

Capitolo 2

Modelli di regressione

2.1 Introduzione

Siano assegnate p variabili $X_1, X_2, \dots, X_p$ concernenti n osservazioni di p caratteri da una popolazione assegnata. Ci si pone qui il problema di esaminare come alcune variabili influenzano (o sembrano influenzare) le altre. In alcuni casi esiste una relazione semplice, anche se in genere ciò costituisce un'eccezione; al contrario, spesso, la relazione funzionale che lega le variabili, nell'ipotesi in cui esista, è troppo complicata da individuare oppure da descrivere in termini semplici. In questi casi si cerca allora di approssimare tale relazione funzionale mediante funzioni matematiche semplici (ad esempio dei polinomi) dipendenti da opportune variabili e che approssimano o descrivono il fenomeno in esame in un certo insieme limitato di valori delle variabili dipendenti. Esaminando tale funzione possiamo cercare di comprendere meglio sia la relazione fra le variabili sia il contributo delle variabili sia considerate separatamente che congiuntamente.

Anche quando non esiste alcuna relazione fisica fra le variabili, a volte possiamo essere interessati ad ottenere una relazione funzionale tra tali variabili la quale, pur non avendo alcun significato fisico, può risultare estremamente utile per effettuare delle previsioni.

In questo capitolo considereremo il caso di una relazione fra le variabili lineare in alcuni parametri incogniti. Tali parametri sono da stimare, sotto opportune ipotesi, in base ad un assegnato insieme di dati.

In genere si considerano due tipi di variabili: le *variabili indipendenti* o *endogene* o *predittori*, e le *variabili esogene* o *risposte*. Le prime vengono usualmente indicate con X e si riferiscono a variabili il cui valore può essere scelto ad arbitrio oppure il cui valore può essere osservato ma non controllato (come ad esempio la temperatura esterna in una data zona). Il risultato della variazione delle variabili indipendenti produce un effetto sulle altre variabili, che pertanto vengono chiamate dipendenti, e che vengono usualmente indicati con Y . Mediante l'analisi di regressione si è interessati ad analizzare la variazione delle variabili dipendenti al variare delle variabili indipendenti.

Va sottolineato che la distinzione fra variabili dipendenti ed indipendenti non è sempre netta e dipende spesso dagli obiettivi dell'indagine statistica: una variabile può essere vista come indipendente in un certo contesto e come dipendente in un altro. In pratica, nel contesto dell'indagine statistica, il ruolo delle variabili dipendenti e di quelle indipendenti deve essere chiaramente individuato.

2.2 Il modello lineare

Il caso più semplice concerne lo studio della relazione fra la variabile risposta Y e la variabile indipendente X . Nel seguito supporremo che il legame funzionale fra le coppie di osservazioni possa essere descritto da una relazione del tipo:

$$Y = \beta_0 + \beta_1 x + \varepsilon, \quad (2.1)$$

che viene chiamato *modello di regressione lineare semplice* o, più in generale del tipo:

$$Y = \beta_0 + \beta_1 x + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon, \quad (2.2)$$

che viene chiamato *modello di regressione lineare multipla*. Pertanto il valore osservato y è la realizzazione di una variabile aleatoria (risposta) Y che può essere riguardata come la somma di $\beta_0 + \beta_1 x$ più una quantità casuale di errore ε , dove β_0, β_1 sono i *parametri* del modello e per la variabile aleatoria ε si possano assumere le seguenti ipotesi:

$$\mathbb{E}(\varepsilon) = 0 \quad (2.3)$$

$$\text{Var}(\varepsilon) = \sigma^2. \quad (2.4)$$

Pertanto la media e la varianza di Y condizionata rispetto a x si scrivono:

$$\mathbb{E}(Y|x) = \beta_0 + \beta_1 x \quad \text{Var}(Y|x) = \sigma^2. \quad (2.5)$$

L'equazione (2.1) costituisce il *modello* della relazione funzionale fra Y e X che si supponga essere valido; in un secondo momento ci si porrà il problema di valutare la validità di tale assunzione (in molti problemi statistici è necessario assumere un modello matematico per effettuare un'indagine).

Il parametro β_1 viene chiamato coefficiente di regressione e misura la variazione della media della risposta Y in corrispondenza di una variazione unitaria della variabile indipendente X . Se il range di X include anche l'origine, allora β_0 misura il valor medio di Y in corrispondenza di $x = 0$; nel caso contrario β_0 non ha alcun significato statistico.

I modelli di regressione vengono utilizzati soprattutto (ma non esclusivamente) a fini esplorativi (descrizione dei dati) e per la stima di modelli a fini previsionali e di controllo statistico di qualità.

Nota 2.1 Quando si parla di modello lineare o non lineare, ci si riferisce alla linearità o non linearità rispetto ai parametri. La potenza maggiore di una variabile indipendente nel modello viene chiamata *ordine* del modello. Per esempio un modello del tipo.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

è un modello lineare (rispetto ai parametri $\beta_0, \beta_1, \beta_2$) del secondo ordine (in X). In generale scriveremo pertanto:

$$Y_i = f_i(\beta_0, \beta_1, \beta_2) + \varepsilon_i$$

per evidenziare il fatto che i parametri sono i $\beta_i, i = 0, 1, 2$.

2.2.1 La stima dei parametri mediante il metodo dei minimi quadrati

Questo metodo, di grande importanza in molte aree della statistica, venne proposto da Gauss agli inizi del diciannovesimo secolo per la stima di misurazioni astronomiche.

Sia assegnato un campione di n coppie di osservazioni (x_i, y_i) , $i = 1, \dots, n$ relative ad una stessa popolazione e supponiamo che, con riferimento alla generica osservazione x_i , la risposta y_i possa essere considerata come una realizzazione della variabile aleatoria:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \dots, n, \quad (2.6)$$

o, più in generale:

$$Y_i = g_i(\beta_0, \beta_1, \beta_2, \dots, \beta_k) + \varepsilon_i$$

dove ε_i , per $i = 1, \dots, n$, sono variabili aleatorie per le quali è possibile assumere le seguenti ipotesi:

$$\mathbb{E}(\varepsilon_i) = 0 \quad 1 \leq i \leq n \quad (2.7)$$

$$\text{Var}(\varepsilon_i) = \sigma^2 \quad 1 \leq i \leq n \quad (2.8)$$

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad 1 \leq i < j \leq n. \quad (2.9)$$

La (2.8) viene chiamata ipotesi di *omoschedasticità*; in caso contrario si è in condizioni di eteroschedasticità. Il metodo di stima dei minimi quadrati si applica soltanto ai parametri di interesse $\beta_0, \beta_1, \dots, \beta_k$ e spesso si utilizza in situazioni in cui risulta difficile poter assumere ipotesi ulteriori rispetto alle (2.7), (2.8) e (2.9). Si noti che in particolare le ipotesi sopra specificate valgono quando le ε_i sono indipendenti ed identicamente distribuite (iid) con media 0 e varianza finita o – caso importante – quando $\varepsilon_i \sim N(0, \sigma^2)$ per ogni $i = 1, 2, \dots, n$.

Usualmente si dispone di n realizzazioni $y_1, \dots, y_n$ di $Y_1, \dots, Y_n$ che derivano da corrispondenti realizzazioni $\varepsilon_1, \dots, \varepsilon_n$ di $\varepsilon_1, \dots, \varepsilon_n$.

Consideriamo dapprima il modello di regressione semplice (2.1). Poniamo $\mathbf{y} = (y_1, \dots, y_n)'$, $\mathbf{x} = (x_1, \dots, x_n)$ e $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ e $f(x_i) = \beta_0 + \beta_1 x_i$; inoltre scriveremo $f_i(\boldsymbol{\beta})$ se vogliamo evidenziare la dipendenza funzionale da $\boldsymbol{\beta}$ e quindi $\mathbf{f}(\boldsymbol{\beta}) = (f_1(\boldsymbol{\beta}), \dots, f_n(\boldsymbol{\beta}))'$.

