Analisi dei dati con modelli di regressione e modelli multilivello/gerarchici

Andrew Gelman
Department of Statistics and Department of Political Science
Columbia University, New York

Jennifer Hill
School of International and Public Affairs
Columbia University, New York

First published in 2007 by Cambridge University Press

Edizione italiana a cura di:
Maria Grazia Pittau e Roberto Zelli,
Facoltà di Scienze Statistiche
Sapienza Università di Roma

Parte I

IA: Regressione non gerarchica.

Questa parte inizia con una panoramica sui modelli di regressione classici e generalizzati. Particolare enfasi viene posta sui principali aspetti pratici relativi all'adattamento e all'interpretazione del modello e all'analisi grafica dei risultati. Questa panoramica permetterà di introdurre il pacchetto statistico R per l'analisi della regressione.

Capitolo 1

Regressione lineare: concetti di base

La regressione lineare è un metodo che sintetizza come i valori medi di una variabile numerica, *outcome*, varino in funzione di sottopopolazioni definite da funzioni lineari di *predittori*.

I testi di base di statistica che trattano la regressione spesso focalizzano l'attenzione su come la regressione possa essere usata per rappresentare eventuali relazioni tra variabili, piuttosto che sul confronto dei valori medi. Focalizzando l'attenzione sulla regressione in termini di confronto di valori medi, noi saremo espliciti circa i suoi limiti per definire queste relazioni causalmente, oggetto su cui si ritornerà nel capitolo 9. La Regressione può essere usata per prevedere un *outcome* data una funzione lineare dei predittori, e i coefficienti della regressione possono essere visti sia in termini di confronto tra valori previsti o in termini di confronto tra valori medi nei dati.

1.1 Predittore singolo

All'interno di questo capitolo ci si concentrerà sui coefficienti senza preoccuparci, almeno per il momento, dei problemi relativi alla stima e all'incertezza. Si inizierà con l'adattare una serie di modelli di regressione per prevedere i punteggi di un test cognitivo valutato su un insieme di bambini di tre-quattro anni in base a determinate caratteristiche delle loro madri, sulla base di un'indagine condotta sulle donne adulte americane e sui loro figli (un sottocampione dell'Indagine Nazionale Longitudinale sui giovani). *Nel caso in cui si abbia un predittore binario, il coefficiente della regressione è la differenza tra le medie dei due gruppi.*

Iniziamo a modellare il punteggio relativo ad un test cognitivo su bambini (`kid.score`) sulla base di un predittore dicotomico (`mom.hs`) relativo al titolo di studio superiore delle loro madri, indicatore che assumerà valore pari a 1 se le madri possiedono un diploma di scuola superiore e valore pari a 0 se non lo possiedono.

Il modello stimato sarà:

$$\texttt{kid.score} = 78 + 12 \cdot \texttt{mom.hs} + \texttt{residuo}, \tag{1.1}$$

ma per il momento ci si concentrerà solo sulla parte deterministica,

$$\widehat{\texttt{kid.score}} = 78 + 12 \cdot \texttt{mom.hs}, \tag{1.2}$$

dove $\widehat{\texttt{kid.score}}$ denota il valore previsto o atteso del punteggio del test in funzione del singolo predittore `mom.hs`.

Questo modello sintetizza la differenza in media del punteggio del test ottenuto dai bambini le cui madri hanno completato le scuole superiori rispetto al punteggio ottenuto da quei bambini le cui madri non hanno completato gli studi superiori. La Figura 1.1 mostra come la linea di regressione passi attraverso la media delle due sottopopolazioni.

L'intercetta, 78, è il punteggio medio (o previsto) ottenuto dai bambini la cui madre non ha completato gli studi superiori. Per poter ottenere questo valore algebricamente, è sufficiente sostituire lo 0 al valore del predittore all'interno dell'equazione. Per poter ottenere invece il punteggio medio ottenuto dai bambini le cui madri possiedono un titolo di studio superiore (o il valore previsto per un singolo bambino), è sufficiente sostituire, all'interno dell'equazione, il valore 1 al predittore, ottenendo $78 + 12 \cdot 1 = 90$.

La differenza tra le medie di queste due sottopopolazioni è uguale al valore del coefficiente del predittore `mom.hs`. Questo coefficiente ci dice che i bambini le cui madri hanno completato gli studi superiori hanno ottenuto un punteggio medio superiore di 12 punti rispetto al punteggio ottenuto dai bambini le cui madri non hanno completato gli studi.

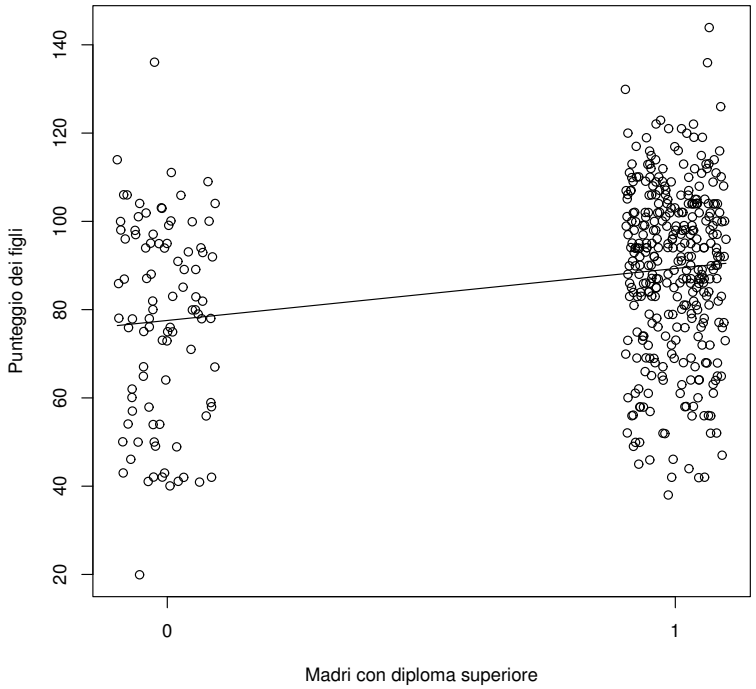


Figura 1.1: *Punteggio di un test effettuato su un campione di bambini in funzione di un indicatore che esprime se le madri dei bambini possiedono o meno un titolo di studio superiore. La linea di regressione stimata passa attraverso la media di ognuna delle due sottopopolazioni definite dal livello di studio delle madri. La variabile indicatrice relativa al titolo di studio delle madri è stata sottoposta ad un procedura cosiddetta di* jittered*; procedura che consiste nell'aggiungere a ciascun valore della variabile un numero casuale in modo tale che i singoli punti non si sovrappongano.*

Regressione con un predittore continuo

Se si regredisce invece in funzione di una variabile continua, in particolare rispetto al punteggio ottenuto dalle madri sottoposte ad un test di valutazione del Quoziente Intellettivo, IQ, il modello stimato diventa

$$\texttt{kid.score} = 26 + 0.6 \cdot \texttt{mom.iq} + \texttt{residuo}, \tag{1.3}$$

ed è riportato in Figura 1.2. I punti sulla linea di regressione si possono interpretare sia come valori previsti del punteggio del test ottenuti dai bambini ad ogni livello di IQ ottenuto dalle madri, sia come punteggio medio delle popolazioni definite da questi stessi punteggi.

Confrontando i punteggi medi del test sui bambini nelle due sottopopolazioni che differiscono di 1 punto in base al Quoziente Intellettivo ottenuto dalle madri ci si aspetta che il gruppo di bambini le cui madri presentano un IQ più elevato raggiungo un punteggio sul test che in media risulta più elevato di 0.6 punti. Ancora più interessante potrebbe essere il confronto tra i gruppi di bambini le cui madri differiscono di 10 punti nel IQ: in questo caso i bambini le cui madri presentano un IQ più consistente hanno un punteggio medio atteso più elevato di 6 punti.

Per capire al meglio il temine costante nella regressione si consideri il caso in cui tutti i predittori assumano valore pari a 0. In questo esempio, il valore dell'intercetta uguale a 26 riflette il valore previsto del test sui bambini le cui madri presentano un valore di IQ pari a 0. Questo non è ovviamente l'esempio più significativo, in quanto non si osserverà mai un Quoziente di Intelligenza nullo. Si discuterà nella prossima sezione di una semplice trasformazione che fornisce una interpretazione più ragionevole della intercetta.

1.2 Predittori multipli

I coefficienti di un modello di regressione con più di un predittore sono più complicati da interpretare in quanto l'interpretazione di ciascun coefficiente è, in parte, dipendente dalle altre variabili presenti nel modello. Un tipico suggerimento è quello di interpretare ciascun coefficiente "tenendo gli altri predittori costanti". Illustreremo questo concetto con un esempio in cui la semplice interpretazione dei coefficienti di regressione non funziona.

Si consideri ad esempio una regressione lineare per prevedere il punteggio di un test su bambini in funzione dell'istruzione delle madri e del punteggio raggiunto sempre dalle madri in un test di tipo IQ. Il modello stimato è il seguente:

$$\texttt{kid.score} = 26 + 6 \cdot \texttt{mom.hs} + 0.6 \cdot \texttt{mom.iq} + \texttt{residuo}, \tag{1.4}$$

ed è mostrato in Figura 1.3. Questo modello impone che la pendenza del modello di regressione del punteggio del test sui bambini sul punteggio IQ delle madri sia lo stesso per i sottogruppi determinati dal livello di istruzione delle madri. La sezione successiva considera modelli le cui pendenze variano a seconda dei gruppi. Ma prima di tutto, iniziamo ad interpretare i coefficienti del modello (1.4):

1. *L'intercetta.* Se si considera un bambino la cui madre presenta un IQ pari a 0 e che non ha completato gli studi superiori (ovvero `mom.hs`=0), allora il punteggio previsto ottenuto dal bambino nel test sarà pari a 26. Questo esempio non rappresenta una previsione ragionevole dal momento che nessuna madre avrà un Quoziente Intellettivo nullo.

2. *Il coefficiente relativo alla licenza superiore.* Se si confrontano bambini le cui madri hanno le stesso Quoziente Intellettivo ma che differiscono in base al fatto che hanno o meno terminato gli studi superiori, allora il modello stima una differenza attesa di 6 punti nei punteggi ottenuti dai due sottogruppi di bambini.

3. *Il coefficiente relativo al Quoziente Intellettivo (IQ) materno.* Se si confrontano invece bambini con lo stesso valore di `mom.hs` ma le cui madri differiscono di 1 punto nel loro IQ, allora ci si aspetta una differenza di 0.6 punti nel punteggio del test ottenuto dai bambini (ovvero una differenza di 10 punti nel punteggio IQ delle madri corrisponde ad una differenza di 6 punti nel punteggio ottenuto dalle due sottopopolazioni di bambini).

