Capitolo 1

Introduzione

Obiettivo del presente volume è introdurre il **modello di regressione lineare**, con cenni alle principali estensioni. I modelli di regressione servono per capire se e come il comportamento di una variabile di interesse può essere spiegato usando altre variabili. Ad esempio, in che modo la portata di un fiume è influenzata dalle precipitazioni nel suo bacino idrografico? Oppure, il prezzo e la spesa in attività promozionali sono variabili che determinano le vendite di un prodotto? Naturalmente, le relazioni di questo tipo non sono mai esatte: note le precipitazioni, la portata di un fiume che ne consegue verrà sempre calcolata con un certo margine di *errore*, dovuto al fatto che vi sono variabili importanti non incluse nel modello (ad esempio la forma del bacino idrografico, o la natura del terreno), oppure alla presenza di fattori non rilevabili o imprevedibili. Questi ultimi vengono riassunti in un termine il cui comportamento viene completamente attribuito al *caso*. Il modello di regressione lineare si compone quindi di una parte strutturale, che spiega la relazione tra la variabile di interesse e le variabili che ne determinano il comportamento, e di un termine casuale, la cui presenza implica che si tratta di un *modello statistico*[1].

Il più semplice modello statistico è quello in cui si compiono osservazioni indipendenti, su n unità statistiche, di una variabile casuale di interesse Y. Si consideri dunque il campione costituito dalle variabili casuali[2] $Y_1, \ldots, Y_n$, indipendenti e identicamente distribuite secondo una comune distribuzione di probabilità, individuata a meno di un ignoto parametro, eventualmente multidimensionale, ossia

$$Y_i \sim f(\cdot;\theta), \;\; \theta \in \Theta \subset \mathbb{R}^d, \;\; d \geq 1, \qquad i = 1, \ldots, n. \tag{1.1}$$

In altre parole, si suppone che la distribuzione della variabile casuale Y sia descritta da una funzione di densità o di probabilità $f(\cdot;\theta)$, con il parametro d-dimensionale θ che assume valori nello spazio parametrico Θ.

Esempio 1.1. Modello normale o gaussiano. Si assume che il campione $Y_1, \ldots, Y_n$ sia costituito da variabili casuali indipendenti e identicamente distribuite secondo una distribuzione $\mathcal{N}(\mu, \sigma^2)$, ossia una distribuzione normale di media μ e varianza σ^2, con $\theta = (\mu, \sigma^2) \in \Theta =$

[1] Si noti che, nel momento in cui si afferma che l'errore è casuale, si sta dicendo che esso è privo di struttura e dunque che tutte le variabili rilevanti (ad esempio le precipitazioni, la forma del bacino idrografico e la natura del terreno) sono state incluse nella parte strutturale del modello.

[2] Per una introduzione ai concetti di variabile casuale, funzione di probabilità, funzione di densità, e agli altri concetti di base della statistica, che qui vengono dati per noti, si vedano, ad esempio, Pace e Salvan (2001), Mood *et al.* (2003) oppure Piccolo (2010).

$\mathbb{R}\times]0,+\infty[$. La densità di Y_i è

$$f(y;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\ \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}. \tag{1.2}$$

I più noti stimatori dei parametri μ e σ^2, ovvero delle componenti del parametro bidimensionale θ, sono

$$\hat{\mu} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \text{e} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Tali stimatori sono corretti, ossia

$$E(\bar{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

e

$$\begin{aligned}
E(S^2) &= E\left(\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right) = E\left(\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\mu+\mu-\bar{Y})^2\right)\\
&= \frac{1}{n-1}\left(E\left(\sum_{i=1}^{n}(Y_i-\mu)^2\right) + 2\,E\left((\mu-\bar{Y})\sum_{i=1}^{n}(Y_i-\mu)\right) + E\left(\sum_{i=1}^{n}(\mu-\bar{Y})^2\right)\right)\\
&= \frac{1}{n-1}\left(n\,\sigma^2 - 2\,E\big((\bar{Y}-\mu)\ n\ (\bar{Y}-\mu)\big) + n\,V(\bar{Y})\right)\\
&= \frac{1}{n-1}\left(n\,\sigma^2 - 2\,n\,V(\bar{Y}) + n\,V(\bar{Y})\right) = \frac{1}{n-1}\left(n\,\sigma^2 - n\,V(\bar{Y})\right)\\
&= \frac{1}{n-1}\left(n\,\sigma^2 - n\,\frac{\sigma^2}{n}\right) = \frac{1}{n-1}\ (n-1)\ \sigma^2 = \sigma^2.
\end{aligned}$$

