Capitolo 5

Regressione logistica

La regressione logistica è il modello standard che viene usato per modellare variabili binarie. La Sezione 5.1 introduce la regressione logistica in un semplice esempio con un solo predittore, nel resto del capitolo lavoreremo su un esempio che considera predittori multipli e interazioni.

5.1 Regressione logistica con un solo predittore

Esempio: modellizzazione delle preferenze politiche in funzione del reddito

I partiti conservatori in genere ricevono più consenso dagli elettori a più alto reddito. Illustriamo la regressione logistica con una semplice analisi che studia questo fenomeno utilizzando i dati dell'indagine Nazionale Americana sugli Studi Elettorali (National Election Study) nel 1992. Per ciascun rispondente i intervistato nel sondaggio, definiamo $y_i = 1$ se il rispondente preferisce George Bush (il candidato Repubblicano per presidente) o 0 se preferisce Bill Clinton (il candidato democratico), escludendo per ora i rispondenti che preferiscono Ross Perot o altri candidati e quelli che dichiarano di non avere alcuna opinione. Stimiamo le preferenze per i candidati in funzione del livello del reddito dei rispondenti, misurato su una scala a 5 punti.[1]

I dati, sotto forma di punti sottoposti ad una procedura di *jittered*, sono rappresentati nella Figura 5.1 insieme alla linea di *regressione logistica* stimata, vincolata ad

[1]Si veda la Sezione 4.7 per ulteriori dettagli sulle categorie in cui è stato suddiviso il reddito e sulle altre variabili misurate in questa indagine.

appartenere all'intervallo [0, 1]. Questa curva viene interpretata come la probabilità che $y = 1$ dato x,—ovvero usando una terminologia matematica $\Pr(y = 1|x)$.

Stimiamo e rappresentiamo graficamente la regressione logistica usando la seguente funzione in R:

R code

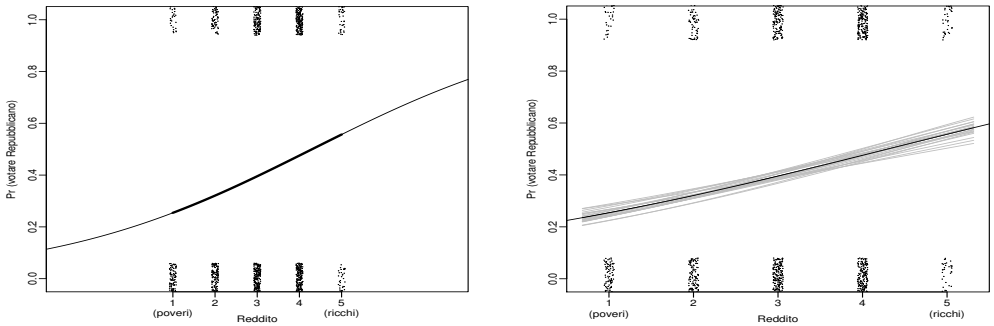
```
fit.1 <- glm (vote ~ income, family=binomial(link="logit"))
display (fit.1)
```

e otteniamo

R output

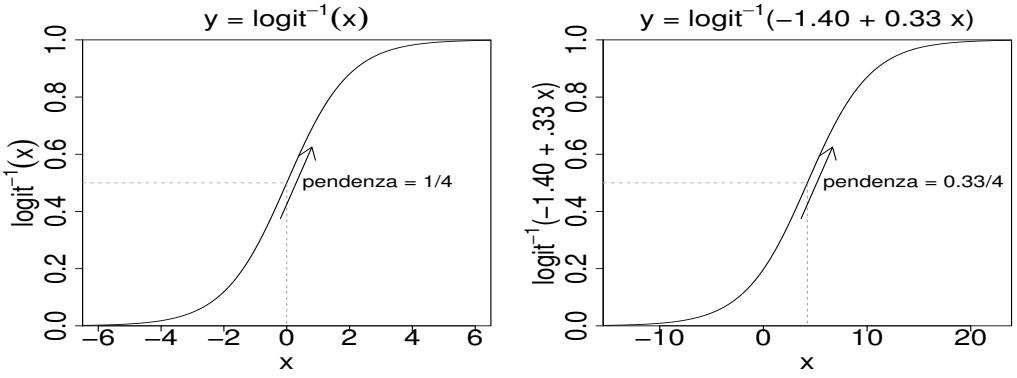
```
            coef.est coef.se
(Intercept)    -1.40    0.19
income          0.33    0.06
  n = 1179, k = 2
  residual deviance = 1556.9, null deviance = 1591.2 (difference = 34.3)
```

Figura 5.1: *Regressione logistica che stima la probabilità di votare a favore di George Bush nelle elezioni presidenziali del 1992, in funzione del livello di reddito reso discreto. I dati dell'indagine sono rappresentati da punti a cui è stata effettuata una procedura di* jittered. *In questo esempio molto poco è stato evidenziato dai punti con relativo* jitter*, ma qui si vuole evidenziare che sia i dati sia il modello stimato si trovano entrambi su una stessa scala. (a) Regressione logistica stimata: la linea più scura indica indica la curva nell'intervallo in cui variano i dati; le linee più sottili verso gli estremi del grafico mostrano come la curva logistica si avvicini ai limiti 0 e 1. (b) Nell'intervallo di variazione dei dati, la linea solida mostra la migliore stima della regressione logistica e le linee chiare l'incertezza intorno alla stima.*



Il modello stimato è quindi $Pr(y_i = 1) = \text{logit}^{-1}(-1.40 + 0.33 \cdot \texttt{reddito})$. Definiamo ora questo modello in termini matematici e quindi ritorneremo a discutere sulla sua interpretazione.

Figura 5.2: *(a) Funzione logistica inversa (inverse-logit function) $\text{logit}^{-1}(x)$: la trasformazione che viene usata nel modello di regressione logistica per passare dai predittori lineari alle probabilità. (b) Un esempio di probabilità stimate a partire da un modello di regressione logistica: $y = \text{logit}^{-1}(-1.40 + 0.33x)$. La forma della curva è la stessa, ma la sua posizione e la scala sono cambiate; si confronti gli assi delle x in entrambi i grafici. Per ciascuna curva, le linee tratteggiate indicano dove la probabilità stimata è pari a 0.5: nel grafico (a) questo accade quando $\text{logit}^{-1}(0.5) = 0$; nel grafico (b) il punto intermedio si ha dove $-1.40 + 0.33x = 0$, ovvero dove $x = 1.40/0.33 = 4.2$.*
La pendenza delle due curve nel punto intermedio è il coefficiente della regressione logistica diviso per 4, ovvero 1/4 per $y = \text{logit}^{-1}(x)$ e $0.33/4$ per $y = \text{logit}^{-1}(-1.40 + 0.33x)$. In questo punto intermedio, la pendenza della curva di regressione logistica è la più ripida.



Il modello di regressione logistico

Potrebbe non avere molto senso stimare un modello di regressione lineare continuo $X\beta + \texttt{errore}$ quando i dati y assumono solo valori pari a 0 e 1. Invece, modelliamo la probabilità che $y = 1$,

$$\Pr(y_i = 1) = \text{logit}^{-1}(X_i\beta), \tag{5.1}$$

sotto l'ipotesi che gli outcome y_i siano indipendenti date queste probabilità. Ci riferiamo a $X\beta$ come al predittore lineare (*linear predictor*).

La funzione $\text{logit}^{-1}(x) = \frac{e^x}{1+e^x}$ trasforma valori continui in valori che stanno nell'intervallo $[0, 1]$, il che risulta necessario dal momento che le probabilità devono essere comprese tra 0 e 1. Questo è illustrato nella Figura 5.1 e, in modo più teorico, nella Figura 5.2. In maniera del tutto simile, il modello (5.1) può essere scritto come

$$\begin{aligned} \Pr(y_i = 1) &= p_i \\ \text{logit}(p_i) &= X_i\beta, \end{aligned} \tag{5.2}$$

dove $\text{logit}(x) = \log(x/(1-x))$ è una funzione che *mappa* l'intervallo $(0, 1)$ sull'intervallo $(-\infty, \infty)$. Si preferisce lavorare con il logit^{-1} dal momento che è naturale

focalizzare l'attenzione sulla trasformazione dei valori continui in probabilità piuttosto che il viceversa. In ogni caso, è necessario capire la formulazione (5.2) per seguire la letteratura ed essere in grado di stimare i modelli logistici in Bugs.

La funzione logistica inversa è curva, di conseguenza la differenza attesa nelle y corrispondente ad una data differenza nelle x non è costante. Come si vede dalla Figura 5.2, il cambiamento più ripido nella pendenza della curva si ha nel mezzo della curva. Per esempio:

- $\text{logit}(0.5) = 0$, e $\text{logit}(0.6) = 0.4$. In questo caso, l'aggiunta di 0.4 su scala logistica corrisponde ad un cambiamento che va dal 50% al 60% sulla scala delle probabilità.
- $\text{logit}(0.9) = 2.2$, e $\text{logit}(0.93) = 2.6$. In questo caso, l'aggiunta di 0.4 su scala logistica corrisponde ad un cambiamento che va dal 90% al 93% sulla scala delle probabilità.

In maniera del tutto simile, l'aggiunta di 0.4 all'estremo inferiore della scala fa variare la probabilità dal 7% al 10%. In generale, ogni cambiamento specifico su scala logistica risulta molto contenuto agli estremi della scala delle probabilità, il che è necessario per poter mantenere le probabilità comprese tra 0 e 1.

5.2 Interpretazione dei coefficienti della regressione logistica

L'interpretazione dei coefficienti nella regressione logistica può essere impegnativa a causa della non linearità. Proveremo a generalizzare la procedura per cercare di interpretare i coefficienti uno alla volta, come abbiamo già fatto per la regressione lineare nel Capitolo 3.1.

Iniziamo col modello $\Pr(\text{Bush support}) = \text{logit}^{-1}(-1.40 + 0.33 \cdot \text{income})$.

In Figura 5.1 il modello è rappresentato graficamente, ma noi vorremmo anche delle sintesi numeriche.

Presentiamo ora un approccio molto semplice, per poi ritornare su questo punto nella Sezione 5.7, dove verranno illustrate sintesi numeriche maggiormente rappresentative.

Valutazione intorno alla media dei dati

La curva della funzione logistica ci impone in qualche modo di scegliere dove valutare i cambiamenti, quando li vogliamo interpretare sulla scala delle probabilità. La media delle variabili di input è spesso un utile punto di partenza.

- Come nella regressione lineare, l'*intercetta* può essere interpretata solo quando i valori assunti dagli altri predittori vengono posti uguali a zero. Quando zero è un valore che non ha senso o addirittura non è neanche presente nel modello (come nell'esempio del voto, in cui il reddito è definito su una scala da 1 a 5), allora la valutazione va effettuata in altri punti maggiormente significativi.

  Per esempio, possiamo valutare Pr(Bush support) nella categoria centrale del reddito e troviamo: $\text{logit}^{-1}(-1.40 + 0.33 \cdot 3) = 0.40$.

  Possiamo valutare Pr(Bush support) nel punto medio del reddito dei rispondenti: $\text{logit}^{-1}(-1.40 + 0.33 \cdot \bar{x})$; che in *R* viene codificato come segue:[2]

  ```
  invlogit (-1.40 + 0.33*mean(income))
  ```
  R code

  o, più genericamente,

  ```
  invlogit (coef(fit.1)[1] + coef(fit.1)[2]*mean(income))
  ```
  R code

  Relativamente a questo dataset, $\bar{x} = 3.1$, che fornisce una stima della probabilità Pr(Bush support) = 0.40 in questo punto centrale.

- Una differenza di 1 nel reddito (su una scala da 1 a 5) corrisponde ad una differenza positiva di 0.33 in termini di probabilità logistica di sostenere Bush. Ci sono due modi possibili per riassumere questo risultato in termini di probabilità.

  – Si può valutare come varia la probabilità quando x varia di una unità intorno al suo valor medio. Poiché in questo esempio, $\bar{x} = 3.1$ possiamo valutare la funzione di regressione logistica in $x = 3$ e $x = 2$; la differenza nella $\Pr(y = 1)$ corrispondente all'incremento unitario nella x è $\text{logit}^{-1}(-1.40 + 0.33 \cdot 3) - \text{logit}^{-1}(-1.40 + 0.33 \cdot 2) = 0.08$. Una differenza di 1 nelle categorie del reddito corrisponde a una differenza positiva pari all'8% nella probabilità di sostenere Bush.

[2] Usiamo una funzione che abbiamo scritto, `invlogit <- function (x) {1/(1+exp(-x))}`.

- Piuttosto che considerare una variazione discreta nella x, possiamo calcolare la derivata della curva logistica nel suo punto medio, in questo caso in $\bar{x} = 3.1$. Differenziando la funzione $\text{logit}^{-1}(\alpha + \beta x)$ rispetto a x otteniamo $\beta e^{\alpha+\beta x}/(1 + e^{\alpha+\beta x})^2$. Il valore del predittore lineare calcolato nel punto medio $\bar{x} = 3.1$ è $-1.40 + 0.33 \cdot 3.1 = -0.39$, e la pendenza della curva—ovvero il "cambiamento" nella $\Pr(y = 1)$ in corrispondenza di "piccoli" cambiamenti nella x—in questo punto risulta pari a $0.33e^{-0.39}/(1 + e^{-0.39})^2 =$ 0.079
- Relativamente a questo esempio, la differenza su scala probabilistica fornisce un valore pari a [redacted] consideriamo una sola cifra decimale, è equivalente a quello relativo al caso di differenziazione discreta; questo è in generale normale, ma in alcuni casi quando la differenza unitaria è elevata, la differenziazione discreta e la derivazione potrebbero fornire risposte leggermente differenti. Saranno, in ogni caso, sempre dello stesso segno.

La regola della "divisione per 4"

La curva logistica risulta maggiormente inclinata intorno al punto intermedio (*cfr* definizione in Figura 5.2 a pagina 107), ovvero nel punto in cui $\alpha + \beta x = 0$ così che $\text{logit}^{-1}(\alpha + \beta x) = 0.5$ (si veda la Figura 5.2). La pendenza della curva—la derivata della funzione logistica—è massimizzata in questo punto e raggiunge il valore $\beta e^0/(1 + e^0)^2 = \beta/4$. Quindi, $\beta/4$ è la massima differenza nella $\Pr(y = 1)$ corrispondente ad una differenza unitaria nella x.

Come regola pratica, possiamo prendere i coefficienti della regressione logistica (oltre che il termine costante) e dividerli per 4 al fine di ottenere un limite superiore nella differenza predittiva corrispondente a una differenza unitaria nella x. Questo limite superiore è una ragionevole approssimazione intorno al punto centrale della curva logistica, laddove le probabilità sono vicino a 0.5. Per esempio, nel modello $\Pr(\text{Bush support}) = \text{logit}^{-1}(-1.40 + 0.33 \cdot \text{income})$, possiamo dividere 0.33/4 ottenendo 0.08: una differenza di 1 nelle categorie del reddito corrisponderà ad una differenza attesa nella probabilità di sostenere Bush che sarà al massimo pari all'8%. Dal momento che i dati in questo esempio si trovano intorno al punto intermedio (si veda la Figura 5.1), l'approssimazione della "divisione per 4" fornisce un valore vicino a 0.13, il valore della derivata valutata nel punto medio della x.

