit?

L'utilizzo della funzione legame logit dà origine alla regressione logistica che è di gran lunga la più popolare e la più usata. I motivi del suo successo sono:

- è computazionalmente più agevole e gode di alcune proprietà teoriche (un dettaglio del quale si tratterà più avanti)
- è più facile interpretare i risultati (effetto lineare delle covariate sul logaritmo degli odds-ratio, o effetto moltiplicativo sugli odds-ratio)
- fornisce stime adeguate anche in presenza di schemi di campionamento retrospettivi

Interpretazione dei parametri nel modello logit

Gli odds $\frac{\pi_i}{1-\pi_i} = \frac{Pr(y_i=1|\boldsymbol{x}_i)}{Pr(y_i=0|\boldsymbol{x}_i)}$ nel modello logit sono pari a

$$\frac{Pr(y_i = 1|\boldsymbol{x}_i)}{Pr(y_i = 0|\boldsymbol{x}_i)} = e^{\beta_0} e^{\beta_1 x_{i1}} \dots e^{\beta_{p-1} x_{i(p-1)}},$$

seguono cioè un modello moltiplicativo. Quindi se aumenta di 1 il valore di x_{i1} l'effetto sugli odds è:

$$\frac{Pr(y_i = 1|x_{i1} + 1, \dots)}{Pr(y_i = 0|x_{i1} + 1, \dots)} \Big/ \frac{Pr(y_i = 1|x_{i1,\dots})}{Pr(y_i = 0|x_{i1,\dots})} = e^{\beta_1}$$

quindi:

- $\beta_1 > 0$: $\frac{Pr(y_i=1|x_i)}{Pr(y_i=0|x_i)}$ cresce del $e^{\beta_1}\%$ per una variazione unitaria di x_i
- $\beta_1 < 0$: $\frac{Pr(y_i=1|x_i)}{Pr(y_i=0|x_i)}$ decresce del $e^{\beta_1}\%$ per una variazione unitaria di x_i
- $\beta_1 = 0$: $\frac{Pr(y_i=1|x_i)}{Pr(y_i=0|x_i)}$ rimane costante al variare di x

Dati prospettivi e retrospettivi e legame logit

Il tema è stato sollevato dapprima nell'ambito bio-statistico ma si applica in realtà a tutti i casi in cui il campione da analizzare vede sovrarappresentata una delle due modalità della variabile dipendente rispetto a quanto accade nella popolazione (nel caso ad esempio che una delle due classi sia più rara).

- Si immagini che nella popolazione la classe per cui $y = 1$ per un individuo con caratteristiche x si presenti con una frequenza pari a $p(x)$ molto bassa e quindi tali casi sono rari.
- Si immagini che sia π_1 la probabilità di includere un individuo nel campione condizionatamente al fatto che sia di tipo 1 e che sia π_0 la probabilità di includere un individuo nel campione condizionata al fatto che sia di tipo 0 .
- Se si fa un campione casuale abbiamo che $\pi_1 = \pi_0$. Tuttavia se decidiamo di sovrarappresentare nel campione individui rari (di tipo 1) allora sarà $\pi_1 >> \pi_0$

Sovracampionamento della classe rara e legame logit

Definiamo la probabilità $p^*(x)$ come la probabilità che un individuo con caratteristiche x sia di tipo 1 condizionatamente alla sua inclusione nel campione e vediamo in che relazione è con la probabilità $p(x)$ che sia di tipo 1 (ovvero non condizionata alla scelta del campione e quindi relativa a un campione casuale in cui però gli 1 saranno rari).

Possiamo applicare il teorema di Bayes:

$$p^*(x) = \frac{\pi_1 p(x)}{\pi_1 p(x) + \pi_0 (1 - p(x))}$$

alcuni semplici passaggi mostrano che per un generico individuo con caratteristiche x:

$$\log\left(\frac{p^*(x)}{1 - p^*(x)}\right) = \log\frac{\pi_1}{\pi_0} + \log\left(\frac{p(x)}{1 - p(x)}\right) = \log\frac{\pi_1}{\pi_0} + x^T\beta.$$

Quindi i parametri β associati alle variabili x non cambiano (cambia solo l'intercetta che sarà pari a $\log\frac{\pi_1}{\pi_0} + \beta_0$).

Modelli a soglia e legame probit

La funzione probit è estremamente popolare in alcuni ambiti applicativi (ad esempio fra gli econometrici).