Il *metodo dei minimi quadrati* fornisce, quale stima $\mathbf{b} = (b_0, b_1)'$ di $\boldsymbol{\beta} = (\beta_0, \beta_1)'$, quel vettore tale che il valore atteso di $\mathbf{Y}$ – cioè $\mathbf{f}(\boldsymbol{\beta}) = (\beta_0 + \beta_1 x_1, \dots, \beta_0 + \beta_1 x_n)'$ in base alla (2.5) – sia il più vicino possibile, secondo la distanza euclidea, alla realizzazione $\mathbf{y} = (y_1, y_2, \dots, y_n)$. In altre parole, si considera il vettore $\mathbf{b}$ che minimizza la cosiddetta *funzione di errore*:

$$\mathcal{E}(\boldsymbol{\beta}) := \|\mathbf{f}(\boldsymbol{\beta}) - \mathbf{y}\|^2 = \sum_{i=1}^n (f_i(\boldsymbol{\beta}) - y_i)^2 \quad (2.10)$$

cioè

$$\begin{aligned} \mathbf{b} &:= \arg \min_{(\beta_0, \beta_1) \in \mathbb{R}} \mathcal{E}(\boldsymbol{\beta}) = \arg \min_{(\beta_0, \beta_1) \in \mathbb{R}} \|\mathbf{f}(\boldsymbol{\beta}) - \mathbf{y}\|^2 \\ &= \arg \min_{(\beta_0, \beta_1) \in \mathbb{R}} \sum_{i=1}^n (f_i(\boldsymbol{\beta}) - y_i)^2 = \arg \min_{(\beta_0, \beta_1) \in \mathbb{R}} \sum_{i=1}^n (\beta_0 + \beta_1 x_i - y_i)^2. \end{aligned}$$

Ne segue che la stima dei minimi quadrati $\mathbf{b}$ deve soddisfare il seguente sistema di equazioni:

$$\frac{\partial}{\partial \beta_j} \sum_{i=1}^n [f_i(\boldsymbol{\beta}) - y_i]^2 = 0 \quad j = 0, 1, \quad (2.11)$$

che vengono chiamate *equazioni normali*; dal sistema (2.11), segue

$$\sum_{i=1}^n [f_i(\boldsymbol{\beta}) - y_i] \frac{\partial f_i(\boldsymbol{\beta})}{\partial \beta_j} = 0 \quad j = 0, 1, 2, \dots, k. \quad (2.12)$$

Se le $f_i(\boldsymbol{\beta})$ sono lineari nei parametri $\boldsymbol{\beta}$, allora le equazioni normali diventano un sistema di equazioni lineari che può essere risolto esplicitamente. Nel caso della regressione semplice si ottiene:

$$\begin{aligned} \sum_{i=1}^n (b_0 + b_1 x_i - y_i) &= 0 \\ \sum_{i=1}^n x_i (b_0 + b_1 x_i - y_i) &= 0, \end{aligned}$$

cioè:

$$\begin{aligned} b_0 n + b_1 \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i \\ b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n x_i y_i, \end{aligned} \quad (2.13)$$

Se i valori osservati x_i non sono tutti uguali, si ottengono le soluzioni:

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{S_{xy}}{SS_x} \quad (2.14)$$

e

$$b_0 = \bar{y} - b_1 \bar{x} \quad (2.15)$$

dove

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{e} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

La retta di equazione $f(x) = b_0 + b_1 x$ viene chiamata *retta di regressione* di $y_1, y_2, \dots, y_n$ su $x_1, x_2, \dots, x_n$. Geometricamente, se si misura la distanza fra un punto (x_i, y_i) e una retta $y = b_0 + b_1 x$ verticalmente mediante $d_i = |y_i - (b_0 + b_1 x_i)|$, allora la retta di regressione è quella che minimizza la somma dei quadrati delle distanze agli n punti $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.

Alcune conseguenze delle scelte effettuate riguardano la media dei residui e la covarianza σ_{ex} fra residui e e la variabile indipendente X . Posto:

$$e_i := b_0 + b_1 x_i - y_i \quad (2.16)$$

si ha infatti in base alla (2.15):

$$\begin{aligned} \bar{e} &= \frac{1}{n} \sum_{i=1}^n e_i = \frac{1}{n} \sum_{i=1}^n (b_0 + b_1 x_i - y_i) = \frac{1}{n} \sum_{i=1}^n (b_0 + b_1 x_i) - \frac{1}{n} \sum_{i=1}^n y_i \\ &= b_0 + b_1 \bar{x} - \bar{y} = 0. \end{aligned} \quad (2.17)$$

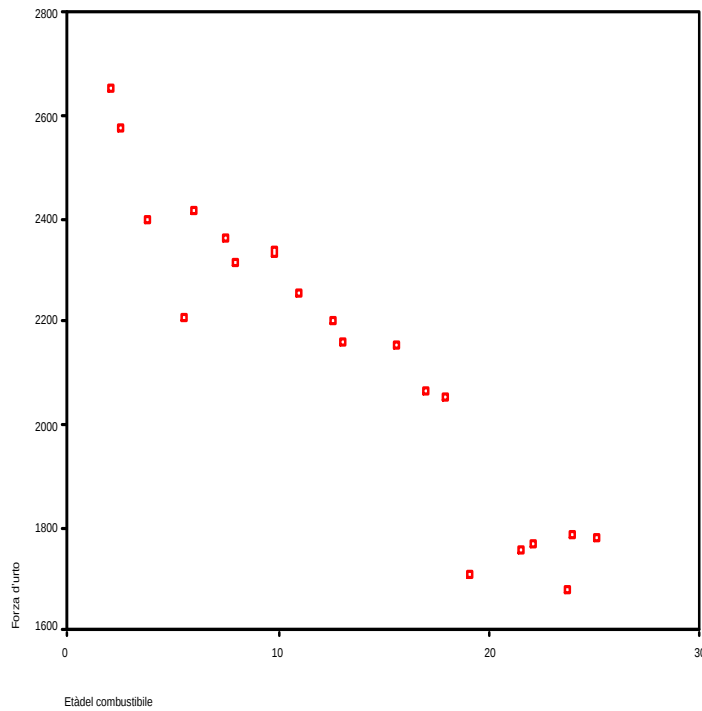


Figura 2.1: *Diagramma a dispersione forza d'urto-età del combustibile.*

Si ha inoltre, in base alla (2.14):

$$\begin{aligned}
 \sigma_{xe} &= \frac{1}{n} \sum_{i=1}^n (e_i - \bar{e})(x_i - \bar{x}) = \frac{1}{n} \sum_{i=1}^n [b_0 + b_1 x_i - y_i - (b_0 + b_1 \bar{x} - \bar{y})](x_i - \bar{x}) \\
 &= \frac{1}{n} \sum_{i=1}^n [b_1(x_i - \bar{x}) - (y_i - \bar{y})](x_i - \bar{x}) = b_1 \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 - \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})
 \end{aligned} \tag{2.18}$$

$$\begin{aligned}
 &= b_1 \sigma_x^2 - \sigma_{xy} = \frac{\sigma_{xy}}{\sigma_x^2} \sigma_x^2 - \sigma_{xy} \\
 &= 0 .
 \end{aligned} \tag{2.19}$$

Si vede in particolare che $\sigma_{xe} = \frac{1}{n} \sum_{i=1}^n e_i x_i$ in quanto $\bar{e} = 0$ e $\sum_i e_i = 0$.

Esempio 2.2 Il motore di un missile viene costruito miscelando insieme un combustibile esplosivo ed uno di origine naturale all'interno di un serbatoio metallico. La forza d'urto della miscela è un'importante caratteristica di qualità del sistema. Si ipotizza che la forza d'urto della miscela possa dipendere dall'età del combustibile di origine naturale (rispetto alla data di produzione). Vengono quindi raccolte $n = 20$ coppie di osservazioni concernenti la forza d'urto (y_i) e l'età del combustibile in settimane (x_i) e riportate nella Tabella 2.1. Il grafico in Figura 2.1 riporta il diagramma a dispersione (scatter plot) di y_i rispetto a x_i . Considerato un modello lineare del tipo:

$$Y = \beta_0 + \beta_1 x + \varepsilon ,$$

i	x_i	y_i	$f(x_i)$	e_i
1	15.5	2158.70	2052.00	106.70
2	23.75	1678.15	1745.51	-67.36
3	8	2316.00	2330.62	-14.62
4	17	2061.30	1996.27	65.03
5	5.5	2207.50	2423.50	-216.00
6	19	1708.30	1921.97	-213.67
7	24	1784.70	1736.22	48.48
8	2.5	2575.00	2534.95	40.05
9	7.5	2357.90	2349.20	8.70
10	11	2256.70	2219.17	37.53
11	13	2165.20	2144.87	20.33
12	3.75	2399.55	2488.51	-88.96
13	25	1779.80	1699.07	80.73
14	9.75	2336.75	2265.61	71.14
15	22	1765.30	1810.52	-45.22
16	18	2053.50	1959.12	94.38
17	6	2414.40	2404.92	9.48
18	12.5	2200.50	2163.45	37.05
19	2	2654.20	2553.52	100.68
20	21.5	1753.70	1829.10	-75.40

Tabella 2.1: Esempio 2.2: valori rilevati (x_i, y_i) , valori stimati e residui $(f(x_i), e_i)$ in accordo al modello di regressione 2.20.

in base al metodo dei minimi quadrati, in base alle (2.15) e (2.14) si ottengono le seguenti stime di β_0 e β_1 :

$$\begin{aligned} b_0 &= 2627.82 \\ b_1 &= -37.15 . \end{aligned}$$

Si ottiene pertanto il modello:

$$f(x_i) = 2627.82 - 37.15x . \quad (2.20)$$

Possiamo interpretare il valore $b_1 = -37.15$ come il decremento medio della forza d'urto in corrispondenza di un aumento di una settimana dell'età del combustibile; il valore $b_0 = 2627.82$ può essere interpretato come la forza d'urto media che si ottiene utilizzando combustibile naturale appena prodotto.

Nella seconda parte della Tabella 2.1 si riportano i valori stimati $f(x_i)$ della variabile risposta ed i residui e_i . 