Interpretazione dei coefficienti in termini di rapporto tra odds

Un altro modo di interpretare i coefficienti della regressione logistica è quello di considerarli in termini di *rapporti tra odds*. Se due eventi hanno probabilità di verificarsi rispettivamente pari a $(p, 1–p)$, il rapporto $p/(1–p)$ è solitamente chiamato *odds*. Una odds pari a 1 è equivalente ad una probabilità di 0.5—ovvero eventi equiprobabili. Odds pari a 0.5 o 2.0 equivalgono a probabilità di $(1/3, 2/3)$. Il rapporto tra due odds—ovvero $(p_1/(1-p_1))/(p_2/(1-p_2))$—è solitamente chiamato *odds ratio*. Quindi, un rapporto di odds pari a 2 corrisponde a una variazione da $p = 0.33$ a $p = 0.5$, o a un cambiamento da $p = 0.5$ a $p = 0.67$.

Il vantaggio di lavorare con i rapporti tra odds (piuttosto che con le probabilità) risiede nella possibilità di poter riscalare i rapporti tra odds senza raggiungere i punti estremi di 0 e 1. Per esempio, passando da una odds di 2 a una di 4 la probabilità aumenta passando da 2/3 a 4/5; raddoppiando ancora le odds la probabilità aumenta fino a 8/9 e così via.

I coefficienti della regressione logistica elevati all'esponente possono essere interpretati come rapporti tra odds. Per semplicità, illustriamo un modello con un solo predittore, del tipo

$$\log\left(\frac{\Pr(y=1|x)}{\Pr(y=0|x)}\right) = \alpha + \beta x. \tag{5.3}$$

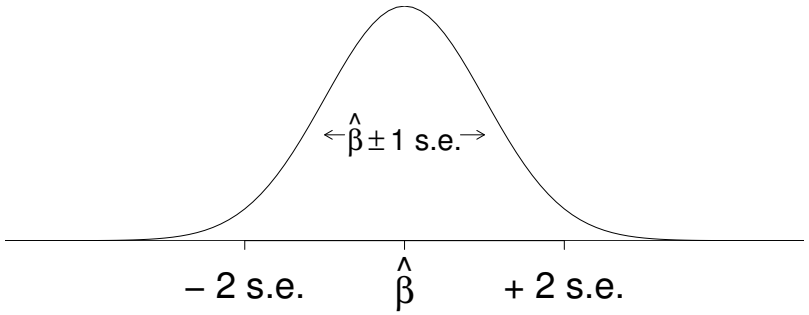
L'aggiunta di 1 alla x (ovvero, cambiando x in $x+1$ nell'equazione (5.3)) ha l'effetto di aggiungere una quantità pari a β in entrambi i lati dell'equazione. Passando all'esponenziale in entrambi i lati dell'equazione, le odds risultano moltiplicate per e^{β}. Per esempio, se $\beta = 0.2$, allora una differenza unitaria nella x corrisponde a una variazione moltiplicativa pari a $e^{0.2} = 1.22$ nelle odds (per esempio, un cambiamento nelle odds da 1 a 1.22, o un cambiamento di p da 0.5 a 0.55).

Riteniamo che il concetto di odds sia in qualche modo difficile da capire, e un rapporto di odds ancora più oscuro. Quindi preferiamo interpretare i coefficienti sulla scala originale dei dati quando possibile. Per esempio, un aggiunta di 0.2 su scala logistica corrisponde ad un cambiamento nella probabilità da $\text{logit}^{-1}(0)$ a $\text{logit}^{-1}(0.2)$.

Inferenza

Stime dei coefficienti ed errori standard. I coefficienti della regressione logistica sono stimati con il metodo della massima verosimiglianza, una procedura che

Figura 5.3: *Distribuzione che rappresenta l'incertezza nelle stime dei coefficienti di regressione (come nella pagina 45). L'intervallo di variazione di questa distribuzione corrisponde ai possibili valori di β consistenti con i dati. Quando si usa questa distribuzione dell'incertezza, si assegna una probabilità pari a circa il 68% che il coefficiente β si discosti dal valore puntuale stimato, $\hat{\beta}$, di una volta $\pm$ il suo errore standard e una probabilità pari a circa il 95% che β stia all'interno di due volte $\pm$ il suo errore standard. Assumendo che il modello di regressione sia corretto, potrebbe accadere che solo per il 5% delle volte il valore stimato $\hat{\beta}$ cada al di fuori dell'intervallo del vero valore $\beta \pm$ due volte il suo errore standard.*



funziona bene anche per stimare modelli su grandi campioni con un numero limitato di predittori (ma si veda 5.8 per potenziali problemi).

Come nel modello lineare, gli errori standard rappresentano l'incertezza delle stime. Possiamo dire, anche se in modo approssimativo, che le stime dei coefficienti che giacciono all'interno dell'intervallo di $\hat{\beta} \pm$ due volte il suo errore standard sono stime consistenti con i dati. La Figura 5.3 mostra la distribuzione normale che rappresenta in modo approssimativo il range dei possibili valori di β. Relativamente all'esempio del voto, il coefficiente del reddito ha una stima $\hat{\beta}$ di 0.33 e un errore standard di 0.06; quindi i dati sono approssimativamente consistenti con i valori di β che cadono nell'intervallo $[0.33 \pm 2 \cdot 0.06] = [0.21, 0.45]$.

Significatività statistica. Come nella regressione lineare, un coefficiente è considerato "statisticamente significativo" se è distante da zero meno di 2 volte l'errore standard. Nell'esempio del voto, il coefficiente del reddito è statisticamente significativo e positivo, il che significa che possiamo sostenere, anche se non con assoluta certezza, che, all'interno della popolazione rappresentata da questa indagine, differenze positive di reddito corrispondono a differenze positive (non negative) nella probabilità di sostenere Bush come presidente.

Come nella regressione lineare, in genere *non* proviamo ad interpretare la significatività statistica dell'intercetta. Il segno dell'intercetta in genere non è di grande interesse, e quindi è usualmente di scarsa rilevanza confrontare il valore dell'intercetta con lo zero e preoccuparci se questa sia significativamente diverso da zero.

Infine, quando consideriamo predittori multipli, seguiremo gli stessi principi che abbiamo seguito nella regressione lineare, come discusso nella Sezione 4.6, quando si deve decidere se includere o meno alcuni input nel modello o se risulta necessario includere una combinazione di input.

Previsioni. Le previsioni della regressione logistica sono probabilità, quindi per ciascun dato puntuale da stimare $\tilde{y}_i$, verrà stimata una probabilità predittiva,

$$\tilde{p}_i = \Pr(\tilde{y}_i = 1) = \text{logit}^{-1}(\tilde{X}_i\beta),$$

piuttosto che una previsione puntuale. Per esempio, per un votante non presente nell'indagine che ha un livello di reddito pari a 5 (si ricordi la scala a 5 punti nelle Figura 5.1), la *probabilità* stimata di sostenere Bush risulta pari a $\Pr(\tilde{y}_i = 1) = \text{logit}^{-1}(-1.40 + 0.33 \cdot 5) = 0.55$. Si noti che non stiamo dicendo che la nostra previsione per l'*outcome* è 0.55, dal momento che l'outcome $\tilde{y}_i$—sostenere o meno Bush—è o 0 o 1.

Stima e rappresentazione del modello in R

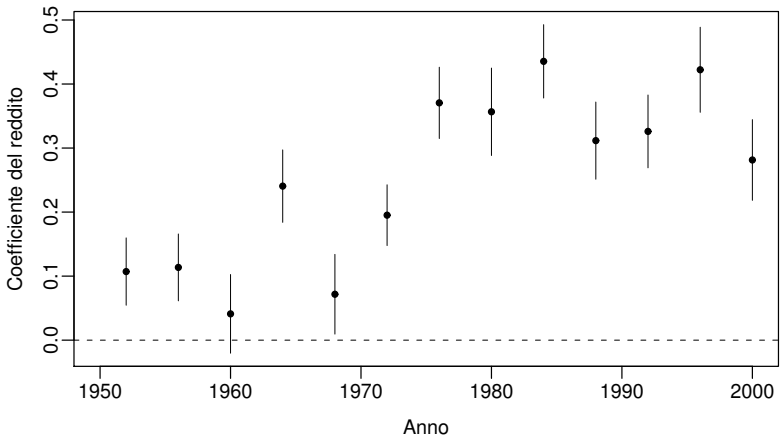
Dopo aver stimato la regressione logistica usando la funzione `glm` (si veda pagina 106), possiamo rappresentare graficamente i dati e la linea stimata (di veda Figura 5.1a) come segue:

R code

```
plot (income, vote)
curve (invlogit (coef(fit.1)[1] + coef(fit.1)[2]*x), add=TRUE)
```

(Il codice in R che usiamo per fare la figura richiede più passi, come la rappresentazione delle etichette sugli assi, la procedura di *jitter* sui punti, aggiustare lo spessore delle linee e così via.) La Figura 5.1b presenta linee tratteggiate che rappresentano l'incertezza dei coefficienti; queste linee possono essere rappresentate aggiungendo i seguenti comandi in R:

Figura 5.4: *Coefficienti del reddito (su una scala da 1 a 5) con relativi limiti di* ± 1 *volta l'errore standard nella regressione logistica che stima le preferenze per un presidente repubblicano, stimati separatamente in base a diverse indagini nella seconda metà del ventesimo secolo. I dati usati nelle stime del 1992 sono quelli della Figura 5.1.*



R code

```
sim.1 <- sim (fit.1)
for (j in 1:10){
  curve (invlogit (sim.1$beta[j,1] + sim.1$beta[j,2]*x),
    col="gray", lwd=.5, add=TRUE)}
```

Mostreremo nel Capitolo 7 l'utilizzo della funzione `sim`.

Rappresentazione dei risultati di diverse regressioni logistiche

Possiamo rappresentare graficamente una serie di regressioni logistiche in un singolo grafico, esattamente come già fatto nella Sezione 4.7 per i coefficienti della regressione lineare. La Figura 5.4 mostra le stime dei coefficienti con le relative incertezza misurate da ± 1 volta l'errore standard del coefficiente del reddito relative allo studio sulle preferenze presidenziali, stimate in base ai dati delle proiezioni pre-elettorali effettuate dagli Studi Nazionali Elettorali (National Election Studies) dal 1952 al 2000. Livelli elevati di reddito sono stati associati in maniera coerente con l'appoggio al partito Repubblicano e questa relazione è diventata più forte con il passare degli anni.

5.3 Formulazione secondo dati-latenti

È possibile interpretare un modello di regressione logistico sia direttamente—come un modello non lineare che stima la probabilità di ottenere un "successo" o una risposta positiva del tipo "si" dati alcuni predittori—sia indirettamente attraverso l'utilizzo di quelle che vengono chiamate variabili non osservate o *latenti*. In questa formulazione, ogni variabile risposta discreta y_i viene associata ad un outcome continuo e non osservato z_i, definito come segue:

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{se } z_i > 0 \\ 0 & \text{se } z_i < 0 \end{cases} \\ z_i &= X_i\beta + \epsilon_i, \end{aligned} \tag{5.4}$$

con errori indipendenti ϵ_i che si distribuiscono secondo una distribuzione di probabilità *logistica*.

La distribuzione logistica è mostrata nella Figura 5.5 ed è definita come

$$\Pr(\epsilon_i < x) = \text{logit}^{-1}(x) \text{ per ogni } x.$$

Quindi, $\Pr(y_i = 1) = \Pr(z_i > 0) = \Pr(\epsilon_i > -X_i\beta) = \text{logit}^{-1}(X_i\beta)$, e quindi i modelli (5.1) e (5.4) sono modelli equivalenti.

La Figura 5.6 mostra il modello per una osservazione i avente livello di reddito $x_i = 1$ (ovvero una persona con livello di reddito più basso), e il cui predittore lineare, $X_i\beta$, è il valore pari a $-1.40+0.33\cdot 1 = -1.07$. La curva mostra la distribuzione della variabile latente z_i, e l'area ombreggiata corrisponde alla probabilità che $z_i > 0$, e quindi $y_i = 1$. In questo esempio, $\Pr(y_i = 1) = \text{logit}^{-1}(-1.07) = 0.26$.

Interpretazione della variabili latenti

Le variabili latenti oltre a poter considerate come un trucco computazionale possono anche essere interpretate in modo sostantivo. Consideriamo, ad esempio, l'indagine pre elettorale, $y_i = 1$ per coloro che sostengono Bush e $y = 0$ per coloro che sostengono Clinton. La variabile continua non osservata z_i può essere interpretata come "utilità" del rispondente o la preferenza per Bush in confronto alla preferenza a Clinton: il segno dell'utilità ci dice quale sia il candidato preferito e il suo ordine di grandezza rivela l'importanza della preferenza.