Gli stimatori $\hat{\mu}$ e S^2 sono anche consistenti. Infatti, oltre a essere corretti hanno varianze che tendono a zero per $n \to \infty$, ossia

$$V(\bar{Y}) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n} \underset{n\to\infty}{\longrightarrow} 0$$

e si dimostra (ad esempio, Mood *et al.*, 2003) che

$$V(S^2) = \frac{2\sigma^2}{n-1}.$$

L'inferenza sul parametro θ si basa sul fatto che $\bar{Y}$ e S^2 sono indipendenti (esercizio 2.1), che $\bar{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ e che $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$, dove χ^2_ν indica la distribuzione chi-quadrato con ν gradi di libertà (si veda Piccolo, 2010, capitolo 14). Si ha dunque la **quantità pivotale**[3]

$$t = \frac{\bar{Y}-\mu}{S/\sqrt{n}} \sim t_{n-1},$$

[3]Si dice *quantità pivotale* una funzione del campione e del parametro di interesse, che non dipende da altri parametri ignoti e che ha distribuzione nota.

dove t_ν indica la distribuzione t di Student con ν gradi di libertà. Dalla quantità t si può ricavare l'intervallo di confidenza di livello $1-\alpha$ per μ, i cui estremi sono

$$\bar{y} \pm t_{n-1;1-\alpha/2} \, \frac{s}{\sqrt{n}},$$

dove $t_{\nu;\alpha}$ rappresenta il quantile di livello α della distribuzione t_ν, mentre $\bar{y}$ e s sono i valori osservati[4] di $\bar{Y}$ e S. Inoltre

$$\left\{ \left| \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \right| > t_{n-1;1-\alpha/2} \right\}$$

è la regione di rifiuto per il sistema di ipotesi

$$\begin{cases} H_0 \colon \mu = \mu_0 \\ H_1 \colon \mu \neq \mu_0. \end{cases}$$

•

Spesso sulle unità statistiche vengono osservate diverse variabili. Ciascuna di esse è suscettibile di essere analizzata con un modello del tipo (1.1); è però anche possibile studiarne l'andamento congiunto.
Ad esempio, potrebbero essere rilevati, per n individui, peso, altezza, sesso ed età, e potrebbe essere di interesse valutare

1. se il peso è diverso, in media, tra maschi e femmine;
2. se la distribuzione del peso è la stessa tra maschi e femmine;
3. se peso e altezza sono correlati;
4. quanto aumenta l'altezza all'aumentare dell'età;
5. se la relazione tra altezza ed età è la stessa per maschi e femmine;
6. ecc.

Ci sono vari strumenti idonei a rispondere ad alcune delle domande sopra: ad esempio, l'indice di dipendenza in media, l'indice di dipendenza, il coefficiente di correlazione, la regressione lineare[5].
Per trattare questi problemi è anche possibile generalizzare il modello (1.1). Indichiamo con $(Y_{i1}, \ldots, Y_{iq})$ la variabile casuale q-dimensionale, $q > 1$, osservata sull'i-ma unità statistica, $i = 1, \ldots, n$. Assumiamo vi sia indipendenza tra le unità statistiche (ossia tra le n variabili casuali q-dimensionali) e che

$$(Y_{i1}, \ldots, Y_{iq}) \sim f(\cdot; \theta), \;\; \theta \in \Theta \subset \mathbb{R}^d, \qquad i = 1, \ldots, n, \tag{1.3}$$

dove $f(\cdot; \theta)$ rappresenta una funzione di probabilità o di densità q-dimensionale e definisce dunque la distribuzione congiunta delle q variabili di interesse.

