Capitolo 6

Modelli Lineari Generalizzati

6.1 Introduzione

I *Modelli lineari generalizzati* rientrano in una metodologia di analisi statistica che include, come casi speciali, i modelli di regressione lineare e logistica. La regressione lineare prevede direttamente valori continui y in base a un *predittore lineare* $X\beta = \beta_0 + X_1\beta_1 + \cdots + X_k\beta_k$.

La regressione logistica prevede $\Pr(y = 1)$ per dati binari in base ad un predittore lineare con trasformazione logistica-inversa. Un modello lineare generalizzato coinvolge:

1. Un vettore di dati $y = (y_1, \ldots, y_n)$
2. Una matrice di predittori X e un vettore di coefficienti β, che formano un predittore lineare $X\beta$
3. *Una funzione link* g, che da luogo ad un vettore di dati trasformati $\hat{y} = g^{-1}(X\beta)$ che vengono usati per modellare i dati
4. Una distribuzione dei dati, $p(y|\hat{y})$
5. Possibilmente altri parametri, come varianze, parametri di *sovradispersione* e valori soglia *cutpoints*, relativi ai predittori, alla funzione link e alla distribuzione dei dati.

Le opzioni in un modello lineare generalizzato sono la trasformazione g e la distribuzione dei dati p.

- Nella *regressione lineare*, la trasformazione è la funzione identità (ovvero, $g(u) \equiv u$) e la distribuzione dei dati è una distribuzione normale, con deviazione standard σ stimata dai dati.

- Nella *regressione logistica*, la trasformazione è la logistica-inversa, $g^{-1}(u) = \text{logit}^{-1}(u)$ (si veda Figura 5.2a a pagina on 107) e la distribuzione dei dati è definita dalla probabilità per dati binari: $\Pr(y=1) = \hat{y}$.

Questo capitolo discute diverse altre classi di modelli generalizzati, che indichiamo nel seguito per opportuna convenienza: :

- Il modello di *Poisson* (Section 6.2) viene usato per dati categorici; ovvero quando ciascun dato puntuale y_i può essere uguale a 0, 1, 2, ... La trasformazione g che viene di solita considerata in questo modello è la trasformazione logaritmica, in questo modo $g(u) = \exp(u)$ trasforma un predittore lineare continuo $X_i\beta$ in valori positivi $\hat{y}_i$. La distribuzione dei dati è una Poisson.

  In genere è una buona idea aggiungere un parametro a questo modello al fine di catturare la *sovradispersione*, ovvero la variazione nei dati oltre quella che viene stimata dallo stesso modello di Poisson.

- Il modello *logistico-binomiale* (Section **??**) viene usato per tutti quei dati puntuali y_i che rappresentano il numero di successi in un numero n_i di prove. (Il numero n_i di prove per ciascun dato i, non è pari al numero di dati complessivo n.) In questo modello la trasformazione g è la logistica-inversa e la distribuzione dei dati la binomiale.

  Come nella regressione di Poisson, il modello binomiale viene notevolmente migliorato se si include un parametro per la *sovradispersione*.

- Il modello *probit* (Section **??**) coincide con il modello di regressione logistica a meno della funzione logistica che viene sostituita dalla funzione di distribuzione cumulato normale o, equivalentemente con la distribuzione normale invece della logistica negli errori dei dati latenti.

- Modelli logit e probit *Multinomiali* (Section 6.5) sono estensioni dei modelli di regressione logistica e probit per dati categorici con più di due categorie, per esempio risposte da indagini del tipo "Molto d'accordo", "D'accordo", "Indifferente", "Disaccordo", "Fortemente in disaccordo". Questi modelli usano la trasformazione logit o probit e la distribuzione multinomiale ma richiedono parametri addizionali per modellare le multiple possibilità dei dati.

  I modelli multinomiali sono ulteriormente classificati come *ordinati* (per esempio "Molto d'accordo", ..., "Fortemente in disaccordo") or *non ordinati* (per esempio Vaniglia, Cioccolato, Fragola, Altro).

- Modelli di regressione *Robusta* (Section 6.6) sono quelli in cui la distribuzione normale o logistica viene sostituita da altre distribuzioni [1] (di solito la famiglia di modelli che usa la distribuzione t di Student permette di modellare distribuzioni di dati caratterizzate dalla presenza di valori estremi.

Questo capitolo studia molti di questi modelli, con anche un esempio di modello di di regressione di Poisson sovradisperso nella Sezione in Section 6.2 e un esempio di modello di regressione logistico ordinato nella Sezione 6.5. Infine, nella Sezione **??** si discute della connessione tra i modelli lineari generalizzati e i modelli di scelta di comportamento usati in psicologia ed economia, usando come esempio la regressione logistica per la scelta dei pozzi in in Bangladesh. Questo capitolo non è da intendersi come una rassegna esaustiva dei modelli lineari generalizzati, ma piuttosto vuole fornire un'idea della della varietà dei modelli di regressione esistenti che potrebbero essere più o meno appropriati per diverse strutture di dati che abbiamo visto nelle applicazioni.

Stima dei modelli lineari generalizzati in R

A causa della varietà delle opzioni, i modelli lineari generalizzati sono in generale più complicati da stimare rispetto ai modelli di regressione lineare e logistica. Il punto di partenza in R è la funzione `glm()`, che abbiamo già usato in modo esteso per la stima del modello di regressione logistico nel Capitolo 5 e rappresenta una generalizzazione della funzione `lm()` per la stima dei modelli lineari. Possiamo usare `glm()` direttamente per la stima del modello logistico-binomiale, probit, e per il modello di regressione di Poisson e anche per correggere, quando necessario, per la sovradispersione. Modelli di regressione logistici e modelli probit ordinati possono essere stimati usando la funzione `polr()`, modelli probit non ordinati usando il pacchetto, sempre in R, `mnp`, e i modelli t possono essere stimati utilizzando il pacchetto, sempre in R, `hett`. (Si veda l'Appendice C per informazioni su questi e su altri pacchetti in R.) Al di là di questo, molti di questi modelli e altre generalizzazioni possono essere essere stimati in Bugs, come vedremo successivamente nell'ambito del contesto dei modelli multilevel.

[1] Nella letteratura statistica, i modelli lineari generalizzati sono stati definiti come i modelli che usano la famiglia esponenziale, un particolare classe di distribuzioni che include, per esempio, la distribuzione t di student. Per i nostri fini, comunque, noi usiamo il termine "modello lineare generalizzato" per tutti quei modelli che hanno un predittore lineare, una funzione link, una distribuzione dei dati senza restringersi alla famiglia esponenziale.

6.2 Regressione di Poisson, *exposure*, e sovradispersione

La distribuzione di Poisson viene usata per modellare la variazione in dati di conteggio (ovvero dati che sono del tipo $0, 1, 2, \ldots$). Dopo una breve introduzione, illustreremo nel dettaglio la regressione di Poisson con un esempio relativo alle perquisizioni effettuate dalla polizia che abbiamo introdotto nella Sezione **??** a New York City.

Incidenti di traffico

Nel modello di Poisson, ogni unità i corrisponde ad uno scenario (tipicamente un luogo o un intervallo temporale) in gli eventi y_i vengono osservati. Per esempio, i potrebbe rappresentare un incrocio in una città e y_i il numero di incidenti in quel particolare incrocio in un determinato anno.

Così come nella regressione lineare e logistica, la variazione di y può essere spiegata dai predittori lineari X. Nell'esempio degli incidenti causati dal traffico, questi predittori potrebbe includere: un termine costante, una misura della velocità media nei pressi dell'incrocio in questione e una variabile indicatrice relativa alla presenza o meno di un semaforo nell'incrocio.

Il modello di regressione di Poisson assume la seguente forma:

$$y_i \sim \text{Poisson}(\theta_i). \tag{6.1}$$

Il parametro θ_i deve essere positivo, in modo tale che sia ragionevole anche stimare un modello di regressione lineare su scale logaritmica:

$$\theta_i = \exp(X_i\beta) \tag{6.2}$$

.

Interpretazione dei coefficienti della regressione di Poisson

Se si considera l'esponente dei coefficienti β, questi possono essere trattati come effetti moltiplicativi. Per esempio, supponiamo che il modello relativo agli incidenti dovuti al traffico sia

$$y_i \sim \text{Poisson}(\exp(2.8 + 0.012X_{i1} - 0.20X_{i2})),$$

dove X_{i1} rappresenta la velocità media (espressa in miglia orarie, ovvero mph) nell strade intorno all'incrocio di riferimento e sia $X_{i2} = 1$ una variabile indicatrice relativa alla presenza o meno di un semaforo nell'incrocio.

Possiamo quindi interpretare i coefficienti come segue:

- Il termine costante rappresenta, come al solito, l'intercetta della regressione, ovvero il valore medio atteso della y quando $X_{i1} = 0$ e $X_{i2} = 0$. Dal momento che non avremo mai una velocità media pari a 0, non ha senso, in questo esempio, interpretare il termine costante.
- Il coefficiente di X_{i1} rappresenta la differenza attesa nella y (o su scala logaritmica) per ogni miglia oraria addizionale velocità. Quindi, l'incremento atteso moltiplicativo risulta pari a $e^{0.012} = 1.012$, ovvero una differenza positiva pari a 1.2% nel tasso di incidenti per mph. Dal momento che la velocità varia in decine di mph, potrebbe essere più ragionevole definire X_{i1} in termini di decine di mph, nel qual caso il coefficiente risulterebbe pari a 0.12, corrispondente ad un incremento del 12% nel tasso di incidenti per miglio orario.
- Il coefficiente di X_{i2} fornisce la differenza media attesa nel tasso di incidenti in seguito alla presenza di un semaforo. Questo valore risulta pari a $\exp(-0.20) = 0.82$ e quindi si ha una riduzione del 18% nel tasso di incidenti se nell'incrocio è presente un semaforo.

Come nel modello di regressione in generale, ciascun coefficiente può essere interpretato in termini di confronto facendo variare di un'unità un solo predittore e tenendo fissi gli altri. Questa interpretazione non è la migliore possibile soprattutto se si estende il modello a possibili scenari più complessi. Per esempio l'installazione di semafori in tutti gli incroci della città non necessariamente implicherebbe una riduzione del tasso di incidenti del 8%.

Regressione di Poisson con exposure

Nella regressione di Poisson, la variabile di conteggio può essere interpretata in relazione ad un termine di riferimento o *exposure*, per esempio, il numero di veicoli che attraversano l'incrocio in un determinato arco di tempo. Nel modello di regressione di Poisson, noi consideriamo y_i come il numero di eventi generati da un processo di Poisson con parametro θ_i e exposure u_i.

$$y_i \sim \text{Poisson}(u_i\theta_i), \tag{6.3}$$

dove, come prima, $\theta_i = \exp(X_i\beta)$. Il logaritmo dell'exposure, $\log(u_i)$, è chiamato *offset* nella terminologia dei modelli lineari generalizzati.

