Capitolo 4

Regressione lineare: prima e dopo la stima del modello

Non sempre è appropriato stimare un modello di regressione lineare usando i dati grezzi. Come verrà discusso nelle Sezioni 4.1 e 4.4, trasformazioni lineari e logaritmiche possono talvolta aiutare nell'interpretazione del modello. Trasformazioni non lineari dei dati possono essere talvolta necessarie al fine di poter soddisfare le proprietà di additività e linearità, migliorando di conseguenza l'adattamento del modello. La Sezione 4.5 presenta altre trasformazioni univariate che sono occasionalmente utili. Abbiamo già discusso delle interazioni nella Sezione 3.3, nella Sezione 4.6 verranno considerate altre tecniche per combinare le variabili di input.

4.1 Trasformazioni lineari

Le trasformazioni lineari non modificano la stima del modello di regressione lineare e non alterano le previsioni: i cambiamenti negli input e nei coefficienti si annullano quando si considerano i valori stimati $X\hat{\beta}$.[1] In ogni caso, trasformazioni accurate e ben scelte possono migliorare l'interpretabilità dei coefficienti, rendendo il modello stimato più facile da interpretare. Abbiamo visto nel capitolo 3.1 come le trasformazioni lineari possano aiutare nell'interpretazione dell'intercetta; in questa sezione invece saranno considerati alcuni esempi relativi all'interpretazione degli altri coefficienti del modello.

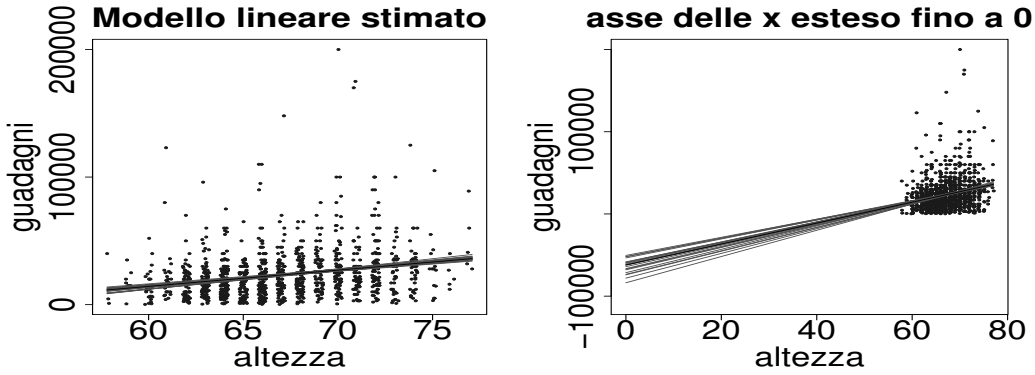
[1] Nei modelli multilevel invece, le trasformazioni lineari possono alterare sia la stima del modello che le sue previsioni, come sarà spiegato nella Sezione 13.6.

Scaling dei predittori e dei coefficienti di regressione. I coefficienti della regressione β_j rappresentano la differenza media nella variabile risposta y sulla base di unità che differiscono di 1 rispetto al j^{mo} predittore e sono identiche altrimenti. In alcuni casi, comunque, una differenza di 1 sulla scala delle ascisse x non sempre rappresenta il confronto più rilevante. Si consideri, ad esempio, un modello stimato su dati relativi ad un'indagine di adulti americani nel 1994 che prevede i loro guadagni (in dollari) in base alla loro altezza (in pollici):

$$\texttt{guadagni} = -61000 + 1300 \cdot \texttt{altezza} + \texttt{residuo}, \tag{4.1}$$

con una deviazione standard residua di 19000. (Un modello lineare non è realmente appropriato per questi dati, come verrà discusso a breve, ma continuiamo a considerare questo semplice esempio per introdurre il concetto delle trasformazioni lineari.)

Figura 4.1: *Regressione dei guadagni in funzione dell'altezza, guadagni* $= -61000+1300\cdot$ *height, la linea solida rappresenta il modello di regressione stimato mentre le linee chiare l'incertezza intorno al modello stimato. Nel grafico a destra la scala delle x è stata estesa fino allo zero al fine di far vedere l'intercetta della linea di regressione.*



Il grafico a destra della figura 4.1 mostra la linea di regressione e la relativa incertezza su una scala in cui l'asse delle x è stato esteso fino allo zero per poter rappresentare anche l'intercetta — il punto sull'asse delle y in cui la linea attraversa lo zero. Il valore stimato dell'intercetta pari a -61000 ha un significato alquanto limitato dal momento che corrisponde al valore stimato del guadagno di una persona di altezza pari a zero.

Ora consideriamo le seguenti forme alternative del modello:

$$\begin{aligned} \texttt{guadagni} &= -61000 + 51 \cdot \texttt{altezza} \text{ (in millimetri)} + \texttt{residuo} \\ \texttt{guadagni} &= -61000 + 81000000 \cdot \texttt{altezza} \text{ (in miglia)} + \texttt{residuo}. \end{aligned}$$

Quanto è importante l'altezza? Mentre \$51 non sembra avere molta importanza, al contrario \$81,000,000 riveste un importanza tutt'altro che trascurabile. Tuttavia, entrambe queste equazioni riflettono la stessa informazione sottostante i dati. Per meglio capire questi coefficienti, abbiamo bisogno di valutare il grado di variabilità dell'altezza della popolazione alla quale il nostro modello fa riferimento. Un modo per avere questa sensibilità è quello di considerare la deviazione standard dell'altezza nei dati che stiamo analizzando, che risulta pari a 3.8 pollici (ovvero 97 millimetri, o 0.000061 miglia). La differenza attesa nei guadagni corrispondente ad una differenza di 3.8 pollici nell'altezza, è pari a $\$1300 \cdot 3.8 = \$51 \cdot 97 = \$81000000 \cdot 0.000061 = \$4900$, valore ragionevolmente grande ma molto più piccolo che la deviazione standard dei residui, non spiegata dalla regressione, che è pari a \$19000 .

Standardizzazione z

Un altro modo di riscalare i coefficienti è quello di *standardizzare* il predittore, sottraendogli la media e dividendo per la deviazione standard, ottenendo in questo modo un punteggio del tipo "z." In questo esempio, l'`altezza` verrà sostituita da `z.altezza` = (`altezza` − 66.9)/3.8, e il coefficiente relativo associato a `z.altezza` sarà pari a 4900. Di conseguenza i coefficienti vengono interpretati in unità di deviazione standard del corrispondente predittore esattamente come lo erano nell'esempio precedente. Inoltre, la standardizzazione dei predittori modifica l'interpretazione dell'intercetta del modello, che corrisponde alla media di y, quando tutti i valori dei predittori sono pari ai rispettivi valori medi.

Noi preferiamo dividere per 2 volte la deviazione standard affinché ci sia una maggiore coerenza con gli input binari, come discusso nella Sezione 4.2.

Standardizzazione attraverso l'utilizzo di scale ragionevoli

Risulta spesso utile tenere alcuni input nelle loro solite scale come pollici, dollari, o anni, ma riscalandoli opportunamente in modo da migliorare l'interpretabilità dei coefficienti. Per esempio, talvolta risulterà opportuno lavorare con la variabile reddito espressa in decine di migliaia di dollari (reddito/\$10000) o con la variabile età espressa in decadi (età/10).

Per fare un altro esempio, in alcune indagini, l'identificazione dei partiti (PID) fa riferimento ad una scala, del tipo 1–7, che va da fortemente Repubblicano a fortemente Democratico. La variabile riscalata (PID − 4)/2, risulta pari a −1 per i Repubblicani, 0 per moderati e +1 per i Democratici, e in questo modo i coefficienti così riscalati sono direttamente interpretabili.

4.2 Centratura e standardizzazione, con particolare riferimento ai modelli con iterazione

La Figura 4.1b illustra la difficoltà di interpretazione dell'intercetta in un modello di regressione quando non ha alcun senso imporre che i predittori siano pari a zero. In particolare, problemi di questo tipo sono comuni quando si devono interpretare i coefficienti nei modelli con interazioni, come abbiamo visto nella Sezione 3.3 con riferimento al seguente modello:

R output

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

             coef.est coef.se
(Intercept)     -11.5    13.8
mom.hs           51.3    15.3
mom.iq            1.1     0.2
mom.hs:mom.iq    -0.5     0.2
  n = 434, k = 4
  residual sd = 18.0, R-Squared = 0.23
```

Il coefficiente di `mom.hs` pari a 51.3 sta a significare che i bambini le cui madri hanno un diploma di scuola superiore raggiungono, in media, un punteggio del test migliore di 51.3 punti? No, infatti il modello include una interazione e, di conseguenza, il valore di 51.3 è pari alla differenza attesa di punteggio tra i bambini con diverso `mom.hs`, ma *all'interno del gruppo dei bambini aventi* `mom.iq` $= 0$. Dal momento che, come osservato diverse volte, il coefficiente `mom.iq` non è mai vicino allo zero (si veda Figura 3.4 a pagina 36), il confronto rispetto allo zero e, di conseguenza, il coefficiente 51.3, è essenzialmente privo di senso.

Analogamente, il coefficiente pari a 1.1, che può essere considerato come "effetto principale" di `mom.iq`, è la pendenza della linea relativa a questa variabile, nel caso in cui `mom.hs` $= 0$. Questo coefficiente è poco rilevante (dal momento che `mom.hs` è uguale a uno per la maggior parte dei casi; si veda Figura 3.1 a pagina 30).

Centratura rispetto alla media dei dati

È possibile semplificare l'interpretazione del modello di regressione inizialmente sottraendo la media di ciascuna variabile di input:

R code

```
c.mom.hs <- mom.hs - mean(mom.hs)

c.mom.iq <- mom.iq - mean(mom.iq)
```

La regressione risultante è facile da interpretare, dal momento che ciascun effetto principale può essere interpretato come differenza rispetto agli altri input fissato al suo valor medio:

R output

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

                 coef.est coef.se
(Intercept)          87.6     0.9
c.mom.hs              2.8     2.4
c.mom.iq              0.6     0.1
c.mom.hs:c.mom.iq    -0.5     0.2
  n = 434, k = 4
  residual sd = 18.0, R-Squared = 0.23
```

La deviazione standard dei residui e R^2 non variano — infatti trasformazioni lineari dei predittori non hanno alcun effetto sulla stima del modello di regressione classico — così come il coefficiente e l'errore standard dell'interazione non cambiano, ma gli effetti principali e l'intercetta cambiano abbastanza e sono ora interpretabili in base al confronto della media dei dati.

Utilizzazione di un punto di centratura convenzionale

Un'altra opzione è quella di effettuare una centratura sulla base di un ragionevole punto di riferimento, per esempio il punto centrale dell'intervallo in cui varia `mom.hs` e la media dell'IQ dell'intera popolazione:

R code

```
c2.mom.hs <- mom.hs - 0.5
c2.mom.iq <- mom.iq - 100
```

In questa parametrizzazione, il coefficiente `c2.mom.hs` è la differenza media predittiva tra un bambino con `mom.hs` = 1 e `mom.hs` = 0, all'interno dell'insieme dei bambini con `mom.iq` = 100. Analogamente, il coefficiente di `c2.mom.iq` corrisponde ad un confronto per il caso `mom.hs` = 0.5, che non corrisponde ad alcun dato reale ma rappresenta il punto medio dell'intervallo.