Esempio 1.2. Normale bidimensionale. Immaginiamo ad esempio di osservare, su n individui, una variabile casuale bidimensionale (Y_{i1}, Y_{i2}), $i = 1, \ldots, n$, come il peso e l'altezza, e supponiamo che tale variabile abbia distribuzione normale bidimensionale con parametro $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. La densità congiunta di Y_{i1} e Y_{i2} è

$$f(y_1, y_2; \theta) = \frac{1}{2\,\pi\,\sigma_1\,\sigma_2\,\sqrt{1-\rho^2}} \; \exp\left\{ -\frac{z}{2\,(1-\rho^2)} \right\}, \tag{1.4}$$

[4] I valori osservati di una variabile casuale vengono anche chiamati *realizzazioni* o *determinazioni*.
[5] Per un'introduzione a tali concetti si veda, ad esempio, Cicchitelli (2012, capitoli 9-11).

dove

$$z = \left(\frac{y_1 - \mu_1}{\sigma_1}\right)^2 - 2\,\rho\,\frac{(y_1 - \mu_1)\,(y_2 - \mu_2)}{\sigma_1\,\sigma_2} + \left(\frac{y_2 - \mu_2}{\sigma_2}\right)^2$$

e ρ è il coefficiente di correlazione tra Y_{i1} e Y_{i2}. Se $\rho = 0$, le due variabili sono incorrelate e questo, sotto assunzione di normalità, equivale a dire che sono indipendenti.
Si dimostra che[6] il modello (1.4) implica che le distribuzioni marginali delle due variabili casuali sono anch'esse normali, con $Y_{ij} \sim \mathcal{N}(\mu_j, \sigma_j^2)$, $j = 1, 2$; ma il vantaggio principale di questo modello è che permette di rappresentare la dipendenza tra Y_{i1} e Y_{i2}. In altri termini, è possibile calcolare $P(Y_1 \in A \cap Y_2 \in B)$ per due insiemi arbitrari A e B. Ad esempio, se il modello riguarda peso e altezza in una popolazione, si potrà valutare la probabilità che un individuo pesi più di $70kg$ e, contemporaneamente, sia più alto di $180cm$.
La relazione tra le due variabili implicata dall'assunzione (1.4) può anche essere scritta nella forma di distribuzione condizionata, come

$$Y_{i1}|Y_{i2} = y_2 \;\sim \mathcal{N}\left(\mu_1 + \rho\,\frac{\sigma_1}{\sigma_2}\,(y_2 - \mu_2), (1 - \rho^2)\,\sigma_1^2\right) \tag{1.5}$$

e, simmetricamente, con la distribuzione condizionata di $Y_{i2}|Y_{i1} = y_1$. Questo permette di evidenziare un legame con i modelli di regressione lineare, che sono il tema principale di questo volume. Infatti, la (1.5) mostra che il valore atteso di Y_{i1} condizionato a $Y_{i2} = y_2$ è una funzione lineare di y_2, e un risultato simmetrico vale per il valore atteso di Y_{i2} condizionato a $Y_{i1} = y_1$. •

I modelli (1.3) e, come caso particolare (1.4), mettono tutte le variabili osservate sullo stesso piano, ossia implicano che tra esse esista una relazione simmetrica. Tuttavia, in molti casi le variabili hanno ruoli diversi. Consideriamo alcuni esempi:

1. Valutare la probabilità che dei potenziali debitori restituiscano il debito sulla base di caratteristiche individuali: reddito, situazione familiare, età, ecc.
2. Prevedere il numero di esami fatti da uno studente iscritto al primo anno sulla base di dati anagrafici, reddito, scuola di provenienza, ecc.
3. Prevedere il reddito di un individuo sulla base del sesso a parità di altre condizioni (titolo di studio, età, ecc.).
4. Valutare la pressione del sangue di un individuo con e senza la somministrazione di un farmaco, tenendo conto delle sue caratteristiche individuali.
5. Valutare come varia la mortalità nella popolazione a seconda della concentrazione di inquinanti atmosferici.
6. Prevedere il numero di sinistri di un assicurato sulla base delle sue caratteristiche individuali e della storia passata.