È possibile giustificare il modello ipotizzando l'esistenza di una variabile continua della quale osserviamo una versione discreta. Si immagini quindi che per la variabile continua Y^* valga il seguente modello:

$$Y_i^* = \beta_0 + \beta_1 x_i + \epsilon_i$$

con $\epsilon_i \sim N(0, \sigma^2)$ conforme alle usuali assunzioni di un modello lineare semplice (l'estensione al caso con più variabili esplicative è immediata). Si ipotizzi ora che

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > k \\ 0 & \text{altrimenti} \end{cases}$$

Una giustificazione dell'impiego del legame probit

Dalle assunzioni fatte risulta che

$$P(Y_i = 1) = Pr(Y_i^* > k) = P(\epsilon \leq -k+\beta_0+\beta_1 x_i) = \Phi((-k+\beta_0+\beta_1 x_i)/\sigma)$$

che è un GLM probit con
$\eta_i = \beta_0^* + \beta_1^* x_i$ ove
$\beta_0^* = (\beta_0 - k)/\sigma$ e $\beta_1^* = \beta_1/\sigma$.

- Se per ϵ non si assume la gaussianità otterremo, analogamente, dei GLM la cui funzione legame è l'inversa della funzione di ripartizione di ϵ. Nel caso di distribuzione logistica si ottiene il modello logit e di distribuzione valore estremo il modello con legame c-log-log.
- Vi possono essere varie motivazioni a giustificare l'impostazione data sopra. La variabile Y^* :
  - può rappresentare la tolleranza nel caso degli insetti sottoposti a diverse dosi di veleno;
  - può rappresentare la differenza fra le utilità che il soggetto i-esimo ricava dalla scelta fra le due alternative (modelli a scelta discreta)

Modelli per risposta binomiale

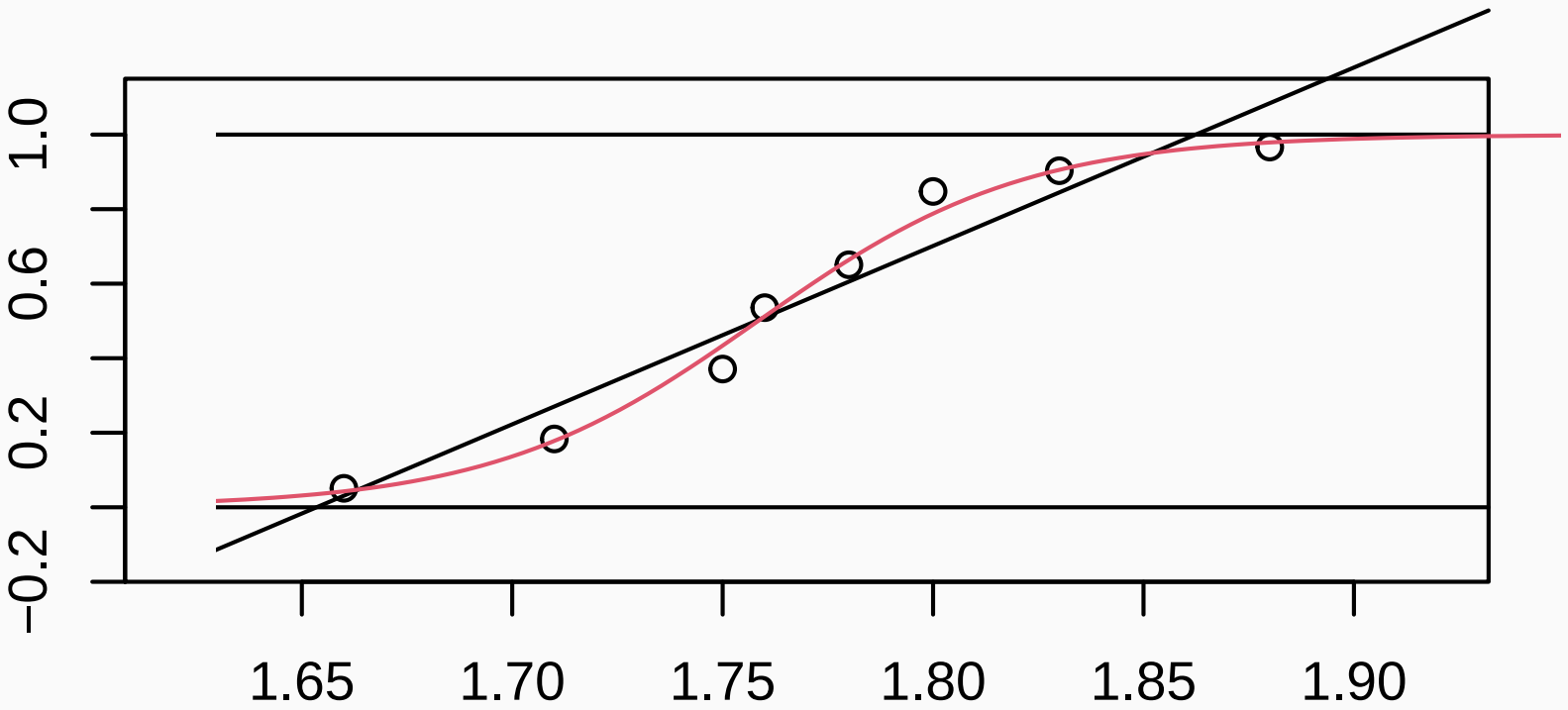
Un secondo esempio: analisi dose risposta