R output

```
lm(formula = kid.score ~ c2.mom.hs + c2.mom.iq + c2.mom.hs:c2.mom.iq)

                  coef.est coef.se
(Intercept)           86.8     1.2
c2.mom.hs              2.8     2.4
c2.mom.iq              0.7     0.1
c2.mom.hs:c2.mom.iq   -0.5     0.2
  n = 434, k = 4
  residual sd = 18.0, R-Squared = 0.23
```

Ancora una volta, la deviazione standard dei residui, R^2, e il coefficiente per l'interazione non cambiano. L'intercetta e l'effetto principale cambiano di poco, dal momento che i punti 0.5 e 100 sono molto vicini alle medie di `mom.hs` e `mom.iq`.

Standardizzazione dei dati sottraendo la media e dividendo rispetto a due volte la deviazione standard

La centratura è di grande aiuto per l'interpretazione degli effetti maggiormente significativi nel modello di regressione, ma purtroppo comporta come conseguenza un problema di scala. Per esempio, il coefficiente di `mom.hs` risulta molto più elevato di quello associato alla variabile `mom.iq`, risultato fuorviante dal momento che si sta considerando il cambiamento completo in una variabile (madre che ha completato o meno la scuola superiore) rispetto alla variazione di 1 punto dell'IQ delle madri, che non è assolutamente elevato (si veda la figura 3.4 a pagina 36).

Un modo naturale per riscalare i predittori è dividere per 2 volte la deviazione standard (spiegheremo brevemente il perché utilizziamo 2 invece che 1) in modo che un cambiamento di una unità nel predittore riscalato corrisponde ad un cambiamento di 1 volta la deviazione standard sotto la media a 1 volta sopra la media. Qui sono i predittori riscalati nell'esempio del test dei bambini:

R code

```
z.mom.hs <- (mom.hs - mean(mom.hs))/(2*sd(mom.hs))
z.mom.iq <- (mom.iq - mean(mom.iq))/(2*sd(mom.iq))
```

Possiamo ora interpretare tutti i coefficienti su una scala più o meno (approssimativamente) comune (ad eccezione dell'intercetta, che ora corrisponde al valore medio atteso della variabile risposta avendo fissato tutti gli altri input al loro valor medio)

R output

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

                 coef.est coef.se
(Intercept)          87.6     0.9
z.mom.hs              2.3     2.0
z.mom.iq             17.7     1.8
z.mom.hs:z.mom.iq   -11.9     4.0
  n = 434, k = 4
  residual sd = 18.0, R-Squared = 0.23
```

Perché riscalare rispetto a due volte la deviazione standard?

Dividiamo per due volte la deviazione standard piuttosto che rispetto a una volta al fine di mantenere una certa coerenza quando consideriamo le variabili di input binarie. Per meglio illustrare questa situazione, si consideri una semplice variabile binaria x che assume valori 0 o 1 con probabilità 0.5. La deviazione standard di x è quindi pari a $\sqrt{0.5 \cdot 0.5} = 0.5$ e, di conseguenza, la variabile standardizzata, $(x - \mu_x)/(2\sigma_x)$, assume i valori ± 0.5 e il suo coefficiente riflette il confronto tra $x = 0$ e $x = 1$. Al contrario, se noi avessimo diviso per 1 volta la deviazione standard, la variabile riscalata avrebbe assunto valori pari a ± 1, e il suo coefficiente sarebbe stato associato ad una variazione pari alla metà della differenza tra i due possibili valori assunti dalla x. Questa identità vale anche quando si considerano input binari in cui le frequenze non sono uguali tra loro, dal momento che $\sqrt{p(1-p)} \approx 0.5$ quando p non è troppo lontano da 0.5.

In un modello di regressione complicato con molti predittori, potrebbe essere conveniente lasciare gli input binari così come sono e trasformare linearmente i predittori continui, possibilmente riscalandoli rispetto alla deviazione standard. In questo caso, la standardizzazione rispetto a due volte la deviazione standard assicura una sorta di conparabilità nei coefficienti. Nell'esempio del test effettuato sui bambini, la differenza predittiva corrispondente a due volte la deviazione standard dell'IQ delle madri è chiaramente molto più elevata del confronto tra le madri che hanno concluso o meno gli studi superiori.

Moltiplicare ciascun coefficiente di regressione per due volte la deviazione standard del suo predittore

Per modelli senza interazione, una procedura che coincide con il centrare e il riscalare è quella di lasciare i predittori esattamente come sono e creare dei nuovi coefficienti di regressione riscalati moltiplicando ciascun β per due volte la deviazione standard del corrispondente input x. Questa procedura fornisce in qualche modo il senso dell'importanza di ciascuna variabile nel modello lineare, tenendo sotto controllo tutte le altre variabili. Come precedentemente osservato, riscalare per due volte la deviazione standard (piuttosto che rispetto a una volta) permette a questi coefficienti riscalati di poter essere confrontati con quelli non riscalati dei predittori binari.

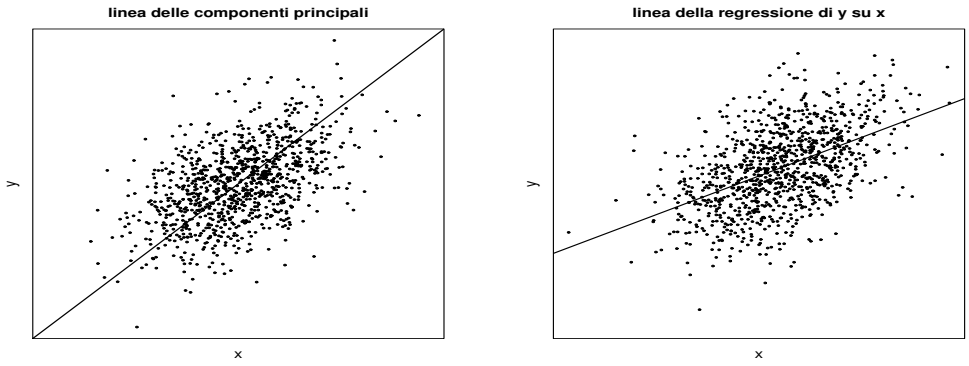
4.3 Correlazione e "regressione rispetto alla media"

Si consideri una regressione con un solo predittore (in aggiunta al termine costante); ovvero $y = a+bx+$errore. Se x e y sono entrambe standardizzate — nel senso che entrambe sono definite come `x <- (x-mean(x))/sd(x)` e `y <- (y-mean(y))/sd(y)` — allora l'intercetta della retta di regressione è uguale a zero e la pendenza della curva è semplicemente pari al coefficiente di correlazione tra x e y. Quindi, la pendenza della curva di regressione deve essere sempre compresa tra 1 e -1, ovvero, detto in un altro modo, se la pendenza della regressione è maggiore di 1 o minore di -1, la varianza di y deve essere superiore alla varianza di x. In generale, la pendenza di una regressione con un solo predittore è pari a $b = \rho\sigma_y/\sigma_x$, dove ρ è il coefficiente di correlazione tra le due variabili e σ_x e σ_y sono le deviazioni standard di x e y.

La linea delle componenti principali e la linea di regressione

Alcuni degli aspetti poco chiari della regressione possono essere facilmente compresi nel caso più semplice in cui si considerano variabili standardizzate. La Figura 4.2 mostra un esempio di variabili simulate standardizzate con coefficiente di correlazione (e quindi pendenza della linea di regressione) pari a 0.5. Il grafico a sinistra mostra la *linea delle componenti principali*, ossia la linea che passa il più vicino possibile alla nuvola dei punti, nel senso che minimizza la somma delle distanze Euclidee al quadrato tra i punti e la linea. La linea delle componenti principali in questo caso è semplicemente $y = x$. Il grafico sulla destra, sempre nella Figura 4.2 mostra invece la *linea di regressione*, linea che minimizza la somma dei quadrati della distanza *verticale* tra i punti e la linea — la familiare linea dei minimi quadrati, $y = \hat{a} + \hat{b}x$,

Figura 4.2: *Dati simulati da una distribuzione normale bivariata con coefficiente di correlazione 0.5. La linea di regressione, che rappresenta la migliore previsione di y dato x, ha una pendenza pari alla metà della pendenza della linea delle componenti principali, che passa il più vicino possibile alla nuvola dei punti.*



con $\hat{a}, \hat{b}$ scelti in modo tale da minimizzare $\sum_{i=1}^{n}(y_i - (\hat{a} + \hat{b}x_i))^2$. In questo caso, $\hat{a} = 0$ e $\hat{b} = 0.5$; la linea della regressione ha una pendenza uguale a 0.5.

Quando si chiede di tracciare una linea di regressione di y su x sulla base di un scatterplot (senza alcuna linea sovrapposta), gli studenti tendono a disegnare la linea delle componenti principali come mostrato nella Figura 4.2a. Comunque, al fine di prevedere y sulla base di x, o per stimare la media di y per ogni dato valore di x, la linea di regressione è infatti migliore, anche se a priva vista non sembrerebbe.

La superiorità della linea di regressione per stimare la media di y dato x può essere vista da uno studio attento della Figura 4.2. Per esempio, si considerino i punti all'estrema sinistra di entrambi i grafici. Questi punti che giacciono sulla linea delle componenti principali sono circa per metà sopra e per metà sotto la linea di regressione. Quindi, la linea delle componenti principali sottostima y per valori bassi di x. In modo del tutto simile, uno studio attento nella parte destra del grafico mostra che la linea delle componenti principali sovrastima y per elevati valori di x. Al contrario, la linea di regressione ancora una volta fornisce delle previsioni non distorte, nel senso che passa per i valori medi di y condizionati a x.

Regressione rispetto alla media

Ricordiamo che quando x e y sono entrambe standardizzate (ovvero sono entrambe definite su una stessa scala comune, come nella Figura 4.2), la linea di regressione ha una pendenza che è sempre minore di 1. Quindi, quando x si trova al di sopra della sua media di una quantità pari a 1 volta la deviazione standard, il valore

previsto di y si trova al di sopra della media di una quantità compresa tra 0 e 1 volta la deviazione standard. Questo fenomeno nei modelli lineari —ovvero quando si prevede che y sia più vicino alla media (in unità di deviazione standard) rispetto a x — è chiamata *regressione rispetto alla media*, situazione che si presenta in diversi contesti.

Per esempio, se una donna è più alta di 10 pollici rispetto alla media delle donne, e la correlazione tra l'altezza delle madri e l'altezza dei figli maschi è 0.5, allora l'altezza stimata del figlio è di 5 pollici superiore alla media. Quindi ci si aspetta che il figlio sia più alto della media, ma non troppo—ovvero una "regressione" (non in senso statistico) sulla media.

Un simile calcolo può essere effettuato per ogni coppia di variabili che non sono perfettamente correlate. Per esempio, siano x_i e y_i il numero di partite vinte da una squadra di baseball i in due stagioni successive. Non essendo correlate al 100% ci si aspetta che le squadre migliori nella prima stagione (ovvero quelle con più alto valore di x) non siano le migliori anche nella seconda stagione, (ovvero con valori di y più vicini alla media per tutte le squadre). In modo analogo, ci si aspetta che una squadra che ha riportato nella prima stagione scarsi risultati migliori nella seconda stagione.

Un'interpretazione "naive" della regressione sulla media è che l'altezza, o i punteggi del baseball, o variabili che indicano i fenomeni di interesse convergano nel tempo verso la media. Questa visione è ogni caso non corretta dal momento che viene completamente ignorato l'errore nella regressione quando si vuole prevedere y a partire dalla x. Per ciascun punto x_i, il corrispondente valore previsto y_i verrà regredito sulla media, ma il valore osservato di y_i non è esattamente lo stesso del valore previsto. Alcuni punti possono cadere molto vicini alla media e alcuni parecchio distanti, come si può vedere nella Figura 4.2b.

4.4 Trasformazioni Logaritmiche

Quando l'additività e la linearità non sono ipotesi ragionevoli (si veda Sezione 3.6) risulta talvolta necessario considerare delle trasformazioni non lineari. È consuetudine prendere il logaritmo della variabile risposta quando tutti i valori sono positivi. Per variabili risposta, questa trasformazione diventa chiara quando si vogliono effettuare delle previsioni sulla scala originale. Il modello di regressione non impone alcun vincolo che faccia in modo che anche i valori previsti siano sempre positivi. Invece, se si prendono i logaritmi della variabile, si stima il modello, si fanno previsioni su scala logaritmica e quindi si riportano i valori stimati sulla scala originale (attraverso l'utilizzo della trasformazione esponenziale), i valori così previsti sono

necessariamente positivi in quanto $\exp(a) > 0$ per ogni valore reale di a.

Ancora più importante risulta sottolineare il fatto che un modello lineare su scala logaritmica corrisponde ad un modello moltiplicativo sulla scala originale. Si consideri il seguente modello di regressione lineare,

$$\log y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + \epsilon_i.$$

Passando all'esponenziale da entrambe le parti dell'equazione,

$$\begin{aligned} y_i &= e^{b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + \epsilon_i} \\ &= B_0 \cdot B_1^{X_{i1}} \cdot B_2^{X_{i2}} \cdots E_i, \end{aligned}$$

dove $B_0 = e^{b_0}$, $B_1 = e^{b_1}$, $B_2 = e^{b_2}$, ...sono i coefficienti di regressione di cui si è stato preso l'esponenziale (e quindi sempre positivi) e $E_i = e^{\epsilon_i}$ è il termine di errore anch'esso trasformato attraverso la funzione esponenziale (quindi sempre positivo). Sulla scala originale dei dati y_i, i predittori $X_{i1}, X_{i2}, \ldots$ compaiono in modo moltiplicativo.

Esempio sull'altezza e il guadagno

Illustreremo la regressione logaritmica considerando dei modelli per stimare i guadagni a partire dall'altezza. L'espressione (4.1) a pagina 65 mostra una regressione lineare dei guadagni sull'altezza. Tuttavia, sembra avere più senso modellare i guadagni su una scala logaritmica (in questo modello sono state escluse tutte quelle persone che riportavano un guadagno nullo).

Andremo quindi a stimare un modello di regressione sul logaritmo dei guadagni per poi utilizzare la trasformazione esponenziale per ottenere i valori previsti sulla scala originale.

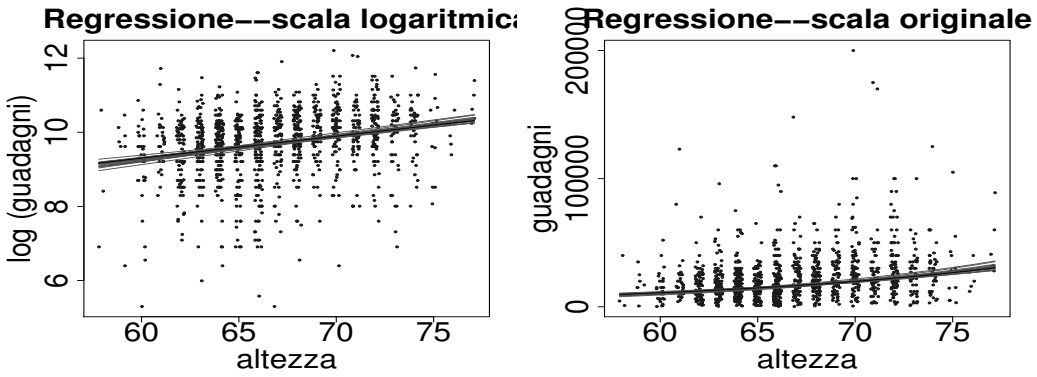
Interpretazione diretta dei coefficienti su scala logaritmica. Prendendo il logaritmo dei guadagni e regredendolo rispetto all'altezza,

R code