Tutti questi esempi ricadono in un medesimo schema: si ha una variabile di interesse, che chiameremo Y, che ha il ruolo di **variabile risposta** (probabilità di restituzione, numero di esami fatti, reddito, ecc.). Vi sono poi altre variabili concomitanti, che chiameremo $X_1, \ldots, X_p$, che hanno il ruolo di **variabili esplicative**. Si vuole determinare come la prima è influenzata dalle seconde.
In termini più formali, lo schema di ragionamento prevede che

$$Y \sim f(\cdot; x_1, \ldots, x_p, \theta), \tag{1.6}$$

cioè la distribuzione di probabilità (unidimensionale) della variabile Y dipende, oltre che dal parametro θ, anche dai valori $x_1, \ldots, x_p$ assunti dalle variabili esplicative. La variabile risposta Y

[6] Per le dimostrazioni si veda ad esempio Rencher e Schaalje (2008); si veda anche la sezione 2.9.

unità statistica	variabili osservate			
1	y_1	x_{11}	...	x_{1p}
2	y_2	x_{21}	...	x_{2p}
⋮	⋮	⋮	...	...
i	y_i	x_{i1}	...	x_{ip}
⋮	⋮	⋮	...	...
n	y_n	x_{n1}	...	x_{np}
	variabile d'interesse	primo regressore	...	p-mo regressore

Tabella 1.1: Campione tipo. Il modello di regressione prevede di ragionare condizionatamente ai valori assunti dai regressori. In altri termini, mentre $y_1, \ldots, y_n$ sono considerate determinazioni delle variabili casuali $Y_1, \ldots, Y_n$, i valori x_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$, dei regressori sono considerati fissati.

viene anche chiamata *variabile dipendente* o *spiegata*. Le variabili esplicative $X_1, \ldots, X_p$ vengono anche chiamate *predittori*, *variabili esogene*, *regressori*, *variabili indipendenti* o *covariate*.
Il modello ha una struttura asimmetrica, ossia una variabile ha un ruolo diverso dalle altre: l'intento è spiegare, almeno in parte, la variabilità della risposta sulla base delle esplicative.

Osservazione 1.1. Relazione e causalità. Nel costruire modelli di regressione è forte la tentazione di interpretare la relazione trovata alla stregua di un rapporto causa effetto, nel senso che le variabili esplicative producono un effetto sulla variabile d'interesse. Una tale interpretazione non è giustificata dal modello di per sé, nel senso che, per quanto il modello possa essere buono (nel senso di adeguato ai dati), questo fatto da solo non permette di trarre conclusioni sul rapporto causa effetto tra le grandezze coinvolte. Nel seguito si tratterà della costruzione, stima e validazione di modelli, tralasciando la questione dell'interpretazione causa effetto, che è un problema distinto. Nell'interpretare i risultati, si terrà dunque presente il loro significato: si valuta la relazione statistica esistente tra le variabili, non la relazione sostanziale o funzionale in termini di rapporti causa-effetto[7]. •

Gli scopi dell'analisi possono essere diversi. Principalmente possiamo distinguere lo scopo previsivo, quando cioè si vuole uno strumento per prevedere il valore della variabile d'interesse noti i valori delle variabili esplicative (perché ad esempio queste sono più facili da misurare o si osservano in anticipo rispetto alla risposta) e lo scopo interpretativo, quando l'interesse primario è stabilire quali tra le esplicative abbiano una più forte relazione con la risposta e in che direzione vada tale relazione. Emblematici del primo scopo sono gli esempi 1 e 6, dove l'obiettivo è scegliere, tra i potenziali clienti, a chi concedere un prestito; emblematico del secondo scopo è l'esempio 3, in cui l'obiettivo è determinare se vi è disparità di trattamento tra i sessi.
Per formalizzare il problema, presentato sin qui in termini generali, cominciamo con lo specificare la base informativa dicendo che, di n unità statistiche, si sono osservate diverse caratteristiche, tradotte in $p+1$ variabili (non necessariamente numeriche); si ha cioè un insieme di dati osservati rappresentabile come nella tabella 1.1.

[7] Ad esempio, un modello con Y = "Numero di morti per annegamento" e X = "Livello delle vendite di gelati" probabilmente mostrerebbe che la relazione statistica è piuttosto forte (essendo entrambe le variabili legate alla temperatura ambientale). Ciò non implica affatto che X causa Y. La correlazione non implica causalità.