Solo il segno di z_i, ma non la sua grandezza, può essere determinata direttamente dai dati binari. In ogni caso, possiamo conoscere molto di più sulle variabili z_i analizzando i predittori della regressione. Talvolta, in alcune formulazioni è disponibile direttamente l'informazione diretta sulle z_i; per esempio in un'indagine si

Figura 5.5: *Funzione di densità di probabilità logistica, usata per il termine di errore nella formulazione (5.4) del modello logistico in termini di dati latenti. La curva logistica nella Figura 5.2a è la funzione di densità cumulata di questa distribuzione. Il massimo valore di questa densità si ha in 0.25, che corrisponde al valore massimo della pendenza pari a 0.25 nella funzione logit inversa della Figura 5.2a.*

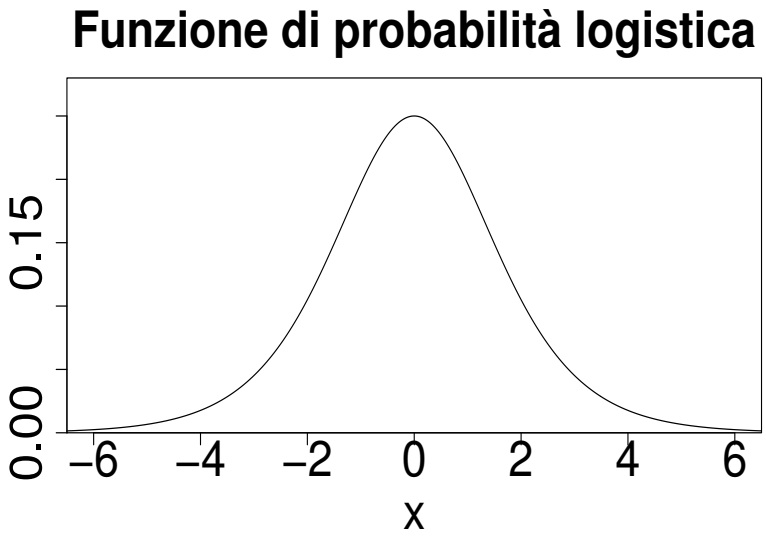
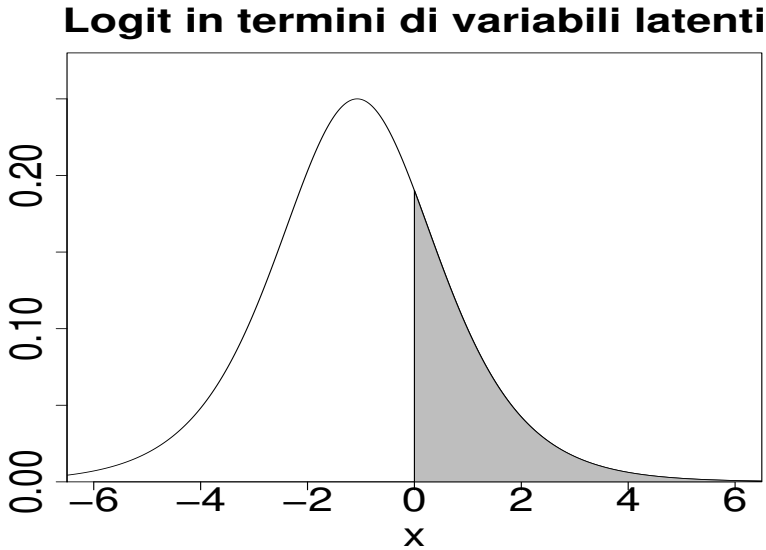


Figura 5.6: *Funzione di densità di probabilità della variabile latente z_i nel modello (5.4) se il predittore lineare, $X_i\beta$, assume il valore -1.07. L'area ombreggiata indica la probabilità che $z_i > 0$, ovvero la probabilità che $y_i = 1$ nella regressione logistica.*



potrebbe porre una domanda del tipo "termometro di gradimento", ovvero "riporta i tuoi sentimenti su George Bush su una scala da 1 a 10, dove 1 rappresenta il punteggio più negativo e 10 il più positivo".

Non-identificabilità del parametro di varianza latente

La funzione di densità di probabilità nella Figura 5.5 sembra essere a forma campanulare, molto simile alla funzione di densità normale tradizionalmente usata per gli errori nella regressione lineare. Infatti, la distribuzione logistica è molto vicina alla distribuzione normale con media 0 e deviazione standard 1.6—identità che discuteremo in modo più approfondito nel seguito nel contesto della "regressione probit". Per ora ci limitiamo ad osservare che il modello logistico (5.4) per la variabile latente z è sufficientemente ben approssimato dal modello di regressione normale,

$$z_i = X_i\beta + \epsilon_i, \quad \epsilon_i \sim \text{N}(0, \sigma^2), \tag{5.5}$$

con $\sigma = 1.6$. Questo porta quindi ad una domanda ovvia, perché non stimare σ?

Non è possibile stimare il parametro σ nel modello (5.5) in quanto risulta non identificato quando considerato congiuntamente al parametro di regressione β. Se tutti gli elementi di β sono moltiplicati per una costante positiva e σ viene anch'essa moltiplicata per la stessa costante, il modello non cambia. Per esempio, supponiamo di stimare il modello

$$z_i = -1.40 + 0.33x_i + \epsilon_i, \quad \epsilon_i \sim \text{N}(0, 1.6^2).$$

Questo modello è equivalente al modello

$$z_i = -14.0 + 3.3x_i + \epsilon_i, \quad \epsilon_i \sim \text{N}(0, 16^2),$$

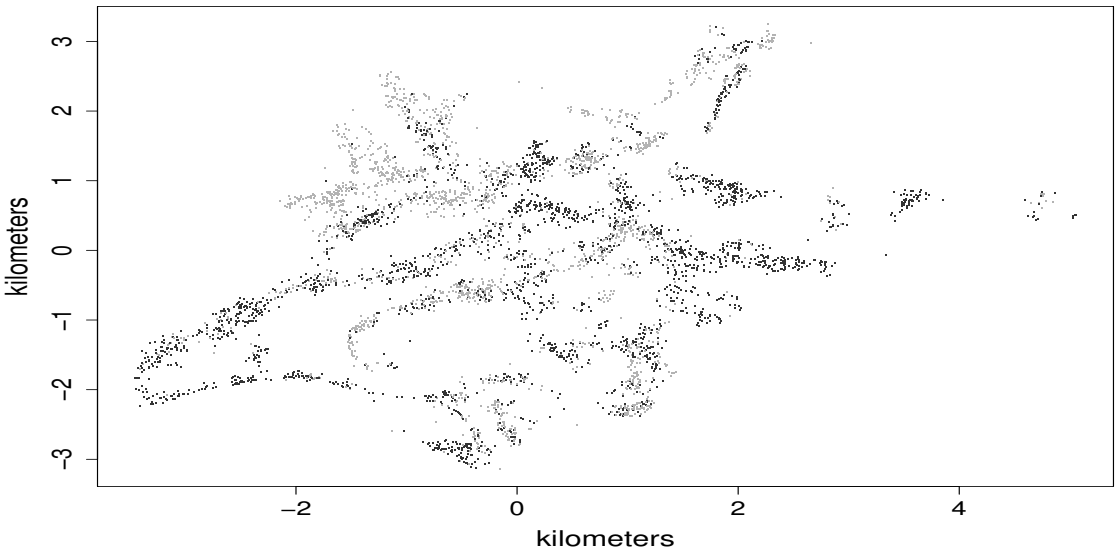
o

$$z_i = -140 + 33x_i + \epsilon_i, \quad \epsilon_i \sim \text{N}(0, 160^2).$$

Appena ci si muove da uno di questi modelli al successivo, z è moltiplicata per 10, ma il *segno* di z non cambia. Quindi, tutti i modello hanno le stesse identiche implicazioni per i dati osservati y: per ciascun modello, $\Pr(y_i=1) \approx \text{logit}^{-1}(-1.40 + 0.33x_i)$ (solo un'approssimazione dal momento che la distribuzione logistica non è esattamente normale).

Di conseguenza, il modello (5.5) ha una indeterminatezza essenziale quando stima dati binari, e questo problema viene risolto in genere fissando il parametro della varianza σ pari ad una valore fisso, per esempio 1.6, che risulta essenzialmente equivalente alla distribuzione logistica unitaria.

Figura 5.7: *Pozzi nell'area di Araihazar upazila in Bangladesh. I punti chiari e scuri rappresentano i pozzi con livelli di arsenico superiori o meno dello standard di sicurezza pari a 0.5 (espressi in unità di centinaia di microgrammi per litro). (I pozzi sono localizzati dove vivono le persone. Le aree vuote tra i pozzi sono in genere aree coltivate). I pozzi sicuri e non sono collocati nelle aree in modo del tutto casuale, il che fa pensare che le persone che si riforniscono di acqua in pozzi non sicuri possono recarsi a rifornirsi di acqua bei vicini pozzi sicuri.*



5.4 Costruzione di un modello logistico: i pozzi in Bangladesh

Illustreremo ora i passi necessari per la costruzione, la comprensione e la valutazione della stima di un modello di scelta binaria attraverso un esempio di tipo economico (o forse psicologico, o di sanità pubblica): modellare la decisione di alcune famiglie del Bangladesh di cambiare fonte di approvvigionamento di acqua potabile.

Base di partenza

Molti pozzi che vengono usati in Bangladesh e in altre nazioni del Sud dell'Asia per l'utilizzo di acqua da bere sono contaminati da arsenico naturale. Si stima che questo problema colpisca almeno 100 milioni di persone. L'arsenico è un veleno cumulativo,

e l'esposizione all'arsenico aumenta il rischio di cancro e di altre malattie, rischio che si stima essere proporzionale all'esposizione.

Ogni località può includere pozzi con un differente livello di arsenico, come si vede dalla Figura 5.7 che rappresenta la mappa di tutti i pozzi in un insieme di villaggi in una piccola zona del Bangladesh. Il fatto negativo è che anche se un pozzo limitrofo è sicuro, non è detto che lo sia anche il pozzo da cui si attinge l'acqua. D'altro canto, una cosa positiva riguarda il fatto che che se anche il proprio pozzo presenta un livello di arsenico molto elevato, si può facilmente trovare un pozzo sicuro in un'area molto vicina dove poter prendere l'acqua da bere—sempre che uno sia in grado di percorrere a piedi la distanza e i vicini abbiano intenzione di dividere la loro acqua. (L'ammontare di acqua necessaria per bere è abbastanza contenuto e l'aggiunta di nuove persone che prendono acqua da un pozzo non ne esaurisce la capacità, e l'acqua in superficie in quest'area è soggetta alla contaminazione da parti di microbi, da cui il desiderio di prendere l'acqua da pozzi molto profondi).

Nella zona mostrata nella Figura 5.7, un gruppo di ricercatori degli Stati Uniti e del Bangladesh ha misurato il livello di arsenico presente in tutti i pozzi e quindi caratterizzato i pozzi in "sicuri" (se il livello di arsenico risulta inferiore a 0.5 in termini di unità di microgrammi per litro, lo standard in Bangladesh per l'arsenico nell'acqua da bere) e in "insicuri" (livello di arsenico nell'acqua superiore a 0.5). Le persone che prelevavano l'acqua da bere in pozzi insicuri sono state incoraggiate a spostarsi verso altri pozzi, sia privati che pubblici o anche di nuova costruzione.

Dopo alcuni anni, i ricercatori sono ritornati nell'area per verificare chi effettivamente avesse cambiato pozzo per il rifornimento dell'acqua da bere. Abbiamo eseguito un'analisi di regressione logistica per capire quali fossero i fattori predittivi per il cambiamento del pozzi tra coloro che inizialmente si rifornivano di acqua in pozzi insicuri. Come nella notazione della sezione precedente, la nostra variabile risposta risulta

$$y_i = \begin{cases} 1 & \text{se la famiglia } i \text{ si rifornisce d'acqua in un nuovo pozzo} \\ 0 & \text{se la famiglia } i \text{ ha continuato ad utilizzare il suo vecchio pozzo.} \end{cases}$$

Consideriamo i seguenti inputs:

- un termine costante;
- la distanza (in metri) verso il pozzo sicuro più vicino;
- il livello di arsenico del pozzo della famiglia intervistata;
- se ciascun membro della famiglia risulta o meno impegnato in attività comunitarie;

- livello di istruzione del capo famiglia.

Inizialmente stimiamo un modello usando solamente la distanza dal pozzo sicuro più vicino, e dopo introduciamo la concentrazione di arsenico, l'appartenza ad una organizzazione e il livello di istruzione.

Regressione logistica con un solo predittore

Stimiamo la regressione logistica in R:

R code
```
fit.1 <- glm (switch ~ dist, family=binomial(link="logit"))
```

E otteniamo i seguenti risultati,

R output
```
glm(formula = switch ~ dist, family=binomial(link="logit"))
            coef.est coef.se
(Intercept)   0.6060  0.0603
dist         -0.0062  0.0010
  n = 3020, k = 2
  residual deviance = 4076.2, null deviance = 4118.1
  (difference = 41.9)
```

Il coefficiente della distanza `dist` è -0.0062, sembrerebbe inizialmente basso ma, ricordando che la distanza è misurata in metri, questo coefficiente corrisponde ad una differenza tra una casa che dista 90 metri dal pozzo sicuro più vicino rispetto ad una casa che dista 91 metri.

La Figura 5.8 mostra la distribuzione della distanza `dist` nei dati. Dall'analisi dell'istogramma sembrerebbe più ragionevole riscalare la distanza ed esprimerla in termini di 100-metri:

R code
```
dist100 <- dist/100
```

stimiamo il modello di regressione logistico con la variabile riscalata e otteniamo,

Figura 5.8: *Istogramma della distanza dal pozzo sicuro più vicino, rilevata in ciascun pozzo non sicuro nella zona di Araihazar (si veda la Figura 5.7).*

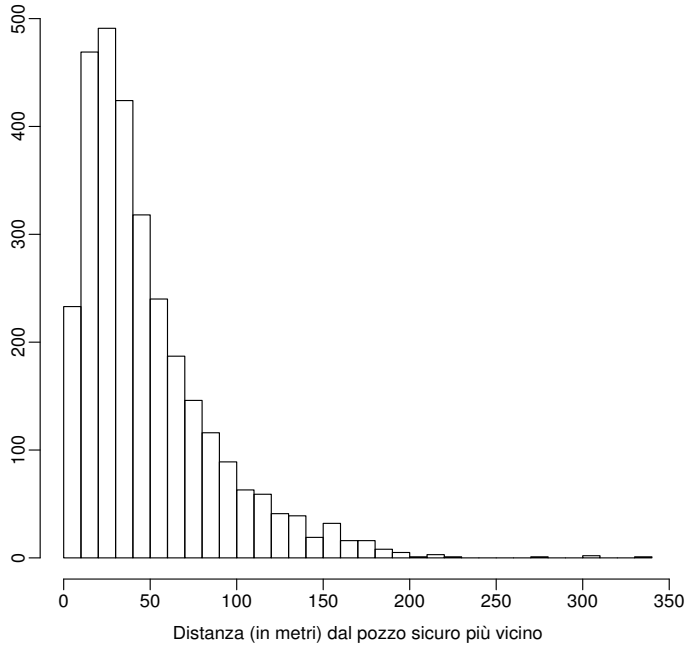
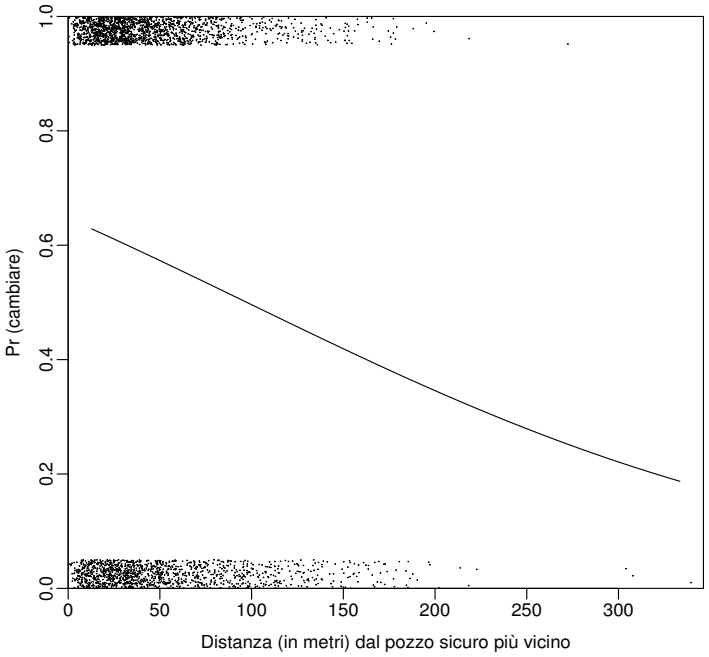