```
log.earn <- log (earn)
earn.logmodel.1 <- lm (log.earn ~ height)
display (earn.logmodel.1)
```

otteniamo il seguente modello stimato:

Figura 4.3: *Grafico della regressione dei guadagni sull'altezza su scala logaritmica. Le linee solide rappresentano il modello di regressione logaritmico stimato, log(guadagni) =* $5.78 + 0.06 \cdot$ *altezza. Si confronti questo grafico con quello relativo al modello lineare (si veda Figura 4.1a).*



R output

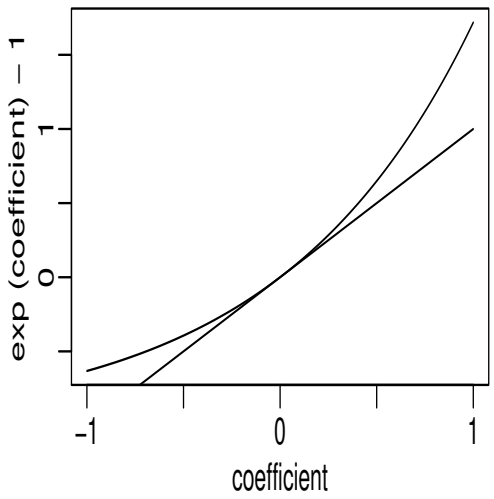
```
lm(formula = log.earn ~ height)

            coef.est coef.se
(Intercept)     5.74    0.45
height          0.06    0.01
  n = 1192, k = 2
  residual sd = 0.89, R-Squared = 0.06
```

Il coefficiente stimato $\beta_1 = 0.06$ implica che una differenza di 1 pollice nell'altezza corrisponde ad una differenza positiva attesa di 0.06 nel logaritmo del guadagno. In questo modo i guadagni sono moltiplicati per $\exp(0.06)$; ma $\exp(0.06) \approx 1.06$ (più precisamente, è pari a 1.062). Quindi, una differenza di 1 unità nel predittore corrisponde ad una differenza attesa positiva di circa il 6% nella variabile risposta. In modo del tutto simile, se β_1 fosse -0.06, allora una differenza positiva di 1 pollice nell'altezza corrisponderebbe ad una differenza *negativa* di circa il 6% nei guadagni.

Questa approssimazione diventa tanto più debole quanto più aumentano i valori dei coefficienti. La Figura 4.4 mostra come questa relazione diventi meno significativa al crescere dei valori dei coefficienti. Il grafico è stato ristretto ai coefficienti nell'intervallo $(-1, 1)$ in quanto, su scala logaritmica, i coefficienti di regressione sono tipicamente (anche se non sempre) minori di 1. Un coefficiente pari a 1 su scala logaritmica implica che un cambiamento di una unità nel predittore equivale ad un cambiamento di $\exp(1) = 2.7$ nella variabile risposta, e se i predittori sono parametrizzati in modo ragionevole, è inusuale vedere effetti di queste dimensioni.

Figura 4.4: *Interpretazione dei coefficienti trasformati esponenzialmente in un modello di regressione logaritmico in termini di differenza relativa (linea curva), con relativa approssimazione* $exp(x) = 1 + x$ *valida quando i coefficienti di* x *presentano valori contenuti (linea retta).*



Perché utilizzare il logaritmo naturale piuttosto che il logaritmo in base 10

Noi preferiamo utilizzare il logaritmo naturale (ovvero il logaritmo in base e) in quanto, come descritto sopra, i coefficienti sono direttamente interpretabili approssimativamente in termini di differenze proporzionali: con un coefficiente pari a 0.06, una differenza di 1 nella x corrisponde approssimativamente ad una differenza pari al 6% nella y, e così via.[2]

Un altro approccio è quello di prendere il logaritmo in base 10, che possiamo scrivere come $\log_{10}$. La relazione tra le due scale è che $\log_{10}(x) = \log(x)/\log(10) = \log(x)/2.30$. Il vantaggio nello scegliere $\log_{10}$ è che il valore previsto è in qualche modo più facile da interpretare; per esempio, quando consideriamo la regressione dei guadagni, $\log_{10}(10000) = 4$ e $\log_{10}(100000) = 5$, e con un po' di esperienza è possibile leggere velocemente i valori intermedi — per esempio, se $\log_{10}(\text{guadagni}) = 4.5$, allora guadagni ≈ 30000.

Lo svantaggio nell'utilizzo di $\log_{10}$ è che i risultanti coefficienti sono difficili da interpretare. Per esempio, se noi definiamo:

R code

```
log10.earn <- log10 (earn)
```

la regressione sull'altezza risulta pari a:

R output

```
lm(formula = log10.earn ~ height)

            coef.est coef.se
(Intercept) 2.493    0.197
height      0.026    0.003
  n = 1187, k = 2
  residual sd = 0.388, R-Squared = 0.06
```

Il coefficiente di 0.026 ci dice che una differenza di 1 pollice nell'altezza corrisponde ad una differenza di 0.026 nel $\log_{10}(\text{guadagni})$; ovvero una differenza moltiplicativa di $10^{0.026} = 1.062$. Questo equivale al cambiamento del 6% di prima, ma ora non può essere visto semplicemente guardando al coefficiente come se questo fosse su scala logaritmica naturale.

[2] Il logaritmo naturale è scritto talvolta come "ln" o "$\log_e$" ma noi scriveremo semplicemente "log" in quanto questo sarà il nostro riferimento.

Costruzione di un modello di regressione su scala logaritmica

Aggiunta di un altro predittore. Un incremento pari a un pollice nell'altezza corrisponde a un incremento del 6% nel guadagno — incremento che sembra molto elevato! Ma gli uomini sono in genere più alti delle donne e quindi tendono ad avere guadagni più consistenti. Per esempio la differenza del 6% può essere "spiegata" dalla differenza nel sesso. Ci si potrebbe chiedere a questo punto se persone dello stesso sesso più alte guadagnano di più di quelle più basse.

Possiamo rispondere a questa domanda includendo nel modello di regressione la variabile sesso — in questo caso un predittore `male` che risulta pari a 1 per gli uomini e 0 per le donne:

R output