I coefficienti della regressione β riassumono la relazione esistente tra i predittori e θ_i che, nel nostro esempio, è pari al tasso di incidenti per veicolo.

L'inclusione del log(exposure) come predittore nella regressione di Poisson. Introdurre il logaritmo dell'exposure come *offset* nel modello (6.3), è come se si includesse in un modello di regressione un predittore il cui coefficiente è costantemente pari a 1. Si potrebbe anche includerlo come un normale predittore e stimare il suo coefficiente sulla base dei dati osservati. In alcune situazioni, la stima del coefficiente in base ai dati risulta migliore ma in altre situazione è preferibile includerlo come *offset* in modo tale che l'interpretazione della stima di θ sia più diretta e immediata.

Differenze tra il modello binomiale e il modello di Poisson

Il modello di Poisson è simile al modello al modello binomiale per dati di conteggio (si veda la Sezione **??**) ma viene applicato in situazioni leggermente differenti:

- Quando i valori osservati della variabile risposta y_i possono essere interpretati come il numero di "successi" su n_i prove, allora il modello standard a cui si fa riferimento è il modello binomiale/logistico (come descritto nella Sezione **??**) o la sua generalizzazione al caso in cui vi è sovradispersione.
- Quando invece i valori osservati della variabile risposta y_i non hanno un limite naturale—ovvero non si basano su un numero di prove indipendenti—allora il modello standard a cui si fa riferimento è il modello di Poisson ovvero di regressione logaritmico (come viene descritto in questa sezione) ovvero la sua generalizzazione nel caso in cui vi sia overdispersione.

Esempio: fermi della polizia in base a diversi gruppi etnici

Relativamente all'analisi dei fermi della polizia

- Le unità i sono i quartieri e i gruppi etnici ($i = 1, \ldots, n = 3 \times 75$).
- La variabile risposta y_i è il numero di fermi che vengono fatti in base ad un dato gruppo etnico in un dato quartiere.
- L'exposure u_i è il numero di arresti di persone appartenenti ad un dato gruppo etnico in un dato quartiere nell'anno precedente in base ai dati raccolti dal Dipartimento dei Servizi della Giustizia Criminale (DCJS).

- Gli input sono i quartieri e gli indici di etnicità.
- I predittori sono dati dalla costante, da 74 indicatori di quartiere (per esempio, i quartieri 2–75, con il quartiere 1 come punto di riferimento), e 2 indicatori di etnicità (per ispanici e bianchi, con i bianchi come punto di riferimento).

Illustreremo l'adattamento del modello in tre fasi. Inizialmente stimiamo un modello considerando solamente il logaritmo dell'exposure (*offset*) e un termine costante:

R output

```
glm(formula = stops ~ 1, family=poisson, offset=log(arrests))
            coef.est coef.se
(Intercept)     -3.4     0.0
  n = 225, k = 1
  residual deviance = 44877, null deviance = 44877
  (difference = 0)
```

Quindi, introduciamo gli indicatori di etnicità:

R output

```
glm(formula = stops ~ factor(eth), family=poisson,
    offset=log(arrests))
             coef.est coef.se
(Intercept)     -3.30    0.00
factor(eth)2     0.06    0.01
factor(eth)3    -0.18    0.01
  n = 225, k = 3
  residual deviance = 44133, null deviance = 44877
  (difference = 744.1)
```

I due indicatori di etnicità sono altamente significativi, ma la devianza del modello è diminuita di 744, più del doppio rispetto a quello che ci saremmo aspettati nel caso in cui l'etnicità non avesse avuto nessun potere esplicativo,.

Confrontando la categoria di riferimento 1 (neri), si vede che la categoria 2 (ispanici) presenta un 6% in più di fermi, e la categoria 3 (bianchi) ha circa il 18% in meno di fermi in meno sulla base delle probabilità di arresti riportate dal Dipartimento della Giustizia Criminale.

Aggiungiamo ora le 75 variabili relativi ai quartieri:

R output

```
glm(formula = stops ~ factor(eth) + factor(precinct),
    family=poisson, offset=log(arrests))
                  coef.est coef.se
(Intercept)          -4.03    0.05
factor(eth)2          0.00    0.01
factor(eth)3         -0.42    0.01
factor(precinct)2    -0.06    0.07
factor(precinct)3     0.54    0.06
. . .
factor(precinct)75    1.41    0.08
  n = 225, k = 77
  residual deviance = 2828.6, null deviance = 44877
  (difference = 42048.4)
  overdispersion parameter = 18.2
```

La diminuzione della devianza del modello da 44,000 a 2800 è enorme—decisamente superiore alla diminuzione di 74 che avremmo avuto nel caso in cui il fattore quartiere non avesse avuto nessun impatto. Dopo aver controllato per i quartieri i coefficienti di etnicità sono leggermente cambiati, i neri e gli ispanici (categoria 1 e 2) hanno circa la stessa probabilità di essere fermati, mentre questa probabilità risulta minore di ben 42% per quanto riguarda i bianchi (categoria 3), ovviamente in relazione alla *baseline* che è la probabilità di essere arrestati, come riportato dai dati dell'anno precedente forniti dal DCJS.[2]

Quindi, il fatto di aver controllato per i quartieri ha aumentato le disparità nelle probabilità di essere fermati dalla polizia tra i bianchi e le minoranze. Analizzeremo meglio questo aspetto nella Sezione **??**.

Consideriamo ora i coefficienti relativi ai quartieri —la probabilità di essere fermati dalla polizia dopo aver controllato per l'etnicità, è di circa il 6% in meno nel quartiere 2, $\exp(0.54) = 1.72$ volte più alta nel quartiere 3, ..., e $\exp(1.41) = 4.09$ volte più alta nel quartiere 75, se confrontato con il quartiere 1 che abbiamo preso come riferimento.

[2]Più precisamente il coefficiente esponenziale per i bianchi è $\exp(-0.42) = 0.66$, e quindi la loro probabilità di essere fermati è più bassa del 34%—l'approssimazione $\exp(-\beta) \approx 1-\beta$ risulta accurata solo quando β è vicina allo 0.

L'input exposure

Nell'esempio che stiamo considerando in questa sezione, i fermi effettuati dalla polizia sono messi in relazioni con in numero di arresti effettuati nell'anno precedente, quindi il coefficiente relativo all'indicatore per gli "ispanici" o per i "bianchi" risulta essere maggiore di 1 se le persone appartenenti a questi due gruppi etnici vengono fermate con probabilità superiore ai "neri", quando confrontati coi dati relativi all'anno precedente. In modo del tutto simile, i coefficienti relativi ai quartieri 2–75 risulteranno superiori a 1 per tutti quei quartieri in cui il tasso di arresti supera il tasso di arresti nel quartiere 1, quando li si confronta coi dati dell'anno precedente.

Quindi in questa analisi l'exposure è dato dal tasso di arresti dell'anno precedente all'analisi.

Sovradispersione

La regressione di Poisson non prevede un parametro indipendente per la varianza σ, e di conseguenza può essere sovradisperso. Questa situazione capita spesso nella pratica, come abbiamo già evidenziato a pagina 16 e come faremo vedere in seguito nell'ambito del contesto della regressione. Nell'ipotesi di distribuzione di Poisson la varianza coincide con la media—ovvero la deviazione standard coincide con la radice quadrata della media. Nel modello (6.3), $\mathrm{E}(y_i) = u_i\theta_i$ and $\mathrm{sd}(y_i) = \sqrt{u_i\theta_i}$. Definiamo i residui standardizzati come,

$$\begin{aligned} z_i &= \frac{y_i - \hat{y}_i}{\mathrm{sd}(\hat{y}_i)} \\ &= \frac{y_i - u_i\hat{\theta}_i}{\sqrt{u_i\hat{\theta}_i}}, \end{aligned} \tag{6.4}$$

dove $\hat{\theta}_i = e^{X_i\hat{\theta}}$. Nell'ipotesi in cui il modello di Poisson sia vero, allora i residui z_i dovrebbero essere approssimativamente indipendenti (non esattamente indipendenti dal momento che tutti i residui z_i sono stati calcolati con lo stesso valore di $\hat{\beta}$), ciascuno con media 0 e deviazione standard 1. Nel caso di sovradispersione, invece, ci si aspetta che i residui standardizzati z_i siano più grandi, in valori assoluto, riflettendo appunto la variazione extra presente nelle previsioni effettuate in base al modello di Poisson.

È possibile verificare la presenza di sovradispersione nella regressione classica di Poisson calcolando la somma dei quadrati degli n residui standardizzati, $\sum_{i=1}^{n} z_i^2$, e confrontando questo valore con la distribuzione χ^2_{n-k}, in quanto sotto l'ipotesi del modello, i residui dovrebbero seguire questa distribuzione (consideriamo $n-k$ gradi

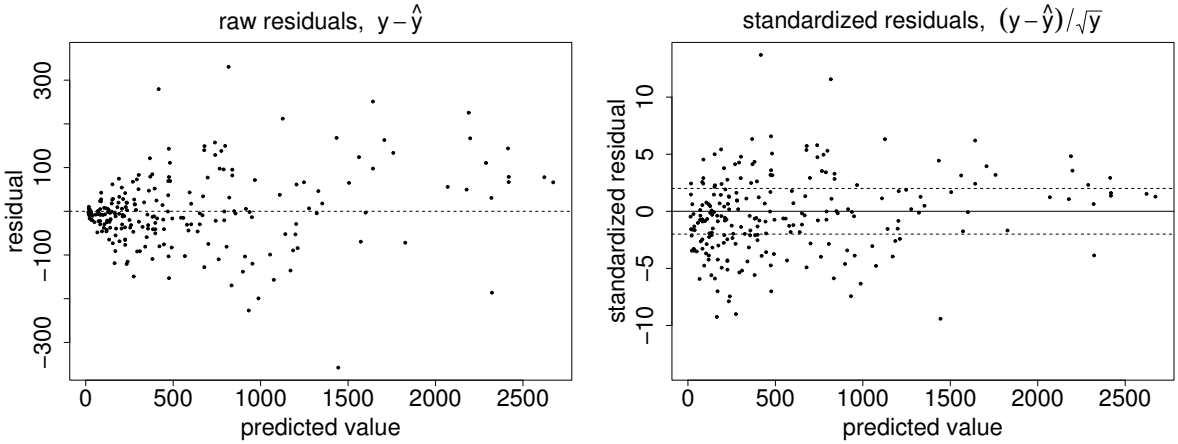

Figura 6.1: *Test per la sovradispersione in un modello di Poisson: (a) residui in funzione dei valori stimati dal modello, (b) residui standardizzati in funzione dei valori stimati. Coerentemente col modello, la varianza dei residui aumenta al crescere dei valori stimati. I residui standardizzati dovrebbero avere media 0 e deviazione standard 1 (quindi sul grafico sono rappresentate le bande di confidenza al livello del 95%). La varianza dei residui standardizzati risulta superiore a 1, indicando elevata sovradispersione.*

di libertà piuttosto che n per tener conto del fatto che abbiamo stimato k coefficienti di regressione).