Dopo aver ottenuto il modello, ci si devono porre alcune importanti questioni:

1. Quanto il modello così ottenuto si adatta ai dati?
2. Il modello può essere opportunamente usato a fini predittivi?
3. Le assunzioni fatte (come ad esempio, varianza costante e non correlazione degli errori) sono verificate?

2.2.2 Il coefficiente di determinazione R^2

Un importante elemento per rispondere alla prima domanda è il coefficiente di determinazione che misura la frazione di variabilità spiegata dal modello rispetto alla variabilità della risposta Y . Consideriamo la devianza della variabile risposta Y :

$$SS_Y := \sum_{i=1}^n (y_i - \bar{y})^2 .$$

Per ciascun valore osservato y_i di Y , possiamo scrivere:

$$\begin{aligned} (y_i - \bar{y})^2 &= [(y_i - f(x_i)) + (f(x_i) - \bar{y})]^2 \\ &= (y_i - f(x_i))^2 + (f(x_i) - \bar{y})^2 + 2(y_i - f(x_i))(f(x_i) - \bar{y}) ; \end{aligned}$$

allora si ha:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - f(x_i))^2 + \sum_{i=1}^n (f(x_i) - \bar{y})^2 + 2 \sum_{i=1}^n (y_i - f(x_i))(f(x_i) - \bar{y}) .$$

Il termine misto $\sum_{i=1}^n [(y_i - f(x_i))(f(x_i) - \bar{y})]$ è nullo. Infatti, sostituendo le quantità:

$$e_i = y_i - f(x_i) , \quad f(x_i) = b_0 + b_1 x_i \quad \text{e} \quad \bar{y} = b_0 - b_1 \bar{x}$$

si ha:

$$\sum_{i=1}^n (y_i - f(x_i))(f(x_i) - \bar{y}) = \sum_{i=1}^n e_i(b_0 - b_1x_i - b_0 - b_1\bar{x}) = b_1 \sum_{i=1}^n e_i(x_i - \bar{x}) .$$

Poichè in base alle (2.17) e (2.19) si ha $\bar{e} = 0$ e $\sigma_{ex} = 0$, la relazione precedente può essere riscritta:

$$\sum_{i=1}^n e_i(x_i - \bar{x}) = \sum_{i=1}^n (e_i - \bar{e})(x_i - \bar{x}) = N\sigma_{xe} = 0 .$$

La devianza di Y è pertanto data dalla somma di due contributi:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - f(x_i))^2 + \sum_{i=1}^n (f(x_i) - \bar{y})^2 . \quad (2.21)$$

La relazione precedente mostra che la variazione di Y intorno alla media è in parte dovuta alla funzione di regressione $\sum_{i=1}^n (f(x_i) - \bar{y})^2$ (la *devianza di regressione*); la restante parte $\sum_{i=1}^n (y_i - f(x_i))^2 = \sum_{i=1}^n e_i^2$ – che viene chiamata *devianza residua* o *di errore* – è dovuta al fatto che non tutte le osservazioni giacciono sulla retta di regressione. Possiamo scriverla sinteticamente:

$$SS_Y = SS_E + SS_R . \quad (2.22)$$

Si noti che dividendo per n ambo i membri della precedente relazione (2.22) si ottiene la stessa relazione espressa in termini di varianze:

$$\sigma_y^2 = \sigma_\varepsilon^2 + \sigma_f^2 . \quad (2.23)$$

Dalla (2.22) si vede che un modo di valutare la bontà del modello di regressione è quello di vedere quanta parte della variabilità di Y è attribuibile alla regressione. Ovviamente si auspica che la devianza di regressione sia molto maggiore della devianza di errore.

Dalla (2.23) possiamo ottenere il rapporto

$$R^2 := \frac{SS_R}{SS_E} = \frac{\sigma_f^2}{\sigma_y^2}$$

che viene chiamato *indice o coefficiente di determinazione* in quanto esprime la parte della varianza totale (devianza) di Y *determinata* o *spiegata* dalla relazione di regressione.

Il rapporto

$$\frac{\sigma_\varepsilon^2}{\sigma_y^2} = 1 - \frac{\sigma_f^2}{\sigma_y^2} = 1 - R^2$$

viene chiamato *coefficiente di alienazione* ed esprime la frazione della variabilità *non spiegata* dalla regressione. L'indice di determinazione varia fra 0 e 1, in particolare:

$R^2 = 0$: quando la devianza di regressione è nulla, cioè nel caso in cui $b_1 = 0$;

$R^2 = 1$: nel caso in cui la devianza residua è nulla, cioè quando in punti (x_i, y_i) sono allineati.

Questa proprietà, secondo cui R^2 misura la dispersione delle y_j e delle x_i intorno alle corrispondenti rette di regressione, induce a ritenere tale indice come misura del grado di accostamento del legame effettivo fra X e Y ad una relazione lineare, assumendo come variabile indipendente una volta la X ed una volta la Y .

Calcolo pratico del coefficiente R^2 . Per il calcolo del coefficiente di determinazione, tenendo conto che:

$$\begin{aligned}\sigma_f^2 &= \frac{1}{N} \sum_{i=1}^n (f(x_i) - \bar{y})^2 = \frac{1}{N} \sum_{i=1}^n (b_0 + b_1 x_i - b_0 - b_1 \bar{x})^2 = b_1^2 \frac{1}{N} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \frac{\sigma_{xy}^2}{\sigma_x^4} \sigma_x^2 = \frac{\sigma_{xy}^2}{\sigma_x^2},\end{aligned}$$

si ha:

$$R^2 = \frac{\sigma_f^2}{\sigma_y^2} = \frac{\sigma_{xy}^2}{\sigma_y^2 \sigma_x^2} = r^2.$$

dove r^2 è il quadrato del coefficiente di correlazione.

Alcune note sul coefficiente di determinazione Il coefficiente di determinazione R^2 va comunque usato con qualche cautela. In generale, un valore alto di R^2 non implica necessariamente che il modello di regressione sia adeguato (cioè che la forma funzionale scelta esprima bene la struttura di dipendenza dei dati). Alcune note:

1. E' sempre possibile aumentarne il valore aggiungendo termini al modello. Per esempio, se fra gli n punti osservati $x_1, \dots, x_n$ non vi sono misure ripetute (cioè valori diversi della risposta Y in corrispondenza degli stessi valori della x), un polinomio di grado $n-1$ fornirà un adattamento perfetto ($R^2 = 1$) agli n valori. Al contrario, quando vi sono misure ripetute, R^2 non sarà mai uguale a uno in quanto il modello non può spiegare la variabilità dovuta all'errore puro.
2. Il valore di R^2 dipende anche dall'intervallo di variazione della variabile di regressione. In generale, se il modello specificato è corretto, il valore di R^2 cresce all'aumentare della dispersione dei valori di x e, al contrario, decresce al diminuire della variabilità della variabile di regressione.

2.3 Un'altra formulazione del modello di regressione

Presentiamo ora un'altra formulazione del modello di regressione (2.1) che può risultare utile in alcuni casi:

$$\begin{aligned}Y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} + \beta_1 x_i + \varepsilon_i = (\beta_0 + \beta_1 \bar{x}) + \beta_1 (x_i - \bar{x}) + \varepsilon_i \\ &= \beta'_0 + \beta_1 (x_i - \bar{x}) + \varepsilon_i,\end{aligned}\tag{2.24}$$

in cui l'origine del regressore X viene traslato nel valor medio $\bar{x}$ di $x_1, \dots, x_n$; in questo caso la nuova intercetta all'origine è $\beta'_0 = \beta_0 + \beta_1 \bar{x}$.

Le equazioni normali per il modello (2.24) sono le seguenti:

$$\begin{aligned}nb'_0 &= \sum_{i=1}^n y_i \\ b'_1 \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n y_i (x_i - \bar{x})\end{aligned}\tag{2.25}$$

da cui seguono le seguenti stime (dei minimi quadrati):

$$b'_0 = \bar{y} \quad (2.26)$$

$$b'_1 = \frac{\sum_{i=1}^n y_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{SS_x} \quad (2.27)$$

In particolare l'intercetta nella nuova origine è ovviamente data da $\bar{y}$, mentre (ancora ovviamente) $b'_1 = b_1$.

Notiamo che il nuovo sistema di equazioni normali (2.25) risulta più semplice da risolvere rispetto al precedente (2.13) in quanto ora ciascuna equazione presenta solo un'incognita. In base alla (2.24), il modello di regressione si scrive:

$$f(x) = \bar{y} + b'_1(x_i - \bar{x}) \quad (2.28)$$

e, benchè sia formalmente equivalente a quello precedente $f(x) = b_0 + b_1x$, evidenzia immediatamente che esso ha validità nell'intervallo dei valori assegnati centrato in $\bar{x}$.