```
lm(formula = log.earn ~ height + male)

            coef.est coef.se
(Intercept)     8.15    0.60
height          0.02    0.01
male            0.42    0.07
  n = 1192, k = 3
  residual sd = 0.88, R-Squared = 0.09
```

Dopo aver controllato per la variabile sesso, un incremento di altezza di entità pari ad un pollice corrisponde ad una differenza stimata attesa pari al 2%: in base a questo modello due persone dello stesso sesso che differiscono di 1 pollice nell'altezza differiscono, in media, del 2% nei guadagni. La differenza attesa rispetto al sesso, comunque, è enorme: confrontando un uomo e una donna della stessa altezza, i guadagni degli uomini sono $\exp(0.42) = 1.52$ volte quelli delle donne; che equivale ad un 52% in più. (Non possiamo convertire semplicemente 0.42 con 42% dal momento che questo coefficiente non è così vicino allo zero; si veda Figura 4.2.)

Codifica degli input. Incidentalmente, abbiamo chiamato la nuova variabile di input `male` in modo tale che essa possa essere direttamente interpretabile. Avendo chiamato la variabile sesso(`sex`), per esempio, si potrebbe tornare indietro alla codifica per verificare se 0 e 1 si riferiscono agli uomini o alle donne, o viceversa[3].

[3]Un altro approccio potrebbe essere quello di considerare la variabile sesso(`sex`) come un fattore a due livelli, `male` e `female`; si veda pagina 89. Il nostro punto qui è che se la variabile viene codificata numericamente, è conveniente assegnare al nome `male` la codifica 1.

Controllo della significatività statistica. La differenza tra i sessi è molto grande e ben nota, ma anche la differenza dovuta all'altezza è ugualmente interessante — una differenza pari al 2%, per guadagni di \$50000, diventa un rilevante \$1000 per pollice. Per valutare la significatività statistica, possiamo verificare se il coefficiente stimato è lontano dallo zero per un valore superiore a 2 volte l'errore standard. In questo caso, con una stima di 0.02 e un errore standard di 0.01 abbiamo bisogno di utilizzare un numero di cifre decimali pari a 3 per essere sicuri (usando l'opzione `digits` nella funzione `display()`):

R output

```
lm(formula = log.earn ~ height + male)

            coef.est coef.se
(Intercept)    8.153   0.603
height         0.021   0.009
male           0.423   0.072
  n = 1192, k = 3
  residual sd = 0.88, R-Squared = 0.09
```

Il coefficiente dell'altezza risulta in effetti significativo. Un altro modo per valutare la significatività è calcolare direttamente l'intervallo di confidenza al 95% sulla base di simulazioni inferenziali, come sarà discusso nella Sezione 7.2.

La deviazione standard dei residui e R^2. L'ultimo modello presentato ha una deviazione standard dei residui uguale a 0.88, il che implica che approssimativamente nel 68% dei casi la distanza tra il valore effettivo dei guadagni in termini logaritmici e il corrispondente valore stimato sarà minore di ± 0.88. Sulla scala originale, circa il 68% dei guadagni sarà all'interno dell'intervallo della previsione a meno di un fattore pari a $\exp(0.88) = 2.4$. Per esempio una persona alta 70 pollici ha un valore stimato del guadagno di $8.153 + 0.021 \cdot 70 = 9.623$, con un'associata deviazione standard di previsione di circa 0.88. Quindi vi è una probabilità pari a circa il 68% che questa persona abbia un guadagno (in log) entro l'intervallo $[9.623 \pm 0.88] = [8.74, 10.50]$, che corrisponde ad avere un guadagno nell'intervallo $[\exp(8.74), \exp(10.50)] = [6000, 36000]$. Questo intervallo molto ampio suggerisce che il modello di regressione stimato non è un buon modello per la stima dei guadagni—non possiamo essere soddisfatti di una stima che potrebbe discostarsi dal vero valore di un fattore pari a 2.4—questo risultato viene ulteriormente confermato da un R^2 il cui valore pari a 0.09 ci dice che solo il 9% di varianza dei dati viene spiegata dal modello. Questo valore dell'R^2 molto basso è visibile anche dalla

Figura 4.2 dove risulta evidente che il campo di variazione dei valori previsti dalla regressione è molto più piccolo se confrontato con il campo di variazione dei dati.

Inclusione di un'iterazione. Consideriamo ora un modello con un'interazione tra altezza e sesso, in modo tale che il confronto predittivo in base all'altezza possa essere differenziato tra uomini e donne:

R code
```
earn.logmodel.3 <- lm (log.earn ~ height + male + height:male)
```

che fornisce,

R output
```
             coef.est coef.se
(Intercept)     8.388   0.844
height          0.017   0.013
male           -0.079   1.258
height:male     0.007   0.019
  n = 1192, k = 4
  residual sd = 0.88, R-Squared = 0.09
```

Ovvero,

$$\log(\text{earnings}) = 8.4 + 0.017 \cdot \text{height} - 0.079 \cdot \text{male} + 0.007 \cdot \text{height} \cdot \text{male}. \quad (4.2)$$

Proviamo ad interpretare ciascuno dei coefficienti in questo modello.

- L'*intercetta* è il valore previsto del logaritmo dei guadagni quando sia l'altezza (`height`) che il sesso maschile (`male`) sono uguali a zero. Poiché l'altezza non sarà mai vicina a zero, l'intercetta di questo modello non ha un'interpretazione diretta.
- Il coefficiente dell'altezza (`height`) è la differenza attesa nel logaritmo del guadagno corrispondente a una differenza nell'altezza pari a 1 pollice, se `male` è pari a zero. Quindi, la differenza attesa stimata per pollice di altezza è pari nelle donne a 1.7%. La stima dista da zero meno di due volte l'errore standard, il che indica che i dati sono anche compatibili con lo zero o con una differenza negativa.

- Il coefficiente relativo al sesso maschile (`male`) rappresenta la differenza attesa nel logaritmo del guadagno tra gli uomini e le donne, quando l'altezza (`height`) è uguale a 0. Dal momento che l'altezza è mai vicina allo zero, il coefficiente relativo al sesso maschile (`male`) non ha un'interpretazione diretta in questo modello. (Abbiamo già incontrato questo modello; per esempio, consideriamo la differenza tra le intercette delle due linee di regressione nella Figura 3.4b a pagina 36.)
- Il coefficiente relativo all'interazione tra altezza e sesso maschile (`height:male`) è la differenza nelle pendenze delle linee che stimano il logaritmo dei guadagni, confrontando uomini e donne. Quindi, un incremento di altezza pari a un pollice corrisponde a un incremento dello 0.7% nei guadagni a vantaggio degli uomini, con una differenza predittiva stimata per pollice di altezza pari a $1.7\% + 0.7\% = 2.4\%$.

Il coefficiente dell'interazione non è statisticamente significativo, ma è plausibile che la correlazione tra altezza e guadagni sia più importante per gli uomini che per le donne, per cui decidiamo di tenere l'interazione nel modello, seguendo il principio discusso nella Sezione 4.6.

Trasformazioni lineari che rendono i coefficienti meglio interpretabili. Possiamo rendere i parametri dell'interazione più facilmente interpretabili riscalando il predittore dell'altezza in modo tale che abbia media pari a 0 e deviazione standard pari a 1.

R code
```
z.height <- (height - mean(height))/sd(height)
```

In base a questi dati, la media dell'altezza (`mean(height)`) e la deviazione standard del peso (`sd(height)`) sono rispettivamente 66.9 e 3.8 pollici. Stimando il modello rispetto all'altezza standardizzata (`z.height`), al sesso maschile (`male`), e alla loro interazione otteniamo,

R output
```
lm(formula = log.earn ~ z.height + male + z.height:male)
              coef.est coef.se
(Intercept)       9.53    0.05
z.height          0.07    0.05
male              0.42    0.07
z.height:male     0.03    0.07
  n = 1192, k = 4
  residual sd = 0.88, R-Squared = 0.09
```

Interpretiamo ora i quattro coefficienti:

- L'*intercetta* rappresenta il valore stimato del logaritmo del guadagno se `z.height` e `male` sono entrambi uguali a zero. Quindi, una donna alta 66.9 pollici (170 cm) avrà un guadagno stimato in termini logaritmici pari a 9.53, ovvero un guadagno stimato pari a $\exp(9.53) = 14000$.
- Il coefficiente relativo a `z.height` rappresenta la differenza attesa nel logaritmo del guadagno corrispondente a un incremento di altezza pari a 1 volta la deviazione standard, quando `male` è uguale a zero. Quindi, la differenza attesa di guadagno (in log) in seguito ad un incremento di altezza di 3.8 pollici è pari al 7% per le donne.
- Il coefficiente di `male` rappresenta la differenza attesa nel logaritmo del guadagno tra uomini e donne, quando `z.height` è uguale a zero. Quindi, un uomo alto 66.9 pollici avrà un guadagno stimato su scala logaritmica che sarà 0.42 volte più elevato del guadagno di una donna della stessa altezza. Questo corrisponde ad un rapporto di $\exp(0.42) = 1.52$, ovvero gli uomini hanno un guadagno stimato superiore del 52% rispetto a quello delle donne.
- Il coefficiente dell'interazione `z.height:male` è invece la differenza nelle pendenze delle linee che mettono in relazione i logaritmi dei guadagni e l'altezza, confrontando gli uomini e le donne. Ovvero, una differenza di 3.8 pollici di altezza corrisponde a un incremento del 3% nei salari degli uomini rispetto ai salari delle donne, così che l'incremento atteso complessivo per gli uomini è pari a $7\% + 3\% = 10\%$.

Si sarebbe anche potuto centrare il predittore relativo al sesso, ma in questo caso era abbastanza semplice interpretare `male` $= 0$, che corrisponde alla categoria base (in questo esempio le donne).

Ulteriori difficoltà nell'interpretazione

Per avere un'idea su altre possibili difficoltà nell'interpretare i coefficienti di regressione, consideriamo il modello più semplice relativo al logaritmo del guadagno senza termine di interazione. L'interpretazione del coefficiente relativo all'altezza (height) è relativamente semplice: confrontando due adulti dello stesso sesso, la persona più alta guadagnerà il 2% in più per un incremento di altezza pari a un pollice (si veda il modello a pagina 78). E questo sembra un confronto ragionevole.

Relativamente al coefficiente del sesso (sex), si potrebbe dire: confrontando due adulti della stessa altezza ma di sesso differente, ci si aspetta un incremento di

guadagno degli uomini, rispetto alle donne, del 52%. Ma ci chiediamo: potrebbe essere un confronto ragionevole? Per esempio, se confrontiamo un uomo alto 66 pollici e una donna alta 66 pollici, in realtà stiamo confrontando una donna alta con un uomo basso. Quindi, in un certo senso, non differiscono solo nel sesso. Per esempio, un confronto più appropriato potrebbe essere quello di confrontare una donna di "altezza media" con un uomo di "altezza media". La soluzione a questo tipo di problema dipende innanzitutto dalle motivazioni che hanno spinto a costruire il modello. Per il momento focalizziamo l'attenzione sui problemi tecnici relativi all'adattamento di modelli ragionevoli ai dati. Ritorneremo su questi punti relativi all'interpretazione nei Capitoli 9 e 10.

Modello doppio logaritmo: trasformazione della variabile risposta e della variabile di input

Se applichiamo la trasformazione logaritmica sia alla variabile risposta che alla variabile di input, i coefficienti possono essere interpretati come la variazione proporzionale attesa nella variabile y a seguito di una variazione proporzionale nella variabile x. Per esempio:

R output