Evidenziando il ruolo delle n unità statistiche, il modello (1.6) può essere scritto come

$$Y_i \sim f(y_i; x_{i1}, \ldots, x_{ip}, \theta), \qquad i = 1, \ldots, n, \qquad \text{indipendenti.} \tag{1.7}$$

Una prima semplificazione della (1.7) si ha supponendo che

$$h(Y_i) = g(x_{i1}, \ldots, x_{ip}; \theta) + \varepsilon_i, \qquad \varepsilon_i \sim f_\varepsilon(\cdot; \theta), \qquad i = 1, \ldots, n, \qquad \text{indipendenti,} \tag{1.8}$$

dove $h(\cdot)$ è una funzione nota, $g(\cdot)$ è una funzione da stimare (nota a meno del parametro θ) e ε è la componente casuale, anche detta errore. Dunque, il modello è formato da due componenti additive: una *componente deterministica* (ossia non casuale) $g(x_{i1}, \ldots, x_{ip}; \theta)$ e una *componente casuale* ε_i. La relazione tra la risposta e le esplicative viene spiegata dalla componente deterministica, che rappresenta la parte strutturale di tale relazione. La componente deterministica, tuttavia, non spiega completamente il comportamento della variabile dipendente: la parte non spiegata viene rappresentata da ε_i, che è chiamato errore in quanto $\varepsilon_i = h(Y_i) - g(x_{i1}, \ldots, x_{ip}; \theta)$, ossia è pari alla differenza tra ciò che il modello si prefigge di spiegare e la spiegazione che riesce effettivamente a dare[8].
Il modello lineare è un'ulteriore particolarizzazione dell'espressione (1.8) in cui

$$h(Y_i) = \beta_1 g_1(x_{i1}) + \ldots + \beta_p g_p(x_{ip}) + \varepsilon_i, \quad \varepsilon_i \sim f_\varepsilon(\cdot; \theta), \quad i = 1, \ldots, n, \quad \text{indipendenti,} \tag{1.9}$$

dove sia $h(\cdot)$ che $g_1(\cdot), \ldots, g_p(\cdot)$ sono funzioni note e $\beta_1, \ldots, \beta_p$ sono parametri da stimare. Ad esempio, sono modelli lineari

$$Y_i = \beta_1 + \beta_2 x_{i2}^2 + \beta_3 \sqrt{x_{i3}} + \varepsilon_i$$

e

$$\log(Y_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i.$$

Introduciamo nel seguito alcuni esempi di insiemi di dati, utili per dare un'idea dei metodi che verranno trattati nei capitoli successivi.

Esempio 1.3. Ciliegi neri. Per $n = 31$ alberi di ciliegio nero si sono osservati il diametro del tronco (misurato a una fissata altezza da terra), l'altezza e il volume di legno ottenuto dopo l'abbattimento dell'albero stesso (si vedano Ryan *et al.*, 1976, e Atkinson, 1985). I dati sono riportati nella tabella all'interno della figura 1.1.
Si hanno quindi tre variabili, tutte quantitative continue. In linea di principio ciascuna delle tre variabili potrebbe fungere da variabile risposta. Tuttavia, il modello di effettivo interesse è quello in cui il volume è spiegato in funzione di diametro e altezza. Infatti, queste ultime quantità sono facili da rilevare, mentre per calcolare il volume occorre abbattere l'albero; dunque, disporre di un modello per prevedere il volume ligneo prima dell'abbattimento sembra sensato. Nella figura 1.1 si danno alcune rappresentazioni grafiche che suggeriscono vi sia effettivamente un legame tra le variabili in gioco, legame che potrebbe essere riassunto da una legge del tipo (1.9) come

$$\text{volume} = \beta_1 + \beta_2\,(\text{diametro}) + \beta_3\,(\text{altezza}) + \text{errore},$$

oppure, prendendo spunto dalla geometria[9], come

$$\text{volume} = \beta_1\,(\text{diametro})^{\beta_2}\,(\text{altezza})^{\beta_3}\,(\text{errore}),$$

che può essere scritta nella forma (1.9) passando ai logaritmi. •

Diametro (in pollici)	Altezza (in piedi)	Volume (in piedi^3)
8.3	70	10.3
8.6	65	10.3
8.8	63	10.2
10.5	72	16.4
10.7	81	18.8
10.8	83	19.7
11.0	66	15.6
11.0	75	18.2
11.1	80	22.6
11.2	75	19.9
11.3	79	24.2
11.4	76	21.0
11.4	76	21.4
11.7	69	21.3
12.0	75	19.1
12.9	74	22.2
12.9	85	33.8
13.3	86	27.4
13.7	71	25.7
13.8	64	24.9
14.0	78	34.5
14.2	80	31.7
14.5	74	36.3
16.0	72	38.3
16.3	77	42.6
17.3	81	55.4
17.5	82	55.7
17.9	80	58.3
18.0	80	51.5
18.0	80	51.0
20.6	87	77.0