2.4 Regressione multipla per distribuzioni triple

Per introdurci al caso generale della regressione multipla (2.2), consideriamo dapprima tre variabili X_1 , X_2 e X_3 , in cui possiamo considerare la dipendenza di X_1 su X_2 e X_3 , di X_2 su X_1 e X_3 di X_3 su X_1 e X_2 e quindi, rispettivamente, le tre equazioni di regressione:

$$X_1 = b_{10} + b_{12}x_2 + b_{13}x_3 + \varepsilon_1 \quad (2.29)$$

$$X_2 = b_{20} + b_{21}x_1 + b_{23}x_3 + \varepsilon_2 \quad (2.30)$$

$$X_3 = b_{30} + b_{31}x_1 + b_{32}x_2 + \varepsilon_3 . \quad (2.31)$$

Nella prima la X_1 è la variabile *dipendente* e X_2 , X_3 sono le variabili *indipendenti* o *esplicative*; nel secondo caso X_2 è la variabile dipendente e X_1 , X_3 sono le variabili indipendenti; nel terzo caso X_3 è la variabile dipendente e X_1 , X_2 sono le variabili indipendenti.

Rappresentando la distribuzione tripla in uno spazio cartesiano tridimensionale, di coordinate x_1, x_2, x_3 , le tre equazioni (2.29), (2.30) e (2.31) rappresentano ciascuna un piano che è detto *piano di regressione*: nel primo caso si ha il modello di regression di X_1 su X_2 e X_3 ; analogamente per gli altri casi.

Per determinare i parametri di tale piano di regressione utilizza il metodo dei minimi quadrati visto in precedenza. Per fissare le idee, consideriamo la regressione di X_3 su X_1, X_2 . Per semplicità, scriviamo

$$f(x_1, x_2) := b_0 + b_1x_1 + b_2x_2 .$$

Consideriamo gli scarti:

$$e_i := x_{3i} - f(x_{1i}, x_{2i}) = x_{3i} - (b_0 + b_1x_{1i} + b_2x_{2i}) \quad (2.32)$$

In base al metodo dei minimi quadrati, bisogna rendere minima la somma dei quadrati degli scostamenti dei valori previsti $f(x_{1i}, x_{2i})$ dai valori empirici x_{3i} :

$$\sum_{i=1}^N [x_{3i} - f(x_{1i}, x_{2i})]^2 = \sum_{i=1}^N [x_{3i} - (b_0 + b_1x_{1i} + b_2x_{2i})]^2 . \quad (2.33)$$

Analogamente a quanto visto nel caso bivariato, al fine di minimizzare la (2.33), si impongono le condizioni:

$$\bar{\epsilon} = 0 \quad \sigma_{1\epsilon} = 0 \quad \sigma_{2\epsilon} = 0 , \quad (2.34)$$

dove $\bar{\epsilon}$ è la media aritmetica delle (2.32) e:

$$\sigma_{1\epsilon} := \frac{1}{N} \sum_i (x_{1i} - \bar{x}_1)(\epsilon_i - \bar{\epsilon}) \quad \sigma_{2\epsilon} := \frac{1}{N} \sum_i (x_{2i} - \bar{x}_2)(\epsilon_i - \bar{\epsilon}) .$$

Dopo alcuni passaggi si perviene al sistema:

$$\begin{aligned} b_1 \sigma_{11} + b_2 \sigma_{12} &= \sigma_{13} \\ b_1 \sigma_{21} + b_2 \sigma_{22} &= \sigma_{23} \\ b_0 &= \bar{x}_3 - b_1 \bar{x}_1 + b_2 \bar{x}_2 \end{aligned} \quad (2.35)$$

dove, in base alle notazioni precedentemente introdotte, $\sigma_{ii} = \sigma_i^2$ e $\sigma_{ij} = \sigma_{ji}$. Si ottengono le relazioni:

$$b_1 = \frac{\sigma_{13}\sigma_2^2 - \sigma_{23}\sigma_{12}}{\sigma_2^2\sigma_1^2 - \sigma_{12}^2} = \frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2} \frac{\sigma_3}{\sigma_1} \quad (2.36)$$

$$b_2 = \frac{\sigma_{23}\sigma_1^2 - \sigma_{13}\sigma_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} = \frac{r_{23} - r_{13}r_{12}}{1 - r_{12}^2} \frac{\sigma_3}{\sigma_2} , \quad (2.37)$$

dove r_{13} , r_{23} e r_{12} sono i tre coefficienti di correlazione introdotti nella (1.12); successivamente si ricava b_0 direttamente dalla terza equazione del sistema (2.35). Si ottiene pertanto:

$$f(x_1, x_2) = b_0 + b_1 x_1 + b_2 x_2 \quad (2.38)$$

che passa per il punto di coordinate $(\bar{x}_1, \bar{x}_2, \bar{x}_3)$. La (2.38) consente di conoscere come varia il carattere dipendente al variare dei due caratteri indipendenti X_1, X_2 .

Se nella (2.38) fissiamo il valore di $X_2 = x_2^\circ$, otteniamo una funzione della sola variabile x_1 :

$$f(x_1, x_2^\circ) = b'_0 + b_1 x_1 \quad (2.39)$$

dove $b'_0 = b_0 + b_2 x_2^\circ$. La (2.38), che esprime come varia in media il carattere X_3 al variare di X_1 per un valore fissato $X_2 = x_2^\circ$ di X_2 , si chiama *retta di regressione parziale di X_3 rispetto a X_1 fissato $X_2 = x_2^\circ$* .

Si noti che facendo assumere a X_2 vari valori, si ottengono altrettante rette di regressione parziale di X_3 su X_1 che differiscono per il termine noto; esse pertanto sono parallele. Il termine b_1 viene chiamato *coefficiente di regressione di X_3 rispetto a X_1 tenendo costante X_2* .

Nota 2.3 Si noti che tale rette hanno ovviamente pendenza costante, ma non è detto che, al variare di x_2° esse giacciono sullo stesso piano.

2.4.1 Coefficiente di correlazione multipla

Anche in questo caso, in analogia al caso delle distribuzioni doppie, la devianza totale si scompone nella devianza di regressione e nella devianza dell'errore o residua. L'indice:

$$R^2 := \frac{\sigma_f^2}{\sigma_y^2} = \frac{\sum_{i=1}^N [f(x_{1i}, x_{2i}) - \bar{y}]^2}{\sum_{i=1}^N [y_i - \bar{y}]^2} \quad (2.40)$$

prende il nome di *coefficiente di determinazione* e rappresenta la parte della devianza totale di Y che è spiegata (o determinata) dalla supposta relazione lineare con X_1 e X_2 .

Il coefficiente R^2 assume valore massimo $R^2 = 1$ quando risulta $f(x_{1i}, x_{2i}) = y_i$, cioè quando i punti appartengono al piano di regressione; Il coefficiente R^2 assume valore minimo $R^2 = 0$ quando si ha $f(x_{1i}, x_{2i}) = \bar{y}$ per ogni $i = 1, \dots, N$, cioè se il piano è parallelo al piano $X_1 X_2$ ad una distanza uguale a $\bar{y}$.

La radice quadrata di R^2 viene chiamata *coefficiente di correlazione multipla*.

- Evidentemente risulta $R^2 = 1$ quando gli scarti a secondo membro della (2.40) sono tutti nulli, cioè quando $y_i = f(x_{1i}, x_{2i})$ per ogni $i = 1, \dots, N$. In tal caso i punti immagine nello spazio dei dati empirici giacciono tutti su di uno stesso piano che ovviamente è il piano di regressione;
- Risulta invece $R^2 = 0$ quando i punti empirici sono massimamente dispersi rispetto al piano di regressione e questo accade quando il tipo di funzione scelto (lineare) attraverso cui ci si propone di interpretare il comportamento del fenomeno y in funzione di x_1 e x_2 , non fornisce alcun utile informazione al riguardo poichè al variare della coppia di valori (x_{1i}, x_{2i}) , il valore teorico $f(x_{1i}, x_{2i})$ è costante, per $i = 1, \dots, N$.

Il calcolo diretto di R^2 è laborioso per più di un regressore, nel caso in esame si dimostra la relazione:

$$R^2 = 1 - \frac{\begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{12} \\ r_{21} & 1 \end{vmatrix}} = \frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{12}^2}. \quad (2.41)$$

Ovviamente deve risultare $r_{12} \neq 0$.