```
lm(formula = log.earn ~ log.height + male)

            coef.est coef.se
(Intercept)     3.62    2.60
log.height      1.41    0.62
male            0.42    0.07
  n = 1192, k = 3
  residual sd = 0.88, R-Squared = 0.09
```

Per ogni differenza nell'altezza pari all'1%, la differenza attesa nel guadagno è pari all'1.41%. L'altro input `male`, è categorico, per cui non ha senso prenderne il logaritmo.

In economia, il coefficiente di un modello doppio logaritmico si chiama "elasticità"; si veda l'Esercizio 4.6 per un esempio.

Considerare i logaritmi quando non è necessario

Se una variabile ha un *range* molto stretto (per esempio se il rapporto tra il valore più alto e quello più basso è vicino ad 1) allora non vi è troppa differenza tra la stima

di un modello su scala logaritmica o su scala originale. Per esempio, se la deviazione standard del logaritmo dell'altezza (`log.height`) nel nostro esempio è 0.06, significa che l'altezza relativa ai nostri dati varia approssimativamente solo di un fattore pari al 6%. In una tale situazione si preferisce, almeno per semplicità, rimanere sulla scala originale. In ogni caso, la trasformazione logaritmica potrebbe avere un senso in quanto i coefficienti risultano molto spesso più facilmente interpretabili quando sono su scala logaritmica. La scelta della scala in situazioni di questo tipo è legata essenzialmente all'interpretabilità: se risulta più semplice capire il modello in termini di aumento proporzionale dei guadagni per pollice, o in termini di incremento proporzionale in altezza.

Quando si hanno input con elevata variabilità (per esempio, l'altezza dei bambini o il peso degli animali), potrebbe avere senso lavorare con i logaritmi da subito, in quanto si avrebbe sia una migliore interpretazione che un miglior adattamento del modello.

4.5 Altre trasformazioni

Trasformazione radice quadrata

La radice quadrata è talvolta utile per comprimere valori elevati in maniera più moderata rispetto a quanto farebbe la trasformazione logaritmica. Consideriamo ancora l'esempio relativo all'altezza e ai guadagni.

Sembrerebbe inappropriato stimare un modello lineare direttamente sui dati grezzi. Detto in modo diverso rispetto a prima, ci si aspetterebbe che la differenza tra persone che non guadagnano nulla rispetto alle persone che guadagnano $10000 sia superiore alla differenza tra, per esempio, chi guadagna $80000 rispetto a chi guadagna $90000. Ma in base al modello lineare, questi incrementi sono considerati uguali (come nel modello (4.1)), dove un aumento di un pollice nell'altezza corrisponde ad un aumento di $1300 nei guadagni, qualsiasi sia il livello di partenza (iniziale).

D'altra parte, la trasformazione logaritmica sembra essere troppo rigida con questi dati. In termini logaritmici, la differenza tra chi guadagna $5000 rispetto a $10000 è equivalente alla differenza tra chi guadagna $40000 e chi ne guadagna $80000. Considerando come scala la radice quadrata, in ogni caso, la differenza tra gruppi di persone che guadagnano 0 e $10000 è all'incirca la stessa differenza che esiste quando si confrontano gruppi di persone che guadagnano $10000 e $40000 ovvero gruppi che guadagnano $40000 e $90000. (Queste differenze si muovono da 0 a 100, 200, e 300 sulla scala che fa riferimento alla radice quadrata.) Si veda il

Capitolo 25 per ulteriori informazioni su questo esempio.

Sfortunatamente, i modelli che si basano su trasformazioni che fanno riferimento alla radice quadrata perdono la chiara interpretazione che si ha quando si considera la scala originale e i modelli che si stimano su scala logaritmica. Le previsioni negative e di una certa entità diventano sulla scala originale quadratiche e assumono valori positivi e molto elevati, introducendo quindi nel modello una non monotonicità. In generale, riteniamo molto più utile usare la radice quadrata nei modelli per effettuare previsioni piuttosto che per interpretare i coefficienti.

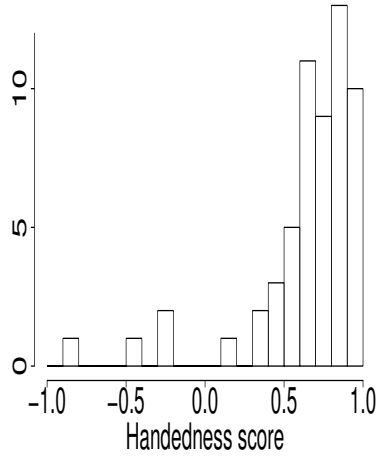
Trasformazioni idiosincratiche

Talvolta, risulta conveniente sviluppare particolari trasformazioni che si adattino a problemi specifici. Per esempio, relativamente ai dati originali dell'altezza-guadagno non è detto che sia possibile prendere semplicemente la trasformata logaritmica del guadagno, dal momento che ci sono molte osservazioni con valori pari a zero. Invece, un modello ragionevole potrebbe essere costruito in due fasi: (1) modellare la probabilità che i guadagni eccedano lo zero (per esempio attraverso un modello di regressione logistica; si veda il Capitolo **??**); (2) stimare un modello di regressione, condizionatamente ai guadagni positivi, che è quello che abbiamo fatto nell'esempio precedente. Si potrebbe anche modellare il reddito complessivo, ma gli economisti sono più interessati in genere ai soli guadagni.

In ogni caso, le rappresentazioni grafiche e le simulazioni dovrebbero definitivamente essere usate nel sintetizzare l'inferenza del modello, dal momento che i coefficienti delle due parti del modello si combinano in modo non lineare nella previsione congiunta dei guadagni. Discuteremo ancora di questa tipologia di modelli nelle Sezioni 6.7 e 7.4.

Che tipo di trasformazione potrebbe quindi essere appropriata per una variabile quale per esempio "patrimonio" che può essere negativa, positiva o nulla? Una possibilità è quella di ricodificarla in modo da comprimere il campo di variazione, assegnando per esempio il valore 0 per patrimoni compresi tra $[-\$100, \$100]$, 1 per patrimoni tra \$100 e \$1000, 2 per patrimoni tra \$1000 e \$10000, -1 per valori compresi tra $-\$100$ e $-\$10000$, e così via. Una mappatura di questo tipo potrebbe essere espressa in modo più appropriato su una scala continua, ma a fini esplicativi è più conveniente usare una scala discreta.

Figura 4.5: *Istogramma relativo al punteggio della variabile "handedness" su un campione di studenti. I punteggi variano da* -1 *(completamente mancino) a* $+1$ *(completamente destro) e sono basati su risposte a 10 quesiti del tipo "Con quale mano scrivi?" e "Con quale mano tieni il cucchiaio?" Il* range *continuo dei possibili valori ottenuti mette in evidenza il limite di trattare la variabile "handedness" come una variabile dicotomica. Da Gelman e Nolan (2002).*



Utilizzo di predittori continui piuttosto che discreti

Molte variabili che sembrano binarie o discrete possono essere utilmente considerate come continue. Per esempio, piuttosto che la variabile definita come "handedness" che significa utilizzo della mano destra o sinistra e che viene considerata come una variabile binaria che assume valore -1 per i mancini e $+1$ per i solo destri, si potrebbe utilizzare una scala continua per sintetizzare l'utilizzo della mano che va da -1 a 1 (si veda la figura 4.3).

In generale, noi evitiamo di discretizzare le variabili continue (eccetto come modo di semplificare una trasformazione complicata, come descritto in precedenza, o per modellare la non linearità, come verrà descritto poi). Un errore tipico è quello di prendere una misura numerica e sostituirla con un punteggio binario "successo/fallimento". Per esempio, supponiamo che si voglia provare a prevedere i vincitori di una campagna elettorale, piuttosto che prevedere i voti su una scala continua. Un modello di questo tipo potrebbe non funzionare bene, poiché si perde molta dell'informazione disponibile nei dati (per esempio, la distinzione tra un candidato che riceve il 51% o il 65% dei voti). Il modello potrebbe potrebbe "sprecare la sua potenzialità" nella vaga speranza di prevedere in modo quasi puntuale il vincitore. Anche se il nostro obiettivo è quello di prevedere i vincitori, potremmo fare

di meglio provando a prevedere le percentuali di voto su una scala continua e quindi trasformarle in previsioni di vincitori, come nell'esempio delle elezioni congressuali nella Sezione 7.3.

Utilizzo di predittori discreti piuttosto che continui

In alcuni casi, comunque, risulta più appropriato discretizzare una variabile continua quando una relazione di tipo monotona o quadratica sembra non essere appropriata. Per esempio, nel modellare le preferenze politiche, potrebbe aver senso includere l'età con quattro variabili indicatrici: 18–29, 29–44, 45–64, e 65+, che permette di tener conto dei diversi comportamenti generazionali. Inoltre, variabili che assegnano dei valori numerici a categorie ordinate ma che distano tra loro non in modo uniforme sono spesso i migliori candidati per la discretizzazione.

Per esempio, il capitolo precedente descrive il punteggio stimato di un test effettuato su bambini sulla base delle informazioni sulle loro madri. Un'altra variabile di input che può essere usata in questi modelli è l'occupazione delle madri che è definita su una scala ordinata a 4 valori:

- `mom.work` = 1: madre che non ha lavorato nei primi tre anni di vita del bambino;
- `mom.work` = 2: madre che ha lavorato nel secondo o nel terzo anno di vita del bambino;
- `mom.work` = 3: madre che ha lavorato part-time nel primo anno di vita del bambino;
- `mom.work` = 4: madre che ha lavorato full-time nel primo anno di vita del bambino.

La stima di un semplice modello usando predittori discreti risulta,

R output

```
lm(formula = kid.score ~ as.factor(mom.work), data = kid.iq)
                     coef.est coef.se