La distribuzione χ^2_{n-k} ha un valor medio pari a $n-k$, e quindi il rapproto

$$\text{estimated overdispersion} = \frac{1}{n-k}\sum_{i=1}^{n} z_i^2, \tag{6.5}$$

risulta pari ad una misura di sintesi della sovradispersione nei dati se confrontati con i valori stimati dal modello.

La regressione classica di Poisson che stima i fermi effettuati dalla polizia si basa su $n = 225$ unità statistiche e $k = 77$ predittori lineari. La Figura 6.1 rappresenta i residui $y_i - \hat{y}_i$ e i residui standardizzati $z_i = (y_i - \hat{y}_i)/\text{sd}(\hat{y}_i)$, in funzione dei valori previsti dal modello di regressione di Poisson. Come ci si aspetta dal modello di Poisson, la varianza dei residui aumenta al crescere dei valori stimati, mentre la varianza dei residui standardizzati rimane approssimativamente costante. In ogni caso, i residui standardizzati hanno una varianza maggiore di 1, che indica una seria sovradispersione.

Per programmare un test per la sovradispersione in R si consideri:

R code
```
yhat <- predict (glm.police, type="response")
```

```
z <- (stops-yhat)/sqrt(yhat)
cat ("overdispersion ratio is ", sum(z^2)/(n-k), "\n")
cat ("p-value of overdispersion test is ", pchisq (sum(z^2), n-k), "\n")
```

La somma dei residui standardizzati è pari $\sum_{i=1}^{n} z_i^2 = 2700$, che deve essere confrontato con un valore atteso di $n - k = 148$. Il fattore stimato di sovradispersione è $2700/148 = 18.2$, con un associato p-value par a 1, quindi la probabilità che una variabile casuale estratta da una distribuzione χ^2_{148} assuma un valore pari a 2700, risulta pari a zero. In sintesi, i dati relativi ai fermi effettuati dalla polizia sono sovradispersi con un fattore di dispersione pari a 18, che è un valore grandissimo, sebbene anche un valore pari a 2 è da considerarsi elevato, ed è anche statisticamente significativo.

Adeguamento delle procedure di inferenza in caso di sovradispersione

In questo esempio, la correzione di base per tener conto della sovradispersione è moltiplicare tutti gli errori standard della regressione per $\sqrt{18.2} = 4.3$. Fortunatamente, l'inferenza sui parametri non risente troppo di questa correzione. Il parametro di primario interesse è α_3—il logaritmo della quota dei fermi effettuati sugli individui bianchi rispetto alle persone di colore —la cui stima era pari -0.42 ± 0.01 prima della correzione (si confronti l'output della regressione a pagina 161) e diventa pari a -0.42 ± 0.04 dopo la correzione. Tornando indietro alla scala originale, gli individui bianchi hanno una stima della quota dei fermi rispetto ai neri pari a 66%, con un relativo intervallo di confidenza al 50% di $e^{-0.42 \pm (2/3)0.04} = [0.64, 0.67]$ e un intervallo di confidenza al 95% pari a $e^{-0.42 \pm 2 \cdot 0.04} = [0.61, 0.71]$.

Stima di un modello di Poisson sovradisperso o un modello binomiale-negativo

Molto semplicemente, possiamo stimare un modello sovradisperso attraverso la famiglia `quasipoisson`:

R output

```
glm(formula = stops ~ factor(eth) + factor(precinct),
    family=quasipoisson, offset=log(arrests))
                 coef.est coef.se
(Intercept)          -4.03    0.21
factor(eth)2          0.00    0.03
```

```
factor(eth)3          -0.42    0.04
factor(precinct)2     -0.06    0.30
factor(precinct)3      0.54    0.24
. . .
factor(precinct)75     1.41    0.33
  n = 225, k = 77
  residual deviance = 2828.6, null deviance = 44877
  (difference = 42048.4)
  overdispersion parameter = 18.2
```

Possiamo scrivere questo modello come,

$$y_i \sim \text{overdispersed Poisson}\,(u_i \exp(X_i\beta),\, \omega),$$

dove ω è il parametro di sovradispersione (pari a 18.2 in questo modello). In modo più rigoroso, un modello di "Poisson overdisperso" non è un singolo modello ma piuttosto racchiude un insieme di modelli per dati di conteggio caratterizzati da una varianza dei dati ω volte superiore alla media e si riduce al modello di Poisson quando $\omega = 1$.

Un modello molto usato quando si presentano situazioni di questo tipo è il cosiddetto modello binomiale-negativo:

$$y_i \sim \text{Negative-binomial}\,(\text{mean} = u_i \exp(X_i\beta),\, \text{overdispersion} = \omega).$$

Purtroppo, la distribuzione binomiale-negativa non viene di solito espressa in termini della sua media e del parametro di sovradispersione ma piuttosto in termini di parametri a e b, e la media della distribuzione è pari a a/b e la sovradispersione è $1 + 1/b$.

6.3 Modello logistico-binomial

Nel capitolo 5 si è discusso della regressione logistica per dati binari (Si/No o 0/1). Il modello logistico può essere utilizzato anche per dati di conteggio, utilizzando la distribuzione binomiale (si veda 8) per modellare il numero di "successi" su un determinato numero di prove, e la probabilità di successo stimata attraverso un modello di regressione logistico.

Il modello binomiale per dati di conteggio, un'applicazione a dati relativi alla pena di morte

Illustriamo il modello logistico binomiale nel contesto di uno studio relativo alla proporzione di condanne alla pena di morte che sono state revocate in ciascuno dei 34 stati durante un intervallo di tempo di 23 anni, 1973–1995. L'unità di analisi sono i $34 \times 23 = 784$ stati-anno (in realtà noi abbiamo solo $n = 450$ nell'analisi dal momento che in alcuni stati la pena di morte è stata reintrodotta diverse volte a partire dal 1973). Per ciascun stato-anno i, indichiamo con n_i il numero di condanne a morte in quel determinato stato e anno e con y_i il numero di condanne che sono state revocate dalla Corte Suprema. Il nostro modello assume la forma,

$$\begin{aligned} y_i &\sim \text{Binomial}(n_i, p_i) \\ p_i &= \text{logit}^{-1}(X_i\beta), \end{aligned} \tag{6.6}$$

dove X è una matrice di predittori. Inizialmente, introduciamo nel modello

- Un termine costante
- 33 indicatori per gli stati
- Un trend temporale per gli anni (ovvero, una variabile che assume valore 1 per l'anno 1973, 2 per il 1974, 3 per il 1975, e così via).

Questo modello potrebbe anche essere scritto come,

$$\begin{aligned} y_{st} &\sim \text{Binomial}(n_{st}, p_{st}) \\ p_{st} &= \text{logit}^{-1}(\mu + \alpha_s + \beta t), \end{aligned}$$

con i pedici s per lo stati e t per il tempo (ovvero, anno-1972). In ogni caso, si preferisce in genere la forma (6.6) data la sua maggiore flessibilità, anche se è comunque utile riuscire a lavorare con entrambe le formulazioni.

Sovradispersione

Quando la regressione logistica viene utilizzata per modellare dati di conteggio, è possibile, e quasi sempre si verifica, che i dati presenti una variabilità superiore a quella spiegata dal modello. Il problema della sovradispersione deriva dal fatto che il modello di regressione logistico non ha un parametro di varianza σ.

Più specificatamente, se i dati y hanno una distribuzione binomiale con parametri n e p, allora la media di y è np e la sua deviazione standard $\sqrt{np(1-p)}$.

Come nel modello (6.4), definiamo i residui standardizzati per ciascun dato i come,

$$\begin{aligned} z_i &= \frac{y_i - \hat{y}_i}{\text{sd}(\hat{y}_i)} \\ &= \frac{y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}}, \end{aligned} \tag{6.7}$$

dove $p_i = \text{logit}^{-1}(X_i\hat{\beta})$. Se il modello binomiale è vero, allora gli z_i dovrebbero essere approssimativamente indipendenti con media 0 e deviazione standard 1.

Come nel modello di Poisson, possiamo calcolare la sovradispersione stimata $\frac{1}{n-k}\sum_{i=1}^{n} z_i^2$ (si confronti il modello (6.5) a pagina 164) e verificare formalmente la presenza di sovradispersione confrontando $\sum_{i=1}^{n} z_i^2$ con la distribuzione teorica del χ^2_{n-k}. (Il valore di n qui rappresenta il numero di dati e non ha nessuna relazione con n_i nei modelli (6.6) e (6.7) che si riferiscono al numero dei casi nello stato anno i.)

In pratica, la sovradispersione è quasi sempre presente nel modello logistico (o la regressione di Poisson, come discusso nella Sezione 6.2) quando applicato a dati di conteggio. Nelle famiglie di distribuzioni più generali conosciute come sovradisperse, la deviazione standard può avere la forma $\sqrt{\omega np(1-p)}$, dove $\omega > 1$ è noto *paramtero di sovradispersione*. Il modello sovradisperso si riduce al modello di regressione binomiale logistico quando $\omega = 1$.

Inferenza nel caso di sovradispersione

Esattamente come nel modello di regressione di Poisson, una semplice correzione che può essere fatta per tener conto della sovradispersione è quella di moltiplicare gli errori standard di tutte le stime dei coefficienti per la radice quadrata della stima della sovradipersione (6.5). Senza questa aggiustamento, gli intervalli di confidenza sarebbero troppo stretti, con un conseguente livello di fiducia "gonfiato".

Modelli di regressione binomiali sovradispersi possono essere stimati in R attraverso la funzione `glm()` con l'optione `quasibinomial(link=logit)` family. Una distribuzione corrispondente è la beta-binomiale.

Modelli per dati binari visti come caso particolare di modelli per dati di conteggio

Il modello di regressione logistico per dati binari come introdotto nel capitolo 5 può essere visto come un caso speciale della forma binomiale (6.6) con $n_i \equiv 1$ per tutti

gli i. Il problema della sovradispersione al livello dei dati individuali non capita mai nei modelli per dati binari, e questo è il motivo per cui non abbiamo introdotto il problema della sovradispersione nel Capitolo 5.

Modelli per dati di conteggio come caso caso particolare di modelli per dati binari

Al contrario, il modello binomiale (6.6) può essere espresso nella forma per dati binari (5.1) considerando ciascuno degli n_i casi come un dato separato. La dimensione campionaria di questa regressione estesa è pari a $\sum_i n_i$, e i dati sono del tipo 0 e 1: a ciascuna unità i corrisponde a un vettore di y_i valori uguali a 1 e $n_i - y_i$ valori pari a zero. Infine, la matrice X viene estesa in modo tale da avere $\sum_i n_i$ righe, dove invece della i^{ma} riga nella matrice originale X troviamo n_i righe identiche nella matrice estesa. In questa nuova parametrizzazione, la sovradispersione potrebbe essere inclusa in un modello gerarchico attraverso una variabile definita sui dati osservati (nell'esempio della pena di morte, la variabile assume i valori $1, \ldots, 450$) e attraverso l'inclusione di di un coefficiente variabile o un termine di errore a questo livello.