Figura 5.9: *Rappresentazione grafica del modello di regressione logistica stimato,* $\Pr(\textit{cambiare pozzo}) = \textit{logit}^{-1}(0.61 - 0.62 \cdot \texttt{dist100})$*, con i dati sottoposti alla solita procedura di jitter. Il predittore* `dist100` *rappresenta la distanza dal pozzo sicuro più vicino espressa in centinaia di metri (*`dist`*/100).*



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)
```

Rappresentazione grafico del modello stimato

Per rappresentare il grafico dei dati, abbiamo creato in R una funzione per poter effettuare una procedura di jitter alla variabile risposta binaria lasciando comunque i punti compresi tra 0 e 1:

R code
```
jitter.binary <- function(a, jitt=.05){
  ifelse (a==0, runif (length(a), 0, jitt),
  runif (length(a), 1-jitt, 1))
}
```

Possiamo quindi rappresentare i dati e il modello stimato:[3]

R code
```
switch.jitter <- jitter.binary (switch)
plot (dist, switch.jitter)
curve (invlogit (coef(fit.1)[1] + coef(fit.1)[2]*x), add=TRUE)
```

Il risultato è rappresentato nella Figura 5.9. La probabilità di cambiare pozzo è pari a circa il 60% per quelle persone che vivono nelle vicinanze di un pozzo sicuro, ma scende al 20% quando si considerano persone che vivono distanti da un pozzo sicuro più di 300 metri. Questo risultato trova la sua ragione nel fatto che la probabilità di cambiare pozzo risulta maggiormente elevata per quelle persone che abitano in un zona prossima ad un pozzo sicuro.

Interpretazione dei coefficienti della regressione logistica

Possiamo interpretare i coefficienti stimati facendo uso della funzione logistica inversa e delle sue derivate, come abbiamo fatto nell'esempio della Sezione 5.1. Il nostro modello è il seguente,

$$\Pr(\text{switch}) = \text{logit}^{-1}(0.61 - 0.62 \cdot \text{dist100}).$$

1. Il termine costante può essere interpretato quando `dist100` $= 0$, in questo caso la probabilità di cambiare pozzo risulta pari a $\text{logit}^{-1}(0.61) = 0.65$. Quindi, il modello stima una probabilità pari al 65% se una persona vive esattamente nella stessa zona in cui si trova un pozzo sicuro.

2. Possiamo valutare la differenza stimata rispetto al predittore `dist100` calcolando la derivata del valor medio del predittore `dist100` calcolata nel punto medio della variabile, che in base ai nostri dati risulta pari a 0.48 (ovvero 48 metri; si veda la Figura 5.8). Il valore del predittore lineare risulta pari a

[3] Un'altra opzione grafica che avrebbe potuto rappresentare in modo altrettanto adeguato la differenza tra le famiglie che decidevano di cambiare pozzo o no, sarebbe potuta essere la costruzione di due istogrammi separati ma sovrapposti della `dist` per coloro che cambiavano pozzo e per coloro che decidevano di continuare ad utilizzare il proprio pozzo.

$0.61 - 0.62 \cdot 0.48 = 0.31$, e quindi la pendenza della curva in questo punto è pari a $-0.62e^{0.31}/(1 + e^{0.31})^2 = -0.15$. Quindi, ogni volta che aggiungiamo 1 al predittore `dist100`—ovvero aggiungiamo 100 metri alla distanza dal pozzo sicuro più vicino—si ha una differenza negativa nella probabilità di cambiare pozzo che risulta pari a il 15%.

3. Più velocemente, utilizzando la regola della "divisione per 4" otteniamo $-0.62/4 = -0.15$. Il risultato è esattamente lo stesso a quello ottenuto con la derivata, quando approssimiamo a due cifre decimali, in quanto la curva passa nel valore della probabilità pari al 50% proprio nel mezzo dei dati (si veda la Figura 5.9).

Oltre che considerare l'entità dei coefficienti stimati, possiamo sempre controllare per la loro significatività statistica. Il coefficiente della distanza è stimato abbastanza bene, con un errore standard pari a 0.10, molto basso se confrontato con la stima del coefficiente -0.62. L'intervallo di confidenza al 95% approssimato, pari a $[-0.82, -0.42]$, mostra chiaramente che il coefficiente risulta significativamente diverso da zero.

Aggiunta di una seconda variabile di input

Consideriamo sempre l'esempio relativo al cambiamento del pozzo, e aggiungiamo, come nuovo input nella regressione, il livello di arsenico presente nel pozzo esistente. Considerando il livello di arsenico presente nell'acqua che si beve in Bangladesh, il rischio sulla salute conseguente alla presenza di arsenico nell'acqua è, all'incirca, proporzionale all'esposizione, e quindi ci si aspetta che il cambiamento da un pozzo all'altro avvenga verso i pozzi con minore concentrazione di arsenico. La Figura 5.10 mostra il livello di concentrazione di arsenico nei pozzi non sicuri prima del cambiamento.

R code

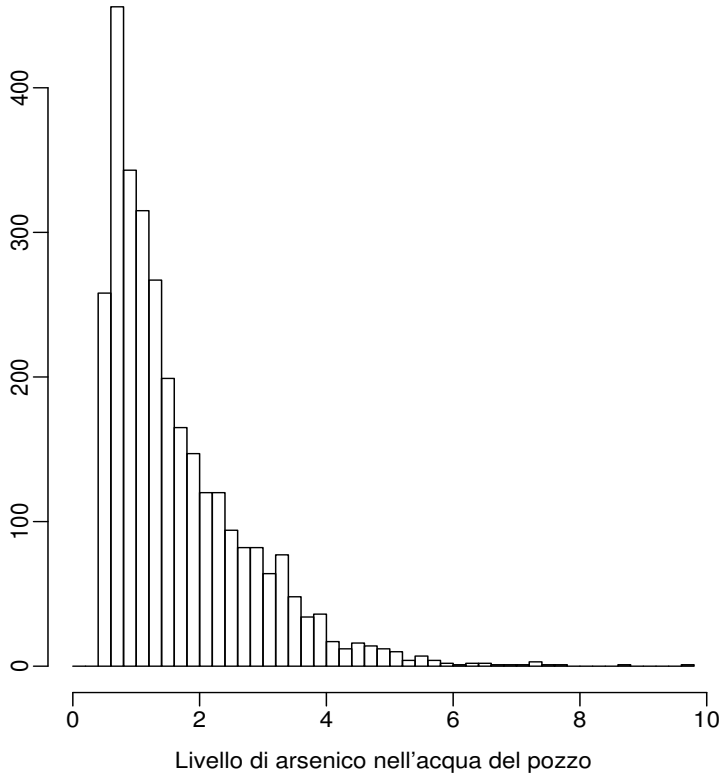
```
fit.3 <- glm (switch ~ dist100 + arsenic,
family=binomial(link="logit"))
```

che fornisce,

R output

```
            coef.est coef.se
(Intercept)     0.00    0.08
dist100        -0.90    0.10
arsenic         0.46    0.04
```

Figura 5.10: *Istogramma relativo al livello di arsenico nei pozzi considerati non sicuri (ovvero quei pozzi il cui livello di arsenico eccede il valore di 0.5) nell'area di Araihazar, in Bangladesh (si veda la Figura 5.7).*



```
n = 3020, k = 3
residual deviance = 3930.7, null deviance = 4118.1
(difference = 187.4)
```

Quindi, confrontando due pozzi aventi lo stesso livello di arsenico, una distanza di 100 metri dal pozzo sicuro più vicino corrisponde ad una differenza *negativa* di 0.90 nella probabilità su scala logistica di cambiare pozzo. In modo del tutto simile, una differenza di 1 nella concentrazione di arsenico corrisponde ad una differenza *positiva* di 0.46 nella probabilità su scala logistica di cambiare pozzo. Entrambi i coefficienti sono statisticamente significativi, essendo entrambi lontani da zero più di due volte l'errore standard. I segni di entrambi i coefficienti sembrano avere senso: cambiare pozzo è più facile se esiste un pozzo sicuro vicino, e se il pozzo di una famiglia ha un livello di arsenico elevato, vi è una maggiore motivazione per cambiare.

Per una interpretazione più veloce, dividiamo i coefficienti per 4: quindi una distanza superiore a 100 metri corrisponde ad una probabilità di cambiare pozzo di circa il 22%, e un'unità in più di concentrazione di arsenico corrisponde ad una differenza positiva di circa l'11% di cambiare pozzo.

Confrontando questi due coefficienti, sembrerebbe inizialmente che la distanza influisca maggiormente sulla probabilità di cambiare pozzo rispetto al livello di arsenico presente nel pozzo. In realtà un'affermazione di questo tipo è fuorviante, in quanto la distanza (`dist100`) presenta nei dati analizzati una varibilità inferiore all'arsenico (`arsenic`): la deviazione standard della distanza dal pozzo più vicino è 0.38 (espressa in 100 metri), mentre i livello di arsenico ha una deviazione standard di 1.10 sulla scala che viene usata nei nostri dati. Quindi, i coefficienti della regressione logistica corrispondenti a differenze pari a una volta la deviazione standard sono pari a $-0.90 \cdot 0.38 = -0.34$ per la distanza e $0.46 \cdot 1.10 = 0.51$ per il livello di arsenico. La divisione per 4 fornisce una stima approssimativa della differenza pari a una volta la deviazione standard nella distanza o nel livello di arsenico corrispondente ad una differenza negativa dell'8% o a una differenza positiva del 13%, rispettivamente, nella Pr(switch).

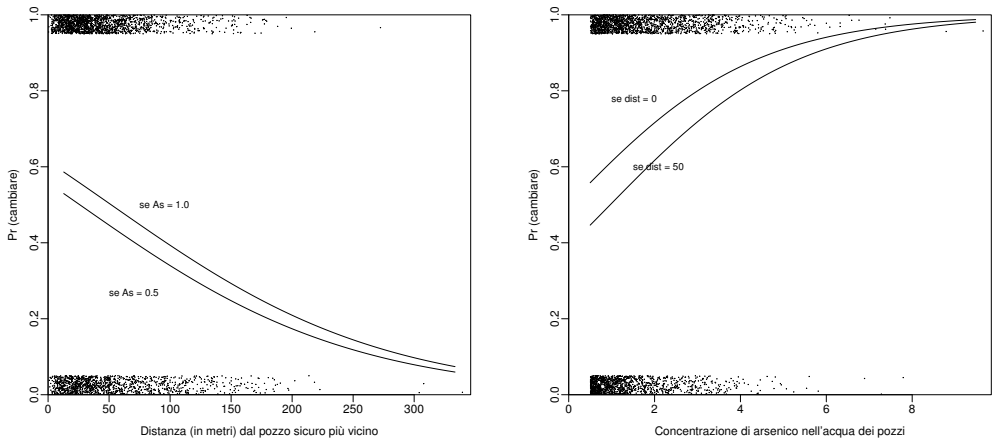
Confronto delle stime dei coefficienti quando si aggiunge un predittore

Il coefficiente relativo alla distanza (`dist100`) cambia da -0.62 nel modello originale a 0.9 quando il livello di arsenico viene aggiunto nel modello. Questo cambiamento accade perché i pozzi che sono lontani dal pozzo sicuro più vicino sono verosimilmente caratterizzati da livelli di arsenico particolarmente elevati.

Rappresentazione del modello stimato con due predittori

Il modo più naturale per rappresentare il grafico della regressione di y in base ai due predittori potrebbe essere una superficie a tre dimensioni con l'asse verticale che rappresenta la $\Pr(y = 1)$ come funzione di predittori rappresentati sui due assi orizzontali.

Figura 5.11: *Stima della regressione logistica della probabilità di spostarsi da un pozzo insicuro verso uno sicuro in funzione di due variabili, nel grafico (a) questa probabilità è rappresentata in funzione della distanza dal pozzo sicuro più vicino e in (b) in funzione del livello di arsenico presente nel pozzo in esame. Per ciascun grafico, le altre variabili di input sono tenute costanti a differenti valori rappresentativi.*



In ogni caso, questi grafici sono abbastanza difficili da leggere, quindi preferiamo fare dei grafici separatamente per ciascuna delle due variabili; si veda la Figura 5.11. Come con le linee nella Figura 3.4, possiamo rappresentare graficamente la variabile di input più importante nell'asse delle x e usare linee multiple per rappresentare il modello stimato per i differenti valori dell'altro input. Per fare la Figura 5.11a, abbiamo prima rappresentato graficamente i punti (sottoposti alla solita procedura di jittered), imponendo che lo zero fosse incluso nell'asse delle x del grafico essendo la base naturale per una qualunque misura di distanza:

```
plot (dist, switch.jitter, xlim=c(0,max(dist)))
```
R code

Aggiungiamo poi le curve stimate:

R code
```
curve (invlogit (cbind (1, x/100,  .5) %*% coef(fit.3)), add=TRUE)
curve (invlogit (cbind (1, x/100, 1.0) %*% coef(fit.3)), add=TRUE)
```

Abbiamo bisogno di dividere x per 100 in quanto il grafico è nella scala in metri mentre il modello è definito in termini di `dist100` = `dist`/100.

L'oggetto creati da `cbind(1,x/100,.5)` è una matrice costruita da una colonna di valori tutti pari a `1`, il vettore `x` (usato internamente della funzione `curve`), e un vettore di valori pari a `.5`. Nel costruire la matrice, R automaticamente espande gli scalari `1` e `.5` alla lunghezza del vettore `x`. Per le due linee, abbiamo considerato i livelli di arsenico pari a 0.5 e 1.0 essendo 0.5 il valore minimo della concentrazione di arsenico (qui stiamo considerando solo pozzi non sicuri), e una differenza di 0.5 rappresenta un confronto ragionevole, data la distribuzione del livello di arsenico nei dati (di veda la Figura 5.10).

Comandi simili sono stati usati per la Figura 5.11b, che mostra probabilità di cambiare pozzo in funzione della concentrazione di arsenico con la distanza tenuta costante:

R code
```
plot (arsenic, switch.jitter, xlim=c(0,max(arsenic)))
curve (invlogit (cbind (1, 0, x) %*% coef(fit.3)), add=TRUE)
curve (invlogit (cbind (1,.5, x) %*% coef(fit.3)), add=TRUE)
```

5.5 Regressione logistica con interazioni

Continuiamo a modellare la probabilità di spostarsi da un pozzo non sicuro a uno sicuro aggiungendo anche l'interazione tra i due input finora considerati:

R code
```
fit.4 <- glm (switch ~ dist100 + arsenic +
              dist100:arsenic,
              family=binomial(link="logit"))
display (fit.4)
```

ottenendo,

R output