(Intercept)              82.0     2.3
as.factor(mom.work)2      3.8     3.1
as.factor(mom.work)3     11.5     3.6
as.factor(mom.work)4      5.2     2.7
  n = 434, k = 4
  residual sd = 20.2, R-Squared = 0.02
```

Questa parametrizzazione del modello permette di considerare diversi punteggi medi per i bambini in corrispondenza a ciascuna categoria di occupazione materna. La categoria di riferimento (o "baseline") è `mom.work` = 1 che corrisponde a bambini le cui madri non rientrano a lavorare nei primi tre anni di vita del bambino; il punteggio medio stimato per questi bambini è pari al valore dell'intercetta 82.0. Il punteggio medio per i bambini delle altre categorie si ottiene aggiungendo al valore medio della categoria di riferimento il corrispondente coefficiente. Questa parametrizzazione permette di vedere che i bambini le cui madri lavorano part-time nel primo anno di vita del neonato riportano il punteggio più elevato 82.0+11.5. Queste famiglie tendono ad essere anche le famiglie che sono maggiormente avvantaggiate in termini di molte altre variabili socio-demografiche, così che l'interpretazione causale non è garantita.

Indici e variabili indicatrici

Gli *indici* dividono la popolazione in diverse categorie. Per esempio:

- `male` = 1 per gli uomini e 0 per le donne
- `age` = 1 per età da 18–29, 2 per età da 30–44, 3 per età da 45–64, 4 per età 65+
- `state` = 1 per Alabama, ..., 50 per Wyoming
- `contea` indici per le 3082 contee degli Stati Uniti.

Le *variabili indicatrici* sono predittori del tipo 0/1 che si basano sugli indici. Per esempio:

- `sex.1` = 1 per le donne e 0 altrimenti
  `sex.2` = 1 per gli uomini e 0 altrimenti
- `age.1` = 1 per età 18–29 e 0 altrimenti
  `age.2` = 1 per età 30–44 e 0 altrimenti
  `age.3` = 1 per età 45–64 e 0 altrimenti
  `age.4` = 1 per età 65+ e 0 altrimenti
- 50 indicatori per gli `stati`
- 3082 indicatori per le `contee`.

Come dimostrato nella sezione precedente, l'inclusione di queste variabili come predittori nella regressione, permette di avere medie differenti corrispondenti a ciascuna categoria definita dalla variabile.

Quando usare gli indici o le variabili indicatrici. Quando un input ha solo due livelli, si preferisce ricodificarlo con una singola variabile appropriatamente rinominata; per esempio, come discusso prima con l'esempio dei guadagni, l'indice sesso maschile (`male`) è più descrittivo di sesso.1 (`sex.1`) e sesso.2 (`sex.2`).

R permette di includere queste variabili come *factors* i cui singoli valori sono chiamati *levels*; per esempio, il sesso potrebbe avere due livelli maschi (`male`) e femmine (`female`). In questo libro, comunque, restringiamo l'attenzione a variabili definite numericamente, il che risulta particolarmente conveniente per le notazioni matematiche e per la configurazione dei modelli in Bugs.

Quando un input ha livelli multipli, preferiamo creare delle variabili indice (per esempio, l'età (`age`) che può assumere i livelli 1, 2, 3, 4), che potrebbero dar luogo a variabili indicatrici se necessario. Come verrà discusso nel Capitolo 11, i modelli *multilevel* offrono un approccio più generale per questi predittori categorici.

Identificabilità

Un modello risulta *non identificabile* se contiene dei parametri che non possono essere stimati in modo univoco—ovvero, detto in altre parole, che presentano degli errori standard che tendono all'infinito. Questo tipo di parametri sono parametri *non identificati*. L'esempio più familiare e importante è legato alla perfetta collinearità nei predittori della regressione. Un insieme di predittori è collineare se esiste una loro combinazione lineare che è uguale a 0 per tutte le osservazioni.

Se un indice assume J valori, allora ci sono J variabili indicatrici ad esso associate. In un modello di regressione classico si possono includere solo $J-1$ valori relativamente a ciascun insieme di indicatori— se infatti includessimo tutti i J, essi risultano collineari con il termine costante. (Si potrebbe ovviare al problema includendo l'insieme completo dei J valori ed escludendo il termine costante, ma lo stesso problema si ripresenterebbe quando si volesse includere un nuovo insieme di indicatori. Per esempio, si potrebbero non includere entrambe le categorie della variabile sesso e tutte e quattro le categorie della variabile età. Più semplicemente si include il termine costante e tutte le categorie degli indicatori ad eccezione di una).

Per ogni variabile categorica, la categoria che viene esclusa dal modello è nota come valore di riferimento (default), o condizione di *baseline* in quanto è la categoria implicata quando tutti gli altri $J-1$ indicatori sono posti uguali a zero. Per default

in R è il primo livello di un fattore che viene posto come condizione di riferimento; altre opzioni includono l'ultimo livello come punto di riferimento, la selezione del punto di riferimento, il vincolo che i coefficienti abbiano somma zero. Esistono delle discussioni nella letteratura sulla regressione su quale sia il modo migliore di definire la condizione di riferimento, ma non preoccupiamoci troppo di questo problema in questa sede in quanto nei modelli multilevel è possibile includere tutte le J variabili indicatrici allo stesso tempo.

In pratica, scopriririete che una regressione non è identificata perché il vostro programma vi restituirà un errore o un valore "NA" per la stima di un coefficiente (a volte il programma semplicemente lo escluderà e non segnalerà altro che è stato rimosso).

4.6 Costruzione di modelli di regressione a fini previsivi

Un modello deve essere costruito prima che venga stimato e valutato, e noi mettiamo il paragrafo relativo alla "costruzione del modello" verso la fine del capitolo. Perché? La situazione migliore sarebbe quella di un modello teorico definito prima che qualsiasi analisi dei dati venga effettuata. In pratica, l'analisi dei dati in genere viene effettuata a partire da un modello il più semplice possibile che poi via via viene complicato nel corso dell'analisi considerando tutte le problematiche che si presentano durante lo studio dei dati.

Tipicamente ci sono molti modi ragionevoli di costruire un modello. I modelli si differenziano in base agli obiettivi inferenziali o in base a come i dati sono stati raccolti. Le scelte chiave riguardano il modo in cui le variabili di input vengono associate per creare i predittori, nonché quali predittori includere nel modello. Nella regressione classica, questi sono punti estremamente rilevanti, in quanto se uno include troppi predittori nel modello le stime dei parametri diventano talmente variabili da essere totalmente inaffidabili. Alcuni di questi punti sono meno importanti nella regressione multilevel ma certamente non spariscono del tutto. Questa sezione pone l'attenzione sul problema relativo alla costruzione di modelli finalizzati alla previsione. La costruzione di modelli che può dar luogo a inferenze causali è un tema separato che verrà affrontato nei Capitoli 9 e 10.

Principi generali

I nostri principi generali per costruire i modelli di regressione a fini previsivi possono essere riassunti come segue:

1. Includere tutte la variabili di input che, per ragioni sostantive, potrebbero essere considerate importanti nel prevedere l'outcome.

2. Non sempre è necessario includere questi input come se fossero predittori separati—per esempio, si può considerare la media dei diversi input o la somma al fine di considerare un "punteggio medio" che può essere utilizzato come singolo predittore nel modello.

3. Per input che hanno effetti consistenti si dovrebbero includere anche le loro interazioni.

4. Suggeriamo la seguente strategia, per decidere se una variabile debba essere esclusa o meno dal modello in base al segno atteso e alla significatività statistica (tipicamente misurata ad un livello del 5%; ovvero un coefficiente risulta essere "statisticamente significativo" se la sua stima è lontano da zero di più due volte il corrispondente errore standard):

   (a) Se un predittore non è statisticamente significativo e il suo segno corrisponde a quello atteso, allora in genere si consiglia di tenerlo nel modello. Potrebbe non aiutare in modo drammatico le previsioni ma, allo stesso tempo, non le danneggia.

   (b) Quando un predittore non è statisticamente significativo e il suo segno non corrisponde a quello atteso (per esempio, l'"incumbency factor"—ovvero il vantaggio di popolarità che ha un candidato, in politica, quando si ripresenta alle successive elezioni politiche—presenta un effetto negativo sul voto), allora si consiglia di rimuovere il predittore dal modello (ovvero fissare il coefficiente del predittore pari a zero).

   (c) Se invece un predittore è statisticamente significativo e non riporta il segno atteso, allora risulta difficile credere che abbia un senso. (Per esempio questo caso potrebbe presentarsi nel caso in cui si considera l'India in cui i politici delle precedenti legislature sono fortemente impopolari, si veda Linden, 2006.). Si provi allora a cercare variabili "latenti" e ad includerli nel modello.

   (d) Se un predittore è statisticamente significativo e ha un segno pari a quello atteso, allora ha ovviamente senso includerlo nel modello.

Queste strategie non risolvono completamente i problemi che si possono incontrare nella costruzione di modelli lineari ma almeno possono aiutarci nell'evitare di commettere errori quali per esempio eliminare importanti informazioni dal modello. In ogni caso, queste linee guida devono essere alla base di un'analisi profonda circa le possibili relazioni tra le variabili *prima* che il modello venga stimato. È sempre

più facile giustificare il segno di un coefficiente a posteriori piuttosto che ragionare in modo approfondito circa il suo segno atteso prima della stima del modello. D'altra parte, ogni spiegazione che viene data una volta che il modello è stato stimato può essere ovviamente ancora sempre valida. L'ideale sarebbe quello di adattare le nostre teorie alla luce delle nuove informazioni che ci arrivano dopo la stima del modello.

Esempio: previsione della raccolta di cespugli di mesquite (alberi leguminosi del Nord America)

Illustreremo alcune idee relative alla validazione del modello con un esempio reale e che non è in qualche modo costruito artificialmente in quanto presentato in modo isolato dal suo contesto applicato. Questo esempio non può essere considerato come un "successo" e i risultati ottenuti sono, in qualche modo, inconcludenti, e quindi rappresenta un tipo di analisi di fronte alla quale uno studente potrebbe trovarsi quando esplora un nuovo insieme di dati.

I dati sono stati raccolti in modo da sviluppare un metodo per stimare la produzione totale (biomassa) di foglie di una particolare categoria di alberi leguminosi del Nord America (mesquite) in funzione di parametri della pianta facilmente misurabili, prima che avvenga l'effettiva raccolta. Sono stati considerati due insiemi di misure separati, un gruppo di 26 cespugli di mesquite e un gruppo di altri 20 cespugli misurati in un periodo di tempo diverso. Tutti questi dati sono riferiti ad una medesima area geografica (il ranch), ma nessuno dei due gruppi è stato costruito secondo uno schema campionario strettamente casuale.

La variabile risposta è il peso totale (in grammi) di materiale fotosintetico che deriva dal raccolto dei cespugli di mesquite. Le variabili di input sono:

`diam1`: diametro della copertura (tetto) creata dalle foglie (l'area fogliacea del cespuglio) in metri, misurata lungo l'asse più lungo del cespuglio;
`diam2`: diametro della copertura (tetto) creata dalle foglie misurata nell'asse più corto;
`canopy.height`: altezza copertura (tetto) creata dalle foglie;
`total.height`: altezza totale del cespuglio;
`density`: densità unitaria della pianta (numero di steli primari per unità di pianta);
`group`: gruppo di misurazione (0 per il primo gruppo, 1 per il secondo gruppo).

È ragionevole prevedere il peso della foglia utilizzando un qualche modello di regressione. Molte formulazioni sono possibili. L'approccio più semplice è quello di regredire il peso (`weight`) su tutti i predittori, ottenendo le seguenti stime:

R output

```
lm(formula = weight ~ diam1 + diam2 + canopy.height + total.height +
    density + group, data = mesquite)

              coef.est coef.se
(Intercept)       -729     147
diam1              190     113
diam2              371     124
canopy.height      356     210
total.height      -102     186
density            131      34
group             -363     100
  n = 46, k = 7
  residual sd = 269, R-Squared = 0.85
```

Per avere il senso dell'importanza di ciascun predittore, è utile conoscere il campo di variazione delle variabili che stiamo considerando:

R output

	min	q25	median	q75	max	IQR
diam1	0.8	1.4	2.0	2.5	5.2	1.1
diam2	0.4	1.0	1.5	1.9	4.0	0.9
canopy.height	0.5	0.9	1.1	1.3	2.5	0.4
total.height	0.6	1.2	1.5	1.7	3.0	0.5
density	1.0	1.0	1.0	2.0	9.0	1.0
group	0.0	0.0	0.0	1.0	1.0	1.0
weight	60	220	360	690	4050	470

"IQR" nell'ultima colonna è la differenza interquartilica (*interquartile range*)—ovvero la differenza tra il 75mo e il 25mo percentile di ciascuna variabile.

Ma per esempio, potrebbe essere più ragionevole stimare il modello su una scala logaritmica, in modo tale da avere effetti moltiplicativi piuttosto che additivi:

R output