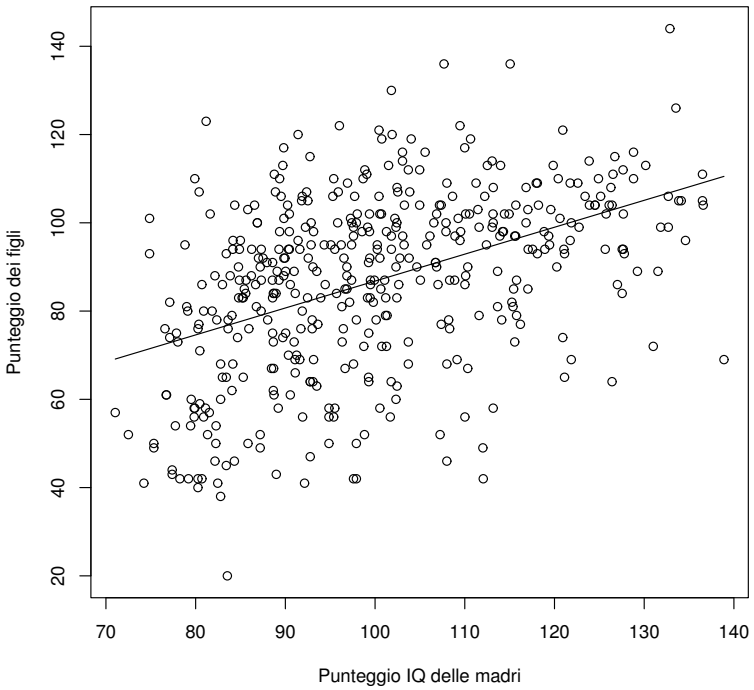


Figura 1.2: *Rappresentazione grafica dei punteggi relativi a un test effettuato su un campione di bambini in funzione del quoziente intellettivo (IQ) delle madri e relativa curva di regressione. Ciascun punto sulla linea può essere interpretato sia come un valore previsto del test effettuato sui bambini in corrispondenza del relativo punteggio IQ delle madri, sia come un punteggio medio relativo ad una sottopopolazione di bambini le cui madri presentano i valori misurati di IQ.*

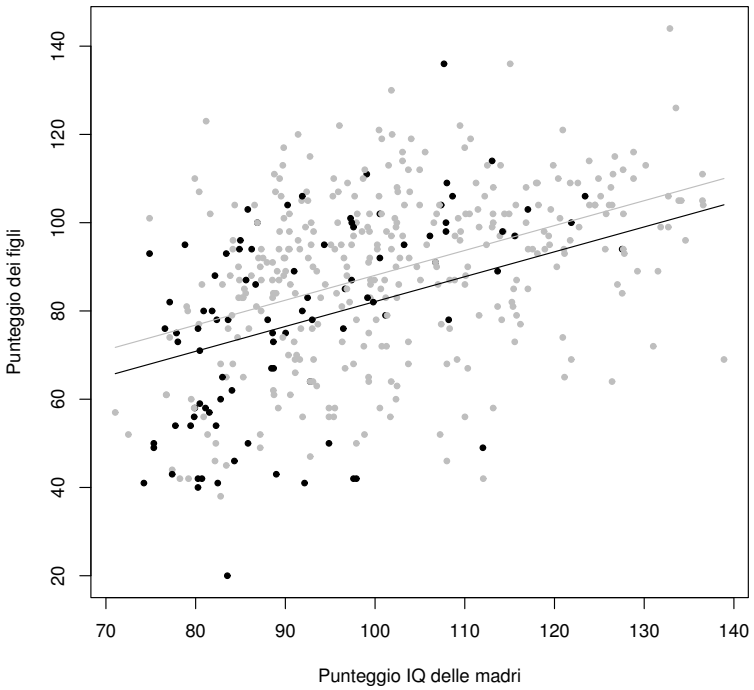


Figura 1.3: *Rappresentazione grafica dei punteggi relativi a un test effettuato su un campione di bambini. I punti in chiaro rappresentano i bambini le cui madri possiedono una licenza superiore mentre i punti in nero i bambini le cui madri hanno un titolo di studio inferiore. Le due linee rappresentano le curve di regressione del punteggio del test sui bambini in funzione del titolo di studio delle madri e del loro punteggio ottenuto nell'indicatore IQ. La linea chiara rappresenta la regressione relativa alla sottopopolazione delle madri che hanno completato la scuola superiore, mentre quella scura quella relativa alla sottopopolazione di bambini le cui madri non hanno un titolo di studio superiore.*

Non sempre è possibile far variare un predittore tenendo costanti gli altri

I coefficienti della regressione vengono interpretati in termini di confronto tra individui che differiscono in un predittore tenendo fissi gli altri. In alcuni casi, è possibile manipolare i predittori per modificarne alcuni tenendo costanti altri. Tuttavia questa interpretazione non è strettamente necessaria. Questo concetto diventerà più chiaro nel seguito quando verranno presentate delle situazioni in cui risulta logicamente impossibile cambiare il valore di un predittore tenendo costanti i valori assunti dagli altri predittori. Per esempio, se un modello includesse come predittori sia IQ che IQ^2 come predittori, non avrebbe nessun senso considerare e valutare cambiamenti in IQ tenendo costante IQ^2. Oppure, come verrà discusso nella sezione successiva, se un modello include `mom.hs` e `mom.IQ` e anche la loro interazione `mom.hs*mom.IQ` non ha nessun senso uno di questi tre predittori tenendo costanti gli altri.

Interpretazioni predittive e controfattuali

In un contesto più generale di regressione multipla, è necessario essere più espliciti nell'interpretazione dei coefficienti. In particolare, è possibile distinguere tra due diverse interpretazioni dei coefficienti di un modello di regressione.

- *Interpretazione di tipo predittivo* considera come il risultato della variabile risposta vari, in media, quando confrontando due gruppi di unità che differiscono di un valore pari a 1 nel predittore più rilevante tenendo costanti tutti gli altri predittori. Nel caso del modello lineare, il coefficiente è la differenza attesa nella variabile y tra queste due unità. E questa è l'interpretazione che abbiamo finora considerato.

- *Interpretazione controfattuale* considera invece cambiamenti all'interno degli individui piuttosto che confrontare gruppi di individui. Di conseguenza, il coefficiente rappresenta il cambiamento atteso nella variabile y causato dall'aggiunta di 1 nel predittore più rilevante e lasciando gli altri predittori invariati. Per esempio, "l'incremento del valore del Quoziente Intellettivo materno da 100 a 101 porterebbe ad un incremento atteso pari a 0.6 nel punteggio del test sui bambini". Questo tipo di interpretazione è propria dell'inferenza causale.

La maggior parte dei libri di statistica di base sulla regressione mettono in guardia rispetto a quest'ultima interpretazione nonostante ammettano interpretazioni simili quali "un cambiamento di 10 nel test IQ delle madri è associato ad un cambiamento di 6 punti nel punteggio ottenuto dai figli". Questa interpretazione, che è appunto l'interpretazione controfattuale è probabilmente la più familiare e, a volte, la più facile da capire. Comunque, come verrà discusso nel Capitolo 9, l'interpretazione controfattuale può risultare talvolta inappropriata, a meno di ipotesi molto forti.

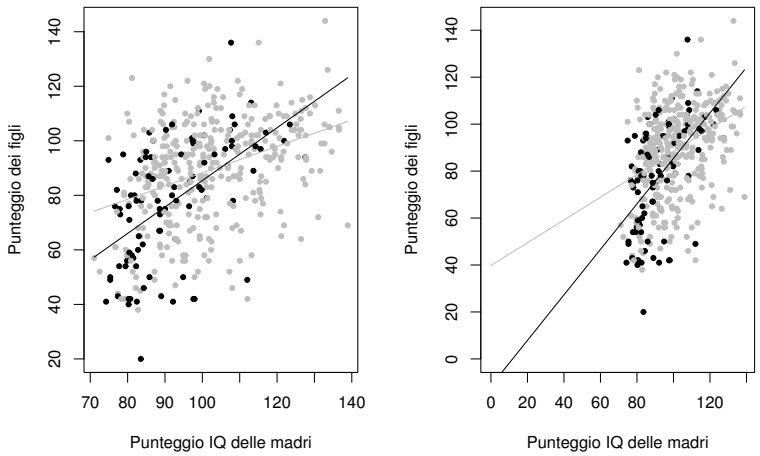


Figura 1.4: *(a) Linee di regressione del test attitudinale sui bambini in base al punteggio IQ materno. I diversi simboli rappresentano i bambini le cui madri hanno completato le scuole superiori (cerchi chiari) e non (cerchi scuri). L'interazione permette di avere diverse pendenze nei due gruppi, con linee chiare e scure corrispondenti ai punti chiari e scuri. (b) Stesso grafico ma l'asse orizzontale viene esteso a partire dallo zero per identificare le intercette delle linee.*

1.3 Interazioni

Nel modello (1.4), si è imposto che la pendenza della curva di regressione del punteggio del test sui bambini in funzione del Quoziente Intellettivo materno fosse la stessa tra i due sottogruppi di bambini definiti in base al completamento o meno degli studi superiori da parte delle madri. Tuttavia, un'ispezione dei dati riportata nella Figura 1.3 suggerisce che le pendenze differiscono sensibilmente. Un possibile rimedio è la possibilità di includere nel modello un'*interazione* tra `mom.hs` e `mom.IQ`, dando luogo ad un nuovo predittore definito dal prodotto di queste due variabili. L'introduzione di questo nuovo predittore nel modello permette alla pendenza di variare tra i due gruppi. Il nuovo modello stimato risulta pari a:

$$\texttt{kid.score} = -11 + 51 \cdot \texttt{mom.hs} + 1.1 \cdot \texttt{mom.iq} - 0.5 \cdot \texttt{mom.hs} \cdot \text{mom.iq} + \texttt{residuo}$$

ed è rappresentato graficamente nella Figura 1.4a, in cui sono presenti due linee di regressione separate per ognuno dei sottogruppi individuati in base all'istruzione materna.

La Figura 1.4b mostra la linea di regressione su una scala il cui asse delle x è stato esteso fino allo zero per evidenziare le intercette – i punti sull'asse delle y in corrispondenza di $x = 0$. Questo evidenzia il fatto che questi valori non solo sono privi di significato in termini di interpretazione, ma sono anche talmente lontani

dal campo di variazione dei nostri dati da non poter essere considerati realistici in termini di stima delle sottopopolazioni.