6.4 Regressione Probit: dati latenti normalmente distribuiti

Il modello *probit* è del tutto simile al modello logit, ma la funzione logistica deve essere sostituita con la distribuzione normale (si veda la Figura 5.5). Possiamo scrivere questo modello come,

$$\Pr(y_i = 1) = \Phi(X_i\beta),$$

dove Φ è la funzione di distribuzione cumulata della normale. Nella formulazione tipica dei dati latenti,

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{se } z_i > 0 \\ 0 & \text{se } z_i < 0 \end{cases} \\ z_i &= X_i\beta + \epsilon_i \\ \epsilon_i &\sim \mathrm{N}(0, 1), \end{aligned} \tag{6.8}$$

ovvero, una distribuzione normale per gli errori latenti con media 0 e deviazione standard 1.

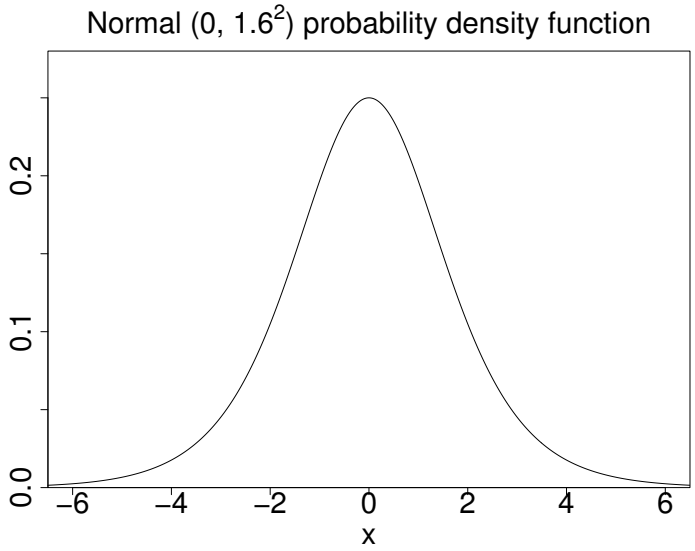

Figura 6.2: *Funzione di densità Normale con media 0 e deviazione standard 1. A fini pratici, risulta quasi impossibile distinguerla dalla densità logistica (Figura 5.5 a pagina 116). Quindi possiamo interpretare i coefficienti dei modelli probit in termini di coefficienti di una regressione lofistica divisi per una costante pari a 1.6.*

In termini più generali, il modello può includere un termine di varianza degli errori, in questo modo l'ultima linea del modello(6.8) diventa,

$$\epsilon_i \sim \mathrm{N}(0, \sigma^2),$$

ma in questo modo σ risulta non identificato, in quanto il modello è invariante se si moltiplica σ per una qualche costante c e il vettore dei coefficienti β per la stessa costante c. Quindi sono necessarie alcune restrizioni sui parametri, e l'approccio standard è quello di fissare $\sigma = 1$ pari ad 1 (6.8).

Probit o logit?

Come evidente nella Figura 6.2 (la si confronti con Figuta 5.5 a pagina 116), il modello probit è simile al modello logit con la deviazione standard dei residui ϵ pari ad 1 piuttosto che 1.6. Di conseguenza, i coefficienti nella regressione probit sono simili ai coefficienti della regressione logistica divisi per 1.6. Per esempio, questa è la versione probit del modello di regressione logistico che abbiamo stimato a pagina 120 relativo al problema della scelta dei pozzi in Bangladesh:

```
glm(formula = switch ~ dist100, family=binomial(link="probit"))
```
R output

```
             coef.est coef.se
(Intercept)     0.38    0.04
dist100        -0.39    0.06
  n = 3020, k = 2
  residual deviance = 4076.3, null deviance = 4118.1 (difference = 41.8)
```

Relativamente agli esempi che abbiamo visto, la scelta tra un modello probit o logit è essenzialmente un problema di preferenze o convenienza, per esempio legato all'interpretazione degli errori latenti normali dei modelli probit. Quando stimiamo i coefficienti di una regressione probit, è sufficiente moltiplicare questi coefficienti per 1.6 per ottenere gli equivalenti della regressione logistica. Per esempio, il modello che abbiamo appena stimato, $\Pr(y = 1) = \Phi(0.38 - 0.39x)$, è essenzialmente equivalente al modello logistico $\Pr(y = 1) = \text{logit}^{-1}(1.6(0.38 - 0.39x)) = \text{logit}^{-1}(0.61 - 0.62x))$, che è esattamente il modello stimato a pagina 120.

6.5 Variabili risposta categoriche ordinate e non ordinate

I modelli di regressione logistici e probit possono essere estesi a categorie multiple, che possono essere o meno ordinate. Esempi di variabili risposta categoriche ordinate sono Democratici, Indipendenti, Repubblicani; Si, Forse, no; Sempre, Frequentemente, Spesso, Raramente, Mai. Esempi di variabili risposta categoriche non ordinate sono Liberali, Laburisti, Conservatori; Calcio, Pallacanestro, Baseball, Hockey; Treno, Autobus, Automobile, Piedi; Bianco, Nero, Ispanico, Asiatico, altri. Analizzeremo prima di tutto le categorie ordinate in base ad un esempio, e quindi passeremo ad analizzare brevemente i modelli di regressione per variabili categoriche non ordinate.

Il modello logistico multinomiale

Consideriamo una variabile categorica y che assume i seguenti valori $1, 2, \ldots, K$. Il modello logistico ordinato può essere scritto in due modi equivalenti. Inizialmente esprimiamo questo modello in termini di regressioni logistiche:

$$\begin{aligned}
\Pr(y > 1) &= \text{logit}^{-1}(X\beta) \\
\Pr(y > 2) &= \text{logit}^{-1}(X\beta - c_2) \\
\Pr(y > 3) &= \text{logit}^{-1}(X\beta - c_3) \\
&\ldots \\
\Pr(y > K-1) &= \text{logit}^{-1}(X\beta - c_{K-1}). \qquad (6.9)
\end{aligned}$$

I parametri c_k (che sono chiamati valori soglia o *cutpoints*, per ragioni che spiegheremo tra breve) hanno il vincolo che devono essere ordinati in ordine crescente: $0 = c_1 < c_2 < \cdots < c_{K-1}$, in quanto le probabilità nel modello (6.9) sono strettamente decrescenti (assumendo che tutti i K outcomes abbiano una probabilità di verificarsi non nulla). Dal momento che c_1 è pari a 0, il modello con K categorie ha $K-2$ parametri liberi c_k oltre i β; in modo del tutto simile a quanto succede nella classica regressione logistica in cui $K=2$ per un solo valore di β da stimare.

I valori soglia $c_2, \ldots, c_{K-1}$ possono essere stimati col metodo di massima verosimiglianza, simultaneamente ai coefficienti β. Per alcuni insiemi di dati comunque i parametri potrebbero non essere identificati, come succede nella regressione logistica per dati binari (si veda la Sezione 5.8).

Sottraendo le espressioni in (6.9) l'una all'altra si ottengono le probabilità individuali della variabile risposta:

$$\begin{aligned} \Pr(y=k) &= \Pr(y>k-1) - \Pr(y>k) \\ &= \text{logit}^{-1}(X\beta - c_{k-1}) - \text{logit}^{-1}(X\beta - c_k). \end{aligned}$$

Interpretazione in termini di variabili latenti con valori soglia

Il modello logistico categorico risulta più semplice da capire se espresso in termini di variabili latenti (5.4), generalizzato al caso di K categorie:

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{se } z_i < 0 \\ 2 & \text{se } z_i \in (0, c_2) \\ 3 & \text{se } z_i \in (c_2, c_3) \\ & \cdots \\ K-1 & \text{se } z_i \in (c_{K-2}, c_{K-1}) \\ K & \text{if } z_i > c_{K-1} \end{cases} \\ z_i &= X_i\beta + \epsilon_i, \end{aligned} \tag{6.10}$$

con errori indipendenti ϵ_i che hanno una distribuzione logistica come in (5.4).

La figura 6.3 rappresenta il modello a variabili latent e mostra quanto la distanza tra due valori soglia adiacenti c_{k-1}, c_k influisca sulla probabilità che $y = k$. Possiamo anche vedere che quando il valore assunto dal predittore lineare $X\beta$ è elevato, y assumerà il suo valore più elevato, se invece il valore assunto dal predittore è basso, allora verosimilmente y assumerà il suo valore più basso.

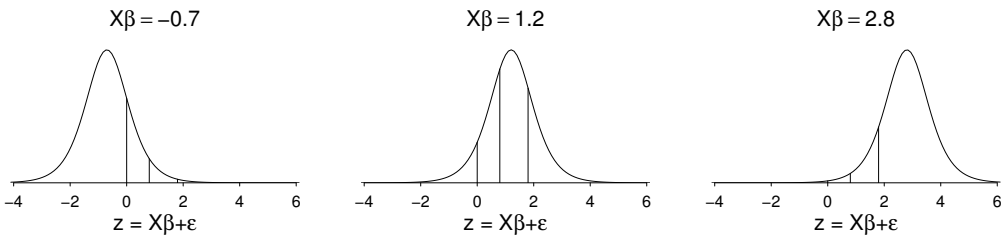

Figura 6.3: *Rappresentazione dei valori soglia in modello logistico categorico ordinato. In questo esempio, ci sono $K = 4$ categorie e i valori soglia sono $c_1 = 0, c_2 = 0.8, c_3 = 1.8$. I tre grafici mostrano la distribuzione della variabile outcome latente z in corrispondenza a tre diversi predittori lineari $X\beta$. Per ciascun valore assunto dai predittori, i valori soglia mostrano i valori di y pari a 1, 2, 3, or 4.*

Esempio: voti storable

Illustriamo l'analisi dei dati categorici in riferimento ad uno studio sperimentale economico relativo al problema dei "storable votes." Questo esempio è in qualche modo complicato, e mette in evidenza sia i vantaggi che i limiti del modello logistico ordinato. In questo esperimento, ad un insime di studenti viene chiesto di simulare alcune situazioni di voto. In ciascuna situazione, un insieme di k studenti deve votare su due argomenti e ogni giocatore ha a disposizione un totale di 4 voti. Relativamente al primo argomento, ciascun giocatore ha la possibilità di assegnare 1, 2, o 3 voti, lasciando i rimanenti voti al secondo argomento. La parte vincente di ciascun argomento è decisa dal voto di maggioranza, i giocatori della parte vincente guadagnano dei benefici positivi che si suppone siano estratti da una distribuzione uniforme $[1, 100]$.