```
lm(formula = log(weight) ~ log(diam1) + log(diam2) + log(canopy.height) +
    log(total.height) + log(density) + group, data = mesquite)

                  coef.est coef.se   IQR of predictor
```

```
(Intercept)            5.35    0.17    --
log(diam1)             0.39    0.28    0.6
log(diam2)             1.15    0.21    0.6
log(canopy.height)     0.37    0.28    0.4
log(total.height)      0.39    0.31    0.4
log(density)           0.11    0.12    0.3
group                 -0.58    0.13    1.0
  n = 46, k = 7
  residual sd = 0.33, R-Squared = 0.89
```

Invece di avere "ogni metro di differenza nel peso del tetto creato dalle foglie è associato a un incremento di 356 grammi di peso della foglia", si ha "una differenza di $x\%$ nel tetto creato dalle foglie è associata ad una differenza positiva (approssimata) di $0.37x\%$ nel peso della foglia" (valutato allo stesso livello di tutte le altre variabili considerate nel confronto).

Fino ad ora, abbiamo considerato nel modello tutti i predittori. Un approccio più "minimalista" è quello di iniziare a costruire un modello molto semplice che abbia senso. Ragionando in termini geometrici, potremmo pensare di stimare il peso delle foglie a partire dal volume del tetto creato dalle foglie, che in modo approssimativo potremmo costruire come,

$$\text{canopy.volume} = \text{diam1} \cdot \text{diam2} \cdot \text{canopy.height}.$$

Questo modello è eccessivamente semplificato: le foglie sono nella maggior parte dei casi sopra il cespuglio e non nel suo interno, e quindi questa misura della superficie dell'area potrebbe essere inappropriata. Ritorneremo su questo punto a breve.

Ha ancora senso lavorare su scala logaritmica:

R output

```
lm(formula = log(weight) ~ log(canopy.volume))

                  coef.est coef.se
(Intercept)           5.17    0.08
log(canopy.volume)    0.72    0.05
  n = 46, k = 2
  residual sd = 0.41, R-Squared = 0.80
```

Quindi, il peso delle foglie è approssimativamente proporzionale al volume del tetto creato dalle foglie (`canopy.volume`) di un coefficiente di proporzionalità pari a 0.72. Peraltro è sorprendente che non sia vicino a 1. La spiegazione classica per

questo risultato è che la variazione nel `canopy.volume` è indipendente dal peso delle foglie, e quindi tende ad *attenuare* il coefficiente di regressione—che quindi decresce in valore assoluto dal suo valore "naturale" di 1 a un valore più basso. In modo del tutto simile, regressioni relative a dati di "dopo" in funzione di dati "in avanti" hanno tipicamente un valore del coefficiente minore di 1. (Per un altro esempio, si veda la Sezione 7.2 che presenta un'analisi di previsione elettorale congressuale in cui il voto relativo alle precedenti elezione ha un coefficiente pari soltanto a 0.58).

La regressione con solo `canopy.volume` ci soddisfa in quanto relativamente molto semplice, e presenta un sorprendente R-quadro pari all'80%. Tuttavia, le previsioni risultano peggiori, e di molto, di quelle relative al modello con tutti i predittori. Potremmo allora tornare indietro al modello che considera tutti i predittori. Definiamo:

$$\begin{aligned} \text{canopy.area} &= \text{diam1} \cdot \text{diam2} \\ \text{canopy.shape} &= \text{diam1}/\text{diam2}. \end{aligned}$$

L'insieme di (canopy.volume, canopy.area, canopy.shape) è quindi semplicemente una diversa parametrizzazione delle tre dimensioni del tetto creato dalle foglie. Includendo questi nuovi predittori nel modello si ottiene:

R output

```
lm(formula = log(weight) ~ log(canopy.volume) + log(canopy.area) +
    log(canopy.shape) + log(total.height) + log(density) + group)

                  coef.est coef.se
(Intercept)           5.35    0.17
log(canopy.volume)    0.37    0.28
log(canopy.area)      0.40    0.29
log(canopy.shape)    -0.38    0.23
log(total.height)     0.39    0.31
log(density)          0.11    0.12
group                -0.58    0.13
  n = 46, k = 7
  residual sd = 0.33, R-Squared = 0.89
```

Il modello stimato si presenta in modo del tutto analogo a quello precedente su scala logaritmica (avendo effettuato soltanto una trasformazione lineare dei predittori), ma le stime dei coefficienti ci sembrano maggiormente interpretabili:

- Il volume del tetto creato dalle foglie e dell'area sono associate positivamente al peso. Nessuno dei due risulta statisticamente significativo ma decidiamo

comunque di tenerli nel modello in quanto crediamo che entrambi abbiano senso: (1) un tetto con un volume maggiore potrebbe avere più foglie e, (2) condizionatamente al volume, un tetto con un'area più grande potrebbe avere una maggiore esposizione al sole.

- Il coefficiente negativo di `canopy.shape` implica che i cespugli che sono più circolari nella sezione trasversale hanno foglie più pesanti (dopo aver controllato per volume e area). Non è molto chiaro se "credere" a questo oppure no. Il coefficiente non è statisticamente significativo; quindi si potrebbe tenere o meno questo coefficiente nel modello.

- L'altezza totale risulta associata positivamente con il peso, il che sembra avere senso se i cespugli vengono piantati gli uni vicino agli altri—i cespugli più alti prendono più sole. Il coefficiente non è statisticamente significativo, ma sembrerebbe avere senso e si potrebbe "credere" a questa spiegazione e quindi tenere il predittore nel modello.

- Non risulta essere molto chiaro, invece, come interpretare il coefficiente relativo a `density`. Dal momento che non è statisticamente significativo, si potrebbe pensare di escluderlo dal modello.

- Per una quale ragione, il coefficiente di `group` risulta essere molto grande e statisticamente significativo, quindi teniamo il predittore nel modello. Potrebbe essere una buona idea capire in che modo i due gruppi differiscono tra loro, in modo da includere nel modello una misura più rilevante per la quale `group` risulta una proxy.

Questo ci porta ad avere un modello del tipo,

R output

```
lm(formula = log(weight) ~ log(canopy.volume) + log(canopy.area)
             + group)

                   coef.est coef.se
(Intercept)            5.22    0.09
log(canopy.volume)     0.61    0.19
log(canopy.area)       0.29    0.24
group                 -0.53    0.12
  n = 46, k = 4
  residual sd = 0.34, R-Squared = 0.87
```

oppure

R output

```
lm(formula = log(weight) ~ log(canopy.volume) + log(canopy.area)
          + log(canopy.shape) + log(total.height) + group)