Particolare attenzione deve essere posta nell'interpretazione dei risultati di questo modello. In genere, si determina il significato dei coefficienti (o, talvolta, funzioni dei coefficienti) esaminando la media o i valori previsti del punteggio del test all'interno e tra specifici sottogruppi. Alcuni coefficienti sono interpretabili solo per certi sottogruppi.

1. *Intercetta* rappresenta il valore previsto del punteggio del test per i bambini le cui madri non hanno completato gli studi superiori e hanno un valore dell'IQ pari a 0 – il che non rappresenta uno scenario significativo. (Come verrà discusso nelle Sezioni 4.1–4.2, le intercette possono essere meglio interpretate se la variabili input sono centrate prima che vengano incluse come predittori nel modello)

2. *Il coefficiente di* `mom.hs` può essere considerato come la differenza tra il valore del test riportato dai bambini le cui madri non hanno completato gli studi secondari e con IQ pari a 0 il punteggio ottenuto dai bambini le cui madri hanno un titolo di studio superiore e a parità di punteggio IQ. Questo può essere facilmente verificato sostituendo i numeri appropriati e confrontando le equazioni. Dal momento che non è plausibile immaginare madri con Quoziente Intellettivo pari a zero, questo coefficiente non è facilmente interpretabile.

3. *Il coefficiente di* `mom.iq` può essere visto come il confronto dei punteggi medi del test tra i bambini le cui madri non hanno completato gli studi superiori, ma le cui madri differiscono di 1 punto nell' IQ. Questo coefficiente altro non è che la pendenza della linea più scura nella Figura 1.4.

4. *Il coefficiente del termine di interazione* rappresenta invece la *differenza* nella pendenza per `mom.iq`, confrontando i bambini le cui madri hanno o meno completato gli studi superiori: rappresenta quindi la differenza tra le due linee di regressione, la più chiara e la più scura nella Figura1.4.

Un modo equivalente per capire il modello è guardare alle linee di regressione separatamente per i bambini le cui madri hanno completato o meno gli studi superiori:

$$\begin{aligned}
\text{no hs: } \texttt{kid.score} &= -11 + 51 \cdot 0 + 1.1 \cdot \texttt{mom.iq} - 0.5 \cdot 0 \cdot \texttt{mom.iq} \\
&= -11 + 1.1 \cdot \texttt{mom.iq} \\
\text{hs: } \texttt{kid.score} &= -11 + 51 \cdot 1 + 1.1 \cdot \texttt{mom.iq} - 0.5 \cdot 1 \cdot \texttt{mom.iq} \\
&= 40 + 0.6 \cdot \texttt{mom.iq}.
\end{aligned}$$

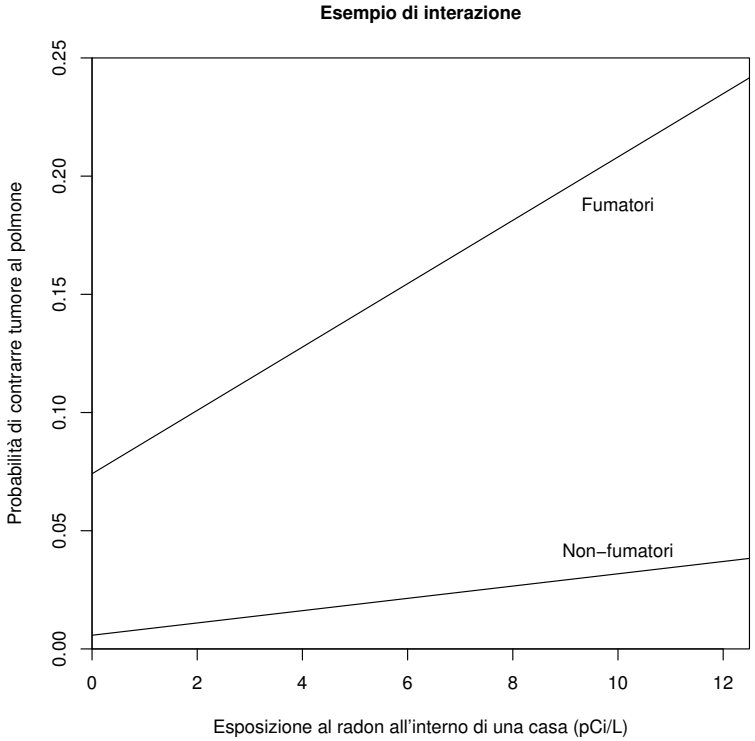


Figura 1.5: *Interazione tra fumo e livello di esposizione al radon in funzione della probabilità di contrarre tumore polmonare nella popolazione maschile. Gli effetti del radon sono molto più severi per i fumatori. Le linee di regressione sono stimate sulla base di uno studio caso-controllo; si veda Lin et al. (1999) per approfondimenti.*

La pendenza stimata pari a 1.1 per i bambini le cui madri non hanno completato gli studi superiori (no hs) e pari a 0.6 per quei bambini le cui madri hanno un titolo di studio superiore (hs) sono direttamente interpretabili. Le intercette risentono invece ancora del problema di essere interpretabili solo per i valori dell'IQ materno pari a 0.

Quando si deve guardare alle interazioni?

Le interazioni posso essere di cruciale importanza nella stima dei modelli di regressione. In pratica, gli input che hanno un effetto molto importante sulla variabile risposta hanno in genere anche una tendenza ad avere forti interazioni anche sugli altri input (comunque, effetti più contenuti ma determinanti non precludono la possibilità di larghe interazioni). Per esempio, il fumo ha un effetto determinante sulla presenza di un tumore. Negli studi epidemiologici di altri tipi di carcinoma, risulta cruciale considerare il fumo sia come effetto principale che come interazione. La Figura 1.5 illustra questo concetto con un esempio legato all'esposizione al radon

all'interno della casa: elevati livelli di radon sono associati ad una maggiore probabilità di contrarre cancro polmonare—relazione che risulta maggiormente accentuata per gli individui fumatori piuttosto che per i non fumatori. L'inclusione delle interazioni permette di adattare un modello a differenti sottoinsiemi di dati. Questi due approcci sono legati, come si vedrà nel seguito nel contesto dei modelli *multilevel*.

Interpretazione dei coefficienti di regressione in presenza di iterazioni

I modelli con interazione sono spesso più facilmente interpretabili se noi prima trasformiamo preliminarmente i dati centrando le variabili di input rispetto alla media o rispetto ad altri valori di riferimento. Si discuterà di questo in modo più approfondito nella Sezione 4.2 nel contesto delle trasformazioni lineari.

1.4 Inferenza Statistica

Quando si illustrano specifici esempi, risulta utile il ricorso a nomi propri per descrivere le variabili che vengono utilizzate. Quando invece si illustra una teoria più generale nonché la manipolazione dei dati, si preferisce far riferimento ad una più generale notazione matematica. Questa sezione introduce questa notazione e discute gli aspetti probabilistici del modello di regressione.

Unità, outcome, predittori e input

Ci si riferisce ai dati individuali puntuali come *unità*—quindi, la risposta alla domanda, "Qual è l'unità di analisi?" sarà qualcosa tipo "persone" o "scuole " o "elezioni congressuali," *non* qualcosa tipo "libbra" o "miglia." I modelli *multilevel* hanno la caratteristica di avere più di un insieme di unità (per esempio, sia persone che scuole), come verrà discusso nel seguito.

Ci si riferisce alle variabili X nella regressione come *predittori* o "variabili predittori," e alla y come *outcome* o "variabile risposta." *Non* useremo i termini come variabili "dipendenti" e "indipendenti", in quanto questi termini verranno usati per descrivere le proprietà delle distribuzioni di probabilità.

Infine, useremo il termine *input* per indicare l'informazione sulle unità strettamente legata alle variabili X. Gli input **non** sono la stessa cosa dei predittori. Per esempio, si consideri il modello che presenta l'interazione tra istruzione materna e punteggio IQ materno:

$$\texttt{kid.score} = 58 + 16 \cdot \texttt{mom.hs} + 0.5 \cdot \texttt{mom.iq} - 0.2 \cdot \texttt{mom.hs} \cdot \texttt{mom.iq} + \texttt{residuo}.$$

Questo modello ha quattro *predittori*—licenza superiore delle madri, punteggio IQ materno, titolo di studio superiore materno × IQ, e il termine costante—ma ha solo due *input*, istruzione materna e IQ.

Regressione secondo la notazione vettore-matrice

Seguendo la usuale notazione indicheremo l'outcome relativo all'i^{mo} individuo come y_i e la previsione deterministica come $X_i\beta = \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$, indicizzando i dati individuali come $i = 1, \ldots, n8$. Nel nostro esempio più recente, y_i è il punteggio ottenuto nel testo dall'individuo i^{mo}, e ci sono $k = 4$ predittori nel vettore X_i (l'i^{ma} riga della matrice X): X_{i1}, un *termine costante* che è definito uguale a 1 per tutti gli individui; X_{i2}, lo stato relativo al completamento o meno degli studi secondari (codificato come 0 o 1); X_{i3}, il punteggio materno del tes di IQ; e X_{i4}, l'interazione tra il punteggio materno ottenuto nel test e lo stato di completamento degli studi. Il vettore β dei coefficienti ha anch'esso lunghezza pari a $k = 4$. Gli errori del modello vengono etichettati come ϵ_i e si assume che seguano una distribuzione normale con media 0 deviazione standard σ, che possiamo scrivere come $\mathrm{N}(0, \sigma^2)$. Il parametro σ rappresenta la variabilità con cui gli *outcome* si discostano dai loro corrispondenti valori previsti sulla base del modello stimato. Useremo la notazione $\tilde{y}$ per i valori non osservati che saranno previsti dal modello, dati i predittori $\tilde{X}$; si veda la Figura 1.6.

Due modi differenti per scrivere lo stesso modello

Il classico modello di regressione può essere scritto in termini matematici come

$$\begin{aligned} y_i &= X_i\beta + \epsilon_i \\ &= \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i, \quad \text{per } i = 1, \ldots, n, \end{aligned}$$

dove gli errori ϵ_i hanno una distribuzione normale con media 0 e deviazione standard σ.