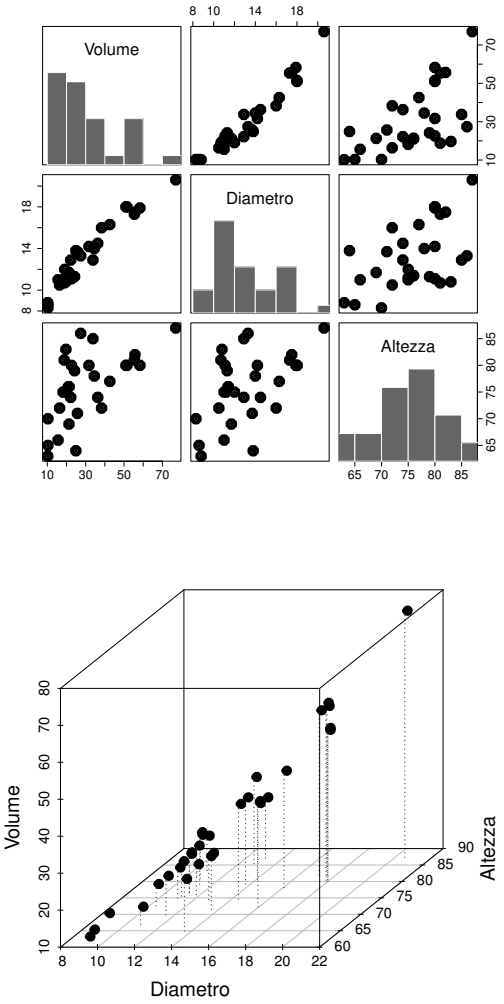
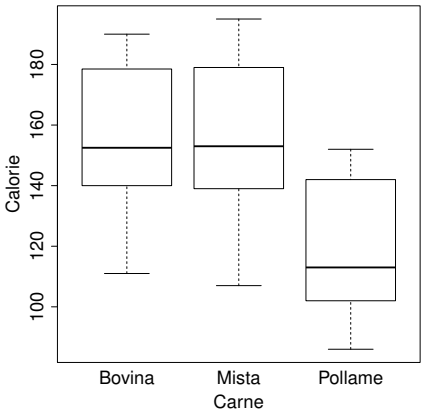


Figura 1.1: Dati sui ciliegi neri.

Bovina	Mista	Pollame
186	191	102
149	172	102
158	139	107
148	179	142
141	195	152
157	138	129
135	182	106
181	147	87
184	175	113
139	153	86
152	135	146
153	173	132
131	190	94
132	146	99
176	136	135
190	107	143
175	140	144
111		
190		
149		



Carne	Calorie
Bovina	186
Mista	191
Pollame	102
Bovina	181
Mista	182
Pollame	106
Bovina	176
Mista	190
Pollame	94
⋮	⋮
Bovina	149
Mista	135
Pollame	152
Bovina	135
Mista	140
Pollame	146
Bovina	132
Mista	138
Pollame	144

Figura 1.2: Contenuto in calorie di $n = 54$ confezioni di *hot-dog*. Da sinistra a destra: dati in forma di elenchi; rappresentazione grafica; dati nella forma di cui alla tabella 1.1.