Al fine di aumentare i loro benefici, i giocatori dovrebbero seguire una strategia che preveda di assegnare un numero di voti superiore per quegli argomenti in cui i benefici sono più alti. Questo esperimento viene fatto in modo tale che i giocatori sono informati sulla distribuzione dei possibili benefici ma sono a conoscenza dei potenziali benefici che possono ottenere da ciascun argomento solo poco prima di ciascun voto. Quindi, quando i giocatori scelgono come allocare i loro voti in base al primo argomento conoscono soltanto i potenziali benefici relativi alla prima votazione. In seguito vengono informati dei loro potenziali benefici relativamente al secondo voto, ma a questo punto la scelta è automatica dal momento che i giocatori devono allocare i voti rimanenti. Le strategie dei giocatori possono essere riassunte in termini di scelta dei voti iniziali, $y = 1$, 2, o 3, dati i loro potenziali benefici, x.

La Figura 6.4 rappresenta i risultati di 6 studenti tra i 100 che hanno partecipato all'esperimento, questi 6 sono stati scelti per rappresentare 6 diversi comportamenti di dati. Non è sorprendente notare che le strategie sono essenzialmente

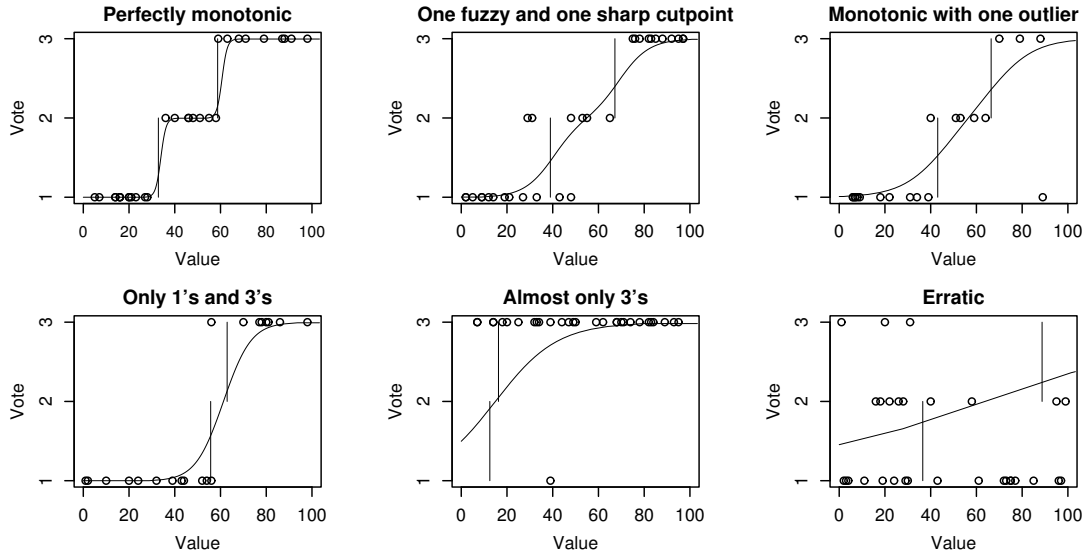

Figura 6.4: *Dati relativi ad alcuni individui coinvolti nello studio dei voti. Le linee verticale rappresentano i valori soglia stimati, mentre le curve rappresentano i valori attesi stimati in base al modello di regressione logistico ordinato. I due grafici a sinistra mostrano un buon adattamento, ma non si hanno risultati simili per i restanti grafici.*

monotoniche, ovvero gli studenti tendono ad allocare il maggior numero di voti laddove i potenziali benefici sono maggiori, ma è anche interessante scoprire diverse varietà di strategie monotoniche.

Come evidente nella Figura 6.4, la maggiorparte dei comportamenti degli individui può essere riassunta da tre parametri—il valore soglia tra 1 voto e 2, il valore soglia tra 2 e 3 e l'incertezza tra queste divisioni. I due valori soglia caratterizzano la strategia monotonica scelta, e la divisione marcata tra i valori soglia indica la coerenza con cui questa strategia è stata seguita.

Tre diverse parametrizzazioni del modello logistico ordinato. Potrebbe essere conveniente modellare la variabile risposta attraverso un modello logit ordinato, con una parametrizzazione leggermente differente rispetto a quella del modello (6.10) che meglio si presta a capire le strategie monotone. Il modello è

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{se } z_i < c_{1.5} \\ 2 & \text{se } z_i \in (c_{1.5}, c_{2.5}) \\ 3 & \text{se } z_i > c_{2.5} \end{cases} \\ z_i &\sim \text{logistic}(x_i, \sigma^2). \end{aligned} \tag{6.11}$$

In questo modello, i valori soglia $c_{1.5}$ e $c_{2.5}$ sono sulla scala dei dati x da 1–100, e la scala σ degli errori ϵ corrisponde all'incertezza dei valori soglia.

Questo modello ha lo stesso numero di parametri come nel caso della parametrizzazione convenzionale (6.10)—abbiamo due coefficienti di regressione in meno, mentre compaiono un ulteriore valore soglia e la varianza dell'errore. Quindi abbiamo il modello (6.10) con $K = 3$ categorie e un predittore x,

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{se } z_i < 0 \\ 2 & \text{se } z_i \in (0, c_2) \\ 3 & \text{se } z_i > c_2 \end{cases} \\ z_i &= \alpha + \beta x + \epsilon_i, \end{aligned} \tag{6.12}$$

con errori indipendenti $\epsilon_i \sim \text{logistic}(0, 1)$.

Ancora un'altra versione del modello che considera ancora due diversi valori soglia ma rimuove il termine costante, α; quindi,

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{se } z_i < c_{1|2} \\ 2 & \text{se } z_i \in (0, c_{2|3}) \\ 3 & \text{se } z_i > c_{2|3} \end{cases} \\ z_i &= \beta x + \epsilon_i, \end{aligned} \tag{6.13}$$

con errori indipendenti $\epsilon_i \sim \text{logistic}(0, 1)$.

I tre modelli sono di fatto equivalenti, con z_i/β nel modello (6.13) e $(z_i - \alpha)/\beta$ nel modello (6.12) che corrispondono a z_i nel modello (6.11) e i parametri legati dalle seguenti relazioni:

Modello (6.11)	Modello (6.12)	Modello (6.13)
$c_{1.5}$	$-\alpha/\beta$	$-c_{1\|2}/\beta$
$c_{2.5}$	$(c_2 - \alpha)/\beta$	$-c_{2\|3}/\beta$
σ	$1/\beta$	$1/\beta$

E' preferibile la parametrizzazione (6.11) in quanto possiamo interpretare in modo diretto i valori soglia $c_{1.5}$ e $c_{2.5}$ come valori soglia sulla stessa scala dell'input x, mentre σ corrisponde ad una misura di transizione nel passaggio da 1 a 2 e da 2 a 3. E' comunque conveniente stimare il modello considerando entrambe le parametrizzazioni standard (6.12) e (6.13), in modo tale da poter passare da un modello all'altro senza troppe complicazioni.

Stima del modello in R. Possiamo stimare i modelli logit (o probit) attraverso la funzione `bayespolr`, una funzione che si basa sulla funzione `polr` ("proportional

odds logistic regression") presente nella libreria `MASS` di R[3].

Illustriamo la stima del modello con riferimento ai dati dell'esperimento sul voto storable:

```
polr (factor(y) ~ x)
```
R code

which yields,

```
Coefficients:
         x
0.07911799

Intercepts:
     1|2      2|3
1.956285 4.049963
```
R output

In base all'output del modello è evidente che abbiamo stimato un modello del tipo (6.13), con stime pari a $\hat{\beta} = 0.079$, $\hat{c}_{1|2} = 1.96$ e $\hat{c}_{2|3} = 4.05$.

Se trasformiamo il modello (6.11) in base alla tabella relativa ai tre modelli otteniamo $\hat{c}_{1.5} = 1.96/0.079 = 24.8$, $\hat{c}_{2.5} = 4.03/0.079 = 51.3$, e $\hat{\sigma} = 1/0.079 = 12.7$.

Rappresentazione grafica del modello stimato. La Figura 6.4 mostra i valori soglia $c_{1.5}, c_{2.5}$ e il valore atteso dei voti $\mathrm{E}(y)$ visto in funzione della x, valori stimati in base ai dati degli studenti. In base al modello (6.11), i valori attesi dei voti possono essere scritti come,

$$\begin{aligned}
\mathrm{E}(y|x) &= 1 \cdot \Pr(y=1|x) + 2 \cdot \Pr(y=2|x) + 3 \cdot \Pr(y=3|x) \\
&= 1 \cdot \left(1 - \mathrm{logit}^{-1}\left(\frac{x - c_{1.5}}{\sigma}\right)\right) + \\
&\quad + 2 \cdot \left(\mathrm{logit}^{-1}\left(\frac{x - c_{1.5}}{\sigma}\right) - \mathrm{logit}^{-1}\left(\frac{x - c_{2.5}}{\sigma}\right)\right) + \\
&\quad + 3 \cdot \mathrm{logit}^{-1}\left(\frac{x - c_{2.5}}{\sigma}\right), \qquad (6.14)
\end{aligned}$$

dove $\mathrm{logit}^{-1}(x) = e^x/(1+e^x)$ rappresenta la curva logistica come nella Figura 5.2a a pagina 107. L'espressione (6.14) sembra molto complicata ma è in realtà semplice ~~da stimare se la si esprime in ter~~mini di una funzione in R:

[3]Rispetto a `polr` la funzione `bayespolr` considera delle informazioni a priori in modo tale che il modello possa essere stimato anche in presenza di problemi legati alla separabilità e non identificabilità.

R code
```
expected <- function (x, c1.5, c2.5, sigma){
  p1.5 <- invlogit ((x-c1.5)/sigma)
  p2.5 <- invlogit ((x-c2.5)/sigma)
  return ((1*(1-p1.5) + 2*(p1.5-p2.5) + 3*p2.5))
}
```

I dati, i valori soglia e le curve della Figura 6.4 possono essere quindi rappresentati graficamente utilizzando i seguenti comandi in R:

R code
```
plot (x, y, xlim=c(0,100), ylim=c(1,3), xlab="Value", ylab="Vote")
lines (rep (c1.5, 2), c(1,2))
lines (rep (c2.5, 2), c(2,3))
curve (expected (x, c1.5, c2.5, sigma), add=TRUE)
```

Una volta che abbiamo rappresentato graficamente le stime per questi individui, il prossimo passo sarà quello di studiare la distribuzione dei parametri nella popolazione, al fine di capire la variazione delle strategie che sono state utilizzate dagli studenti. In questo contesto, i dati hanno una struttura gerarchica–30 osservazioni per ciascun studente—e quindi torneremo su questo esempio nella Sezione relativa ai modelli gerarchici lineari generalizzati.