2.4.2 Correlazione parziale

Da quanto visto in precedenza si evince che b_0, b_1, b_2 dipendono dalla correlazione di X_3 con X_1 e di X_3 con X_2 , ma anche dalla correlazione fra X_1 e X_2 . Soltanto se $r_{12} = 0$ allora b_1 e b_2 dipendono esclusivamente dalla correlazione di X_1 con X_2 e di X_1 con X_3 , anzi si ha in questo caso dalle (2.36) e (2.37) segue:

$$b_{12} = \frac{\sigma_3}{\sigma_1} r_{13}, \quad b_{13} = \frac{\sigma_3}{\sigma_2} r_{23}.$$

Per comprendere il significato del coefficiente di correlazione parziale, preliminarmente ricordiamo che il coefficiente di regressione lineare fra due variabili X_1 e X_2 può essere visto

come la media geometrica dei due coefficienti di regressione b_{12} e b_{21} rispettivamente della retta di regressione di X_1 su X_2 e di quella di X_2 su X_1 :

$$r = \sqrt{b_{12} \cdot b_{21}}$$

Supponiamo di aver determinato l'equazione di regressione di X_1 rispetto a X_2 e X_3 e quella di X_2 rispetto a X_1 e X_3 :

$$\begin{aligned} X_1 &= f(x_2, x_3) + \epsilon_1 = b_{12}x_2 + b_{13}x_3 + b_{10} + \epsilon_1 \\ X_2 &= f(x_1, x_3) + \epsilon_2 = b_{21}x_1 + b_{23}x_3 + b_{20} + \epsilon_2, \end{aligned}$$

dove b_{12} fornisce la variazione media di X_1 rispetto ad una variazione unitaria di x_2 per x_3 fissato, e b_{21} fornisce la variazione media di x_2 per un aumento unitario di x_1 , con x_3 costante. Infine ϵ_1, ϵ_2 sono variabili aleatorie con media zero e varianza finita e costante.

Si può quindi misurare la concordanza o la discordanza fra X_1 e X_2 , per X_3 costante, con la media geometrica dei due coefficienti di regressione parziale $b_{12,3}$ e $b_{21,3}$ che viene chiamata *coefficiente di correlazione parziale fra X_1 e X_2 con X_3 costante*. I valori $b_{12,3}$ e $b_{21,3}$ coincidono numericamente rispettivamente con b_{12} e b_{21} ma hanno significato statistico un po' diverso. Infatti, facendo assumere a X_3 valori diversi, si ottengono altrettante rette di regressione di X_1 su X_2 (ed analogamente di X_2 su X_1) che differiscono per l'ordinata all'origine (termine noto) ma che hanno sempre lo stesso valore di $b_{12,3}$ (risp. $b_{21,3}$), cioè lo stesso coefficiente angolare e pertanto sono parallele. Si intende per coefficiente di regressione parziale di X_1 su X_2 (risp. di X_2 su X_1) tenendo costante X_3 proprio questo valore comune.

Si ha pertanto:

$$r_{12,3} = \sqrt{b_{12,3} \cdot b_{21,3}},$$

dove $b_{12,3}$ è il coefficiente di regressione parziale di X_1 rispetto a X_2 , fissato X_3 e $b_{21,3}$ è il coefficiente di regressione parziale di X_2 rispetto a X_1 , fissato X_3 . Tenendo conto delle relazioni fra coefficiente di regressione e coefficienti di correlazione nella regressione multipla, che – come visto nelle (2.36) e (2.37) – sono dati rispettivamente da:

$$\begin{aligned} b_{12,3} &= \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \frac{\sigma_1}{\sigma_2} \\ b_{21,3} &= \frac{r_{21} - r_{13}r_{23}}{1 - r_{13}^2} \frac{\sigma_2}{\sigma_1} \end{aligned}$$

da cui la media geometrica è data da (si tenga conto della simmetria della matrice R , per cui $r_{12} = r_{21}$):

$$r_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}. \quad (2.42)$$

Esempio 2.4 Un'azienda produttrice di bevande analcoliche desidera analizzare le prestazioni del servizio di assistenza e manutenzione dei distributori automatici della propria rete distributiva. In particolare si è interessati alla previsione del tempo richiesto da un operatore per effettuare l'approvvigionamento e l'ordinaria manutenzione dei distributori automatici in un generico punto vendita (edificio, o complesso di edifici, che in genere contiene più distributori automatici). Le attività da svolgere comprendono la ricarica dei distributori con le bevande

previsite, oltre ad alcune semplici operazioni di ordinaria manutenzione. Il responsabile dello studio, ha suggerito che le due variabili più importanti che influenzano il tempo richiesto per effettuare il servizio completo in ciascun punto vendita comprendono il numero di casse di prodotti (lattine o bottiglie in plastica) richiesti dal distributore e la distanza percorsa a piedi dall'operatore. Il responsabile ha raccolto 25 osservazioni concernenti le seguenti variabili:

- Y : tempo richiesto per l'espletamento del servizio, in minuti (variabile dipendente),
 X_1 : numero di casse di prodotti,
 X_2 : distanza percorsa dall'operatore, in metri,

i	Y	X_1	X_2
1	16.68	7	171
2	11.50	3	67
3	12.03	3	104
4	14.88	4	24
5	13.75	6	46
6	18.11	7	101
7	8.00	2	34
8	17.83	7	64
9	79.24	30	445
10	21.50	5	184
11	40.33	16	210
12	21.00	10	66
13	13.50	4	78
14	19.75	6	141
15	24.00	9	137
16	29.00	10	237
17	15.35	6	61
18	19.00	7	40
19	9.50	3	11
20	35.10	17	235
21	17.90	10	43
22	52.32	26	247
23	18.75	9	137
24	19.83	8	194
25	10.75	4	46

I dati sono riportati nel file *manutenzione.xls*.

1. Costruire il modello di regressione di Y su X_1 e X_2 ;
2. il coefficiente di correlazione lineare semplice r_{Y1} di Y e X_1 ;
3. il coefficiente di correlazione parziale di Y e X_1 tenendo costante la distanza percorsa X_2 , cioè $r_{Y1.2}$;

4. il coefficiente di correlazione multipla di Y su X_1 e X_2 ;

Il problema richiede la stima dei parametri dell'equazione di regressione multipla $f(x_1, x_2) = b_0 + b_1x_1 + b_2x_2$ in base al metodo dei minimi quadrati. I valori dei parametri che rendono minima la quantità:

$$\sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n [y_i - (b_0 + b_1x_{i1} + b_2x_{i2})]^2$$

si ottengono dal seguente sistema (2.35) dove:

$$\begin{aligned} \sigma_y^2 &= M(Y^2) - \bar{y}^2 & \sigma_{y1} &= M(YX_1) - \bar{y}\bar{x}_1 \\ \sigma_1^2 &= M(X_1^2) - \bar{x}_1^2 & \sigma_{y2} &= M(YX_2) - \bar{y}\bar{x}_2 \\ \sigma_2^2 &= M(X_2^2) - \bar{x}_2^2 & \sigma_{12} &= M(X_1X_2) - \bar{x}_1\bar{x}_2 . \end{aligned}$$

Possiamo pertanto impostare la seguente tabella:

i	y_i	x_{i1}	x_{i2}	y_i^2	x_{i1}^2	x_{i2}^2	y_ix_{i1}	y_ix_{i2}	x_ix_{i2}
1	16.68	7	171	278.22	49	29134	116.76	2847.08	1195
2	11.50	3	67	132.25	9	4497	34.50	771.14	201
3	12.03	3	104	144.72	9	10740	36.09	1246.69	311
4	14.88	4	24	221.41	16	595	59.52	362.83	98
5	13.75	6	46	189.06	36	2090	82.50	628.65	274
6	18.11	7	101	327.97	49	10117	126.77	1821.58	704
7	8.00	2	34	64.00	4	1124	16.00	268.22	67
8	17.83	7	64	317.91	49	4097	124.81	1141.26	448
9	79.24	30	445	6278.98	900	198032	2377.20	35262.43	13350
10	21.50	5	184	462.25	25	34005	107.50	3964.69	922
11	40.33	16	210	1626.51	256	43975	645.28	8457.30	3355
12	21.00	10	66	441.00	100	4294	210.00	1376.17	655
13	13.50	4	78	182.25	16	6041	54.00	1049.27	311
14	19.75	6	141	390.06	36	19830	118.50	2781.15	845
15	24.00	9	137	576.00	81	18646	216.00	3277.21	1229
16	29.00	10	237	841.00	100	55944	290.00	6859.22	2365
17	15.35	6	61	235.62	36	3716	92.10	935.74	366
18	19.00	7	40	361.00	49	1619	133.00	764.44	282
19	9.50	3	11	90.25	9	120	28.50	104.24	33
20	35.10	17	235	1232.01	289	55082	596.70	8237.83	3990
21	17.90	10	43	320.41	100	1821	179.00	763.83	427
22	52.32	26	247	2737.38	676	60954	1360.32	12917.18	6419
23	18.75	9	137	351.56	81	18813	168.75	2571.75	1234
24	19.83	8	194	393.23	64	37461	158.64	3838.06	1548
25	10.75	4	46	115.56	16	2090	43.00	491.49	183
somme	559.60	219	3119	18310.63	3055	624837	7375.44	102739.45	40812
medie	22.38	9	125	732.43	122	24993	295.02	4109.58	1632

e pertanto si ottengono le seguenti varianze e covarianze:

$$\begin{aligned}\sigma_y^2 &= 231.38 & \sigma_{y1} &= 98.93 \\ \sigma_1^2 &= 45.46 & \sigma_{y2} &= 1317.21 \\ \sigma_2^2 &= 9431.28 & \sigma_{12} &= 539.70 .\end{aligned}$$

e successivamente gli scarti quadratici medi: $\sigma_y = 15.21$, $\sigma_1 = 6.74$ e $\sigma_2 = 97.11$. A questo punto possiamo risolvere il sistema di equazioni, ottenendo i valori:

$$\begin{aligned}b_0 &= 2.3412 \\ b_1 &= 1.6159 \\ b_3 &= 0.0144\end{aligned}$$

e pertanto l'equazione di regressione multipla si scrive:

$$f(x_1, x_2) = 2.3412 + 1.6159x_1 + 0.0144x_2 . \quad (2.43)$$

Dai dati sopra ricavati, si ottengono immediatamente le altre grandezze. Il coefficiente di correlazione semplice di Y e X è dato da:

$$r_{y1} = \frac{\sigma_{1y}}{\sigma_y \sigma_1} = 0.965 .$$

Il coefficiente di correlazione parziale fra il tempo richiesto per l'espletamento del servizio Y ed il numero di casse di prodotti X_1 tenendo costante la distanza percorsa a piedi dall'operatore X_2 è dato da:

$$r_{y1.2} = \frac{r_{y1} - r_{y2} r_{12}}{\sqrt{1 - r_{12}^2} \sqrt{1 - r_{y2}^2}} . \quad (2.44)$$