Esempio 1.4. *Hot-dog* e calorie. Di $n = 54$ confezioni di *hot-dog* è stato rilevato il numero di calorie per confezione (Moore e McCabe, 1989). Le confezioni si distinguono per il tipo di carne, che può essere bovina, di pollame o mista. I dati sono riportati nella tabella di sinistra della figura 1.2, ma non sono nella forma della tabella 1.1; tuttavia, essi possono essere riscritti in tale forma, come mostrato nella tabella di destra della figura 1.2. Da quest'ultima è evidente che sono coinvolte due variabili: una quantitativa, le calorie, e una qualitativa, il tipo di carne. Ci si chiede se e in che misura l'apporto calorico sia diverso per i tre tipi di carne. Si formula cioè un modello del tipo (1.8), ossia

$$\text{calorie} = g(\text{carne}; \theta) + (\text{errore}),$$

dove $\theta = (\mu_1, \mu_2, \mu_3)$,

$$g(\text{carne}; \theta) = \begin{cases} \mu_1 & \text{se carne} = \text{Bovina} \\ \mu_2 & \text{se carne} = \text{Pollame} \\ \mu_3 & \text{se carne} = \text{Mista} \end{cases}$$

e μ_1, μ_2 e μ_3 sono i parametri da stimare. Il modello può essere scritto nella forma (1.9), come verrà illustrato nel capitolo 5. Notiamo che il problema potrebbe essere affrontato usando l'indice di dipendenza in media[10]. •

[8] In termini ingegneristici, potremmo dire che $h(Y_i)$ è un segnale disturbato, dove la componente deterministica è il segnale, mentre la componente casuale è il disturbo.

[9] Il volume di un cilindro è $\pi\,(\text{diametro}/2)^2\,(\text{altezza})$.

[10] Tale indice viene anche chiamato rapporto di correlazione. Si veda, ad esempio, Cicchitelli (2012, capitolo 9).

Esempio 1.5. Effetti del fumo sul peso dei neonati. Il peso alla nascita di un bambino dipende dalla durata della gravidanza? E, a parità di durata della gravidanza, dipende dal fatto che la madre fumasse durante la gestazione?
Al fine di rispondere a questi quesiti, si sono rilevati (Rosenberg e Buescher, 2002) per $n = 32$ neonati, il peso alla nascita (in grammi), la durata della gravidanza (in settimane), e la circostanza se la madre sia fumatrice o meno (S=si/N=no). I dati sono mostrati nella figura 1.3, in forma tabellare e grafica.
Per quanto riguarda la prima questione, il problema si può esprimere nella forma (1.8) con il modello

$$\text{peso} = g(\text{durata}; \theta) + (\text{errore}),$$

che, assumendo un semplice legame lineare, diventa

$$\text{peso} = \beta_1 + \beta_2\,(\text{durata}) + (\text{errore}).$$

Per rispondere al secondo quesito, il modello dovrebbe essere strutturato come

$$\text{peso} = g(\text{durata}, \text{fumo}; \theta) + (\text{errore}),$$

dove

$$g(\text{durata}, \text{fumo}; \theta) = \begin{cases} g_1(\text{durata}; \theta) & \text{se fumo} = \text{S} \\ g_2(\text{durata}; \theta) & \text{se fumo} = \text{N}. \end{cases}$$

Nella versione più semplice del modello, si potrebbe avere

$$g(\text{durata}, \text{fumo}; \theta) = \begin{cases} \beta_1 + \beta_2\,(\text{durata}) & \text{se fumo} = \text{S} \\ \beta_3 + \beta_4\,(\text{durata}) & \text{se fumo} = \text{N}. \end{cases}$$

L'inferenza per un modello di questo tipo è illustrata nella sezione 5.4. •

Esempio 1.6. Tempi olimpici. Consideriamo i tempi (in secondi) impiegati dai vincitori di medaglia d'oro per percorrere i 100 metri nelle olimpiadi moderne fino al 2016 (tabella 1.2).
Considerando dapprima i soli risultati maschili, rappresentiamo i tempi contro l'anno (figura 1.4, a sinistra): è evidente l'andamento decrescente che, si ritiene generalmente, riflette il miglioramento nei materiali e nelle tecniche di allenamento. Ci si può allora chiedere se l'ipotesi che negli anni ci sia un miglioramento dei risultati sia supportata dai dati, di quanto si migliora ogni anno o quale potrebbe essere il risultato alla prossima olimpiade.
Consideriamo poi i risultati olimpici nei $100m$ per uomini e donne (figura 1.4, a destra). Le due serie mostrano un andamento simile, ma su livelli diversi. Ci si chiede allora quale sia in media la differenza tra tempi maschili e femminili, se la differenza vari negli anni, se il miglioramento annuo (se c'è) sia lo stesso per uomini e donne, se le due serie possano essere spiegate in un unico modello. Questo esempio sarà ripreso nei capitoli 2, 3 e 5. •