Having displayed these estimates for individuals, the next step is to study the distribution of the parameters in the population, to understand the range of strategies applied by the students. In this context, the data have a multilevel structure—30 observations for each of several students—and we pursue this example further in Section **??** in the chapter on multilevel generalized linear models.

Approcci alternativi per stimare dati categorici ordinati

I dati categorici ordinati possono essere stimati in modi diversi, tra questi ricordiamo:

- Attraverso un modello logistico ordinato con $K-1$ parametri di soglia, esattamente come abbiamo illustrato fino ad ora.
- Attraverso un modello del tutto simile ma nella sua forma probit.

- Attraverso una semplice regressione lineare (possibilmente preceduta da una semplice trasformazione della variabile risposta). Questa possibilità è indicata essenzialmente quando si è in presenza di una grande numero di categorie e queste possono essere considerate come equispaziate. E inoltre si presuppone che venga effettivamente considerato un ragionevole range di categorie. Per esempio, se le categorie sono definite su una scala da 1 a 10 ma in pratica sono sempre uguali a 9 0 10, allora, in questo caso il modello lineare non risulterà adatto.

- Diverse regressioni logistiche—ovvero un modello di regressione logistico per $y = 1$ *versus* $y = 2, \ldots, K$; quindi se $y \geq 2$, una regressione logistica per $y = 2$ *versus* $y = 3, \ldots, K$; e così via, se $y \geq K - 1$ per $y = K-1$ *versus* $y = K$. In modo del tutto simile se si decide di considerare un modello di tipo probit. Diverse regressioni logistiche (o probit) hanno come grande vantaggio quello di avere una maggiore flessibilità nello stimare i dati ma, come conseguente, lo svantaggio di perdere l'interpretazione dei valori soglia in termini di variabili latenti.

- Infine, attraverso regressioni di tipo robit che verranno discusse nella Sezione 6.6. Queste regressioni rispetto a quelle logistiche permettono di tenere in considerazione dati molto particolari come per esempio gli outliers nella parte in alto a destra della Figura 6.4.

Regressione per dati categorici non ordinati

Come abbiamo discusso all'inizio della Sezione 6.5, talvolta risulta più appropriato modellare outcome discreti come non ordinati. Un esempio di questo tipo si ha considerando il problema della scelta del pozzo in Bangladesh. Come descritto nella Sezione 5.4, le famiglie con pozzi non salubri avevano la possibilità di cambiare pozzo e di andare a rifornirsi presso un posso più salubre. In realtà le possibili alternative risultano più complicate e possono essere riassunte come segue: (0) non fare nulla, (1) spostarsi verso un pozzo privato già esistente, (2) spostarsi verso un pozzo già esistente ma pubblico, (3) costruire un nuovo pozzo. Se queste diverse possibilità vengono ricodificate come 0, 1, 2, 3, allora possiamo modellare $\Pr(y \geq 1), \Pr(y \geq 2|y \geq 1), \Pr(y = 3|y \geq 2)$. Sebbene queste quattro opzioni possano essere considerate in qualche modo ordinate, non ha nessun senso applicare il modello logit multinomiale ordinato o il modello probit, dal momento che sono diversi i fattori che verosimilmente influenzano le possibili decisioni. Piuttosto avrebbe più senso stimare separatamente un modello logit (o probit) per ciascuna delle componenti che influenzano la decisione: (a) vuoi cambiare pozzo o non fare nulla? (b) se decidi di cambiare pozzo, se orientato verso un pozzo già esistente oppure ne vuoi costruire

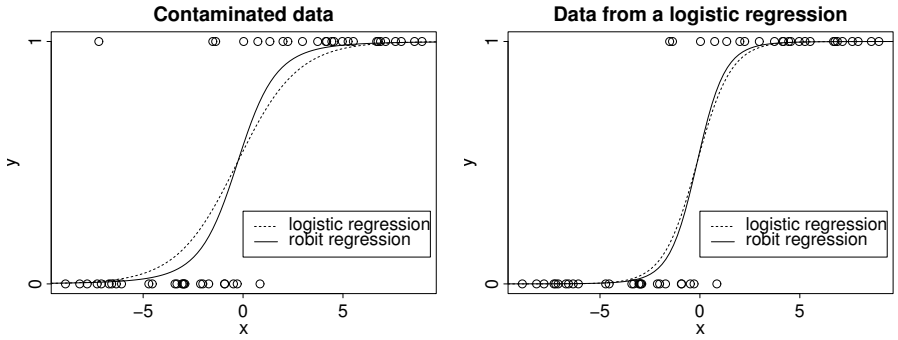

Figura 6.5: *Dati fittizi stimati attraverso un modello di regressione logistico: (a) un insieme di dati con un "outlier" (il valore inatteso $y = 1$ vicino alla parte superiore sinistra); (b) dati simulati da un modello di regressione logistico, senza outliers. In ciascun grafico, le linee tratteggiate e solide rappresentano rispettivamente i modelli di regressione logit e probit. In ciascun caso la linea di regressione robit è più inclinata— specialmente nel caso in cui si è in presenza di un outlier—in quanto questo modello da una minore importanza ai valori anomali.*

uno nuovo? (c) se decidi di utilizzare un pozzo già esistente, sei orientato verso un pozzo pubblico o privato? Relativamente a queste problematiche si veda la ricca bibliografia alla fine di questo capitolo.

6.6 Regressione robusta attraverso il modello t

La distribuzione t invece della Normale

Nel momento in cui un modello di regressione presenta occasionalmente degli errori troppo grandi, potrebbe essere più appropriato usare come distribuzione degli errori una distribuzione t di Student piuttosto che una normale. L'equazione di base del modello di regressione rimane invariata—$y = X\beta + \epsilon$—ma abbiamo una distribuzione diversa per modellare gli ϵ e un metodo leggermente diverso per stimare β (si utilizzano stime di massima verosimiglianza) e una distribuzione diversa per le previsioni. Stime di regressione ottenute con il modello t sono definite stime *robuste* in quanto le stime dei coefficienti sono meno influenzate da singoli outliers. Regressioni con errori t possono essere stimati attraverso la funzione `tlm()` presente nel pacchetto `hett` in R.

Robit invece di logit o probit

La regressione logistica, o equivalentemente la regressione probit, sono modelli flessibili e convenienti per modellare dati binari ma risentono della presenza di "outlier" o valori anomali. Gli "outlier" sono in genere considerati come osservazioni estreme, ma nel contesto di dati discreti un "outlier" rappresenta qualcosa di più di un valore *inatteso*. La Figura 6.5a illustra questo problema, attraverso dei dati simulati da un modello di regressione logistico, con un punto estremo che si è spostato da 0 a 1. Nel contesto del modello logistico, un osservazione $y = 1$ potenzialmente inverosimile dato il valore della x, potrebbe comunque verificarsi in dati reali. Quindi questo grafico rappresenta rappresenta una situazione in cui, sulla base dei dati, noi stimiamo un modello logistico nonostante il modello non sia esattamente appropriato.

Per avere ancora un esempio di regressione logistica con un valore del tutto anomalo, si faccia riferimento alla parte in alto a destra della Figura 6.4. Questo esempio ha tre "outcomes", ma per semplicità noi ci focalizziamo su "outcomes".

La regressione logistica può essere resa più robusta attraverso la generalizzazione della formulazione in termini di dati latenti (5.4):

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{se } z_i > 0 \\ 0 & \text{se } z_i < 0 \end{cases} \\ z_i &= X_i\beta + \epsilon_i, \end{aligned}$$

che fornisce agli errori latenti ϵ una t distribuzione del tipo:

$$\epsilon_i \sim t_\nu \left(0, \frac{\nu - 2}{\nu}\right), \tag{6.15}$$

con il parametro relativo ai gradi di libertà $\nu > 2$ stimato in base ai dati e la distribuzione t riscalata in modo tale che la sua deviazione standard sia pari a 1.

Il modello t utilizzato come distribuzione degli errori ϵ_i permette delle previsioni occasionali e inattese—un valore di z positivo in base an valore fortemente negativo del predittore lineare $X\beta$ o viceversa. La Figura 6.5a illustra meglio questa eventualità sulla base di un insieme di dati simulato ma in qualche modo "contaminato": la linea continua mostra la probabilità $\Pr(y = 1)$ in funzione delle x per il modello robit stimato, che risulta leggermente più inclinata della corrispondente ottenuta con un tradizionale modello di regressione logistico. La distribuzione t effettivamente riduce il peso del dato discordante in modo tale che il modello si adatti bene alla maggior parte dei dati.

La Figura 6.5b mostra cosa effettivamente accade con dati che provengono da un modello logistico: in questo caso il modello robit e logit sono ragionevolmente molto simili dal momento che non sono presenti incongruità.

Da un punto di vista matematico, il modello robit può essere considerato come una generalizzazione del modello probit e una generalizzazione approssimata del modello logit. Il modello robit tende al probit quando i gradi di libertà $\nu = \infty$, mentre il modello logit è molto vicino al modello robit quando $\nu = 7$.

6.7 Costruzione di modelli lineari generalizzati più complessi

I modelli di regressione che abbiamo finora considerato riescono a trattare parecchi problemi pratici. Nel caso di dati continui, inizialmente conviene stimare un modello di regressione lineare con errori distribuiti normalmente, considerando trasformazioni appropriate e interazioni come discusso nel Capitolo 4, quindi potrebbe essere conveniente utilizzare una distribuzione t di Student nel caso di dati che presentano occasionalmente errori molto grandi. Nel caso di dati binari possiamo utilizzare un modello logit o probit o, in caso un robit, ancora una volta trasformando le variabili di input e considerando degli opportuni grafici per i residui come discusso nel capitolo 5. Per dati categorici, il punto di partenza sono le distribuzioni binomiali e di Poisson, e nel caso di variabili discrete con più di due categorie possiamo stimare un modello logit multinomiale oppure una regressione di tipo probit. Nel seguito descriviamo brevemente alcune situazioni che possono verificarsi in pratica e che richiedono l'utilizzo di altri modelli diversi da quelli visti finora.

Dati misti, sia continui che discreti

I guadagni sono un esempio di variabile risposta che presenta aspetti sia continui che discreti. Nelle regressioni relative ai salari e alle altezze nel Capitolo 4, noi abbiamo analizzato i dati rimuovendo i dati relativi agli individui con guadagni pari a zero. In generale, è più appropriato modellare una variabile di questo tipo in due fasi: inizialmente attraverso un modello di regressione logistico per stimare la probabilità che la variabile y sia positiva: $\Pr(y > 0)$ e quindi un modello di regressione lineare sui valori positivi della variabile y considerando la trasformazione logaritmica, ovvero $\log(y)$, condizionatamente ai valori di $y > 0$. Come ragionevole conseguenza, anche le previsioni in un modello di questo tipo devono essere fatte in due fasi distinte.