```
                coef.est coef.se
(Intercept)        -0.15    0.12
dist100            -0.58    0.21
arsenic             0.56    0.07
dist100:arsenic    -0.18    0.10
  n = 3020, k = 4
  residual deviance = 3927.6, null deviance = 4118.1
  (difference = 190.5)
```

Per capire meglio i numeri della tavola precedente usiamo i seguenti "trucchi":

- Valutiamo le previsioni e le interazioni nel valore medio dei dati, che sono pari a 0.48 per `dist100` e 1.66 per `arsenic` (ovvero, una distanza media di 48 metri dal pozzo sicuro più vicino, e un livello medio di arsenico di 1.66 nei pozzi non sicuri).
- Dividiamo i coefficienti per 4 per ottenere la differenza predittiva approssimata sulla scala delle probabilità.

A questo punto siamo in grado di interpretare i coefficienti della regressione.

- *Termine costante:* $\text{logit}^{-1}(-0.15) = 0.47$ rappresenta la probabilità stimata di cambiare pozzo, quando la distanza dal pozzo sicuro più vicino è 0 e il livello di arsenico nel pozzo sotto analisi è 0. Questa condizione non si verifica mai (dal momento che il livello di arsenico nei pozzi non sicuri eccede sempre il valore di 0.5), di conseguenza non tenteremo di interpretare il termine costante. Invece, proveremo a valutare la previsione nei punti medi della distanza (`dist100` = 0.48) e del livello di arsenico (`arsenic` = 1.66), in corrispondenza dei quali la probabilità stimata di cambiare pozzo è pari a: $\text{logit}^{-1}(-0.15 - 0.58 \cdot 0.48 + 0.56 \cdot 1.66 - 0.18 \cdot 0.48 \cdot 1.66) = 0.59$.
- *Coefficiente della distanza*: questo valore corrisponde ad un eventuale confronto tra due pozzi che differiscono di una unità nella variabile `dist100` quando il livello di arsenico è pari a 0 in entrambi i pozzi in esame. Quindi, ancora una volta, non proveremo ad interpretare questo coefficiente.

  Invece, possiamo a considerare il valore medio dell'arsenico `arsenic` = 1.66, nel qual caso la distanza ha un coefficiente di $-0.58 - 0.18 \cdot 1.66 = -0.88$ su scala logistica. Per interpretare velocemente questo coefficiente sulla scala delle probabilità, lo dividiamo per 4: $-0.88/4 = -0.22$. Quindi, quando il livello della concentrazione di arsenico è pari al suo livello medio, ad una distanza di 100 metri corrisponde una differenza *negativa* approssimata pari a circa il 22% di cambiare pozzo.

- *Coefficiente dell'arsenico*: questo coefficiente corrisponde al confronto tra due pozzi che differiscono di una unità nel livello di concentrazione dell'arsenico (`arsenic`), quando entrambi i pozzi che stiamo considerando distano zero dal pozzo sicuro più vicino.

  Se effettuiamo invece il confronto nel punto medio della distanza, `dist100` = 0.48, il coefficiente dell'arsenico ha un valore pari a $0.56 - 0.18 \cdot 0.48 = 0.47$ su scala logistica. Per interpretare velocemente questo coefficiente sulla scala delle probabilità lo dividiamo per 4: $0.47/4 = 0.12$. Quindi, quando si considera il livello medio della distanza, ad ogni unità addizionale di arsenico corrisponde una differenza *positiva* approssimata pari a circa il 12% di rifornirsi di acqua in un altro pozzo.

- *Coefficiente del termine di interazione*: questo coefficiente può essere interpretato in due modi diversi. Se guardiamo verso una direzione, per ogni unità addizionale di arsenico, un valore pari a -0.18 viene aggiunto al coefficiente della distanza. Abbiamo già visto che il coefficiente della distanza valutato in corrispondenza del livello medio di arsenico risulta pari a -0.88, quindi possiamo interpretare l'interazione dicendo che la distanza intesa come predittore risulta maggiormente importante per quelle famiglie i cui pozzi hanno un livello di arsenico più elevato.

  Se lo vogliamo interpretare seguendo una seconda direzione possiamo dire che per ogni 100 metri di distanza in più dal pozzo sicuro più vicino, il valore di -0.18 viene aggiunto al coefficiente dell'arsenico, Abbiamo visto che il coefficiente dell'arsenico uguale a 0.47 al livello medio della distanza dal pozzo sicuro più vicino, quindi possiamo interpretare l'interazione dicendo che l'importanza dell'arsenico come predittore risulta meno importante per quelle famiglie che vivono molto distanti da pozzi sicuri.

Centratura delle variabili di input

Come discusso precedentemente nel contesto della regressione lineare, prima di introdurre le interazioni nella stima del modello bisognerebbe centrare le variabili di input in modo da rendere i coefficienti più facilmente interpretabili. Le variabili di input centrate sono pari a:

R code

```
c.dist100 <- dist100 - mean(dist100)
c.arsenic <- arsenic - mean(arsenic)
```

Abbiamo deciso di non standardizzare totalmente queste variabili—ovvero di non dividere per la loro rispettiva deviazione standard—in quanto risulta maggiormente conveniente valutare le differenze sulla scala originale dei dati (la distanza in termini di 100 metri e l'arsenico in unità di concentrazione).

Nuova stima del modello con interazione usando gli input centrati

Possiamo quindi effettuare una nuova stima del modello in cui vengono considerate le variabili di input centrate, rendendo i coefficienti più facilmente interpretabili:

R code

```
fit.5 <- glm (switch ~ c.dist100 + c.arsenic +
              c.dist100:c.arsenic,
              family=binomial(link="logit"))
```

Si noti che abbiamo centrato gli *input* e non i *predittori*. Quindi, non centriamo l'interazione (`dist100*arsenic`); piuttosto, includiamo nel modello l'interazione delle due variabili di input centrate. La rappresentazione del modello stimato `fit.5` fornisce,

R output

```
                   coef.est coef.se
(Intercept)         0.35     0.04
c.dist100          -0.88     0.10
c.arsenic           0.47     0.04
c.dist100:c.arsenic -0.18     0.10
  n = 3020, k = 4
  residual deviance = 3927.6, null deviance = 4118.1
  (difference = 190.5)
```

Interpretiamo i risultati su questa nuova scala:

- *Termine costante*: $\text{logit}^{-1}(0.35) = 0.59$ rappresenta la probabilità stimata di cambiare pozzo, quando `c.dist100` = `c.arsenic` = 0, ovvero quando la distanza dal pozzo sicuro più vicino e i livello di arsenico sono pari ai loro valori medi. (Abbiamo ottenuto lo stesso risultato nel modello precedente ma con uno sforzo interpretativo maggiore).

- *Coefficiente della distanza*: rappresenta il coefficiente della distanza (su scala logistica) quando il livello di arsenico è pari al suo livello medio. Per interpretare velocemente questo coefficiente su scala probabilistica lo dividiamo per 4: $-0.88/4 = -0.22$. In questo modo possiamo dire che quando il livello di concentrazione dell'arsenico si trova al suo livello medio, ad una distanza di 100 metri corrisponde una differenza stimata *negativa* pari a circa il 22% nella probabilità di cambiare pozzo.

- *Coefficiente per l'arsenico*: rappresenta il coefficiente dell'arsenico quando la distanza dal pozzo sicuro più vicino è pari alla distanza media. Per interpretare velocemente questo coefficiente su scala probabilistica, lo dividiamo al solito per 4: $0.47/4 = 0.12$. Quindi, a livello medio della distanza, ad ogni unità addizionale di concentrazione di arsenico corrisponde una differenza stimata *positiva* pari a circa il 12% nella probabilità di cambiare pozzo.

- *Coefficiente per il termine di interazione*: questo coefficiente non è cambiato in seguito alla centratura e quindi ha la stessa interpretazione di prima.

Le previsioni per le nuove osservazioni sono rimaste invariate. La procedura di centratura lineare dei predittori cambia l'interpretazione dei coefficienti del modello ma non il modello sottostante.

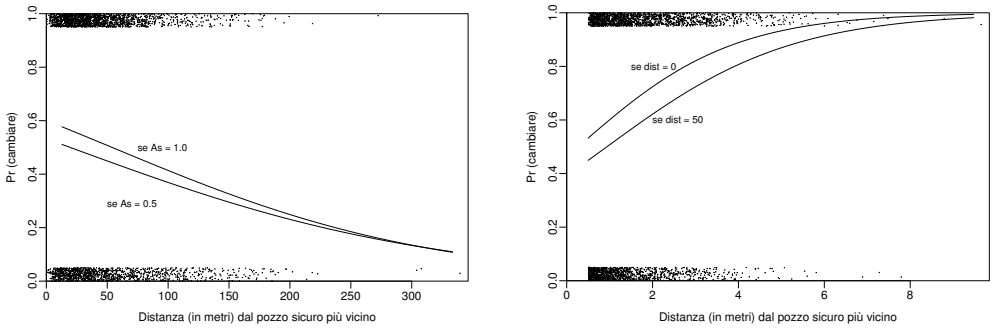
Significatività statistica dell'interazione

Come abbiamo visto nella tavola della regressione precedente, il coefficiente del termine di interazione `c.dist100:c.arsenic` ha un valore stimato pari a -0.18 con un associato errore standard di 0.10. La stima non risulta quindi esterna all'intervallo intorno allo zero di ampiezza $\pm$ due volte l'errore standard e, di conseguenza, non risulta significativa. In ogni caso, il segno negativo sembra del tutto ragionevole—è infatti plausibile che il livello dell'arsenico diventi un predittore meno importante per quelle famiglie che vivono lontano da un pozzo sicuro e anche il valore del coefficiente sembrerebbe del tutto ragionevole. Quindi decidiamo di tenere il termine di interazione nel modello, seguendo appunto quelle che sono i nostri principi circa i predittori nel modello e la loro significatività statistica, come descritto nella Sezione 4.6.

Rappresentazione grafica del modello con interazione

Il modo più chiaro per visualizzare il modello con interazione è quello di rappresentare graficamente le curve di regressione in funzione di ciascun predittore. Il risultato

Figura 5.12: *Stima del modello di regressione logistica con interazione che stima la probabilità di passare da un pozzo insicuro a uno sicuro in funzione della distanza dal sicuro più vicino e del livello di arsenico presente nel pozzo in esame. Si confronti il modello senza interazione della Figura 5.11*



è quello che appare nella Figura 5.12, per il primo grafico sono stati usati in R i seguenti comandi (comandi simili sono stati usati per il secondo grafico):

R code

```
plot (dist, switch.jitter, xlim=c(0,max(dist)))
curve (invlogit (cbind(1,x/100, .5, .5*x/100) %*% coef(fit.4)),
       add=TRUE)
curve (invlogit (cbind(1,x/100,1.0,1.0*x/100) %*% coef(fit.4)),
       add=TRUE)
```

Come appare evidente dalla Figura 5.12, l'interazione non risulta molto grande nell'intervallo della maggior parte dei dati. L'andamento maggiormente significativo si evince dalla Figura 5.12a, dove le due curve si intersecano intorno ad un valore pari a circa 300 metri. Questo grafico mostra che, evidentemente, le differenze nella probabilità di cambiare pozzo associate al livello di arsenico, sono elevate quando ci si trova vicino a un pozzo sicuro, ma l'effetto va via via diminuendo tanto più elevata diventa la distanza dal pozzo sicuro più vicino. Questa interazione sembra del tutto ragionevole, anche se vi è una qualche incertezza sulla grandezza dell'interazione (dalla tabella di prima si evince un coefficiente stimato pari a -0.18 con un relativo errore standard di 0.10) e, come mostrato dalla Figura 5.12a, ci sono solo pochi dati nell'area in cui l'interazione comporta una differenza di probabilità consistente. L'interazione compare anche nella Figura 5.12b, questa volta in un grafico che rappresenta la probabilità di cambiare pozzo in funzione del livello di arsenico presente, probabilità valutata a due diversi livelli di distanza.

Aggiunta di predittori sociali

Possiamo chiederci: chi utilizza i pozzi è maggiormente disposto a cambiare se ha una qualche relazione con la comunità o un livello di istruzione superiore? Per verificare se questa ipotesi ha un qualche senso, aggiungiamo altri due input nel modello:

- `assoc` = 1 se la famiglia è membro di una qualche organizzazione all'interno della comunità
- `educ` = anni di istruzione dell'utilizzatore del pozzo.

Lavoriamo con `educ4` = `educ`/4, per le solite ragioni di rendere i coefficienti del modello di regressione maggiormente interpretabili—che ora rappresenta la differenza attesa conseguente all'aggiunta di quattro anni di istruzione.[4]

R output

```
glm(formula = switch ~ c.dist100 + c.arsenic +
    c.dist100:c.arsenic + assoc + educ4,
    family=binomial(link="logit"))
                  coef.est coef.se
(Intercept)          0.20     0.07
c.dist100           -0.88     0.11
c.arsenic            0.48     0.04
c.dist100:c.arsenic -0.16     0.10
assoc               -0.12     0.08
educ4                0.17     0.04
  n = 3020, k = 6
  residual deviance = 3905.4, null deviance = 4118.1
  (difference = 212.7)
```

Per le famiglie che si riforniscono in un pozzo insicuro, l'appartenenza ad una associazione all'interno della comunità risulta sorprendentemente *non* predittiva per il cambiamento, dopo aver controllato per le altre variabili. In ogni caso, persone con livelli di istruzione più elevati sono maggiormente propense a spostarsi: la differenza attesa stimata risulta pari a $0.17/4 = 0.04$, ovvero un 4% di differenza positiva nella probabilità di spostarsi verso un pozzo diverso quando si confrontano persone con un

[4]Gli anni di istruzione dei 3000 intervistati variano da 0 a 17 anni, con un terzo circa dei rispondenti che riporta un livello di istruzione pari a zero. Abbiamo ripetuto l'analisi con una ricodifica discreta della variabile istruzione (0 = 0 anni, 1 = 1–8 anni, 2 = 9–12 anni, 3 = 12+ anni), ma i risultati sono rimasti essenzialmente invariati.

livello di istruzione superiore o inferiore di 4 anni.[5] Il coefficiente relativo al livello di istruzione sembra del tutto ragionevole e risulta anche statisticamente significativo, quindi lo teniamo senza dubbio nel modello. Il coefficiente relativo all'associazione sembra non avere alcun senso e non è statisticamente significativo, quindi lo rimuoviamo dal modello. (Si veda la Sezione 4.6 per una discussione approfondita circa l'inclusione e l'esclusione dei predittori nel modello di regressione.) Il nuovo modello stimato risulta pari a:

R output

```
glm(formula = switch ~ c.dist100 + c.arsenic +
    c.dist100:c.arsenic +
    educ4, family = binomial(link = "logit"))
                   coef.est coef.se
(Intercept)         0.15     0.06
c.dist100          -0.87     0.11
c.arsenic           0.48     0.04
c.dist100:c.arsenic -0.16     0.10
educ4               0.17     0.04
  n = 3020, k = 5
  residual deviance = 3907.9, null deviance = 4118.1
  (difference = 210.2)
```

Aggiunta di ulteriori interazioni

Quando gli input hanno effetti molto grandi, è nostra abitudine includere nel modello anche le possibili interazioni. Procediamo inizialmente col centrare la variabile istruzione:

R code