Nel nostro caso dobbiamo ancora ricavare r_{y2} e r_{12} :

$$r_{y2} = \frac{\sigma_{y2}}{\sigma_y \sigma_2} = 0.892 \quad r_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = 0.824 .$$

da cui, sostituendo nell'equazione (2.44), si ottiene:

$$r_{y1.2} = 0.896 .$$

In base alla (2.41) possiamo infine calcolare il coefficiente di correlazione multipla di Y su X_1 e X_2 :

$$R = 0.96 .$$



2.5 La regressione multipla (caso generale)

Nel caso della regressione lineare multipla si considerano p variabili indipendenti (o predittive) $X_1, \dots, X_p$ ed una variabile dipendente Y . Il modello si scrive pertanto:

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon , \quad (2.45)$$

che viene chiamato *modello di regressione multipla* e descrive un iperpiano nello spazio $\mathbb{R}^p$; le variabili $X_1, \dots, X_p$ vengono anche chiamati *regressori*. Da un punto di vista geometrico, l'equazione $\mathbb{E}(Y) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$ descrive un iperpiano.

I coefficienti $\beta_0, \beta_1, \dots, \beta_p$ vengono chiamati coefficienti regressione. Il parametro β_j , ($j = 1, \dots, p$) fornisce la variazione media della variabile risposta Y in corrispondenza di una variazione unitaria del regressore X_j , supponendo costante il valore delle rimanenti variabili; per questo motivo i β_j ($j = 1, \dots, p$) vengono anche chiamati *coefficienti di regressione parziale*.

Ponendo $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)'$ e $\mathbf{x} = (1, x_1, \dots, x_p)'$, il modello (2.45) si può scrivere:

$$Y = \boldsymbol{\beta}'\mathbf{x} + \varepsilon.$$

Nota 2.5 Il modello (2.45) viene anche utilizzato anche una struttura più complessa. Consideriamo ad esempio il modello polinomiale:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

può essere ricondotto al modello (2.45) ponendo $x_1 = x$, $x_2 = x^2$ e $x_3 = x^3$, così che risulta:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

che è evidentemente un modello di regressione lineare. Anche modelli che includono effetti di interazione possono essere ricondotti al modello (2.45). Considerato:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

ponendo $x_3 = x_1 x_2$ possono essere scritti nella forma (2.45) come nel caso precedente. ♣

Se si effettuano n esperimenti, si ottengono $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ valori. Indicando con x_{ji} il valore della i -esima osservazione ($i = 1, \dots, n$) nella j -esima variabile ($j = 1, \dots, p$), possiamo scrivere:

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \epsilon_i = \boldsymbol{\beta}'\mathbf{x}_i + \epsilon_i.$$

Come in precedenza, si assume che la media degli scarti $\bar{\epsilon}$, o equivalentemente la somma degli scarti $\sum_i \epsilon_i$, risulti uguale a zero.

La stima $\mathbf{b} = (b_0, b_1, \dots, b_p)$ viene ottenuta in base al metodo dei minimi quadrati, cioè in modo tale che risulti minima la somma dei quadrati degli scarti:

$$\begin{aligned} \mathbf{b} &:= \arg \min_{\beta_0, \beta_1, \dots, \beta_p \in \mathbb{R}^{p+1}} \mathcal{E}(\boldsymbol{\beta}) = \arg \min_{(\beta_0, \beta_1, \dots, \beta_p) \in \mathbb{R}^{p+1}} \|\mathbf{f}(\boldsymbol{\beta}) - \mathbf{y}\|^2 \\ &= \arg \min_{\beta_0, \beta_1, \dots, \beta_p \in \mathbb{R}^{p+1}} \sum_{i=1}^n \epsilon_i^2 = \arg \min_{\beta_0, \beta_1, \dots, \beta_p \in \mathbb{R}^{p+1}} \sum_{i=1}^n (f(\mathbf{x}_i) - y_i)^2 \\ &= \arg \min_{\beta_0, \beta_1, \dots, \beta_p \in \mathbb{R}^{p+1}} \sum_{i=1}^n (\beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} - y_i)^2. \end{aligned}$$

In questo caso il sistema di equazioni normali (2.12) si scrive:

$$\begin{aligned}
 \sum_{i=1}^n (b_0 + b_1 x_{0i} + \cdots + b_p x_{pi} - y_i) &= 0 \\
 \sum_{i=1}^n x_{1i} (b_0 + b_1 x_{1i} + \cdots + b_p x_{pi} - y_i) &= 0 \\
 \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots &= 0 \\
 \sum_{i=1}^n x_{pi} (b_0 + b_1 x_{1i} + \cdots + b_p x_{pi} - y_i) &= 0
 \end{aligned}$$

da cui si ottiene

$$\begin{aligned}
 nb_0 + b_1 \sum_{i=1}^n x_{1i} + \cdots + b_p \sum_{i=1}^n x_{pi} &= \sum_{i=1}^n y_i \\
 b_0 \sum_{i=1}^n x_{1i} + b_1 \sum_{i=1}^n x_{1i}^2 + b_2 \sum_{i=1}^n x_{8i} x_{4i} + \cdots + \sum_{i=1}^n x_{1i} x_{pi} &= \sum_{i=1}^n y_i x_{1i} \\
 \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots &= \sum_{i=1}^n y_i x_{ji} \\
 b_0 \sum_{i=1}^n x_{pi} + b_1 \sum_{i=1}^n x_{1i} x_{pi} + b_2 \sum_{i=1}^n x_{2i} x_{pi} + \cdots + \sum_{i=1}^n x_{pi}^2 &= \sum_{i=1}^n y_i x_{pi} .
 \end{aligned} \tag{2.46}$$

Dalla prima equazione delle (2.46) si ricava

$$b_0 = \bar{y} - b_1 \bar{x}_1 - \cdots - b_p \bar{x}_p$$

dopo aver ricavato $b_1, \dots, b_p$ dalle restanti equazioni del sistema (2.46).

Indichiamo con $\sigma_{x_j, e} = \text{Cov}(X_j, \mathbf{e})$ la covarianza fra la j -esima variabile ed i residui. Con procedimenti analoghi al caso univariato, si dimostra che

$$\sigma_{x_1 e} = \sigma_{x_2 e} = \cdots = \sigma_{x_p e} = 0 .$$

2.6 Un approccio matriciale

In generale l'equazione di regressione può essere ricavata utilizzando gli strumenti del calcolo matriciale.

2.6.1 La regressione lineare semplice

Introduciamo il vettore delle osservazioni $\mathbf{y}$, la matrice delle variabili indipendenti $\mathbf{X}$, il vettore dei parametri da stimare $\boldsymbol{\beta}$, il vettore degli errori $\boldsymbol{\varepsilon}$ ed il vettore n -dimensionale $\mathbf{1}$.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{2.47}$$

Le equazioni (2.6) possono quindi essere scritte sinteticamente:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} .$$

Si noti che $\mathbf{y}$, $\boldsymbol{\varepsilon}$ e $\mathbf{1}$ sono vettori a n dimensioni, $\mathbf{X}$ è una matrice di ordine $n \times 2$ e $\boldsymbol{\beta}$ è un vettore a due dimensioni.

$$\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \begin{pmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ \beta_0 + \beta_1 x_2 + \varepsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_n + \varepsilon_n \end{pmatrix}$$

Possiamo scrivere:

$$\begin{aligned} y_1^2 + y_2^2 + \dots + y_n^2 &= \mathbf{y}'\mathbf{y} \\ n\bar{y} &= y_1 + y_2 + \dots + y_n = \mathbf{1}'\mathbf{y} \\ n\bar{y}^2 &= \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 = \frac{1}{n} \mathbf{y}'\mathbf{1}\mathbf{1}'\mathbf{y} . \end{aligned}$$

In generale, per la regressione semplice:

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \quad (2.48)$$

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} \quad (2.49)$$

Ciò significa che le equazioni normali (2.11) possono essere scritte come:

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$$

dove

$$\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \quad (2.50)$$

da cui, nell'ipotesi in cui la matrice $\mathbf{X}'\mathbf{X}$ sia non singolare, si ottiene la stima $\mathbf{b}$ di $\boldsymbol{\beta}$ in base al metodo dei minimi quadrati:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \quad (2.51)$$

cioè, nel caso di regressione semplice:

$$\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \frac{1}{n \sum (x_i - \bar{x})^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} .$$

2.6.2 La regressione lineare multipla

In termini formalmente analoghi si affronta il caso generale. Sia assegnata una variabile statistica Y e supponiamo che essa possa essere descritta come combinazione lineare di p variabili statistiche $X_1, \dots, X_p$ più un termine costante β_0 ed una componente casuale ε :

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon .$$

Dato quindi un insieme di n osservazioni di $Y, X_1, \dots, X_p$ scriveremo:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_p x_{1p} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_p x_{2p} + \varepsilon_2 \\ &\dots = \dots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_p x_{np} + \varepsilon_n \end{aligned} \quad (2.52)$$

Analogamente a quanto visto nel paragrafo precedente, possiamo introdurre il vettore delle osservazioni $\mathbf{y}$, la matrice delle variabili indipendenti $\tilde{\mathbf{X}}$, il vettore dei parametri da stimare $\boldsymbol{\beta}$, il vettore degli errori $\boldsymbol{\varepsilon}$ ed il vettore n -dimensionale $\mathbf{1}$.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \tilde{\mathbf{X}} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad (2.53)$$

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad (2.54)$$

In generale, quindi, $\mathbf{y}$, $\boldsymbol{\varepsilon}$ e $\mathbf{1}$ sono vettori a n dimensioni, $\tilde{\mathbf{X}}$ è una matrice di ordine $n \times (p+1)$ e $\boldsymbol{\beta}$ è un vettore a $p+1$ dimensioni. In base a quanto scritto sopra, il sistema (2.52) in termini matriciali viene scritto in maniera formalmente analoga a quanto visto nella sezione precedente per la regressione semplice:

$$\mathbf{Y} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

La stima $\mathbf{b}$ di $\boldsymbol{\beta}$ in base al metodo dei minimi quadrati si ottiene esattamente come in precedenza:

$$\mathbf{b} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{y} \quad (2.55)$$

che risulta formalmente uguale alla (2.51).

2.6.3 Il coefficiente di determinazione multiplo

Analogamente al caso della regressione lineare semplice, consideriamo un modello $y_i = f(\mathbf{x}_i) + \varepsilon_i$ dove $f(\mathbf{x}_1) = b_0 + b_1 x_{1i} + \dots + b_p x_{pi}$, la devianza totale si scompone nella devianza di regressione e nella devianza dell'errore o residua:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - f(x_i))^2 + \sum_{i=1}^n (f(x_i) - \bar{y})^2. \quad (2.56)$$

L'indice:

$$R^2 = \frac{SS_R}{SS_Y} = 1 - \frac{SS_e}{SS_y} \quad (2.57)$$

prende il nome di *coefficiente di determinazione multiplo* e rappresenta la parte della devianza totale di Y che è spiegata (o determinata) dalla supposta relazione lineare con X_1 e X_2 . La radice quadrata di R^2 viene chiamata *coefficiente di correlazione multipla*.

- $R^2 = 1$: quando gli scarti ε_i sono tutti nulli, cioè quando $y_i = f(\mathbf{x}_i)$ per ogni $i = 1, \dots, n$.
- $R^2 = 0$: quando i punti sono massimamente dispersi rispetto al piano di regressione, cioè quando la relazione lineare attraverso cui ci si propone di interpretare il comportamento del fenomeno Y in funzione di $\mathbf{X}$, non fornisce alcun utile informazione al riguardo.

Il coefficiente di determinazione multiplo corretto Il valore di R^2 aumenta sempre (o comunque non decresce) quando si aggiunge un nuovo regressore al modello; ciò non significa che necessariamente il nuovo modello risulta superiore al precedente. Per questa ragione, nella regressione multipla si considera il coefficiente R^2 *corretto* (*adjusted $\bar{R}^2$*) che penalizza l'inclusione di variabili non necessarie nel nostro modello. Tale coefficiente viene definito sostituendo nella (2.57) i valori di SS_e e SS_y con i corrispondenti valori divisi per i corrispondenti *gradi di libertà* che tengono conto del numero delle variabili del modello:

$$\bar{R}^2 = 1 - \frac{SS_e/(n-p-1)}{SS_y/(n-1)} = 1 - \frac{n-1}{n-p-1}(1-R^2). \quad (2.58)$$

Se i valori di R^2 e $\bar{R}^2$ sono molto differenti, allora verosimilmente il modello risulta sovradimensionato, cioè sono stati inclusi termini che non contribuiscono significativamente all'adattamento dei dati.

2.6.4 La matrice di correlazione parziale.

Nel caso generale, possiamo anche considerare la matrice delle correlazioni parziali, i cui elementi sono i coefficienti di correlazione tra due variabili a parità di tutte le altre, [Riz85].

Il generico elemento (i, j) di tale matrice viene indicato con $r_{ij,p}$, dove con p indichiamo l'insieme di tutte le variabili eccetto X_i e X_j . L'espressione generale di $r_{ij,p}$ è:

$$r_{ij,p} := \frac{-R_{ij}}{\sqrt{R_{ii}R_{jj}}} \quad (2.59)$$

dove R_{ij} è il complemento algebrico o aggiunto¹ di r_{ij} nella matrice di correlazione $\mathbf{R}$.

Ad esempio, nel caso di tre variabili, dalla (2.59) si ha:

$$r_{12,3} = \frac{-R_{12}}{\sqrt{R_{11}R_{22}}} = \frac{-(-) \begin{vmatrix} r_{12} & r_{23} \\ r_{13} & 1 \end{vmatrix}}{\sqrt{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix} \begin{vmatrix} 1 & r_{13} \\ r_{13} & 1 \end{vmatrix}}} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{23}^2)(1 - r_{13}^2)}}$$

che è proprio la (2.42).

Ritorniamo alla matrice dei dati dei vini di Bordeaux e calcoliamo $R_{12,34}$. Dalla (1.14) si ha preliminarmente:

$$\begin{aligned} R_{12} &= (-1)^{1+2} \begin{vmatrix} r_{21} & r_{23} & r_{24} \\ r_{31} & r_{33} & r_{34} \\ r_{41} & r_{43} & r_{44} \end{vmatrix} = - \begin{vmatrix} 0.7147 & 0.6459 & -0.4750 \\ 0.8614 & 1.0000 & -0.4011 \\ -0.4118 & -0.4011 & 1.0000 \end{vmatrix} = -0.1185 \\ R_{11} &= (-1)^{1+1} \begin{vmatrix} r_{22} & r_{23} & r_{24} \\ r_{32} & r_{33} & r_{34} \\ r_{42} & r_{43} & r_{44} \end{vmatrix} = \begin{vmatrix} 1.0000 & 0.6459 & -0.4750 \\ 0.6459 & 1.0000 & -0.4011 \\ -0.4750 & -0.4011 & 1.0000 \end{vmatrix} = 0.4424 \\ R_{22} &= (-1)^{2+2} \begin{vmatrix} r_{11} & r_{13} & r_{14} \\ r_{31} & r_{33} & r_{34} \\ r_{41} & r_{43} & r_{44} \end{vmatrix} = \begin{vmatrix} 1.0000 & 0.8614 & -0.4118 \\ 0.8614 & 1.0000 & -0.4011 \\ -0.4118 & -0.4011 & 1.0000 \end{vmatrix} = 0.2121 \end{aligned}$$

e quindi:

$$r_{12,34} = \frac{-R_{12}}{\sqrt{R_{11}R_{22}}} = \frac{0.1185}{\sqrt{0.4424 \cdot 0.2121}} = 0.3870$$

¹**Definizione.** Sia $\mathbf{A} = (a_{ij})$ una matrice quadrata di ordine $n \times n$. Si definisce *minore* dell'elemento a_{ij} il determinante della sottomatrice ottenuta da $\mathbf{A}$ sopprimendo la i -esima riga e la j -esima colonna di $\mathbf{A}$. Si definisce *complemento algebrico* o *aggiunto* dell'elemento a_{ij} , e si denota col simbolo A_{ij} , il prodotto del minore di a_{ij} per $(-1)^{i+j}$.

Ad esempio, assegnata la matrice $\mathbf{A}(3 \times 3)$:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

i minori degli elementi a_{11} e a_{12} sono dati rispettivamente da:

$$\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}, \quad \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix}$$

ed i complementi algebrici A_{11} e A_{12} sono dati rispettivamente da:

$$A_{11} = \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}, \quad A_{12} = - \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix}$$

Effettuando tutti i calcoli, si ottiene la seguente matrice di correlazione parziale:

$$\mathbf{R}^* = \begin{pmatrix} 1.0000 & 0.3870 & 0.7447 & -0.0188 \\ 0.3870 & 1.0000 & 0.0595 & -0.2774 \\ 0.7447 & 0.0595 & 1.0000 & -0.0797 \\ -0.0188 & -0.2774 & -0.0797 & 1.0000 \end{pmatrix}. \quad (2.60)$$

Dal confronto con la (1.14) si può notare l'influenza delle altre variabili nella correlazione bivariata.