- Si considerino i dati nella tabella sotto

dose	1.66	1.74	1.75	1.76	1.78	1.80	1.86	1.88
n. positivi	3	9	23	30	46	54	59	58
n. di pazienti	59	60	62	56	63	59	62	60
proporzione	0.051	0.150	0.371	0.536	0.730	0.915	0.951	0.967

- I dati sono raccolti su 481 individui che hanno assunto un farmaco. Per ogni dose del farmaco si è osservato se l'individuo ha risposta positiva o meno.
- Vi erano solo 8 dosi diverse e per ciascuna si può osservare il numero e la proporzione di risposte positive

Un secondo esempio: analisi dose-risposta



- La proporzione di risposte positive su m_i pazienti, cresce con la dose del farmaco.
- Una relazione lineare non sarebbe appropriata. Le proporzioni devono stare nel range [0,1]
- $Y_i \sim \text{Bin}(m_i, r(\eta_i))$ specifica un modello non lineare per le diverse scelte di $r(\cdot)$. Se $p_i = \frac{Y_i}{m_i}$ è la frequenza relativa di successi, $p_i \sim \text{Bin}(m_i, r(\eta_i))/m_i$ è una binomiale riscalata.

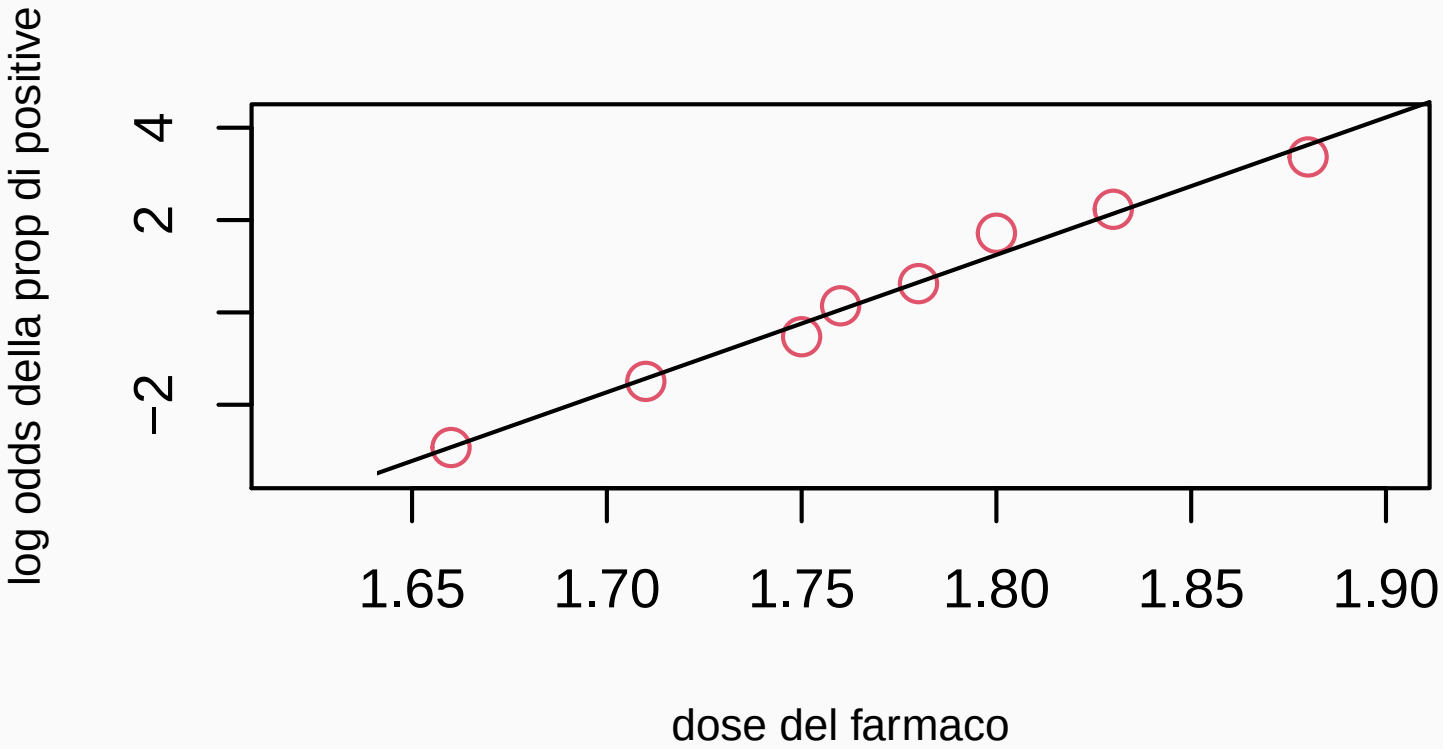
Regressione lineare sui log-odds (regressione logistica)