Quando modelliamo una variabile risposta in diverse fasi, risulta necessario un ulteriore sforzo di programmazione per riportare i valori stimati dei parametri sulla scala originale. Per esempio, in un modello a due stadi per stimare i guadagni in funzione dell'altezza e del sesso, inizialmente, attraverso un modello di regressione logistico, stimiamo se i guadagni sono positivi o nulli:

R code
```
earn.pos <- ifelse (earnings>0, 1, 0)
fit.1a <- glm (earn.pos ~ height + male, family=binomial(link="logit"))
```

che fornisce il seguente output,

R output
```
            coef.est coef.se
(Intercept)    -3.85    2.07
height          0.08    0.03
male            1.70    0.32
  n = 1374, k = 3
  residual deviance = 988.3, null deviance = 1093.2 (difference = 104.9)
```

Quindi stimiamo un modello di regressione lineare sui logaritmi dei valori positivi del guadagno:

R code
```
log.earn <- log(earnings)
fit.1b <- lm (log.earn ~ height + male, subset = earnings>0)
```

che fornisce il seguente output,

R output
```
            coef.est coef.se
(Intercept)     8.12    0.60
height          0.02    0.01
male            0.42    0.07
  n = 1187, k = 3
  residual sd = 0.88, R-Squared = 0.09
```

Quindi, per esempio, una donna alta 66 pollici ha una probabilità di avere un guadagno non nullo pari a $\text{logit}^{-1}(-3.85 + 0.08 \cdot 66 + 1.70 \cdot 0) = 0.81$, ovvero pari all'81%. Se il suo guadagno non è nullo, il suo valore stimato è $\exp(8.12 + 0.02 \cdot 66 + 0.42 \cdot 0) = 12600$. La combinazione di queste due distribuzioni è una mistura che consiste in un picco in 0 e una distribuzione lognormale.

Modelli di dati latenti. Un altro modo di analizzare dati misti è mediante l'utilizzo di dati latenti, per esempio considerando un livello di reddito z_i "latente"—il reddito che la persona i avrebbe nel case in cui avesse un lavoro—che però viene osservato solo quando è positivo, $y_i > 0$. La *Tobit regression* appartiene a questa categoria di modelli ed è molto utilizzato in Econometria.

Scarafaggi e il modello di Poisson con inflazione di zeri *zero-inflated Poisson model*

I modelli binomiali e di Poisson, così come le loro generalizzazioni nel caso in cui i dati siano sovradispersi, possono essere espressi in termini di probabilità sottostante o in termini di tasso di manifestazione di un evento. Talvolta, comunque, il tasso sottostante ha esso stesso degli aspetti di discretezza. Consideriamo uno studio relativo ad una infestazione di scarafaggi all'interno di appartamenti di città. In ciascun appartamento i vengono sistemate delle trappole per diverse giorni. Definiamo con u_i il numero di trappole giornaliere e con y_i il numero di scarafaggi intrappolati. Partendo dall'idea di voler studiare l'infestazione degli scarafaggi in funzione di un certo numero di predittori X (tra cui il reddito, l'etnicità dei proprietari dell'appartamento, indicatori relativi al quartiere e altre misure di qualità dell'appartamento), iniziamo con lo stimare il modello,

$$y_i \sim \text{overdispersed Poisson}\,(u_i e^{X_i\beta},\, \omega). \tag{6.16}$$

E' comunque possibile, che nei dati sia presente un numero di zeri (nel caso in cui nell'appartamento i non sia stato catturato nessun scarafaggio$y_i = 0$) superiore a quello previsto dal modello.[4] Una spiegazione naturale è che alcuni appartamenti hanno effettivamente un tasso di presenza di scarafaggi pari a zero (o molto vicina allo zero), mentre altri hanno zero a causa della discretizzazione dei dati¿ Il modello *con inflazione di zeri* considera il modello (6.16) all'interno di un modello di mistura:

$$y_i \begin{cases} = 0, \text{ if } S_i = 0 \\ \sim \text{overdispersed Poisson}\,(u_i e^{X_i\beta},\, \omega), \text{ if } S_i = 1. \end{cases}$$

Qui, S_i è un indicatore che fornisce la presenza o meno di scarafaggi nell'appartamento i, e potrebbe essere modellato attraverso un modello di regressione logistico:

$$\Pr(S_i = 1) = \text{logit}^{-1}(X_i\gamma),$$

[4]In questo particolare esempio, il modello di Poisson sovradisperso ha fornito una buona stima in termini di zero; si veda pagina **??**. Ma in altre situazioni il modello con inflazioni di zeri fornisce delle prestazioni superiori in termini di stima e adattamento.

dove γ è nuovo insieme di coefficienti di regressione propri per questo modello. Stimare un modello di questo tipo diviso in due fasi non è una cosa banale—i valori di S_i non sono osservati e di conseguenza non si può stimare direttamente γ; inoltre non si può sapere quali siano le osservazioni pari a zero che corrispondono a $S_i = 0$ e quali invece corrispondono a risposte relative alla distribuzione di Poisson, e di conseguenza non possiamo stimare direttamente β. Sono state scritte alcune funzione in R per stimare modelli di questo tipo che possono essere, in ogni caso, stimati in Bugs.

ALtri modelli

I modelli lineari, logistici, di Poisson, misture di questi modelli, le loro generalizzazioni sovradisperse, robuste e multinomiali sono in grado di analizzare e studiare la maggior parte dei problemi pratici. Tuttavia, vale la pena di ricordare che esistono altre forme distribuzionali che possono essere usate quando si hanno forme particolari di dati, come la distribuzione esponenziale, la gamma, i modelli di Weibull per dati relativi a tempi di attesa, e modelli di rischio per dati di sopravvivenza. In generale, i modelli nonparametrici comprendono i modelli generalizzati additivi, le reti neurali e molti altri modelli che vanno oltre i modelli lineari generalizzati, e che sono stati sviluppati per studiare relazioni non necessariamente lineari tra gli input e i dati.

6.8 Modelli di scelta

Finora abbiamo considerato modelli di regressione per stimare una variabile risposta in funzione di alcune variabili di input. Un approccio completamente differente è quello di modellare le decisioni. Questo approccio è talvolta applicabile a dati di scelta come i dati relativi agli esempi dei Capitoli 5 and 6 sulla regressione logistica e sui modelli lineari generalizzati.

Faremo vedere questa idea attraverso l'esempio relativo alla scelta dei pozzi in Bangladesh (si veda la Sezione 5.4). Come possiamo capire la relazione che esiste tra la distanza dal pozzo, il livello di arsenico nel pozzo e la decisione di rifornirsi in un altro pozzo? E' ragionevole aspettarsi che persone i cui pozzi presentano elevati livelli di arsenico siano più propense a scegliere di cambiare pozzo per il rifornimento dell'acqua, ma che valori dovremmo attenderci per i coefficienti? La relazione dovrebbe essere su scala lineare o logaritmica? Il rischio effettivo per la salute degli individui dovrebbe essere misurato in funzione lineare dell'arsenico? A tutte queste domande dovremmo dare una risposta utilizzando un modello per decisioni individuali.

Per costruire un *modello di scelta* dobbiamo innanzitutto specificare una *funzione di valore*, che rappresenta l'importanza delle preferenze per una decisione rispetto a tutte le altre—in questo caso, la preferenza di cambiare pozzo per piuttosto che continuare ad utilizzare il pozzo vicino alla propria abitazione. La funzione di valore viene riscalata in modo tale che zero rappresenta indifferenza, uno la preferenza di cambiare pozzo e valori negativi rappresentano la preferenza di continuare ad utilizzare il pozzo precendente. Questo modello risulta quindi simile al modello di regressione logistico espresso in termini di dati latenti (si veda pag.115) e risulta essere, come faremo vedere tra breve, un caso particolare dei modelli di scelta.

Regressione logistica o probit visti come modelli di scelta a una dimensione

Alcuni semplici modelli di scelta a una dimensione si riducono a modelli di regressione logit o probit con un singolo predittore, come faremo vedere di seguito nell'ambito del modello di scelta di cambiare pozzo per il rifornimento dell'acqua in funzione della distanza dal pozzo sicuro più vicino. Come ripostato in pag. 120 il modello di regressione logistica risulta pari a:

R output

```
glm(formula = switch ~ dist100, family=binomial(link="logit"))
            coef.est coef.se
(Intercept)     0.61    0.06
dist100        -0.62    0.10
  n = 3020, k = 2
  residual deviance = 4076.2, null deviance = 4118.1 (difference = 41.9)
```

Ora pensiamo al problema in termini decisionali. Per ciascuna famiglia i, definiamo

- a_i = il beneficio che si ha spostandosi da un pozzo non sicuro a uno sicuro
- $b_i + c_i x_i$ = il costo di spostarsi verso un altro pozzo che dista x_i dal vecchio.

Consideriamo ora una teoria di utilità in cui il beneficio (corrispondente alla diminuzione del rischio di malattia) viene espresso nella stessa scala del costo di cambiare pozzo (l'inconveniente di non utilizzare più il proprio pozzo a cui va aggiunto un sforzo addizionale legato—e proporzionale alla distanza—al trasporto dell'acqua.

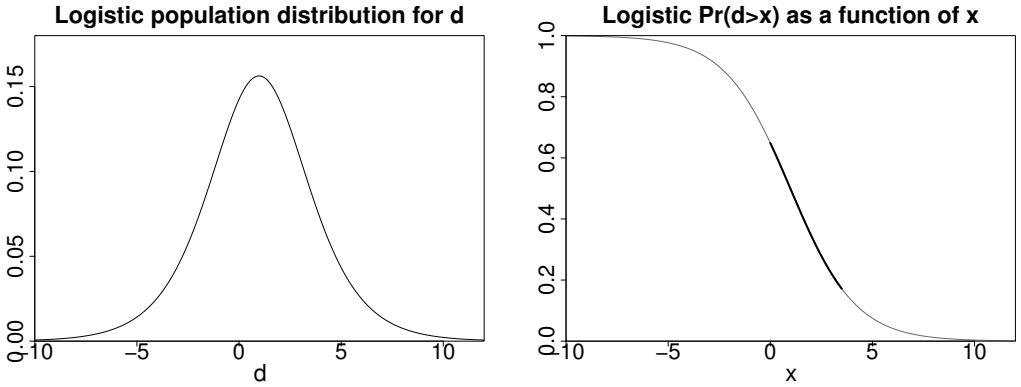

Figura 6.6: *(a) Distribuzione logistica teorica di $d_i = (a_i - b_i)/c_i$ nella popolazione e (b) corrisponde alla curva della regressione logistica della probabilità di cambiare pozzo in funzione della distanza. Entrambe queste curve corrispondono al modello* $\Pr(y_i = 1) = \Pr(d_i > x_i) = \mathit{logit}^{-1}(0.61 - 0.62x)$. *La parte scura della curva (b) corrisponde al range di x (distanza in 100 metri) nei dati relativi alla probabilità di cambiare pozzo, si veda Figura 5.9 a pag. 122.*

Modello Logit. In base al modello di utilità, la famiglia i si deciderà di rifornirsi in un altro pozzo sicuro se il beneficio che ne consegue è superiore al costo: $a_i > b_i + c_i x_i$. In ogni caso, non abbiamo una misura diretta degli a_i, b_i e c_i. Tutta l'informazione che possiamo avere dai dati si traduce essenzialmente nella probabilità di cambiare pozzo in funzione di x_i, ovvero

$$\Pr(\text{switch}) = \Pr(y_i = 1) = \Pr(a_i > b_i + c_i x_i), \tag{6.17}$$

considerando a_i, b_i, c_i come variabili casuali la cui distribuzione è fissata (ma ignota) in base ai valori dei parametri nella popolazione.