Una rappresentazione equivalente è

$$y_i \backsim \mathrm{N}(X_i\beta, \sigma^2), \text{ per } i = 1, \ldots, n,$$

in cui X è una matrice n per k la cui i^{ma} riga è X_i, oppure, usando la notazione multivariata,

$$y \backsim \mathrm{N}(X\beta, \sigma^2 I),$$

1.4	1	0.69	–1	–0.69	0.5	2.6	0.31
1.8	1	1.85	1	1.85	1.94	2.71	3.18
0.3	1	3.83	1	3.83	2.23	2.53	3.81
1.5	1	0.5	–1	–0.5	1.85	2.5	1.73
2.0	1	2.29	–1	–2.29	2.99	3.26	2.51
2.3	1	1.62	1	1.62	0.51	0.77	1.01
0.2	1	2.29	–1	–2.29	1.57	1.8	2.44
0.9	1	1.8	1	1.8	3.72	1.1	1.32
1.8	1	1.22	1	1.22	1.13	1.05	2.66
1.8	1	0.92	–1	–0.92	2.29	2.2	2.95
0.2	1	1.7	1	1.7	0.12	0.17	2.86
2.3	1	1.46	–1	–1.46	2.28	2.4	2.04
–0.3	1	4.3	1	4.3	2.3	1.87	0.48
0.4	1	3.64	–1	–3.64	1.9	1.13	0.51
1.5	1	2.27	1	2.27	0.47	3.04	3.12
?	1	1.63	–1	–1.63	0.84	2.35	1.25
?	1	0.65	–1	–0.65	2.08	1.26	2.3
?	1	1.83	–1	–1.83	1.84	1.58	2.99
?	1	2.58	1	2.58	2.03	1.8	1.39
?	1	0.07	–1	–0.07	2.1	2.32	1.27

Figura 1.6: *Notazione per i modelli di regressione. Il modello consiste nella stima della variabile risposta osservata y dati i predittori X. Come descritto nel testo, il modello può essere applicato per prevedere outcomes non osservabili $\tilde{y}$ (indicate con il punto interrogativo), dati i predittori relativi ai nuovi dati $\tilde{X}$.*

dove y è un vettore di lunghezza n, X è la matrice $n \times k$ di predittori, β è un vettore colonna di lunghezza k, e I è la matrice identità di dimensione $n \times n$. Stimando il modello (in ognuna delle sue forme) attraverso l'utilizzo dei minimi quadrati si ottengono le stime del vettore $\hat{\beta}$ e della deviazione standard $\hat{\sigma}$.

Stima e sintesi della regressione in R

È possibile stimare un modello di regressione utilizzando la funzione `lm()` in R. Illustreremo questa funzione attraverso il modello che include il completamento o meno degli studi secondari delle madri e l'IQ materno come predittori e senza aggiungere, almeno per ora, alcuna interazione. Chiameremo questo modello `fit.3` essendo il terzo modello stimato in questo capitolo:

R code

```
fit.3 <- lm (kid.score ~ mom.hs + mom.iq)
display (fit.3)
```

(Gli spazi nel codice di R non sono strettamente necessari, ma vengono inclusi per rendere la sintassi più leggibile.) Il risultato è,

R output

```
lm(formula = kid.score ~ mom.hs +mom.iq)

            coef.est coef.se
(Intercept)     25.7     5.9 mom.hs           5.9     2.2 mom.iq
0.6     0.1
  n = 434, k = 3
  residual sd = 18.1, R-Squared = 0.21
```

La funzione `display()` è stata scritta da noi (si veda la Sezione C.2 per dettagli) al fine di fornire un output il più chiaro possibile, focalizzando l'attenzione sulle informazioni da noi ritenute maggiormente pertinenti: i coefficienti e le loro deviazioni standard, l'ampiezza campionaria, il numero di predittori, la deviazione standard dei residui e R^2. In contrasto, l'opzione che R fornisce per default,

R code

```
print(fit.3)
```

fornisce poche informazioni, dando solo le stime dei coefficienti senza i corrispondenti errori standard e nessuna informazione circa la deviazione standard dei residui:

R code

```
Call:
lm(formula = kid.score ~ mom.hs + mom.iq)

Coefficients: (Intercept)       mom.hs       mom.iq
   25.73154      5.95012      0.56391
```

Un'altra opzione fornita da R è la funzione `summary()`:

R code

```
summary (fit.3)
```

che però produce una massa di informazioni difficili da assimilare e riporta troppe cifre decimali:

R output

```
Call:
lm(formula = formula("kid.score ~ mom.hs + mom.iq"))

Residuals:
    Min      1Q  Median      3Q     Max
-52.873 -12.663   2.404  11.356  49.545

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.73154    5.87521   4.380 1.49e-05 *** mom.hs
5.95012    2.21181   2.690  0.00742 ** mom.iq       0.56391
0.06057   9.309  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.14 on 431 degrees of freedom Multiple
R-Squared: 0.2141,
Adjusted R-squared: 0.2105
F-statistic: 58.72 on 2 and 431 DF,
p-value: < 2.2e-16
```

Noi preferiamo quindi la nostra funzione `display()`, che in modo conciso riporta solo le informazioni maggiormente rilevanti del modello stimato.

Stima dei minimi quadrati del vettore dei coefficienti di regressione, β

Per il modello $y = X\beta + \epsilon$, la stima dei minimi quadrati è quel vettore $\hat{\beta}$ che minimizza la somma dei quadrati dei residui, $\sum_{i=1}^{n}(y_i - X_i\hat{\beta})^2$, dati X e y. Intuitivamente, il criterio dei minimi quadrati trova il suo utilizzo nel fatto che, dal momento che stiamo provando a prevedere un outcome attraverso l'utilizzo di altre variabili, vorremmo fare questa previsione in modo da minimizzare l'errore della nostra previsione.

La stima ottenuta con i minimi quadrati coincide con la stima di massima verosimiglianza se gli errori ϵ_i sono indipendenti, omoschedastici e normalmente distribuiti (si veda la Sezione 18.1). In ogni caso, le stime dei minimi quadrati possono essere espresse in notazione matriciale come $\hat{\beta} = (X^tX)^{-1}X^ty$. In pratica, il calcolo viene effettuato usando delle opportune scomposizioni matriciali senza calcolare in realtà X^tX e senza invertirla. Per i nostri scopi, è essenzialmente utile considerare il fatto che $\hat{\beta}$ è una funzione lineare dell'outcome y.

Errori standard: incertezza sulle stime del coefficienti

Insieme alle stime di $\hat{\beta}$ si hanno i corrispondenti errori standard, come mostrato nell'output della regressione. Gli errori standard rappresentano la stima dell'incertezza. Si potrebbe dire, anche se in modo approssimativo, che quando le stime dei coefficienti cadono all'interno di un intervallo ± 2 volte l'errore standard di $\hat{\beta}$ possono considerarsi stime consistenti rispetto ai dati. La Figura 1.7 mostra la distribuzione normale che rappresenta approssimativamente il *range* dei possibili valori di β. Per esempio, nel modello a pag.19, il coefficiente di `mom.hs` ha una stima $\hat{\beta}$ pari a 5.9 e un corrispondente errore standard pari a 2.2; quindi i dati sono approssimativamente consistenti con valori di β nell'intervallo $[5.9 \pm 2 \cdot 2.2] = [1.5, 10.3]$. Più precisamente, si potrebbe modellare l'incertezza considerando gli stessi errori standard attraverso l'utilizzo della distribuzione t di Student con gradi di libertà pari al numero di osservazioni meno il numero di coefficienti stimati, ma la distribuzione normale fornisce una buona approssimazione quando i gradi di libertà sono più di 30.

L'incertezza delle stime dei coefficienti non è inoltre esente da possibili correlazioni (ad eccezione di casi speciali di studi con disegni bilanciati). L'informazione sulla correlazione delle stime viene riassunta dalla matrice di covarianza stimata $V_\beta\hat{\sigma}^2$, dove $V_\beta = (X^tX)^{-1}$. Gli elementi sulla diagonale di $V_\beta\hat{\sigma}^2$ sono le varianze stimate delle componenti individuali di β, mentre gli elementi al di fuori della diagonale rappresentano le covarianze delle stime. Quindi, per esempio $\sqrt{V_{\beta\,11}}\,\hat{\sigma}$ è l'errore standard di $\hat{\beta}_1$, $\sqrt{V_{\beta\,22}}\,\hat{\sigma}$ è l'errore standard di $\hat{\beta}_2$, mentre $V_{\beta\,12}/\sqrt{V_{\beta\,11}V_{\beta\,22}}$

è la correlazione delle stime $\hat{\beta}_1, \hat{\beta}_2$.

Noi in genere non analizzeremo la matrice di covarianza; piuttosto l'inferenza sul modello verrà riassunta facendo riferimento ai coefficienti e ai loro errori standard, mentre la matrice di covarianza verrà utilizzata per effettuare simulazioni predittive, come descritto nella Sezione 7.2.

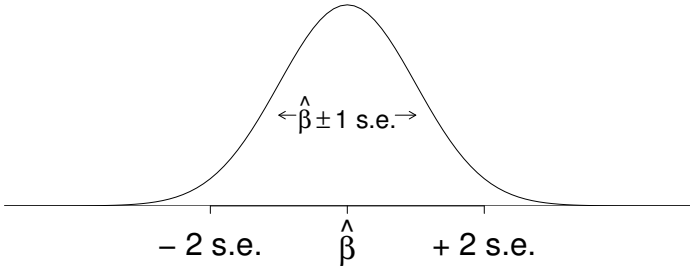


Figura 1.7: *Distribuzione rappresentante l'incertezza nei coefficienti della regressione stimati. Il campo di variazioni di questa distribuzione corrisponde ai possibili valori di β consistenti con i dati. Quando si usa una tale distribuzione per rappresentare l'incertezza si assegna una probabilità pari a circa il 68% che β si discosti dal valore puntuale stimato, $\hat{\beta}$, di 1 volta $\pm$ il suo errore standard e una* chance *pari a circa il 95% che sia all'interno di 2 volte $\pm$ il suo errore standard. Assumendo che il modello di regressione sia corretto, solo per il 5% dei casi si potrebbe verificare di ottenere un valore stimato, $\hat{\beta}$, che si discosti dal vero valore di β più di 2 volte $\pm$ il suo errore standard.*

I residui, r_i

I *residui*, $r_i = y_i - X_i\hat{\beta}$, sono pari alla differenza tra i dati osservati e i valori stimati dal modello. Come conseguenza della procedure di stima dei minimi quadrati utilizzata per β, i residui r_i saranno incorrelati con tutti i predittori del modello. Se il modello include anche un termine costante, allora i residui devono essere incorrelati anche con la costante, il che significa che devono avere media pari a 0. Anche questo come conseguenza della procedure di stima del modello; ma *non* è un'ipotesi del modello di regressione. Discuteremo nel seguito di questo capitolo come i residui possano essere utilizzati per diagnosticare eventuali problemi nel modello.