Per i dati sopra possiamo calcolare odds e log-odds empirici

dose	1.66	1.74	1.75	1.76	1.78	1.80	1.86	1.88
n. positive	3	9	23	30	46	54	59	58
n. of patients	59	60	62	56	63	59	62	60
proportion (p)	0.051	0.150	0.371	0.536	0.730	0.915	0.951	0.967
$p/(1-p)$	0.05	.177	0.59	1.15	2.71	10.80	19.67	29.00
$\log(p/(1-p))$	-2.92	-1.73	-0.53	0.14	0.99	2.38	2.98	3.36

- $\frac{p}{1-p}$ sono gli odds che possono assumere qualsiasi valore tra 0 e ∞
- $\log \frac{p}{1-p}$ sono i log–odds che possono assumere qualsiasi valore reale

Regressione lineare sui log-odds: rappresentazione del modello dose-risposta



- La relazione fra la dose e i log-odds delle proporzioni empiriche è lineare!
- Un incremento unitario della variabile x si traduce in una variazione nel log-odds pari al coefficiente angolare della retta β_1

Inferenza

Stime dei parametri nel modello logit

Il metodo della massima verosimiglianza è adeguato in questo caso in quanto abbiamo un'ipotesi distributiva precisa e assumeremo l'indipendenza per il campione di n osservazioni. Nel caso del modello Bernoulliano, si ha la log verosimiglianza $\log(L(\beta)) = l(\beta)$:

$$
\begin{aligned}
l(\beta) &= \sum_{i=1}^{n} [y_i \log(\pi_i) - y_i \log(1-\pi_i) + \log(1-\pi_i)] \\
&= \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i) \right]
\end{aligned}
$$

che nel caso del modello binomiale diventa:

$$
\begin{aligned}
l(\beta) &= \sum_{i=1}^{n} \left[\log\binom{m_i}{y_i} + y_i \log(\pi_i) - y_i \log(1-\pi_i) + m_i \log(1-\pi_i) \right] \\
&= \sum_{i=1}^{n} \left[\log y_i \left(\frac{\pi_i}{1-\pi_i}\right) + m_i \log(1-\pi_i) \right]
\end{aligned}
$$

Stime dei parametri nel modello logit

Per il modello logit si ha $\log(\frac{\pi_i}{1-\pi_i}) = \boldsymbol{x}_i^T \beta$ e pertanto nel caso Bernoulliano:

$$l(\beta) = \sum_{i=1}^{n} [y_i \boldsymbol{x}_i^T \beta - \log(1 + \exp(\boldsymbol{x}_i^T \beta))] = \sum_{i=1}^{n} [y_i \eta_i - \log(1 + exp(\eta_i))].$$

Si può quindi calcolare la funzione score e porla pari a 0 per ottenere le equazioni di verosimiglianza per β, che risultano pari a

$$s(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} \boldsymbol{x}_i (y_i - \pi_i),$$

da cui le equazioni di verosimiglianza $s(\beta) = 0$. Si tratta di un sistema di equazioni non lineari la cui soluzione va cercata numericamente (vedremo però più avanti alcuni algoritmi adatti allo scopo).

Inferenza sui parametri

1. Per l'inferenza sui singoli parametri è utile ricordare le proprietà asintotiche delle stime di ML:
   per n elevato si ha $\hat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}(\boldsymbol{\beta}, I(\boldsymbol{\beta})^{-1})$, ove $I(\boldsymbol{\beta})$ è la matrice di informazione attesa che nel caso del modello Bernoulliano è pari a

$$I(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \pi_i (1 - \pi_i),$$

   ove $\pi_i = r(\boldsymbol{x}_i^T \boldsymbol{\beta})$.