```
c.educ4 <- educ4 - mean(educ4)
```

e quindi andiamo a stimare un modello con due nuovi predittori che rappresentano rispettivamente l'interazione del livello di istruzione con la distanza dal pozzo sicuro più vicino e del livello di istruzione con il livello di arsenico presente nel pozzo in esame:

[5] In questo esempio, abbiamo sempre parlato di "coefficienti" e "differenze," piuttosto che di "effetti" e "cambiamenti," in quanto la natura osservazionale dei dati rende particolarmente complicato interpretare il modello di regressione in termini causali. Studieremo l'inferenza causale in modo maggiormente approfondito nel Capitolo 9 e discuteremo brevemente del problema dell'arsenico alla fine della Sezione 9.8.

R output

```
glm(formula=switch~c.dist100 + c.arsenic + c.educ4 +
   c.dist100:c.arsenic + c.dist100:c.educ4 +
   c.arsenic:c.educ4, family=binomial(link="logit"))
                   coef.est coef.se
(Intercept)         0.36     0.04
c.dist100          -0.90     0.11
c.arsenic           0.49     0.04
c.educ4             0.18     0.04
c.dist100:c.arsenic -0.12    0.10
c.dist100:c.educ4   0.32     0.11
c.arsenic:c.educ4   0.07     0.04
  n = 3020, k = 7
  residual deviance = 3891.7, null deviance = 4118.1
  (difference = 226.4)
```

Possiamo interpretare queste nuove interazioni cercando di capire come il livello di istruzione modifica la differenza attesa probabilità di cambiare pozzo corrispondente alla distanza e al livello di arsenico.

- *Interazione della distanza e del livello di istruzione*: una differenza di 4 anni di istruzione corrisponde and una differenza di 0.32 nel coefficiente della distanza `dist100`. Come abbiamo visto, `dist100` ha un segno negativo in media; quindi cambiamenti positivi nell'istruzione *riducono* l'associazione negativa della distanza. Questa associazione sembra ragionevole, infatti persone con un livello di istruzione più elevato probabilmente hanno altre risorse piuttosto che camminare a piedi per recarsi al pozzo sicuro e quindi una distanza aggiuntiva non risulta particolarmente problematica.
- *Interazione del livello di arsenico e del livello di istruzione*: una differenza di 4 anni di istruzione corrisponde ad una differenza pari a 0.07 nel coefficiente del livello di arsenico`arsenic`. Abbiamo visto che il coefficiente dell'arsenico ha un segno positivo in media; quindi l'aumento del livello di istruzione *aumenta* l'associazione positiva dell'arsenico. Anche questa associazione sembra ragionevole, infatti persone con un livello di istruzione più elevato sono probabilmente maggiormente informate circa i rischi dell'arsenico e quindi maggiormente sensibilizzate ad incrementi nel livello di arsenico (oppure, al contrario, meno veloci nel cambiare con livelli di arsenico relativamente bassi).

Come prima, la centratura permette di interpretare gli effetti principali come coefficienti quando gli altri input sono tenuti costanti ai loro livelli medi stimati.

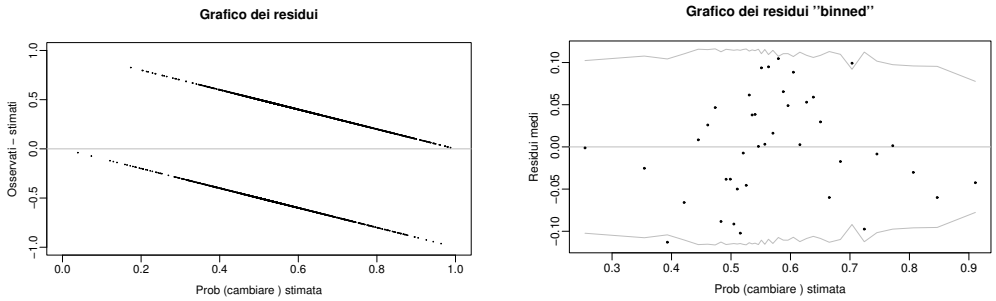
Predittori standardizzati

Bisognerebbe pensare seriamente di standardizzare tutti i predittori e di usare questa regola come una opzione di *default* ogni qual volta si aggiungono delle interazioni nel modello. Gli sforzi che abbiamo fatto per interpretare `dist100` e `educ4` in questo esempio suggeriscono che la standardizzazione—attraverso la sottrazione della media da tutte le variabili di input continue e la divisione per due volte la deviazione standard, come suggerito alla fine della Sezione 4.2—potrebbe essere l'approccio più semplice.

5.6 Valutazione, controllo e confronto dei modelli di regressione logistica

Residui e "binned" residui

Figura 5.13: *(a) Grafico dei residui e (b) grafico dei residui "binned" per il modello relativo al cambiamento del pozzo mostrato a pagina 135. L'andamento marcatamente lineare dei residui grezzi deriva dal fatto che i dati sono discreti e questo induce ad usare il grafico dei residui "binned". Gli intervalli non sono equi-spaziati; invece ciascun intervallo ha uno stesso numero di punti. Le linee chiare nel grafico dei "binned" residual rappresentano i limiti di errori al 95%.*



Possiamo definire i residui per la regressione logistica, come fatto per la regressione lineare come differenza tra i valori osservati e i valori stimati:

$$\text{residui}_i = y_i - \text{E}(y_i|X_i) = y_i - \text{logit}^{-1}(X_i\beta).$$

I dati y_i sono discreti e quindi anche i residui. Per esempio, se $\text{logit}^{-1}(X_i\beta) = 0.7$, allora $\text{residual}_i = -0.7$ o $+0.3$, a seconda che $y_i = 0$ o uguale a 1. Come conseguenza, i grafici dei residui grezzi in genere non sono di grande utilità. Per esempio, la Figura

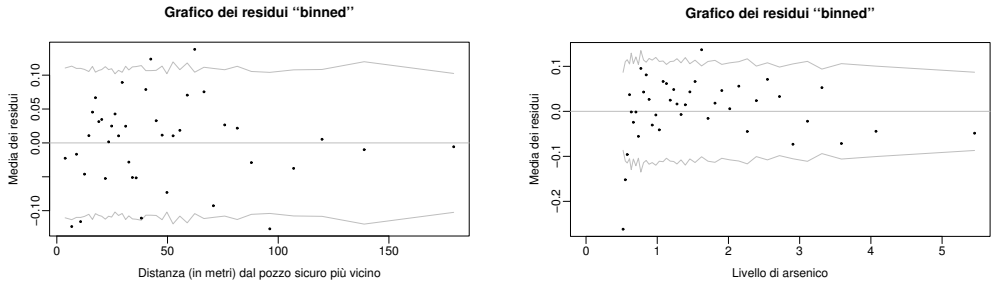
5.13a rappresenta i residui in funzione dei valori stimati per il modello di regressione relativo al cambiamento del pozzo.

Invece, rappresentiamo graficamente i *residui binned* che vengono ottenuti dividendo i dati in categorie (bins) definite in base ai loro valori stimati e quindi rappresentiamo la media dei residui in funzione della media dei valori stimati per ciascuna categoria (bin). Il risultato è rappresentato nella Figura 5.13b; in questo esempio i dati sono stati divisi in 40 categorie aventi stessa ampiezza.[6]

Le linee tratteggiate (calcolate come $2\sqrt{p(1-p)/n}$, dove n è il numero di punti per categoria, pari a $3020/40 = 75$ in questo casi) indicano i limiti dell'intervallo ± 2 volte l'errore standard, all'interno del quale ci si aspetta che cadano circa il 95% dei residui categorizzati ("binned") se il modello fosse vero. Uno dei 40 residui categorizzati nella Figura 5.13b cade fuori dai limiti dell'intervallo, il che non è sorprendente e non indica alcun comportamento particolare.

Rappresentazione grafica dei residui "binned" in funzione degli input di interesse

Figura 5.14: *Grafico dei residui "binned" per il modello relativo al cambiamento del pozzo, in funzione (a) della distanza dal pozzo più vicino e (b) del livello di arsenico. Le linee tratteggiate nel grafico dei residui "binned" indicano i limiti dell'intervallo teorico dell'errore al livello 95% che sarebbe quello appropriato qualora il modello stimato fosse quello vero. Il secondo grafico mostra un problema nella parte inferiore delle categorie dei livelli di arsenico.*



[6]Vi è sempre una certa arbitrarietà nella scelta del numero delle categorie: quello che si vuole è che ciascuna categoria contenga un numero di punti sufficientemente elevato in modo tale che i valori medi non siano troppo irregolari, ma allo stesso tempo avere molte categorie può aiutare a capire se esistono dei comportamenti tipici nei residui. Per questo esempio, 40 categorie forniscono una risoluzione sufficiente avendo, allo stesso tempo, un numero sufficientemente elevato di punti in ciascuna categoria. Un altro approccio potrebbe essere quello di applicare una procedura di *smoothing* non parametrica sui residui come per esempio la curva *lowess* (Cleveland, 1979).

È possibile analizzare i residui anche in maniera maggiormente strutturata, categorizzandoli prima e rappresentandoli graficamente poi in funzione di variabili di input individuali o di combinazioni di input. Per esempio, nell'esempio del modello per il cambiamento del pozzo, la Figura 5.14a mostra la media dei residui in ciascuna categoria in funzione della distanza dal pozzo più vicino, e la Figura 5.14b mostra i residui "binned" in funzione del livello di arsenico.

Quest'ultimo grafico mostra un comportamento strano, con residui molto negativi nelle prime categorie più basse dell'arsenico: le persone che vivono vicino ad un pozzo nella categoria più bassa (che corrisponde ad un livello di arsenico tra 0.51 e 0.53) hanno una probabilità di cambiare pozzo di circa il 20% in meno rispetto a quella stimata, la probabilità media stimata di cambiare pozzo per questi utilizzatori risulta pari al 49%, ma in realtà solo il 32% di loro cambia. Risulta inoltre, in generale, un comportamento dei residui leggermente anomalo, come risulta dalla presenza di residui positivi (in media) per la categorie centrali del livello di arsenico e negativi per le categorie più alte.

Introduzione di una trasformata logaritmica

In base all'esperienza di chi stima molti modelli di regressione, sembrerebbe che il comportamento alternato dei residui come nella Figura 5.14b potrebbe essere un segnale della necessità di trasformare in termini logaritmici il predittore sull'asse delle x—in questo caso il livello di arsenico. Un'altra opzione potrebbe essere quella di aggiungere un termine quadratico alla regressione; ma dal momento che il livello di concentrazione dell'arsenico è sempre positivo ha senso considerare la sua trasformata logaritmica. In ogni caso, non vogliamo modellare la distanza in termini logaritmici dal momento che il grafico dei residui, come si nota dalla Figura 5.14a, indica un buon adattamento del modello lineare.

Definiamo

R code
```
log.arsenic <- log(arsenic)
c.log.arsenic <- log.arsenic - mean (log.arsenic)
```

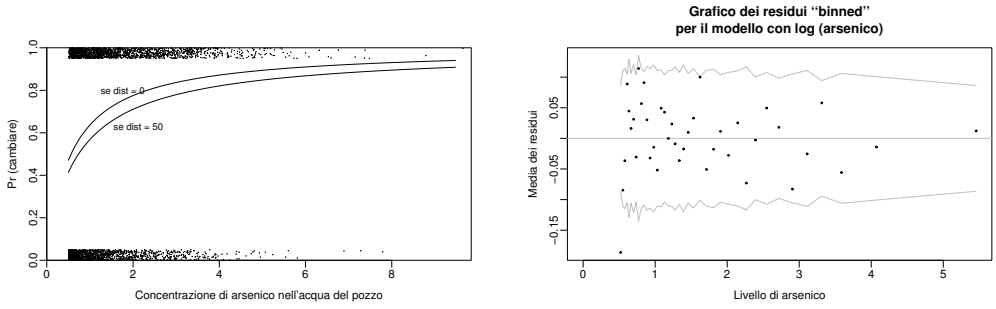
e quindi stimiamo ancora una volta lo stesso modello di prima usando `log.arsenic` invece di `arsenic`:

R output
```
glm(formula = switch ~ c.dist100 + c.log.arsenic + c.educ4 +
    c.dist100:c.log.arsenic + c.dist100:c.educ4 + c.log.arsenic:c.educ4,
```

```
    family = binomial(link = "logit"))
                          coef.est coef.se
(Intercept)                0.35     0.04
c.dist100                 -0.98     0.11
c.log.arsenic              0.90     0.07
c.educ4                    0.18     0.04
c.dist100:c.log.arsenic   -0.16     0.19
c.dist100:c.educ4          0.34     0.11
c.log.arsenic:c.educ4      0.06     0.07
  n = 3020, k = 7
  residual deviance = 3863.1, null deviance = 4118.1 (difference = 255)
```

Questo modello è qualitativamente simile a quello stimato sulla scala originale: l'interazione ha lo stesso segno di prima e anche i segni degli effetti più importanti non sono cambiati.

Figura 5.15: *(a) Probabilità di cambiare pozzo in funzione del livello di arsenico (in corrispondenza di due diversi valori della distanza (*`dist`*) tenendo il livello di istruzione costante al suo valor medio), per il modello che include il livello di arsenico su scala logaritmica. Si confronti questo grafico con quello in Figura 5.11b (il corrispondente grafico che considera come predittore lineare il livello di arsenico), il modello sembra simile ma con una pendenza più ripida all'estremità bassa della curva e una pendenza più graduale all'estremità alta della curva.*
(b) Residui medi per questo modello, categorizzati in base al livello di arsenico. Se si confronta questo grafico con quello della Figura 5.14b, vediamo che esistono ancora dei problemi per i livelli più bassi dell'arsenico ma altrimenti il grafico sembra abbastanza pulito.



La Figura 5.15a mostra la probabilità stimata di cambiare pozzo in funzione del livello dell'arsenico nel pozzo. Se la si confronta con il modello in cui il livello dell'arsenico era stato incluso come predittore lineare (si veda la 5.11 a pagina 127), le figure appaiono schiacciate a sinistra e allungate verso destra.

La Figura 5.15b rappresenta i residui del modello in termini logaritmici, categorizzati in base al livello di arsenico. Se li si confronta con i residui del modello precedente, questi residui sembrano migliori ma esiste tuttavia ancora un problema nella parte in basso del grafico. Gli utilizzatori dei pozzi con livello di arsenico di poco superiore a 0.5 sono meno propensi a spostarsi rispetto a quanto previsto dal modello. A questo punto dell'analisi, non sappiamo se questo fenomeno posso essere spiegato in termini psicologici (misure appena sopra la soglia non sono ritenute troppo pericolose), attraverso errori di misura (per esempio il livello di arsenico in alcuni pozzi misurato pari a 0.51 o 0.52 è stato poi trovato inferiore a 0.5), o anche per qualsiasi altra ragione che non siamo in grado di spiegare.

Tasso di errore e confronto con il modello nullo

Il *tasso di errore* è definito come la proporzione di casi in cui la previsione deterministica—supponiamo $y_i = 1$ se $\text{logit}^{-1}(X_i\beta) > 0.5$ e supponiamo $y_i = 0$ se $\text{logit}^{-1}(X_i\beta) < 0.5$—è sbagliata. In R, possiamo scrivere,

R code