La deviazione standard dei residui $\hat{\sigma}$ e la varianza spiegata R^2

La deviazione standard (aggiustata) dei residui, $\hat{\sigma} = \sqrt{\sum_{i=1}^{n} r_i^2/(n-k)}$, riassume la scala dei residui. Per esempio, nell'esempio del punteggio del test, $\hat{\sigma} = 18$, ci dice

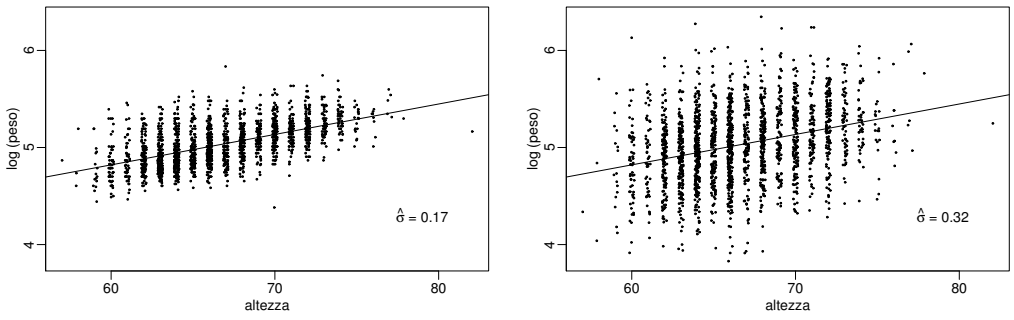
che il modello lineare è in grado di stimare il punteggio raggiunto dai bambini nel test con un'accuratezza di 18 punti. Detto in un altro modo, noi possiamo pensare alla deviazione standard come ad una misura della distanza media tra le osservazioni rispetto ai valori stimati dal modello.

L'adattamento del modello può essere riassunto da $\hat{\sigma}$ (minore è la varianza residua, migliore è l'adattamento del modello) e da R^2, la frazione di varianza "spiegata" dal modello. La varianza "non spiegata" è $\hat{\sigma}^2$, e se noi indichiamo con s_y la deviazione standard dei dati, allora $R^2 = 1 - \hat{\sigma}^2/s_y^2$. Nella regressione dei punteggi del test, R^2 è pari solo al 22%. (Comunque, visto in termini più profondi, è verosimilmente un fatto positivo che questo modello presenti un R^2 piuttosto basso—ad indicare che come non sia possibile prevedere in maniera accurata l'apprendimento dei bambini solo in funzione di alcune caratteristiche delle loro madri).

La quantità $n-k$, pari al numero delle osservazioni meno il numero dei coefficienti stimati, è chiamata *gradi di libertà* per la stima degli errori residuali. Nella regressione classica, k deve essere minore di n—altrimenti i dati potrebbero essere stimati perfettamente e non sarebbe possibile stimare gli errori della regressione.

Difficoltà nell'interpretazione della deviazione standard e della varianza speigata

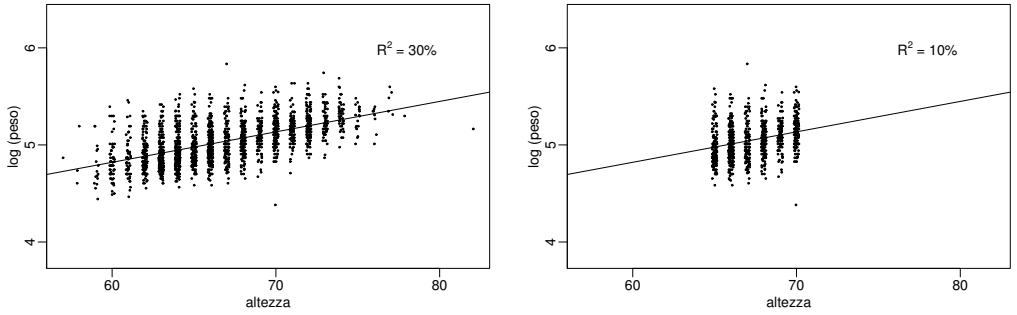
Figura 1.8: *Due ipotetici datasets con la stessa linea di regressione, $y = a + bx$, ma con differenti valori della deviazione standard dei residui, σ. Il grafico sulla sinistra mostra i dati da un campione di adulti; il grafico sulla destra mostra dei dati a cui è stato aggiunto un rumore casuale alla variabile y.*



Come chiariremo nel corso del libro, noi siamo generalmente interessati alla parte "deterministica" del modello, $y = X\beta$, piuttosto che a quella stocastica che incorpora ϵ. Comunque, quando si guarda alla deviazione standard dei residui, $\hat{\sigma}$, noi siamo interessati tipicamente a questa componente–sia come una misura della

varianza non spiegata dei dati—o per la sua rilevanza nella valutazione della precisione delle stime dei coefficienti della regressione β. (Come abbiamo già discusso, gli errori standard dei β sono proporzionali a σ.). La Figura 1.8 illustra due regressione con lo stesso modello deterministico, $y = a + bx$, ma con differenti valori di σ.

Figura 1.9: *Due ipotetici datasets con la stessa linea di regressione, $y = a + bx$ e con la stessa deviazione standard dei residui σ, ma con differenti valori della varianza spiegata R^2. Il grafico sulla sinistra è relativo all'intero campione di dati; il grafico sulla destra è relativo ad un sottocampione di dati ristretto agli individui con con altezza compresa tra 65 e 70 pollici.*



Interpretiamo ora la proporzione di varianza spiegata, R^2, che può essere non banale dal momento che sia il numeratore che il denominatore di R^2 possono cambiare in diversi modi. La Figura 1.9 illustra con un esempio in cui il modello di regressione è identico, ma R^2 diminuisce in quanto il modello è stimato su un sottoinsieme di dati. (Passando dal grafico di sinistra a quello di destra nella Figura 1.9, la deviazione standard dei residui σ è rimasta invariata ma la deviazione standard dei dati grezzi, s_y, diminuisce dal momento che stiamo lavorando su un sottoinsieme di dati; quindi, $R^2 = 1 - \hat{\sigma}^2/s_y^2$ diminuisce.) Anche se R^2 risulta di molto inferiore nel grafico destro della figura, il modello si adatta ai dati esattamente allo stesso modo che nel grafico alla sinistra della figura.

Significatività Statistica

Se la stima di un coefficiente risulta essere distante dallo zero più che 2 volte l'errore standard, allora è detta *statisticamente significativa*. Quando una stima è statisticamente significativa, noi siamo ragionevolmente sicuri che il segno (+ o −) della stima è stabile, e non invece un artefatto dovuto ad un'ampiezza del campione limitata.

Talvolta si crede che se la stima di un coefficiente non è significativa, allora il coefficiente dovrebbe essere escluso dal modello. Noi non siamo d'accordo. È

ragionevole tenere nel modello anche coefficienti anche se non significativi, se questi hanno senso. Discuteremo di questo nella Sezione 4.6.

Incertezza nella deviazione standard dei residui

Sotto le ipotesi del modello stimato, la varianza stimata dei residui, $\hat{\sigma}^2$, ha una distribuzione campionaria centrata nel vero valore, σ^2, e si distribuisce secondo una distribuzione χ^2 con $n-k$ gradi di libertà. Noi faremo uso di questa incertezza nelle nostre simulazioni predittive, come descritto nella Sezione 7.2.

1.5 Rappresentazione grafica dei dati e del modello stimato

Rappresentazione di una linea di regressione in funzione di una variabile di input

Abbiamo rappresentato alcuni aspetti del nostro modello del punteggio del test attraverso la rappresentazione dei dati nelle Figure 1.1–1.3.

Noi possiamo fare un grafico come quello nella Figura 1.2 in questo modo:

R code

```
fit.2 <- lm (kid.score ~ mom.iq)

plot (mom.iq, kid.score, xlab="Punteggio IQ delle madri",
      ylab="Punteggio dei figli")

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

La funzione `plot()` crea uno scatterplot delle osservazioni, e `curve` sovrappone la linea di regressione usando i coefficienti che sono stati salvati da `lm()` *call* (come se fossero estratti usando la funzione `coef()`). L'espressione all'interno di `curve()` può anche essere scritta usando la notazione matriciale in R:

R code

```
curve (cbind(1,x) %*% coef(fit.2), add=TRUE)
```

Rappresentazione di due linee di regressione stimate

Modello senza iterazione. Per il modello con due input, possiamo fare un grafico con due insiemi di punti e due linee di regressione, come nella Figura 1.3:

R code

```
fit.3 <- lm (kid.score ~ mom.hs + mom.iq)
       colors <- ifelse (mom.hs==1, "black", "gray")
       plot (mom.iq, kid.score, xlab="Punteggio IQ delle madri",
       ylab="Punteggio dei figli", col=colors, pch=20)
curve (cbind (1, 1, x) %*% coef(fit.3), add=TRUE, col="black")
curve (cbind (1, 0, x) %*% coef(fit.3), add=TRUE, col="gray")
```

Definendo `pch=20` facciamo in modo che la funzione `plot()` rappresenti i dati usando dei punti di piccole dimensioni, mentre l'opzione `col` impone il colore dei punti, che noi abbiamo imposto essere neri o grigi a seconda del valore del predittore `mom.hs`.[1] Infine, chiamando la funzione `curve()` sovrapponiamo le linee di regressione per i due gruppi definiti dal completamento o meno degli studi superiori delle madri.

Modello con iterazione Possiamo definire un grafico simile a quello precedente anche per il modello con interazione, con la sola differenza che le due linee hanno diverse pendenze:

R code

```
fit.4 <- lm (kid.score ~ mom.hs + mom.iq + mom.hs:mom.iq)

colors <- ifelse (mom.hs==1, "black", "gray")

plot (mom.iq, kid.score, xlab="Punteggio IQ delle madri",
      ylab="Punteggio dei figli", col=colors, pch=20)
```

[1] Una sequenza alternativa di comandi è,

```
plot (mom.iq, kid.score, xlab=Punteggio IQ delle madri, ylab=Punteggio dei
figli, type=n)
points (mom.iq[mom.hs==1], kid.score[mom.hs==1], pch=20, col=black)
points (mom.iq[mom.hs==0], kid.score[mom.hs==0], pch=20, col=gray)
```

Qui, `plot()`, chiamato con l'opzione `type=n`, rappresenta gli assi senza che i punti siamo nel grafico. Quindi ogni volta che richiamiamo la funzione `points()` sovrapponiamo le osservazioni per ogni sottogruppo (gruppi definiti in questo caso dal completamento o meno degli studi superiori delle madri) separatamente—ognuno con simboli differenti.