2. Tale matrice dipende dai parametri incogniti in $\boldsymbol{\beta}$ ma una sua stima consistente si ottiene sostituendo a $\boldsymbol{\beta}$ la sua stima $\hat{\boldsymbol{\beta}}$.
3. L'elemento sulla diagonale $I(\hat{\boldsymbol{\beta}})^{-1}_{jj}$ è quindi una stima della varianza di $\hat{\beta}_j$.
4. Pertanto il rapporto $\frac{\hat{\beta}_j}{\sqrt{I(\hat{\boldsymbol{\beta}})^{-1}_{jj}}}$ relativo all'ipotesi nulla $H_0 : \beta_j = 0$ è asintoticamente distribuito come una Gaussiana standard se vera H_0.

Verifica dell'adeguatezza

- L'adeguatezza del modello nel caso di dati raggruppati in G gruppi può essere valutata confrontando il valore previsto della frequenza relativa per il gruppo i-esimo $\hat{\pi}_i$ con il valore osservato tramite la seguente statistica:

$$\sum_{i=1}^{G} \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)/m_i}$$

- Un'altra idea è quella di considerare la cosidetta *Devianza*, definita, per il caso dei dati raggruppati, come:

$$D = 2\sum_{i=1}^{G}[l_i(\tilde{\pi}_i) - l_i(\hat{\pi}_i)]$$

dove $l_i(\tilde{\pi}_i)$ e $l_i(\hat{\pi}_i)$ rappresentano rispettivamente (i) il valore della log-verosimiglianza in un modello con tanti parametri quanti sono i gruppi e (ii) la log-verosimiglianza del modello stimato. Si tratta di un criterio generale che approfondiremo successivamente. Entrambi gli indici sono asintoticamente (se m_i è grande per ciascuno dei G gruppi) distribuiti come una χ^2_{G-p}.

Infine, sono stati proposti indici che mimano R^2 e calcolabili anche per il modello bernoulliano, ad esempio:

$$R^2 = \frac{1 - (\hat{L}_0/\hat{L})^{2/n}}{1 - \hat{L}_0^{2/n}},$$

che varia tra 0 e 1, e $\hat{L}_0$ è la verosimiglianza nel modello nullo (con il solo parametro β_0)

Sovradispersione

Il problema della sovradispersione in un modello binomiale si applica solo al caso di dati raggruppati (e quindi non nel caso di modello Bernoulliano).

Il modello binomiale stimato prevede che la varianza nell'i-esimo gruppo sia pari a $\hat{\pi}_i(1-\hat{\pi}_i)/m_i$ ove $\hat{\pi}_i = r(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})$.

Accade spesso però che la varianza del fenomeno sia molto maggiore di quella prevista dal modello. Questo implica che i valori di $r_i = \frac{(p_i-\hat{\pi}_i)}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)/m_i}}$ siano spesso più grandi di quanto ci si aspetterebbe.

Tale fenomeno è noto come *sovradispersione* e implica che il modello binomiale non sia quello più adeguato a descrivere i dati.

Le cause del problema potrebbero essere la mancata indipendenza delle unità dovuta a fenomeni di clustering o la presenza di eterogeneità non osservata. Le possibili soluzioni implicano o l'utilizzo di metodi di stima che prescindano dal'assunzione binomiale o la specificazione di modelli per misture di binomiali. Tali temi saranno trattati più avanti con maggiore dettaglio.

Problemi di stima del modello in caso di perfetta separazione

- I valori delle stime di massima versosimiglianza per un modello binomiale sono in genere reperiti facilmente mediante efficienti algoritmi numerici
- Tuttavia vi possono esser problemi di convergenza se è possibile trovare una funzione delle covariate che separi perfettamente $y_i = 1$ e $y_i = 0$, o se per alcune categorie definite da una covariata osservo y solo 0 o solo 1.
- In tal caso la funzione di verosimiglianza non ha un massimo e le stime che si ottengono sono instabili.
- Il principale sintomo è dato quindi da un messaggio che dice "l'algoritmo non ha raggiunto la convergenza" e che "si sono ottenute previsioni di probabilità che sono numericamente pari a 1 o 0". Un altro sintomo è che i valori degli errori standard delle stime sono elevatissimi.
- Vi sono diverse soluzioni. Una possibile è quella che utilizza una penalizzazione della verosimiglianza.
- Tale soluzione si può ottenere considerando una verosimiglianza cui viene aggiunto un termine per eliminare la distorsione (ad esempio con il pacchetto `R brglm`)