                   coef.est coef.se
(Intercept)            5.31    0.16
log(canopy.volume)     0.38    0.28
log(canopy.area)       0.41    0.29
log(canopy.shape)     -0.32    0.22
log(total.height)      0.42    0.31
group                 -0.54    0.12
  n = 46, k = 6
  residual sd = 0.33, R-Squared = 0.88
```

Noi vorremmo includere nel modello sia il volume che l'area, dal momento che ci aspettiamo, per ragioni essenzialmente geometriche, che entrambi possano stimare (con un segno positivo) il volume delle foglie. Sembrerebbe inoltre del tutto ragionevole andare ad analizzare il grafico dei residui per vedere se esistono dei comportamenti particolari nei residui che potrebbero indicare che qualcosa non è stata spiegata bene nel modello stimato.

Infine, potrebbe essere una buona idea andare a considerare le interazioni tra `group` e gli altri predittori. Purtroppo, dal momento che abbiamo solo 46 dati, risulta impossibile stimare in modo accurato queste interazioni: nessuna infatti risulta statisticamente significativa.

Per concludere con questo esempio: si è avuto un minimo di successo nella stima di questo modello in seguito alla trasformazione della variabile risposta e delle variabili di input così da ottenere un modello di previsione ragionevole. In ogni caso, non abbiamo trovato un modo chiaro per scegliere un modello piuttosto che un altro (o una combinazione di diversi modelli). Non esiste neanche un modo facile e immediato per scegliere un modello lineare piuttosto che un modello basato su una trasformazione logaritmica, o cercare di creare un ponte tra loro. Per questo particolare tipo di problema, il modello in termini logaritmici sembra avere più senso, anche se vorremmo aver trovato una ragione che fa riferimento ai dati per preferire questa specificazione.

4.7 Stima di una serie di regressioni

È abbastanza comune stimare un modello di regressione ripetutamente, sia considerando diversi insiemi di dati, sia sottoinsiemi dello stesso insieme di dati. Per esempio, uno potrebbe stimare la relazione tra altezza e guadagni usando delle indagini relative ad anni diversi, a diverse nazioni o all'interno di diverse regioni o stati negli Stati Uniti.

Come discusso nella seconda parte di questo libro, i modelli multilevel rappresentano un modo di stimare un modello di regressione ripetutamente, raggruppando le informazioni parziali relative ai diversi modelli stimati. In questo capitolo consideriamo una procedura più informale per stimare le regressioni separatamente—senza nessun raggruppamento tra anni o gruppi—e quindi proviamo a rappresentare tutte queste stime insieme, che può essere considerato come un precursore dei modelli multilevel.[4]

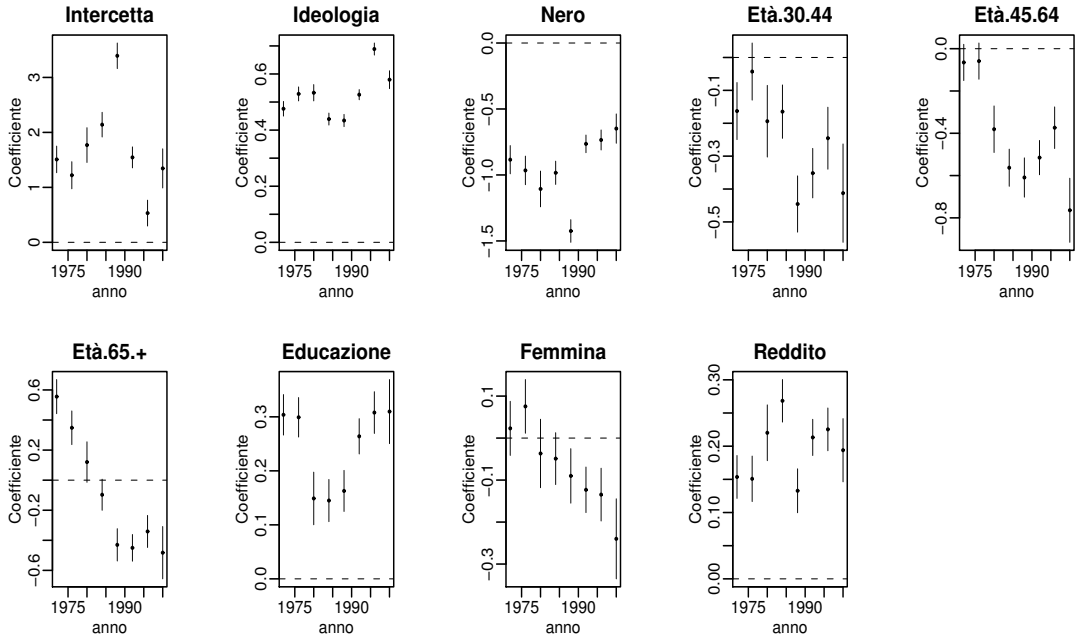
Previsione dell'identificazione dei partiti

Gli scienziati politici sono stati per tanto tempo interessati all'identificazione dei partiti e al loro cambiamento nel tempo. Analizziamo ora una serie di regressioni sezionali che modellano l'identificazione dei partiti in funzione dell'ideologia politica e di alcune variabili demografiche.

Abbiamo considerato l'indagine sugli studi nazionali elettorali (National Election Study) che definisce l'identificazione dei partiti attraverso una variabile definita su una scala da 1 a 7:
1 = fortemente Democratico(strong Democrat),
2 = Democratico (Democrat),
3 = debolmente Democratico (weak Democrat),
4 = indipendente (independent), . . . ,
7 = fortemente Repubblicano (strong Republican),
che trattiamo come una variabile continua.
Includiamo inoltre nel modello i seguenti predittori:
ideologia politica
1 = fortmente libelare (strong liberal),

[4]Il metodo di modellizzazione ripetuto, affiancato da un grafico di stime di serie storiche, a volte chiamato "secret weapon", in quanto è molto facile e potente ma raramente utilizzato come strumento per l'analisi dei dati. Sospettiamo che la ragione per la quale non viene utilizzato è che, una volta che si conosce la struttura temporale di un insieme di dati, è naturale che come passo successivo si modellino direttamente i dati. In pratica, comunque, esiste tutta una serie di problemi per cui le analisi sezionali sono informative, e per cui una rappresentazione di esse in serie storica risulta appropriata al fine di fornire un'idea dell'andamento temporale dei dati.

Figura 4.6: *Stime dei coefficienti (con relativi intervalli di confidenza al 50%) dei modelli di regressione che stimano l'identificazione dei partiti come funzione dell'ideologia politica, dell'etnicità e di altri predittori, modelli stimati separatamente in base ai dati di sondaggi elettorali relativi a campagne elettorali che vanno dal 1976 al 2000. I grafici sono su scale differenti, con le variabili di input ordinate approssimativamente in base alla grandezza dei loro coefficienti. L'insieme dei grafici mostra la rappresentazione grafica dell'inferenza effettuata su una serie di regressioni.*



2=liberale (liberal), ...,
7=fortemente conservativo (strong conservative)),
etnicità
(0=bianco (white),
1=nero (black),
0.5=altro (other)),
età (nelle seguenti categorie: 18–29, 30–44, 45–64, e più di 65+ anni, con la categoria di età più bassa come base di riferimento),
istruzione
(1=licenza media inferiore (no high school),
2=licenza media superiore(high school graduate),
3=titolo di studio di specializzazione non laurea (some college),
4=laurea (college graduate)),
sesso
(0=maschio (male),
1=femmina (female)),
e reddito
(1=0–16mo percentile,
2=17–33mo percentile,
3=34–67mo percentile,
4=68–95mo percentile,
5=96–100mo percentile.

La figura 4.6 mostra le stime dei coefficienti negli anni considerati. L'ideologia e l'etnicità sono i predittori più rilevanti,[5] e rimangano abbastanza stabili nel tempo. Le differenze stimate relative al sesso e all'età cambiano in modo drammatico intorno al 30-mo anno.

4.8 Nota bibliografica

Per letture aggiuntive sulle trasformazioni si veda Atkinson (1985), Mosteller e Tukey (1977), Box e Cox (1964), e Carroll e Ruppert (1981). Bring (1994) presenta un'ampia discussione sulla standardizzazione dei coefficienti di regressione; si veda anche Blalock (1961) e Greenland, Schlessman, e Criqui (1986). Harrell (2001) discute invece sulle strategia per la costruzione dei modelli di regressione.

Per saperne di più sull'esempio dei guadagni e dell'altezza, si veda Persico, Postlewaite, e Silverman (2004) e Gelman e Nolan (2002). Per saperne di più sul-

[5]L'ideologia fa riferimento ad una scala da 1 a 7, di conseguenza i suoi coefficienti devono essere moltiplicati per 4 per avere la differenza attesa quando si confronta un liberale (ideology=2) con un conservativo (ideology=6).

l'esempio relativo all'utilizzo della mano destra o sinistra, si veda Gelman e Nolan (2002, sezioni 2.5 e 3.3.2). L'analisi storica della regressione sulla media è trattata in Stigler (1986), e la sua connessione con i modelli multilevel sono discussi in Stigler (1983).

L'esempio relativo ai cespugli di mesquite della Sezione 4.6 viene da un problema di esame degli anni '80; non abbiamo memoria della provenienza dei dati. Per Per saperne di più sull'esempio relativo all'ideologia nella Sezione 4.7 si veda Bafumi (2005).

4.9 Esercizi

1. Trasformazione logaritmica e regressione: si consideri la seguente regressione:

$$\log(\text{weight}) = -3.5 + 2.0 \log(\text{height}) + \text{error},$$

con errori aventi deviazione standard pari a 0.25. Il peso è espresso in libbre e l'altezza in pollici.

   (a) Riempire gli spazi in bianco: approssimativamente il 68% delle persone avrà peso all'interno di un intervallo di __ e __ dei loro valori stimati in base al modello di regressione.

   (b) Disegnare la linea di regressione e lo scatterplot del log(weight) in funzione di log(height) in modo tale che abbia un senso e sia coerente con il modello stimato. Siate sicuri di identificare gli assi in modo appropriato.

2. La cartella `earnings` contiene i dati dell'indagine sul lavoro, la famiglia e il benessere (Work, Family, and Well-Being Survey) (Ross, 1990). Mettete insieme i dati relativi al guadagno, al sesso, all'altezza e al peso.

   (a) In R, controllate il dataset e sistemate i dati eliminando le codifiche inusuali.

   (b) Stimate il modello di regressione al fine di prevedere il guadagno in funzione dell'altezza. Quale trasformazione bisogna usare per interpretare l'intercetta del modello come guadagno medio delle persone di altezza media?

   (c) Stimate alcuni modelli di regressione al fine di prevedere i guadagni in base ad alcune combinazioni di sesso, altezza e peso. Siate sicuri di considerare trasformazioni e interazioni che abbiano senso. Scegliete il vostro modello preferito e giustificate questa scelta.

   (d) Interpretate tutti i coefficienti del modello scelto.

3. Rappresentiamo graficamente i modelli di regressione linearie e nonlineare: abbiamo utilizzato i dati di peso in libbre (`weight`) ed età (`age`) (in anni) da un campione casuale di adulti americani. Abbiamo creato due nuove variabili `age10` = age/10 e `age10.sq` = $(\text{age}/10)^2$, e i seguenti indicatori `age18.29`, `age30.44`, `age45.64`, e `age65up` per le 4 categorie dell'età. Abbiamo quindi stimato i seguenti modelli di regressione ottenendo i seguenti risultati:

R output

```
 lm(formula = weight ~ age10)
            coef.est coef.se
(Intercept)    161.0     7.3
age10            2.6     1.6
  n = 2009, k = 2
  residual sd = 119.7, R-Squared = 0.00

lm(formula = weight ~ age10 + age10.sq)

            coef.est coef.se
(Intercept)     96.2    19.3
age10           33.6     8.7
age10.sq        -3.2     0.9
  n = 2009, k = 3
  residual sd = 119.3, R-Squared = 0.01

lm(formula = weight ~ age30.44 + age45.64 + age65up)

             coef.est coef.se
(Intercept)     157.2     5.4
age30.44TRUE     19.1     7.0
age45.64TRUE     27.2     7.6
age65upTRUE       8.5     8.7
  n = 2009, k = 4
  residual sd = 119.4, R-Squared = 0.01
```

   (a) Su un grafico del peso in funzione dell'età (ovvero peso sull'asse delle y e età sull'asse delle x) disegnate la linea di regressione del primo modello stimato.

   (b) Sullo stesso grafico rappresentate la linea relativa al secondo modello.

   (c) Su un altro grafico avente gli stessi assi e la stessa scala rappresentate la linea del terzo modello di regressione stimato (che sarà discontinua).

4. Trasformazioni logaritmiche: la cartella `pollution` contiene tassi di mortalità e diversi fattori ambientali relativi a 60 aree metropolitane degli Stati Uniti (si veda McDonald e Schwing, 1973). Per questo esercizio si vuole costruire un modello per prevedere il tasso di mortalità in funzione dell'ossido nitrico, diossido di zolfo e idrocarburi come input. Questo modello è una estrema semplificazione in quanto combina tutte queste possibili cause di mortalità senza prendere in considerazione fattori cruciali quali l'età e il fumo. Utilizziamo la trasformazione logaritmica nella regressione.

   (a) Costruite uno scatterplot del tasso di mortalità in funzione del livello di ossido nitrico. È possibile sostenere che questo modello stima bene i dati? Stimate il modello di regressione e valutate il grafico dei residui della regressione.

   (b) Trovate una trasformazione dei dati che sia più appropriata per il modello di regressione. Stimate il modello su questi dati trasformati e analizzate il nuovo grafico sui residui.

   (c) Interpretate la pendenza della retta in base al modello scelto in (b).

   (d) Stimate ora un modello che stimi il tasso di mortalità in funzione del livello dell'ossido nitrico, del diossido di zolfo e dell'idrocarburo come input. Usate un'appropriata trasformazione quando risulta utile. Rappresentate graficamente il modello di regressione stimato e interpretate i coefficienti.

   (e) Validazione incrociata (cross-validation): stimate il modello scelto sulla prima metà dei dati e quindi stimate i valori relativamente alla seconda metà. (Abbiamo usato tutti i dati per costruire il modello in (d), quindi non è una vera e propria validazione incrociata ma fornisce il senso di come i passi della *cross-validation* devono essere implementati.)

5. Trasformazioni relative a particolari obiettivi: relativamente ad uno studio sulle elezioni congressuali, si vorrebbe avere una misura delle somme di denaro raccolte da ciascuno dei due candidati dei due partiti maggiori in ogni distretto. Supponiamo di conoscere l'ammontare di denaro raccolto da ciascun candidato e chiamiamo questi valori, espressi in dollari, D_i e R_i. Vorremmo combinare questi valori in modo da ottenere una singola variabile che può essere inclusa come variabile di input nel modello per prevedere la quota di voto dei Democratici.

   (a) Discutere i vantaggi e gli svantaggi delle seguenti misure:

   - La differenza semplice, $D_i - R_i$
   - Il rapporto, D_i/R_i
   - La differenza su scala logaritmica, $\log D_i - \log R_i$

- La proporzione relativa, $D_i/(D_i + R_i)$.

(b) Proponete una trasformazione idiosincratica (come nell'esempio a pagina 85) e discutete dei vantaggi e degli svantaggi relativi al suo possibile utilizzo nel modello di regressione.

6. Un economista stima un modello di regressione esaminando la relazione tra il prezzo medio delle sigarette, P, e la quantità consumata, Q, su un campione di grandi dimensioni di contee negli Stati Uniti, assumendo la seguente forma funzionale, $\log Q = \alpha + \beta \log P$. Supponiamo che la stima di β sia 0.3. Interpretare questo coefficiente.

7. Una sequenza di regressione: trovare un problema di regressione che è di vostro interesse e che deve essere stimato ripetutamente (per esempio dati di anni diversi, o di diverse nazioni). Effettuate un'analisi separatamente per ogni anno, per ogni nazione e rappresentate graficamente le stime come nel grafico della Figura 4.6 a pagina 99.

8. Ritorniamo ai dati sulla valutazione dell'insegnamento dell'Esercizio 4. Stimate un modello di regressione per la valutazione dell'insegnamento dati diversi input presenti nel dataset. Considerate l'interazione, la combinazione di predittori, e trasformazioni se appropriate. Considerate diversi modelli, discutete in dettaglio il modello definitivo che avete scelto, e spiegate il motivo per cui avete scelto questo modello piuttosto che altri.