```
curve (cbind (1, 1, x, 1*x) %*% coef(fit.4), add=TRUE, col="black")
curve (cbind (1, 0, x, 0*x) %*% coef(fit.4), add=TRUE, col="gray")
```

Il risultato è il grafico mostrato nella Figura 1.4.

Rappresentazione dell'incertezza nel modello di regressione stimato

Come discusso nella Sezione 7.2, noi possiamo usare la funzione `sim()` in R per creare simulazioni che rappresentino l'incertezza delle stime dei coefficienti del modello di regressione. Qui descriviamo brevemente come usare queste simulazioni. Per semplicità consideriamo ancora una volta il modello con un solo predittore:

R code

```
fit.2 <- lm (kid.score ~ mom.iq)
```

che fornisce,

R output

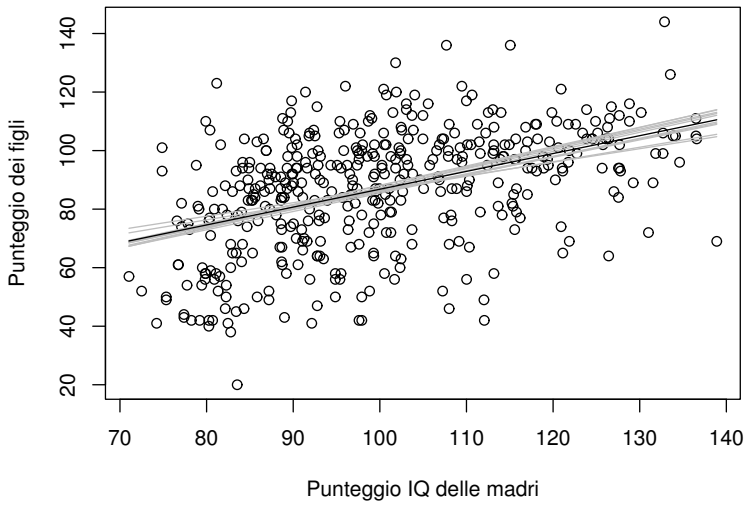
```
            coef.est coef.se
(Intercept)    25.8     5.9 mom.iq          0.6     0.1
  n = 434, k = 2
  residual sd = 18.3, R-Squared = 0.2
```

Il seguente codice crea la Figura 1.10, che mostra la linea di regressione stimata insieme a diverse simulazioni rappresentanti l'incertezza intorno alla linea di regressione:

R code

```
fit.2.sim <- sim (fit.2)
plot (mom.iq, kid.score, xlab="Punteggio
IQ delle madri", ylab="Punteggio dei figli")
for (i in 1:10){
  curve (fit.2.sim$beta[i,1] + fit.2.sim$beta[i,2]*x, add=TRUE,col="gray")
} curve (coef(fit.2)[1] + coef(fit.2)[2]*x, add=TRUE, col="black")
```

Figura 1.10: *Dati e modello di regressione per la stima del punteggio dei figli su un test di apprendimento in funzione del punteggio IQ delle madri, la linea solida rappresenta il modello di regressione stimato e le linee in chiaro indicano l'incertezza intorno al modello di regressione stimato.*



Il ciclo `for (i in i:10)` permette di rappresentare 10 diverse simulazioni. [2] La Figura 1.10 inoltre rappresenta l'incertezza che noi abbiamo circa i valori che vogliamo *prevedere* in base al nostro modello stimato. Questa incertezza aumenta mano mano che ci allontana dalla media del predittore.

Rappresentazione usando un grafico per ogni variabile di input

Ora consideriamo il modello di regressione con una variabile dicotomica che indica il completamento o meno degli studi supeiori da parte delle madri:

R code
```
fit.3<- lm (kid.score ~ mom.hs + mom.iq)
```

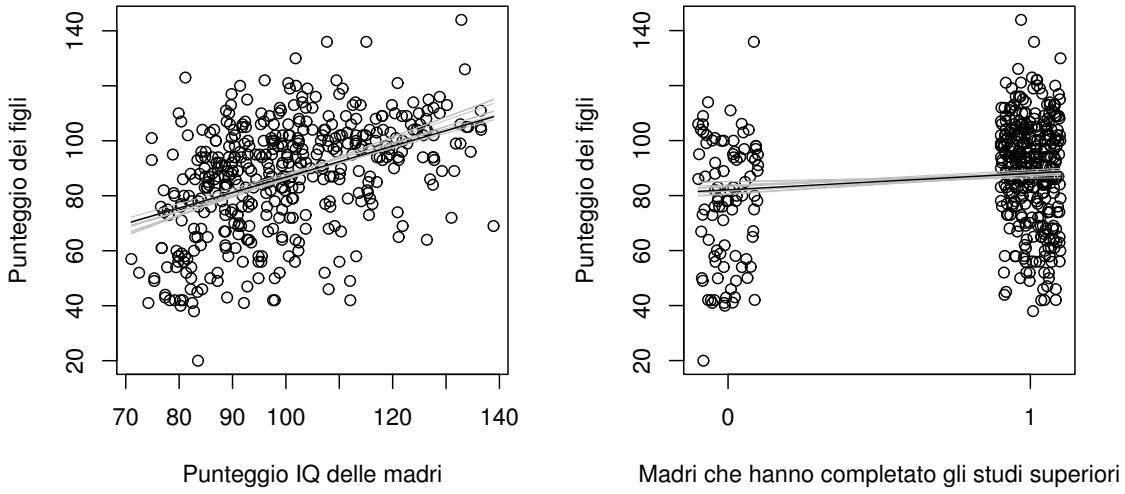
Rappresentiamo il modello della Figura 1.11 come due grafici differenti, uno per ciascuna delle due variabili input con l'atra fissata al suo valore medio:

R code
```
beta.hat <- coef (fit.3)
beta.sim <- sim (fit.3)$beta

par (mfrow=c(1,2))

plot (mom.iq, kid.score, xlab="Punteggio IQ delle madri",
ylab="Punteggio dei figli")

for (i in 1:10){
  curve (cbind (1, mean(mom.hs), x) %*% beta.sim[i,], lwd=.5,
     col="gray", add=TRUE)
}

curve (cbind (1, mean(mom.hs), x) %*% beta.hat, col="black", add=TRUE)

plot (mom.hs, kid.score, xlab="Madri che hanno completato gli
```

[2] Un altro modo di codificare questo ciclo in R è attraverso la funzione `apply()`, per esempio;
`Oneline <- function (beta) {curve (beta[1]+beta[2]*x, add=TRUE, col=gray)}`
`apply (fit.2.sim$beta, 1, Oneline)`
Attraverso l'utilizzo della funzione `apply()` scriviamo il ciclo in modo più elegante e pulito, almeno per gli utilizzatori esperti di R; la forma in termini di ciclo, sebbene meno elegante, risulta più semplice per i non esperti di R.

```
studi superiori",
  ylab="Punteggio dei figli")