```
error.rate <- mean ((predicted>0.5 & y==0) |
                    (predicted<.5 & y==1))
```

Il tasso di errore dovrebbe essere sempre inferiore a 1/2 (altrimenti potremmo semplicemente porre tutti i β uguali a 0 e prendere un modello che si adatti meglio), ma in molti casi ci si aspetta che sia di molto inferiore. Possiamo quindi confrontare il tasso di errore del *modello nullo*, che si ha semplicemente assegnando la stessa probabilità a ciascun y_i. Questo modello è un modello logistico con il solo termine costante, e la probabilità stimata semplicemente la proporzione di valori pari ad 1 nei dati, ovvero $p = \sum_{i=1}^{n} y_i/n$ (si ricordi che $y_i = 1$ o 0). Il tasso di errore nel modello nullo risulta quindi pari a p o a $1-p$, a seconda di quale sia il valore più basso.

Nell'esempio relativo al cambiamento del pozzo, il modello nullo ha un tasso di errore pari al 42% (58% dei rispondenti cambiano pozzo per il rifornimento dell'acqua e il 42% no, quindi il modello senza predittori fornisce per ciascuna persona una probabilità pari al 58% di cambiare pozzo, che corrisponde ad una previsione puntuale di cambiare per ciascuna persona, e che potrebbe essere sbagliata il 42% delle volte). Il nostro modello di regressione logistico (così come calcolato in R) ha un tasso di errore del 36%. Il modello quindi prevede in modo corretto il comportamento del 64% dei rispondenti.

Il tasso di errore non è una misura di sintesi perfetta del cattivo adattamento del modello, in quanto non distingue per esempio tra previsioni pari a 0.6 e 0.9. Ma,

così come R^2 per il modello di regressione lineare, è facile da interpretare e risulta spesso molto utile per la comprensione dell'adattamento del modello. Un tasso di errore uguale uguale al tasso del modello nullo è il peggiore, mentre il miglior tasso di errore possibile è zero. Quindi, il modello relativo al cambiamento del pozzo non risulta particolarmente impressionante dal momento che presenta un tasso di errore del 38%, un semplice 4% in più rispetto a semplicemente supporre che tutte le persone cambieranno pozzo.

Questo tasso di errore non significa che il modello sia totalmente inutile—come i grafici hanno mostrato, il modello stimato presenta una buona capacità predittiva della probabilità di cambiare pozzo. Ma la maggior parte dei dati sono sono vicini al livello medio degli input (le distanze dal pozzo sicuro più vicino sono inferiori a 100 metri e i livelli di arsenico sono tra 0.5 e 1), e quindi la semplice previsione media, Pr(cambiare)=0.58 funziona abbastanza bene. Il modello è informativo vicino agli estremi, ma relativamente pochi dati di trovano in quella posizione e quindi complessivamente la capacità predittiva del modello non risulta elevata.

Devianza

Per i modelli di regressione logistica e altri modelli per dati discreti, non ha alcun senso calcolare la deviazione standard dei residui e R^2, essenzialmente per la stessa ragione per cui i modelli non sono stimati col metodo dei minimi quadrati—l'errore al quadrato non è matematicamente un misura ottimale dell'errore del modello. Invece, è una procedura standard usare la *devianza*, una misura di sintesi statistica del modello di adattamento, definita per la regressione logistica e altri modelli lineari generalizzati e può essere vista in analogia con la deviazione standard dei residui.

Per ora, bisogna conoscere le seguenti proprietà della devianza:

- La devianza è una misura dell'errore; più bassa è la devianza, migliore è l'adattamento ai dati.
- Se si aggiunge al modello una variabile puramente casuale come predittore, ci si aspetta che la devianza diminuisca in media di 1.
- Quando si aggiunge al modello un predittore informativo, ci si aspetta che la devianza diminuisca più di 1. Quando k predittori sono aggiunti al modello, ci si aspetta che la devianza diminuisca di più di k. In generale, è difficile interpretare il livello assoluto della devianza; piuttosto la si usa per confrontare l'adattamento di due modelli.

Per i modelli classici (non-multilevel), la devianza viene definita come -2 volte il logaritmo della funzione di verosimiglianza (a meno di una costante arbitraria

additiva che, dal momento che si confrontano sempre devianze, non vengono mai valutate da sole).

Per esempio, i risultati dell'adattamento del primo modello relativo al cambiamento del pozzo, a pagina 120, riportano una "null deviance" pari a 4118.1 e una "residual deviance" pari a 4076.2. La devianza nulla ("null deviance") corrisponde al modello nullo, con il solo termine costante. Quindi aggiungendo la distanza (`dist`) come predittore nel modello, la devianza risulta diminuita di 41.9. Questo è molto di più della diminuzione attesa di 1 che si avrebbe se il predittore fosse una variabile puramente causale, quindi il predittore ha chiaramente migliorato l'adattamento.

Il modello successivo usa invece `dist100 = dist/100` come predittore. La devianza resta al valore di 4076.2, in quanto trasformazioni lineari non hanno effetti sulle previsioni nei modelli classici della regressione. (Vedremo poi che, comunque, le trasformazioni lineari possono comportare delle differenze nei modelli gerarchici.)

Quando aggiungiamo l'arsenico (`arsenic`) come predittore, la devianza diminuisce di 145.5, raggiungendo un valore pari a 3930.7, ancora una volta superiore al decremento atteso di 1 che si avrebbe avuto se il predittore fosse stato puramente casuale, quindi l'aggiunta di questo predittore ha chiaramente migliorato l'adattamento.

Il modello seguente include l'interazione tra la distanza (`dist`) e l'arsenico (`arsenic`) e riporta una devianza residua di 3927.6, una diminuzione di 3.1 dal precedente modello, solo un pò di più di quanto atteso nel caso di predittore puramente casuale. Questo decremento nella devianza non è statisticamente significativo (in quanto il coefficiente del predittore aggiunto non è inferiore a zero di due volte più o meno l'errore standard) ma, come discusso nella Sezione 5.5, teniamo l'interazione nel modello in quanto questa interazione ha senso nel contesto.

L'aggiunta dei due predittori di tipo sociale nel modello, `assoc` e `educ`, fanno diminuire la devianza a 3905.4, il che implica una migliore previsione rispetto a tutti i modelli. Se rimuoviamo il predittore `assoc` la devianza aumenta di un piccolo ammontare, risultando pari a 3907.9. L'aggiunta della interazione del livello di istruzione con la distanza e con il livello di arsenico riduce la devianza di un pò di più, e si ha un valore pari a 3891.7.

La trasformazione dell'arsenico su scala logaritmica—ovvero si rimuove `arsenic` dal modello e lo si sostituisce con `log.arsenic`, porta ad una diminuzione della devianza a 3863.1, che comporta un grande miglioramento.

Per i modelli *multilevel*, la devianza viene generalizzata ottenendo il criterio di informazione della devianza (deviance information criterion—DIC), come viene descritto nella Sezione 24.3.

5.7 Confronto medio predittivo su scala probabilistica

Come illustrato, per esempio, nella Figura 5.11 a pagina 127, le regressioni logistiche non sono lineari su scala probabilistica—ovvero una data differenza in una delle variabili x *non* corrisponde ad una differenza costante nella $\Pr(y=1)$. Di conseguenza, i coefficienti della regressione logistica non possono essere direttamente interpretati sulla scala dei dati. Le regressioni logistiche sono intrinsecamente più difficili da interpretare delle regressioni lineari.

Grafici come quelli nella Figura 5.11 sono utili, ma per modelli con tanti predittori o, quando la rappresentazione grafica non risulta conveniente, risulta utile avere una sintesi, simile a quella che avevamo nel modello lineare, che fornisca la differenza media, o attesa, della $\Pr(y=1)$ corrispondente ad una differenza unitaria in ciascuna variabile di input.

Esempio: pozzi in Bangladesh

Per un modello con linearità o interazioni, o entrambe, questo *confronto medio predittivo* dipende dai valori delle variabili di input, come illustreremo nell'esempio relativo al cambiamento del pozzo in Bangladesh. Per cercare di avere per il momento una semplice rappresentazione, consideriamo un modello senza interazione,

R code

```
fit.10 <- glm (switch ~ dist100 + arsenic + educ4,
  family=binomial(link="logit"))
```

che fornisce,

R output

```
            coef.est coef.se
(Intercept)    -0.21    0.09
dist100        -0.90    0.10
arsenic         0.47    0.04
educ4           0.17    0.04
  n = 3020, k = 4
  residual deviance = 3910.4, null deviance = 4118.1 (difference = 207.7)
```

considerando la probabilità di cambiare pozzo come una funzione della distanza dal pozzo sicuro più vicino (espressa in 100 metri), del livello di arsenico e dal livello di istruzione (espressa in termini di 4 anni).

Differenza media predittiva nella probabilità di cambiare pozzo, confronto tra famiglie che vivono vicino al pozzo sicuro più vicino e famiglie che distano dal pozzo sicuro più vicino più di 100 metri. Confrontiamo le due famiglie—una con `dist100` = 0 e una con `dist100` = 1—ma identiche in base alle altre variabili di input, `arsenic` e `educ4`. La *differenza predittiva* nella probabilità di cambiare pozzo in queste due famiglie risulta pari a

$$\begin{aligned}\delta(\text{arsenic}, \text{educ4}) = &\text{logit}^{-1}(-0.21 - 0.90 \cdot 1 + 0.47 \cdot \text{arsenic} + 0.17 \cdot \text{educ4}) - \\ &\text{logit}^{-1}(-0.21 - 0.90 \cdot 0 + 0.47 \cdot \text{arsenic} + 0.17 \cdot \text{educ4}). \quad (5.6)\end{aligned}$$

Abbiamo scritto δ in funzione dell'arsenico (`arsenic`) e dell'istruzione (`educ4`) per enfatizzare che dipende dai livelli di queste altre due variabili.

Prendiamo la media di queste due differenze predittive sulle n famiglie nei dati otteniamo,

$$\text{differenza media predittiva: } = \frac{1}{n}\sum_{i=1}^{n}\delta(\text{arsenic}_i, \text{educ4}_i). \quad (5.7)$$

In R:

R code

```
b <- coef (fit.10)
hi <- 1
lo <- 0
delta <- invlogit (b[1] + b[2]*hi + b[3]*arsenic + b[4]*educ4) -
         invlogit (b[1] + b[2]*lo + b[3]*arsenic + b[4]*educ4)
print (mean(delta))
```

Il risultato, pari a -0.20, implica che, le famiglie che sono lontane 100 metri dal pozzo sicuro più vicino hanno una probabilità di cambiare pozzo inferiore del 20% rispetto alle famiglie che vivono in prossimità del pozzo sicuro più vicino, fissato il livello di arsenico e il livello di istruzione.

Differenza media predittiva nella probabilità di cambiare pozzo, confronto tra famiglie il cui pozzo ha un livello di arsenico pari a 0.5 e 1.0 In modo simile, calcoliamo le differenze predittive, quindi ne facciamo la media e confrontiamo le famiglie in corrispondenza di due livelli di arsenico differenti, assumendo che la distanza dal pozzo sicuro più vicino sia la stessa per entrambe le famiglie e assumiamo anche stesso livello di istruzione. Scegliamo come livelli di arsenico

`arsenic` = 0.5 e 1.0; la scelta risiede nel fatto che 0.5 rappresenta il livello più basso per un pozzo insicuro, 1.0 è due volte questo valore, e questo confronto coglie la maggior parte del *range* dei dati (Si veda la Figura 5.10 a pagina 125). Questi sono i calcoli in R:

R code

```
hi <- 1.0
lo <- 0.5
delta <- invlogit (b[1] + b[2]*dist100 + b[3]*hi + b[4]*educ4) -
         invlogit (b[1] + b[2]*dist100 + b[3]*lo + b[4]*educ4)
print (mean(delta))
```

Il risultato, pari a 0.06—ci dice che questo confronto corrisponde ad una differenza di probabilità di cambiare pozzo pari al 6%.

Differenza media predittiva nella probabilità di cambiare pozzo, confronto tra famiglie con 0 e 12 anni di istruzione. In modo del tutto simile calcoliamo ora la media della differenza predittiva della probabilità di cambiare pozzo per famiglie il cui capo famiglia ha rispettivamente 0 e 12 anni di istruzione (ovvero confrontiamo `educ4` = 0 con `educ4` = 3):

R code

```
hi <- 3
lo <- 0
delta <- invlogit (b[1]+b[2]*dist100+b[3]*arsenic+b[4]*hi) -
         invlogit (b[1]+b[2]*dist100+b[3]*arsenic+b[4]*lo)
print (mean(delta))
```

il cui risultato risulta pari a 0.12.

Confronto medio predittivo in presenza di interazioni

Possiamo fare dei calcoli simili per modelli con interazioni. Per esempio, consideriamo la differenza media predittiva confrontando `dist` = 0 con `dist` = 100, per il modello che considera l'interazione distanza × arsenico:

R code

```
fit.11 <- glm (switch ~ dist100 + arsenic + educ4 + dist100:arsenic,
  family=binomial(link="logit"))
```

che fornisce come risultato,

R output

```
                coef.est coef.se
(Intercept)        -0.35    0.13
dist100            -0.60    0.21
arsenic             0.56    0.07
educ4               0.17    0.04
dist100:arsenic    -0.16    0.10
  n = 3020, k = 5
  residual deviance = 3907.9, null deviance = 4118.1 (difference = 210.2)
```

Questi sono i comandi in R per calcolare le differenze medie predittive quando si confronta `dist1` $= 1$ con `dist1` $= 0$:

R code