L'espressione (6.17) può essere scritta come,

$$\Pr(y_i = 1) = \Pr\left(\frac{a_i - b_i}{c_i} > x_i\right),$$

una nuova espressione che risulta particolarmente utile in quanto mette insieme tutte le variabili casuali ed evidenzia come la relazione tra y e x dipenda dalla distribuzione di $(a - b)/c$ nella popolazione.

Per convenienza, indichiamo con $d_i = (a_i - b_i)/c_i$ il beneficio netto di spostarsi verso un pozzo vicino diviso il costo associato alla distanza necessaria per raggiungere il nuovo pozzo. Se d_i ha una distribuzione logistica nella popolazione e se d è indipendente da x, allora $\Pr(y = 1)$ avrà la forma di una regressione logistica in funzione di x come risulta evidente dalla figura 6.6.

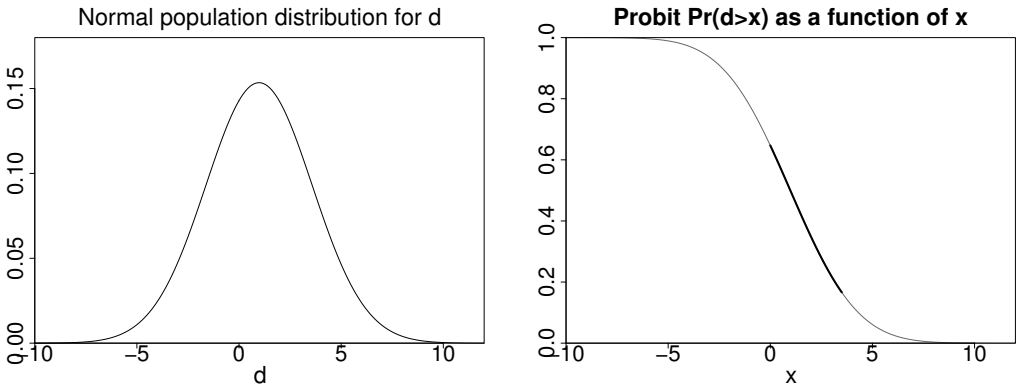

Figura 6.7: *(a) Distribuzione normale teorica di $d_i = (a_i - b_i)/c_i$ con media 0.98 e deviazione standard 2.6 (b) curva della regressione probit della probabilità di cambiare pozzo in funzione della distanza. Entrambe queste curve corrispondono al modello* $\Pr(y_i = 1) = \Pr(d_i > x_i) = \Phi(0.38 - 0.39x)$. *Si confronti con la Figura 6.6.*

Se d_i ha una distribuzione logistica centrata in μ e con parametro di scala σ, allora $d_i = \mu + \sigma\epsilon_i$, dove ϵ_i ha una densità logistica; si veda la Figura 5.2 a pagina 107. Quindi

$$\begin{aligned}\Pr(\text{switch}) \;=\; \Pr(d_i > x) \;&=\; \Pr\left(\frac{d_i - \mu}{\sigma} > \frac{x - \mu}{\sigma}\right) \\ &=\; \text{logit}^{-1}\left(\frac{\mu - x}{\sigma}\right) \;=\; \text{logit}^{-1}\left(\frac{\mu}{\sigma} - \frac{1}{\sigma}\,x\right),\end{aligned}$$

che è semplicemente una regressione logistica con coefficienti μ/σ and $-1/\sigma$. Possiamo quindi stimare una regressione logistica e risolvere per μ e σ. Per esempio nel modello relativo alla probabilità di cambiare pozzo, $\Pr(y = 1) = \text{logit}^{-1}(0.61 - 0.62x)$, corrisponde a $\mu/\sigma = 0.61$ e $-1/\sigma = -0.62$; quindi $\sigma = 1/0.62 = 1.6$ e $\mu = 0.61/0.62 = 0.98$. La Figura 6.6 mostra la distribuzione de d, insieme alla curva di $\Pr(d > x)$ vista come una funzione di x.

Modello Probit. Un modello simile si ottiene partendo da una distribuzione normale del parametro di utilità: $d \sim \text{N}(\mu, \sigma^2)$. In questo caso,

$$\begin{aligned}\Pr(\text{scambiare}) \;=\; \Pr(d_i > x) \;&=\; \Pr\left(\frac{d_i - \mu}{\sigma} > \frac{x - \mu}{\sigma}\right) \\ &=\; \Phi\left(\frac{\mu - x}{\sigma}\right) \;=\; \Phi\left(\frac{\mu}{\sigma} - \frac{1}{\sigma}x\right),\end{aligned}$$

che risulta essere semplicemente una regressione probit. Il modello $\Pr(y = 1) = \Phi(0.38 - 0.39x)$ corrisponde a $\mu/\sigma = 0.38$ e $-1/\sigma = -0.39$; quindi $\sigma = 1/0.39 = 2.6$

and $\mu = 0.38/0.39 = 0.98$. La Figura 6.7 rappresenta questo modello che è del tutto simile al modello logistico della Figura 6.6.

Modelli di scelta, regressione per dati discreti e dati latenti

Modelli di regressione logistica e modelli lineari generalizzati sono usualmente costruiti al fine di stimare le probabilità di diverse variabili risposta y in funzione dei predittori x. Un modello stimato rappresenta un'intera popolazione i cui "errori" entrano nel modello come probabilità che non sono semplicemente 0 o 1 (quindi la differenza tra i dati e le curve stimate appare come nella Figura 5.9 a pagina 122).

I modelli di scelta, invece sono definiti a livello degli individui, come abbiamo visto nell'esempio relativo alla probabilità di cambiare pozzo, dove ciascuna famiglia i ha, oltre gli usuali valori osservati X_i, y_i, altri parametri a_i, b_i, c_i che determinano la funzione di utilità per ciascuna famiglia e, di conseguenza, la decisione di rifornirsi in un altro pozzo.

Regressione logistica o probit vista come un modello di scelta di dimensioni multiple

Possiamo estendere il modello relativo alla probabilità di cambiare pozzo al caso di dimensioni multiple considerando il livello di arsenico presente nei pozzi come ulteriore fattore da che può essere considerato nella decisione di cambiare pozzo.

Siano:

- $a_i \cdot (\text{As})_i$ = il beneficio che si ottiene spostandosi da un pozzo con un livello di arsenico pari a As_i verso un pozzo sicuro e non inquinato. (Il fatto che i benefici siano proporzionali al livello di arsenico è del tutto ragionevole dal momento che il rischio è considerato proporzionale all'esposizione continuativa all'aresenico.)
- $b_i + c_i x_i$ = il costo di cambiare pozzo per il rifornimento dell'acqua che è distante x_i.

La famiglia i dovrebbe quindi decidere di cambiare pozzo se $a_i \cdot (\text{As})_i > b_i + cx_i$— decisione che dipende dal livello di arsenico $(\text{As})_i$ presente nel proprio pozzo, dalla distanza x_i verso il pozzo più vicino e dai parametri di utilità a_i, b_i, c_i.

La Figura 6.8 mostra lo spazio decisionale per una singola famiglia, in funzione del livello di arsenico presente nel proprio pozzo e la distanza dal dal pozzo sicuro più vicino. Dati a_i, b_i, c_i, la decisione in base a questo tipo di modello è deterministica.

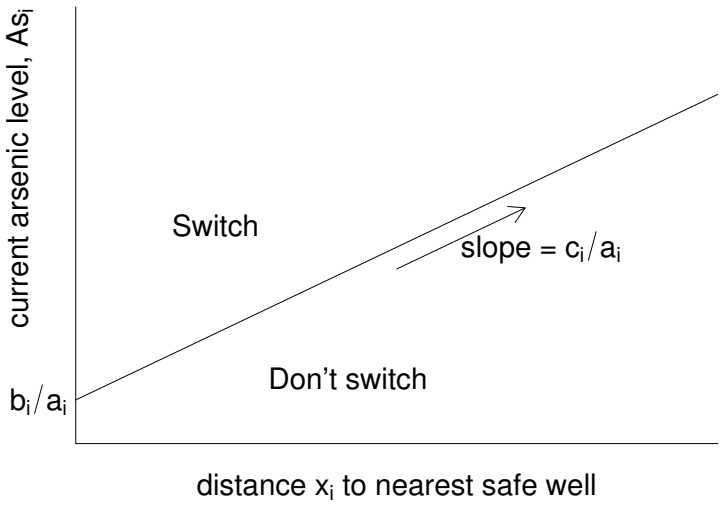

Figura 6.8: *Eventuali possibili decisioni di cambiare pozzo in base al livello di arsenico presente nel pozzo e in base alla distanza dal pozzo sicuro più vicino basate sulla regola decisionale: cambio pozzo se* $a_i \cdot (\text{As})_i > b_i + cx_i$.

In ogni caso, a_i, b_i, c_i non sono direttamente osservabili—quello che noi in realtà osserviamo sono le decisioni ($y_i = 0$ o 1) prese dalle singole famiglie dati i livelli di arsenico As_i e le distanze x_i dal pozzo sicuro più vicino.

Se si considerano particolari distribuzioni per i parametri (a, b, c) si ritorna al modello di regressione logistico, per esempio nel caso in cui a_i e c_i sono costanti e b_i/a_i ha una distribuzione logistica indipendente da $(\text{As})_i$ e x_i. In generale, i modelli di scelta si riconducono alla regressione logistica quando si introducono i fattori in termini additivi, i coefficienti non variano nella popolazione e se è presente un costo fisso (b_i nell'esempio) che ha una distribuzione logistica nella popolazione.

Sono ovviamente possibili altre distribuzioni per (a, b, c), e di conseguenza possiamo stimare i modelli corrispondenti, trattando questi parametri di utilità in termini di dati latenti. Non risulta facile stimare questi modelli in R attraverso la funzione `glm()` (ad eccezione dei casi in cui si ritorna ai modelli logit o probit) ma possono essere stimati in Bugs.

Alcuni intuizioni in base ai modelli di scelta

Un modello di scelta potrebbe fornisci alcune intuizioni circa il problema che stiamo analizzando, anche se formalmente non stimiamo il modello. Per esempio, nella stima del modello logistico abbiamo trovato che la distanza risulta essere un buon