for (i in 1:10){
  curve (cbind (1, x, mean(mom.iq)) %*% beta.sim[i,], lwd=.5,
     col="gray", add=TRUE)
}
curve (cbind (1, x, mean(mom.iq)) %*% beta.hat, col="black", add=TRUE)
```

Figura 1.11: *Dati e modello di regressione la stima del punteggio dei figli su un test di apprendimento in funzione del punteggio IQ e del completamento degli studi superiori delle madri visto come una funzione delle due variabili di input (le linee chiare rappresentano l'incertezza nella regressione). I valori relativi al completamento degli studi superiori sono stati sottoposti alla procedure di* jitter *al fine di rendere i punti tra loro distinti.*



1.6 Ipotesi del modello e diagnostica

Iniziamo a considerare le ipotesi del modello, insieme ad alcune procedure di diagnostica che possono essere utilizzate per verificare se alcune di queste assunzioni sono ragionevoli. Alcune delle più importanti assunzioni, comunque, si basano sulla conoscenza del ricercatore sull'oggetto di studio e non possono essere direttamente testate soltanto a partire dai dati.

Ipotesi sul modello di regressione

Elenchiamo le assunzioni del modello di regressione in ordine di importanza.

1. *Validità.* I dati che si decide di analizzare devono prima di tutto essere coerenti con le domande che ci si pone e a cui si sta cercando di rispondere. Su questo principio, sebbene sembrerebbe ovvio, spesso si sorvola e talvolta viene addirittura ignorato.

   Idealmente, questo significa che la variabile risposta dovrebbe riflettere accuratamente il fenomeno di interesse, il modello dovrebbe includere i predittori più rilevanti e dovrebbe essere in grado di poter essere utilizzato anche per casi più generali rispetto a quello per il quale è stato stimato.

   Per esempio, con riferimento alla variabile risposta, un modello che analizza i guadagni non necessariamente ci dice tanto circa il comportamento dei guadagni complessivi. Un modello cha analizza i punteggi di un test non necessariamente fornisce anche informazioni circa l'intelligenza dei bambini o sullo sviluppo cognitivo.

   La scelta degli input nel modello di regressione è spesso uno dei passi più ambiziosi dell'analisi. Noi in genere veniamo incoraggiati ad includere nel modello tutti i predittori ritenuti più "rilevanti", ma in pratica può essere difficile determinate quali siano quelli effettivamente necessari e come interpretare quei coefficienti che presentano standard error molto elevati. Il capitolo 9 discute la scelta degli input per la regressione usata nell'inferenza causale.

   Un campione rappresentativo di tutte le madri e di tutti i bambini potrebbe non essere appropriato per fare inferenza su madri e bambini che partecipano al programma *Temporary Assistance for Needy Families*. Comunque, una selezione accurata di un sottocampione potrebbe riflettere bene la distribuzione di questa popolazione. In maniera del tutto simile, risultati relativi alla dieta e a particolari esercizi ottenuti da uno studio effettuato su un campione di pazienti a rischio di malattie cardiache potrebbero non essere generalmente validi e applicabili su campioni di individui sani. In questo caso risulta necessario effettuare delle assunzioni su come risultati ottenuti da analisi effettuate su una popolazione a rischio potrebbero essere legati a risultati relativi a una popolazione di individui sani.

   Dati effettivamente utilizzati per le ricerche empiriche raramente soddisfano in maniera precisa tutti, o anche solo alcuni, di questi criteri. Comunque tenere in considerazione questi obiettivi può aiutare ad essere precisi sulla possibilità di rispondere o meno ad alcune tipologie di domande.

2. *Additività e linearità.* La più importante assunzione matematica del modello di

regressione è relativa alla sua componente deterministica che si suppone essere una funzione lineare di predittori tra loro separati: $y = \beta_1 x_1 + \beta_2 x_2 + \cdots$.

Qualora l'additività venga violata, potrebbe avere senso considerare qualche trasformazione dei dati (per esempio, se $y = abc$, allora $\log y = \log a + \log b + \log c$) oppure aggiungere eventuali interazioni. Qualora venga violata l'assunzione di linearità, un predittore potrebbe per esempio essere inserito nel modello nella forma $1/x$ o $\log(x)$ invece che essere inserito linearmente. Oppure una relazione più complicata potrebbe essere espressa includendo nel modello sia x che x^2 come predittori.

Per esempio, è abbastanza comune inserire nel modello di regressione sia l'`età` che l'`età`^{2}. Nelle analisi di tipo medico e di salute pubblica, questo permette di considerare una misura di salute che decresce al crescere dell'età, il cui tasso di decrescita diventa sempre più inclinato man mano che l'età aumenta. Nelle analisi politiche, l'inclusione di entrambi i predittori `età` e `età`^{2} permette di accentuare la pendenza al crescere dell'età e la possibilità di introdurre un andamento a U qualora, ad esempio, i giovani e gli anziani sono maggiormente favorevoli alle tasse rispetto alle persone di mezza età.

In questi tipi di analisi noi in genere preferiamo includere l'età in termini di predittore categorico, come discusso nella Sezione 4.5. Un'altra opzione potrebbe essere quella di usare un'altra funzione non lineare come una *spline* o un altro modello generalizzato additivo. In ogni caso, l'obiettivo è quello di aggiungere predittori in modo tale che il modello additivo lineare sia un'approssimazione ragionevole.

3. *Indipendenza degli errori.* Il modello di regressione semplice assume che gli errori della previsione siano indipendenti. Ritorneremo su questo punto in dettaglio quando discuteremo dei modelli *multilevel*.

4. *Uguale varianza degli errori.* Se la varianza degli errori della regressione non sono le stesse, le stime più efficienti sono quelle che si ottengono con i minimi quadrati ponderati, in cui ogni punto viene ponderato con pesi inversamente proporzionali alla loro varianza (di veda la Sezione 18.4). Nella maggior parte dei casi questo comunque è un problema minore. Varianze ineguali non riguardano l'aspetto più importante del modello di regressione, la forma del predittore $X\beta$.

5. *Normalità degli errori.* L'assunzione sul modello di regressione che in generale è la *meno* importante riguarda il fatto che gli errori debbano essere normalmente distribuiti. Infatti, per l'obiettivo della stima della linea di regressione (intesa come confronto tra dati e valori stimati), l'assunzione di Normalità degli errori

è scarsamente rilevante. Quindi, in contrasto con molti libri di testo sulla regressione, *non* raccomandiamo alcuna diagnostica sulla normalità dei residui.

Se la distribuzione dei residui è di qualche interesse, per esempio a fini di obiettivi di tipo predittivo, questa dovrebbe essere distinta dalla distribuzione dei dati, y. Per esempio, si consideri una regressione effettuata su un singolo predittore discreto, x, il quale assume valori 0, 1, e 2, con un terzo della popolazione in ognuna delle tre categorie. Supponiamo che la vera linea di regressione sia $y = 0.2 + 0.5x$ con relativi errori normalmente distribuiti e aventi deviazione standard pari a 0.1. Allora un grafico dei dati y mostrerà ragionevolmente tre mode abbastanza nette e centrate nei valori 0.2, 0.7, and 1.2. Altri esempi di misture di questo tipo sono comuni in economia, quando per esempio si considerano occupati e disoccupati, o negli studi elettorali, quando si confrontano distretti che hanno in carica legislatori di partiti differenti.

Ulteriori ipotesi sono necessarie se viene data al coefficiente di regressione un'interpretazione causale, come verrà discusso nei Capitoli 9 e 10.

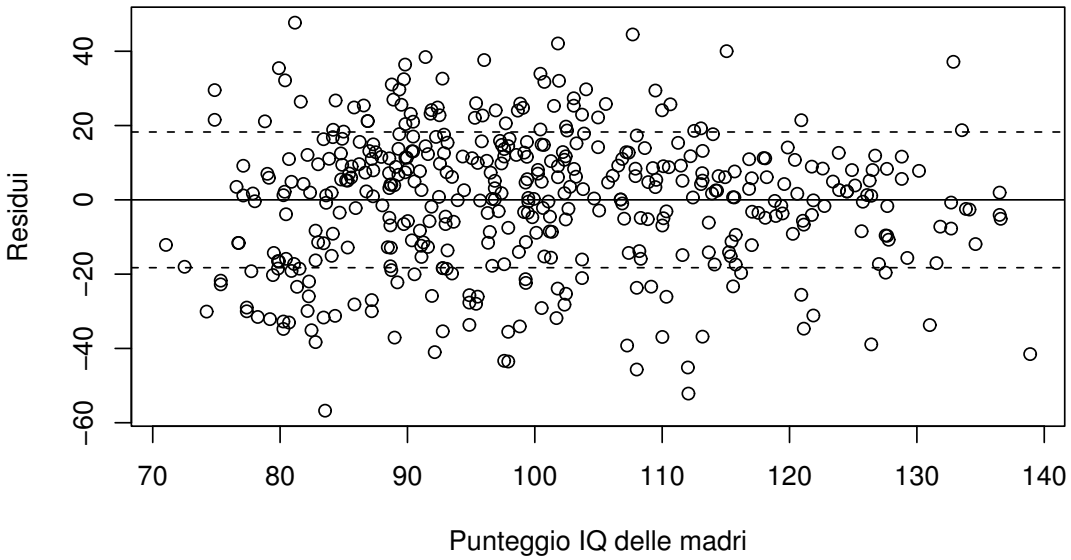
Rappresentazione grafica dei residui al fine di rilevare aspetti dei dati non catturati dal modello

Un modo possibile per diagnosticare la violazione di alcune ipotesi appena viste (soprattutto la linearità) è quello di rappresentare graficamente i residui r_i rispetto ai valori stimati $X_i\hat{\beta}$ o semplicemente rispetto ai predittori individuali x_i; la Figura 1.12 rappresenta i residui del modello relativo al test di apprendimento quando i valori sono visti solamente in funzione del punteggio IQ delle madri. Dal grafico sembra che non ci siano particolari comportamenti nell'andamento dei residui. In altri modelli, i grafici dei residui posso rivelare problemi sistematici come illustrato, per esempio, nel Capitolo 6.

1.7 Previsione e validazione del modello

Talvolta l'obiettivo del nostro modello è quello di fare previsioni usando dei nuovi dati. Nel caso di previsioni relative a dati in serie storica, l'eventuale disponibilità di nuovi dati nel tempo permette al ricercatore di valutare quanto bene il modello lavora in termini di previsioni. Talvolta vengono effettuate delle previsioni fuori-dal campione per il solo fine di valutare il modello, come verrà illustrato nel seguito.

Figura 1.12: *Grafico dei residui del modello di regressione del punteggio del test di apprendimento effettuato sui bambini in funzione del punteggio IQ delle madri. Le linee tratteggiate mostrano le bande* ± 1 *la deviazione standard. I residui non presentano andamenti particolari.*



Previsione

Partendo dal modello (1.4) a pagina 8, è possibile prevedere che un bambino la cui madre possiede un diploma di scuola secondaria e un punteggio IQ pari a 100 riporterà un punteggio nel test di apprendimento pari a $26 + 6 \cdot 1 + 0.6 \cdot 100 = 92$. Se questa equazione rappresenta il modello vero, piuttosto che un modello stimato, allora noi potremmo usare $\hat{\sigma} = 18$ come una stima dell'errore standard della nostra previsione. In realtà la deviazione standard degli errori stimata è leggermente più elevata di $\hat{\sigma}$ a causa dell'incertezza nella stima dei parametri della regressione– un'ulteriore complicazione che fornisce i cosiddetti errori standard di previsione come riportato in molti libri di testo.[3] In R possiamo creare un data frame per i nuovi dati e quindi usare la funzione `predict()`. Per esempio, i comandi

R code

```
x.new <- data.frame (mom.hs=1, mom.iq=100)