```
b <- coef (fit.11)
hi <- 1
lo <- 0
delta <- invlogit (b[1] + b[2]*hi + b[3]*arsenic + b[4]*educ4 +
                   b[5]*hi*arsenic) -
         invlogit (b[1] + b[2]*lo + b[3]*arsenic + b[4]*educ4 +
                   b[5]*lo*arsenic)
print (mean(delta))
```

che fornisce -0.19.

Notazioni generali per confronti predittivi

Consideriamo un input alla volta, e usiamo la notazione u per *l'input di interesse* e v per il vettore di tutti gli altri input. Supponiamo di considerare il confronto tra $u = u^{(1)}$ e $u = u^{(0)}$ tendendo costanti tutti gli altri input (per esempio, abbiamo considerato il confronto tra famiglie che vivono a 0 metri dal pozzo sicuro più vicino rispetto a famiglie che vivono a 100 metri di distanza). Le *differenza predittiva* in termini di probabilità tra i due casi, essendoci una differenza solo in u, sarà

$$\delta(u^{(\text{hi})}, u^{(\text{lo})}, v, \beta) \;=\; \Pr(y{=}1|u^{(\text{hi})}, v, \beta) - \Pr(y{=}1|u^{(\text{lo})}, v, \beta), \tag{5.8}$$

dove la barra verticale nell'espressione di sopra va letta come "condizionatamente a" (per esempio, probabilità che $y = 1$ dato $u^{(\text{hi})}$, v, e β).

La differenza media predittiva si ottiene quindi facendo la media rispetto agli n punti nel dataset usato per la stima del modello di regressione logistico:

$$\Delta(u^{(\text{hi})}, u^{(\text{lo})}) = \frac{1}{n}\sum_{i=1}^{n}\delta(u^{(\text{hi})}, u^{(\text{lo})}, v_i, \beta), \tag{5.9}$$

dove v_i rappresenta il vettore degli altri input(nel nostro esempio, l'arsenico e il livello di istruzione) per ciascuna osservazione i. Queste espressioni generalizzano le formule (5.6) e (5.7).

Per modelli con interazione, la formula per la differenza predittiva (5.8) deve essere calcolata con particolare attenzione, con la consapevolezza di quali siano gli input che entrano nella regressione. La distinzione qui tra variabili di input (in questo caso, distanza, arsenico ed istruzione) predittori (termine costante, distanza, arsenico ed istruzione e interazione distanza $\times$ arsenico) risulta cruciale. Discuteremo del confronto medio predittivo in presenza di interazione successivamente nella Sezione 21.4.

5.8 Identificabilità e separazione

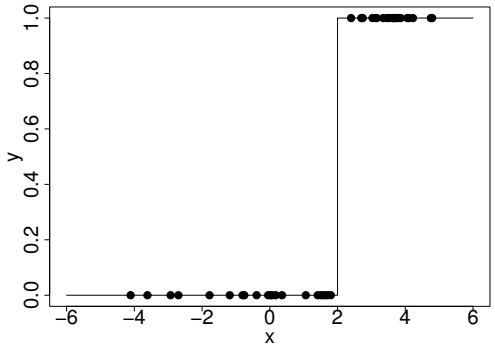
Ci sono due motivi per cui un modello di regressione logistica può non essere identificato (questo significa che i parametri non possono essere stimati dal modello in base ai dati disponibili, come abbiamo già discusso nella Sezione 4.5 nel contesto della regressione lineare):

1. Come nella regressione lineare, se i predittori sono collineari, allora la stima del predittore lineare, $X\beta$, non permette una stima separata per i parametri individuali β. Possiamo analizzare questa forma di non identificabilità in modo del tutto simile a quanto fatto per la regressione lineare, come descritto nella Sezione 4.5.

2. Un problema di identificabilità completamente diverso, chiamato *separazione*, può essere una conseguenza della discretizzazione dei dati.

   - Se un predittore x_j risulta completamente allineato rispetto all'outcome, ovvero $y = 1$ in tutti quei casi in cui x_j supera una data soglia T, e $y = 0$ in tutti i casi in cui $x_j < T$, allora la miglior stima per il coefficiente β_j risulta pari a ∞. La Figura 5.16 mostra un esempio. L'esercizio 11 fornisce un esempio con un predittore lineare.
   - Al contrario, se $y = 1$ per tutti i casi in cui $x_j < T$, e $y = 0$ per tutti i casi in cui $x_j > T$, allora $\hat{\beta}_j$ sarà $-\infty$.

Figura 5.16: *Esempio di dati per cui un modello di regressione logistica risulta non identificabile. L'outcome y è uguale a 0 per tutti quei dati che sono minori di $x = 2$ e pari a 1 in corrispondenza dei dati tali che $x = 2$, quindi il miglior adattamento della linea di regression logistica risulta $y = logit^{-1}(\infty(x - 2))$, che ha una pendenza infinita in $x = 2$.*



- Più in generale, questo problema capita se una combinazione lineare di predittori risulta perfettamente allineata con l'outcome. Per esempio, supponiamo che $7x_1 + x_2 - 3x_3$ sia completamente e positivamente allineato con i dati, con $y = 1$ se e solo se questa combinazione lineare eccede una qualche soglia. Quindi la combinazione lineare $7\hat{\beta}_1 + \hat{\beta}_2 - 3\hat{\beta}_3$ avrà una stima ∞, il che comporterà che almeno uno di tre coefficienti $\beta_1, \beta_2, \beta_3$ avrà una stima pari a ∞ o $-\infty$.

Un modo per analizzare la separazione è quello di utilizzare un approccio di tipo Bayesiano o di verosimiglianza penalizzato (implementato in R nel pacchetto `brlr`) che fornisce un piccolo ammontare di informazione su tutti i coefficienti della regressione, inclusi quelli che non possono essere identificati solamente in base ai dati. (Si veda il Capitolo 18 per approfondimenti sull'inferenza Bayesiana.)

5.9 Nota bibliografica

Secondo Cramer (2003, capitolo 9), la regressione logistica è stata introdotta per i dati binari intorno alla metà del ventesimo secolo ed è diventata sempre più popolare soprattutto quando i progressi tecnologici e computazionali hanno permesso lo sviluppo di strumenti di routine per l'analisi dei dati.

Per saperne di più sul reddito e i voti nelle elezioni presidenziali si veda Gelman, Shor et al. (2005). L'esempio dell'acqua da bere in Bangladesh è descritto in modo approfondito in van Geen et al. (2003) e in Gelman, Trevisani, et al. (2004).

I grafici dei residui "binned" e i relativi strumenti per la verifica del modello logistico stimato sono discussi in Landwehr, Pregibon and Shoemaker(1984), da Gelman, Goegebeur, et al. (2000), da Pardoe and Cook (2002), e da Pardoe (2004).

La devianza è discussa da McCullagh e Nelder (1989); idee ad essa legate includono il criterio di informazione di Akaike (AIC) Akaike, (1973), C_p (Mallows, 1973) e il criterio di informazione della devianza (DIC; Spiegelhalter et al., 2002). Si veda anche Fox per una panoramica applicata e Gelman et al. (2003, sezioni 6.7–6.8) per una prospettiva Bayesiana.

La non identificabilità del modello di regressione logistico e la separazione nei dati discreti sono discussi da Albert (1984), da Lesaffre e Albert (1989), da Heinze e Schemper (2003), e nel libro di Agresti (2002). Zorn, propone una soluzione di tipo Bayesiano seguendo Firth (1993).

5.10 Esercizi

1. La cartella `nes` contiene i dati dell'indagine delle preferenze elettorali e del reddito per le elezioni relative al 1992 analizzate nella Sezione 5.1, insieme ad altre variabili tra cui sesso, etnicità, istruzione, identificazione con il partito e ideologia politica.

   (a) Stimate un modello di regressione logistica per prevedere il sostegno a Bush dati tutti questi input. Si consideri come includere questi input come predittori della regressione e si consideri anche la possibilità di includere eventuali interazioni.

   (b) Si valutino e si confrontino i diversi modelli stimati. Si considerino i coefficienti stimati e i rispettivi errori standard, i grafici dei residui e le devianze.

   (c) relativamente al modello scelto, si discuta e si confronti l'importanza di ciascuna variabile di input nella previsione.

2. Senza usare un computer, disegnare le seguenti curve di regressione logistica:

   (a) $\Pr(y = 1) = \text{logit}^{-1}(x)$

   (b) $\Pr(y = 1) = \text{logit}^{-1}(2 + x)$

   (c) $\Pr(y = 1) = \text{logit}^{-1}(2x)$

(d) $\Pr(y = 1) = \text{logit}^{-1}(2 + 2x)$

(e) $\Pr(y = 1) = \text{logit}^{-1}(-2x)$

3. Siamo interessati a capire in che modo i guadagni dei genitori in una famiglia con figli possano prevedere il livello di educazione dei figli. Ci viene detto che la probabilità che un ragazzo ottenga la licenza superiore è pari al 27% per coloro i cui genitori non hanno nessuna fonte di reddito e cresce all'88% per i ragazzi con genitori che guadagnano \$60,000. Si determini il modello di regressione logistica che risulta consistente con questa informazione. (Per semplicità si assuma che il reddito sia misurato in decine di migliaia di dollari).

4. Si effettui una regressione logistica per un problema che interessa. Questo potrebbe essere fatto su dati relativi a un progetto di ricerca, a corsi precedenti o a dati che si prendono sul web. Scegliete una variabile come outcome, che assume i valori 0 e 1 (dal momento che si sta facendo una regressione logistica).

   (a) Analizzare i dati in R. Usare la funzione `display()` per riassumere i risultati.

   (b) Si stimino diverse versioni del modello. Si provi ad includere diversi predittori, interazioni e trasformazioni degli input.

   (c) Una volta scelta una particolare formulazione del modello si effettuino le seguenti operazioni:

      i. Si descriva come ciascun input influenzi la $\Pr(y = 1)$ nel modello stimato. Si devono considerare le stime dei coefficienti, il campo di variazione degli input e la funzione non lineare logit inversa.

      ii. Quale è il tasso di errore del modello stimato?

      iii. Si considerino le devianze del modello stimato e del modello nullo. Il miglioramento nella stima del modello è effettivo?

      iv. Si usi il modello per fare previsioni per alcuni casi di interesse.

5. In una classe di 50 studenti, viene effettuata una regressione logistica sul superamento del corso (superare l'esame o meno) in funzione del voto ottenuto in un esame intermedio (valori continui con media 60 e deviazione standard 15). Il modello stimato è pari a $\Pr(\text{pass}) = \text{logit}^{-1}(-24 + 0.4x)$.

   (a) Rappresentate graficamente il modello. Inserite in questo grafico uno scatterplot di dati ipotetici coerenti con le trasformazioni date.

   (b) Supponete che i punteggi dell'esame intermedio abbiano una media pari a 0 e una deviazione standard pari a 1. Quale sarebbe l'equazione della della regressione logistica usando questi punteggi trasformati come predittori?

(c) Create un nuovo predittore in modo tale che sia una variabile puramente casuale (per esempio, in R si può creare `newpred <- rnorm (n,0,1)` . Si aggiunga questo predittore al modello. Di quanto diminuisce la devianza?

6. Formulazione del modello logistico in termini di dati latenti: si consideri il modello $\Pr(y = 1) = \text{logit}^{-1}(1 + 2x_1 + 3x_2)$ e si consideri una persona per cui $x_1 = 1$ e $x_2 = 0.5$. Si disegni la distribuzione dei dati latenti per questa persona. Si calcoli la probabilità che $y = 1$ per questa persona e si ombreggi l'area corrispondente nel grafico.

7. Limiti della regressione logistica: si consideri un dataset con $n = 20$ punti, un singolo predittore x che assume i valori $1, \ldots, 20$, e una variabile binaria y. Si costruiscano i valori $y_1, \ldots, y_{20}$ in modo tale che siano inconsistenti con ogni modello logistico in funzione di x. Si stimi una regressione logistica con questi dati, si rappresentino i dati e la curva stimata, e si spieghi il motivo per cui diciamo che questo modello non stima adeguatamente i dati.

8. Si costruisca un modello di regressione logistica: la cartella `rodents` contiene dati sui roditori in un campione di appartementi di New York City.

   (a) Si costruisca un modello di regressione per prevedere la presenza di roditori (la variabile `rodent2` nel dataset) dati gli indicatori per diversi gruppi etnici (`race`). Si combinino le categorie in modo appropriato. Si discutano i coefficienti stimati del modello.

   (b) Si aggiunga al modello altri potenziali rilevanti predittori che descrivono l'appartamento, il palazzo e il distretto dove si vive. Si costruisca il modello usando i principi generali che sono stati spiegati nella Sezione 4.6. Si discutano i coefficienti associati ai diversi gruppi etnici.

9. Rappresentazione della regressione logistica: i dati dell'esempio sui pozzi in Bangladesh descritto nella Sezione 5.4 sono nella cartella `arsenic`.

   (a) Si consideri una regressione per stimare la probabilità di cambiare pozzo usando il logaritmo della distanza dal pozzo sicuro più vicino come predittore.

   (b) Si faccia un grafico simile a quello della Figura 5.9 che rappresenta la Pr(cambiare) in funzione della distanza dal pozzo sicuro più vicino e si rappresentino sul grafico anche i dati.

   (c) Si faccia un grafico dei residui e dei residui "binned" come nella Figura 5.13.

   (d) Si calcoli il tasso di errore del modello stimato e lo si confronti col tasso di errore del modello nullo.

(e) Si creino degli indicatori corrispondenti a `dist` < 100, $100 \leq$ `dist` < 200, e `dist` > 200. Si stimi una regressione logistica per la Pr(cambiare) usando questi indicatori. Con questo nuovo modello, si ripetano i calcoli e i grafici della parte (a) di questo esercizio.

10. Costruzione del modello e confronto: si continui con i dati dell'esempio dei pozzi descritto nell'esempio precedente.

    (a) Si consideri una regressione logistica per stimare la probabilità di cambiare pozzo usando, come predittori, la distanza, log(arsenico), e la loro interazione. Si interpretino i coefficienti stimati e i loro rispettivi errori standard.

    (b) Si facciano dei grafici simili a quelli della Figura 5.9 per far vedere la relazione tra la probabilità di cambiare pozzo, la distanza e il livello di arsenico.

    (c) Seguendo la procedura descritta nella Sezione 5.7, si calcolino le differenze medie predittive corrispondenti a:

        i. Confronto tra `dist` $= 0$ e `dist` $= 100$, con `arsenic` costante.
        ii. Confronto tra `dist` $= 100$ e `dist` $= 200$, con `arsenic` costante.
        iii. Confronto tra `arsenic` $= 0.5$ e `arsenic` $= 1.0$, con `dist` costante.
        iv. Confronto tra `arsenic` $= 1.0$ e `arsenic` $= 2.0$, con `dist` costante.

        Si discutano questi risultati.

11. Identificabilità: la cartella `nes` contiene i dati relativi allo Studio Elettorale Nazionale Americano usato nella Sezione 5.1 per modellare le preferenze di voto in base al reddito. Quando proviamo a stimare un modello simile considerando come predittore l'etnicità incontriamo un problema. Questi sono i risultati della stima per gli anni 1960, 1964, 1968, e 1972:

R output

```
glm(formula = vote ~ female + black + income,
  family=binomial(link="logit"), subset=(year==1960))
            coef.est coef.se
(Intercept) -0.14     0.23
female       0.24     0.14
black       -1.03     0.36
income       0.03     0.06

glm(formula = vote ~ female + black + income,
  family=binomial(link="logit"), subset=(year==1964))
```