Esempio 1.7. Emissioni di CO_2 e ricchezza. L'anidride carbonica (CO_2) è un gas prodotto da animali (respirazione) e alcuni batteri e da vari processi chimici (ad esempio la combustione). Esso è indispensabile alla vita (ad esempio è indispensabile alla fotosintesi), ma è anche tra i responsabili dell'effetto serra (trattenimento dell'energia solare nell'atmosfera). Si ritiene che il suo aumento (per effetto antropico) negli ultimi decenni stia portando a un aumento dell'effetto

Peso	Durata	Fumo
2940	38	S
2420	36	S
2760	39	S
2440	35	S
3301	42	S
2715	36	S
3130	39	S
2928	39	S
3446	42	S
2957	39	S
2580	38	S
3500	42	S
3200	41	S
3346	42	S
3175	41	S
2740	38	S
...	...	...

Peso	Durata	Fumo
...	...	...
3130	38	N
2450	34	N
3226	40	N
2729	37	N
3410	40	N
3095	39	N
3244	39	N
2520	35	N
3523	41	N
2920	38	N
3530	42	N
3040	37	N
3322	39	N
3459	40	N
2619	35	N
2841	36	N

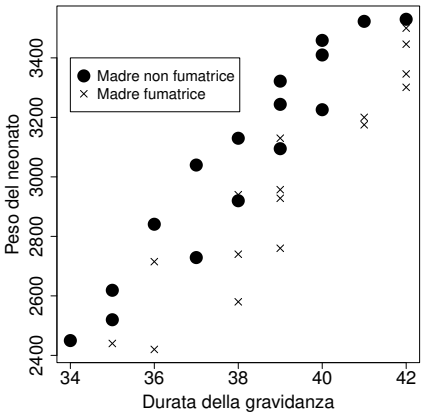


Figura 1.3: Peso dei neonati, durata della gravidanza e carattere fumatrice/non fumatrice della madre: dati e diagramma di dispersione.

serra e quindi delle temperature. Parte delle emissioni dovute ad attività antropica sono contabilizzate e attribuite ai singoli paesi ed è interessante valutare le emissioni in relazione ad altre caratteristiche del paese.

Per $n = 188$ paesi disponiamo (Marland *et al.*, 1999) delle emissioni di CO_2 (espresse in chilotonnellate), del PIL (in milioni di dollari USA), del PIL pro capite (in dollari USA) e della popolazione (in milioni di unità). Un estratto dei dati a disposizione è rappresentato nella tabella 1.3.

Tra le varie domande che ci si può porre:

- C'è una relazione tra emissioni e ricchezza?
- C'è una relazione tra emissioni e ricchezza, a parità di popolazione?
- Ricchezza e popolazione insieme possono spiegare le emissioni meglio che separatamente?

Dalla figura 1.5*(a)* si evince che sussiste una relazione tra emissioni e ricchezza, in particolare una relazione lineare nei logaritmi, come mostra la figura 1.5*(b)*. Per rispondere alla seconda domanda si può ragionare in termini di PIL pro capite (figura 1.5*(c)*). Rispondere all'ultima domanda richiede un modello simile a quello illustrato nell'esempio della sezione 4.11. •

Comune agli esempi visti sinora è il carattere quantitativo continuo della variabile risposta. Questa è infatti una limitazione del modello lineare che, essendo basato sull'assunzione di normalità, richiede variabili risposta di questo tipo. Invece, le variabili esplicative possono essere quantitative (sia continue che discrete) o qualitative e le due tipologie possono anche presentarsi nello stesso modello.

Il modello lineare risulterà inadatto in casi in cui la variabile risposta è quantitativa discreta o qualitativa, come negli esempi seguenti[11].

[11] Fanno eccezione le situazioni in cui la risposta, pur essendo discreta, abbia natura tale che la sua distribuzione sia approssimabile da una normale, come succede ad esempio per una Poisson con media molto elevata.