predict (fit.3, x.new, interval="prediction", level=0.95)
```

forniscono una previsione puntuale e un *intervallo predittivo* del 95%.

Più in generale, possiamo propagare l'incertezza *predittiva* attraverso le simulazioni, come spiegato nella Sezione 7.2.

Usiamo la notazione $\tilde{y}_i$ per l'outcome misurato su un nuovo dato puntale e $\tilde{X}_i$ per il vettore dei predittori (in questo esempio, $\tilde{X}_i = (1, 1, 100)$). Il valore previsto dal modello è $\tilde{X}_i\hat{\beta}$, con un errore standard di previsione leggermente più elevato di $\hat{\sigma}$. La distribuzione normale implica che circa il 50% dei valori deve cadere all'interno dell'intervallo di previsione $\pm 0.67\hat{\sigma}$, circa 68% all'interno dell'intervallo $\pm\hat{\sigma}$, e circa il 95% all'interno dell'intervallo $\pm 2\hat{\sigma}$.

In maniera del tutto simile possiamo prevedere un vettore di $\tilde{n}$ nuovi valori, $\tilde{y}$, data una matrice $\tilde{n} \times k$ di predittori, $\tilde{X}$; si veda Figura 3.13.

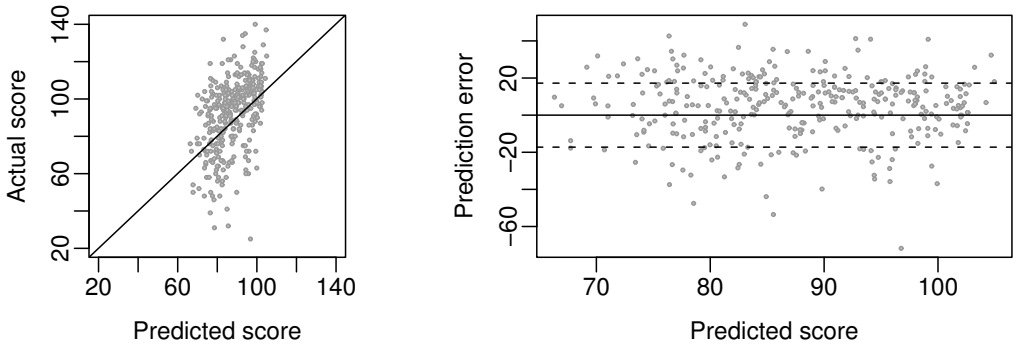
Validazione esterna

Il modo migliore per validare un modello, in tutti i contesti scientifici, è utilizzarlo per effettuare previsioni e confrontarle con i dati osservati.

[3] Per esempio nel modello di regressione lineare con un predittore, l'"errore standard di previsione" intorno alla previsione relativa ad un nuovo dato puntuale con valore previsto $\tilde{x}$ è

$$\hat{\sigma}_{\text{forecast}} = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(\tilde{x} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Figura 1.13: *Grafici che valutano il modello stimato con un campione di bambini più grandi per fare previsioni su bambini più piccoli. Il primo grafico confronta le previsioni per bambini più piccoli in funzione dei valori effettivi. Il secondo grafico confronta invece i residui di queste previsioni in funzione dei valori previsti.*



La Figura 1.13 illustra questo concetto con il modello relativo al punteggio del test, che è stato stimato per un campione di dati raccolti nel 1986 e 1994 per bambini nati prima del 1987. Abbiamo utilizzato il modello per stimare i valori dei bambini nati nel 1987 o anche dopo (dati relativi al periodo 1990–1998). Questo non è un esempio ideale per le previsioni in quanto non ci si aspetta che un modello adatto a bambini più grandi sia in grado di stimare il livello di apprendimento anche per bambini più piccoli, anche se i test sono stati tutti effettuati su bambini di 3 o 4 anni di età. In ogni caso, possiamo usare questo modello per illustrare i metodi per effettuare e valutare le previsioni del modello. Consideriamo in questo capitolo le previsioni puntuali mentre le previsioni basate su procedure di simulazione sono oggetto della Sezione 7.2.

I nuovi dati, $\tilde{y}$, sono i valori per i nuovi 336 bambini stimati sulla base di `mom.iq` e `mom.hs`, usando il modello stimato sui dati relativi ai bambini più grandi. Il primo grafico della Figura 1.13 riporta i valori effettivi $\tilde{y}_i$ rispetto ai valori stimati $\tilde{X}_i\hat{\beta}$, mentre sul secondo grafico sono riportati i residui rispetto ai valori stimati con le relative bande di confidenza $\pm\hat{\sigma}$ (approssimativamente all'interno dei limiti di errore del 68%; si veda Sezione 2.3). Il grafico dei residui mostra che non ci sono problemi nell'adattare un modello stimato per valutare l'apprendimento di bambini più grandi allo studio dell'apprendimento di bambini di età inferiore, sebbene dalla scala si può notare che le previsioni presentano un'elevata variabilità.

Anche se avessimo riscontrato dei problemi nelle previsioni, non necessariamente questo avrebbe significato la presenza di qualcosa di sbagliato nella stima del modello sui dati osservati. Comunque, prima di generalizzare il modello ad altri campioni di bambini bisogna sempre capirlo in modo approfondito.

1.8 Nota Bibliografica

La regressione lineare è stata usata per secoli nelle scienze sociali e fisiche; si veda Stigler (1986). Numerosi manuali di statistica hanno una buona discussione sulla regressione lineare semplice, per esempio Moore e McCabe (1998) e De Veaux et al. (2006). Fox (2002) insegna R in un contesto di regressione applicata. In aggiunta, il sito di R offre diversi *links* verso un tipo di letteratura *open source*.

Carlin e Forbes (2004) forniscono un'eccellente introduzione sui concetti di modellizzazione lineare e regressione, mentre Pardoe (2006) è un testo introduttivo che focalizza l'attenzione su casi aziendali. Per una visione più completa, Neter et al. (1996) e Weisberg forniscono un'introduzione accessibile alla regressione, e Ramsey e Schafer (2001) è un buon complemento, che si concentra su problemi relativi alla comprensione del modello, alla rappresentazione grafica e al disegno sperimentale. Woolridge (2001) presenta il modello di regressione secondo una visione prettamente econometrica. L'importanza di R^2 in termini di varianza spiegata è analizzata da Wherry (1931); si veda anche King (1986) per alcuni esempi relativi alla possibilità di commettere errori comuni utilizzando la regressione e la Sezione 21.9 per alcuni riferimenti più avanzati su R^2 e altri metodi per sintetizzare i modelli stimati. Berk (2004) discute le diverse ipotesi implicite nell'analisi della regressione.

Per ulteriori informazioni sui test relativi all'apprendimento dei bambini e all'occupazione materna si veda Hill *et al.*(2005). Si veda l'Appendice B e Murrell (2005) per saperne di più su come fare i grafici mostrati in questo capitolo e in tutto il libro. La tecnica di *jittering* (usata nella Figura 1.1 e altrove nel libro) proviene da Chambers *et al.* (1983).

1.9 Esercizi

1. La cartella `pyth` contiene la variabile risposta y e due variabili di input x_1, x_2 per 40 punti, con ulteriori 20 dati che riportano le variabili di input ma non la variabile risposta. Si salvi il file nella working directory e lo si importi in R usando la funzione `read.table()`.

   (a) Si utilizzi R per stimare un modello di regressione per prevedere y sulla

base delle variabili x_1, x_2, usando i primi 40 dati nel file. Sintetizzare l'inferenza e valutare l'adattamento del modello stimato.

(b) Rappresentare graficamente il modello stimato come in Figura 1.2.

(c) Fare un grafico dei residui per questo modello. Sembra che le ipotesi del modello vengano rispettate?

(d) Fare delle previsioni per i restanti 20 punti nel file. Quanto ci si sente fiduciosi circa queste previsioni?

Dopo aver fatto questo esercizio, si dia uno sguardo a Gelman e Nolan (2002, sezione 9.4) per vedere da dove provengono questi dati.

2. Supponiamo che, per una data popolazione, si sia in grado di prevedere il logaritmo dei guadagni in funzione del logaritmo dell'altezza come segue:

- Una persona che è alta 66 pollici (1 pollice equivale a 2.54 cm) guadagna, in base a queste previsioni, \$30,000 dollari.
- Ogni incremento dell'1% in altezza corrisponde ad un incremento previsto di 0.8% nel guadagno.
- I guadagni di circa il 95% delle persone cadono all'interno di un fattore pari a 1.1 dei valori previsti.

(a) Scrivere l'equazione di regressione e la deviazione standard della regressione.

(b) Supponiamo che la deviazione standard del logaritmo dell'altezza sia il 5% in questa popolazione. Cosa rappresenta il valore di R^2 in questo modello?

3. In questo esercizio simulate due variabili che sono statisticamente indipendenti l'una dall'altra per vedere cosa succede quando regrediamo una di queste variabili rispetto all'altra.

(a) Inizialmente generate 1000 punti da una distribuzione normale con media 0 e deviazione standard 1 attraverso il seguente comando `var1 <- rnorm(1000,0,1)` in R. Generate un'altra variabile nello stesso modo (chiamatela `var2`). Regredite una variabile rispetto all'altra. Il coefficiente della regressione è statisticamente significativo?

(b) Ora fate una simulazione ripetendo questo processo 100 volte. Questo può essere fatto attraverso un ciclo. Da ogni simulazione, salvate i punteggi z (i coefficienti stimati della variabile `var1` divisi per il proprio errore

standard). Se il valore assoluto di z eccede 2, la stima è statisticamente significativa. Questi sono i comandi in R per la simulazione:[4]

R code

```
z.scores <- rep (NA, 100) for (k in 1:100) {
  var1 <- rnorm (1000,0,1)
  var2 <- rnorm (1000,0,1)
  fit <- lm (var2 ~ var1)
  z.scores[k] <- coef(fit)[2]/se.coef(fit)[2]
}
```

Quanti di questi 100 punteggi sono statisticamente significativi?

(c) La cartella `child.iq` contiene un sottoinsieme di dati relativi a un campione di bambini e di madri come discusso precedentemente nel capitolo. Sono dati relativi al punteggi di un test di apprendimento di bambini di 3 anni di età, dati realtivi all'età della madre e all'età della madre al momento del parto per un campione di 400 bambini. I dati sono un in un file di Stata e possono essere letti e importati in R salvandoli nella working directory e usando il comando

R code

```
library ("foreign")

iq.data <- read.dta ("child.iq.dta")
```

(d) Stimate un modello di regressione sui punteggi dei bambini in funzione dell'età della madre, rappresentate graficamente i dati e il modello stimato, verificate le assunzioni del modello e interpretate il coefficiente della regressione. Quando consigliate alle madri di avere un figlio?

(e) Ripetete l'esercizio includendo anche il livello di istruzione delle madri, interpretate entrambi i coefficienti del modello. Le vostre conclusioni circa l'età del parto sono state modificate?

(f) Ora create una variabile indicatrice che rifletta il fatto che la madre abbia o meno completato gli studi superiori. Considerate l'interazione tra il completamento degli studi superiori e l'età della madre nella famiglia. Create inoltre un grafico che mostra due linee di regressione separate in funzione del completamento degli studi delle madri.

[4] Abbiamo inizializzato il vettore dei punteggi z usando valori mancanti (NAs). Un altro approccio è iniziare con `z.scores <- numeric(length=100)`, che da luogo ad un vettore iniziale di zeri. In generale, preferiamo inizializzare con gli NAs, in quanto si presenta un *bug* nel codice la presenza di valori mancanti come NAs nei risulti ci allerta sul problema.

(g) Infine, stimate un modello di regressione sui punteggi del test ottenuti dai bambini in funzione dell'età della madre al momento del parto e il loro livello di istruzione per i primi 200 bambini e usate questo modello per prevedere i punteggi per i successivi 200 bambini. Rappresentate graficamente il confronto tra i valori previsti e i valori osservati dei successivi 200 bambini.

4. La directory `beauty` contiene i dati tratti da Hamermeshe e Parker (2005) sulla valutazione degli studenti sulla qualità e sulla bellezza degli insegnanti di diversi corsi all'Università del Texas. Le valutazione sull'insegnamento sono state condotte alla fine del semestre, mentre i giudizi sulla bellezza sono stati fatti dopo da 6 studenti che non hanno seguito i corsi e non erano consapevoli della valutazione dei corsi.

   (a) Effettuare una regressione usando `beauty` (la variabile `btystdave`) per prevedere la valutazione del corso, (`courseevaluation`), controllando per altri vari input. Rappresentare graficamente il modello stimato e spiegare il significato di tutti i coefficienti, e delle loro deviazioni standard. Rappresentare graficamente i residui in funzione dei valori stimati.

   (b) Stimare altri modelli, includendo `beauty` e altre variabili di input. Considerare almeno un modello con interazione. Per ciascun modello, dite chi sono i *predittori* e chi sono gli input *inputs* are (si veda Sezione 2.1.) e spiegate il significato di ciascuno dei coefficienti.

   Si veda Felton, Mitchell, e (2003) per approfondimenti.