2.7 Analisi dei residui

Le principali ipotesi che abbiamo fatto per la costruzione dei modelli di regressione sono state le seguenti:

1. La relazione di dipendenza fra la variabile risposta Y e quella indipendente X è (con buona approssimazione) di tipo lineare;
2. Gli errori ε hanno valore atteso uguale a zero e varianza σ^2 costante;
3. Gli errori ε_i sono incorrelati;
4. Gli errori seguono una distribuzione normale.

Tali ipotesi vanno sempre verificate, una volta preso in considerazione un certo modello. Forti violazioni di tali ipotesi, infatti, possono rendere instabile un modello nel senso che un altro campione potrebbe condurre ad un modello molto diverso. Usualmente non è possibile rilevare violazioni delle ipotesi del modello esaminando unicamente statistiche quali l'indice R^2 , che sono invece misure "globali" delle proprietà del modello, e che quindi non assicurano l'adeguatezza del modello ai dati.

In precedenza abbiamo definito il residuo:

$$e_i := f(x_i) - y_i \quad i = 1, \dots, n \quad (2.61)$$

dove y_i è il valore osservato della variabile risposta e $f(x_i)$ è il corrispondente valore stimato dal modello. Poichè un residuo può essere visto come la distanza fra il valore stimato ed il dato, è una misura della variabilità non spiegata dal modello di regressione. E' anche opportuno pensare ai residui e_i come realizzazioni della variabile errore ε_i . Pertanto qualunque allontanamento dalle ipotesi soggiacenti sugli errori verrà evidenziata dall'analisi dei residui. In effetti, l'analisi dei residui è un metodo efficace per scoprire diversi tipi di difetti nel modello in esame.

2.7.1 Rappresentazioni grafiche

La rappresentazioni grafica mediante nube di punti dei residui e_i rispetto ai valori stimati $f(x_i)$ risulta utile per individuare vari tipi di non adeguatezza del modello. Si noti che, per la (2.19), i residui ed i valori stimati sono non correlati (a differenza delle distribuzioni dei residui e_i e dei valori osservati y_i che invece risultano fra loro correlate).

I residui e_i della regressione multipla svolgono un ruolo importante nella valutazione della bontà del modello, così come accade nella regressione lineare semplice. Usualmente si considerano le seguenti rappresentazioni grafiche dei residui:

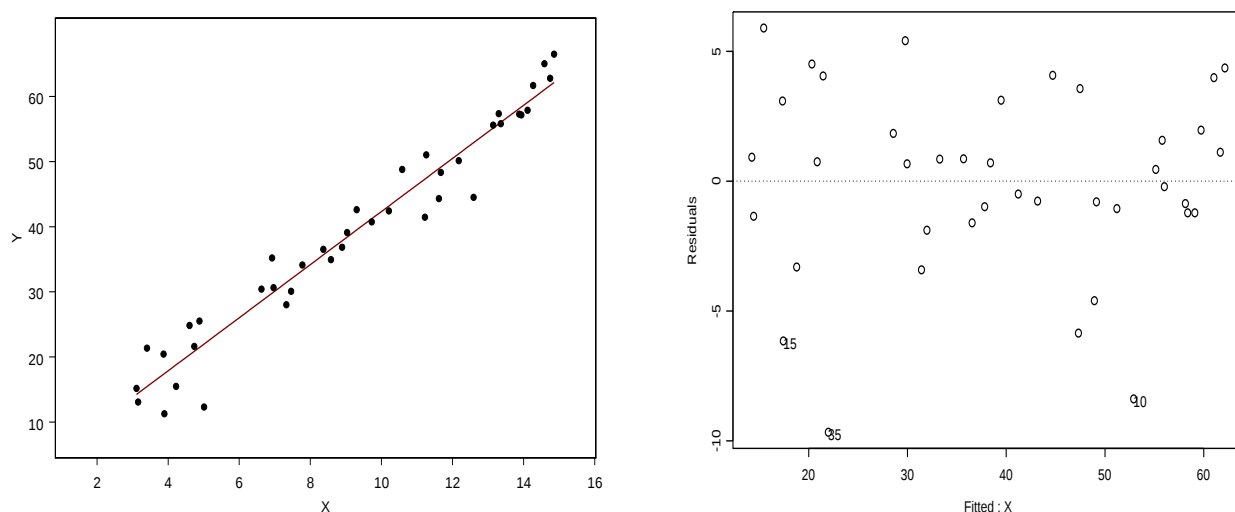


Figura 2.2: *Nube di punti, modello di regressione e diagramma dei corrispondenti residui (in condizioni di omoschedasticità)*

1. *normal probability plot*;
2. residui rispetto a ciascun regressore X_j , $j = 1, \dots, p$;
3. residui rispetto al modello adattato $f(x_i)$, $i = 1, \dots, n$;
4. residui in sequenza temporale (ove nota).

Come rilevato nel paragrafo 2.7, questi grafici rivestono particolare importanza per rilevare l'allontanamento dall'ipotesi di distribuzione normale, dati anomali, eteroschedasticità e per valutare l'errata specificazione funzionale di un regressore. Possono essere considerati sia i residui e_i che sue trasformazioni di scala d_i o r_i .

In Figura 2.2 viene rappresentata una nube di punti, il corrispondente modello di regressione ed il diagramma dei residui (rispetto ai valori previsti dal modello) in condizioni di omoschedasticità; si noti la presenza di tre valori anomali (vedi punto successivo).

In Figura 2.3 forniamo un esempio di adattamento lineare in condizioni di eteroschedasticità; anche in questo caso si noti la presenza di tre valori anomali.

Nell'ambito della regressione multipla, vi sono altre tecniche di analisi dei residui che qui non affrontiamo.

2.7.2 Valori anomali (outliers)

Valori anomali (*outliers*) riguardano valori della risposta che, in qualche modo, si discostano fortemente dal resto della distribuzione. Residui che, in valore assoluto, presentano valori notevolmente più grandi degli altri – diciamo pari a tre, quattro volte allo scarto quadratico medio della loro distribuzione – derivano potenzialmente valori anomali. A seconda della loro collocazione rispetto all'asse x (variabile indipendente), i valori anomali possono avere effetti

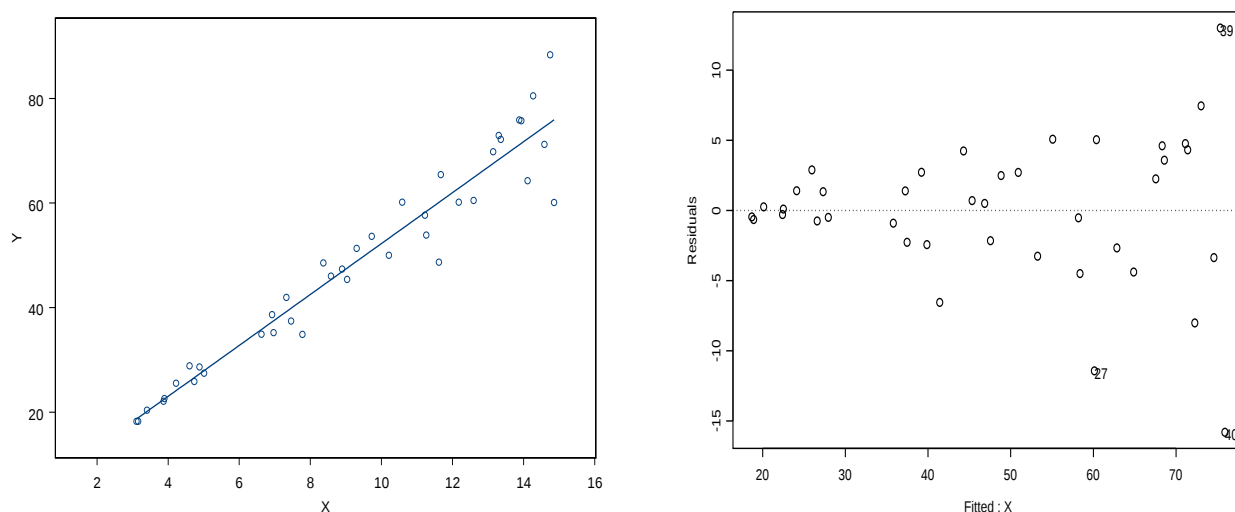


Figura 2.3: *Nube di punti, modello di regressione e diagramma dei corrispondenti residui (in condizioni di eteroschedasticità)*

rilevanti sul modello di regressione. I valori estremi possono essere individuati utilizzando residui standardizzati o studentizzati.

In presenza di valori anomali, è importante valutare se vi sono delle ragioni che possono spiegare tali valori. A volte valori anomali si verificano come risultato di eventi inusuali ma comunque ben spiegabili: analisi o misurazioni errate, guasti negli strumenti di rilevazione, registrazioni errate dei dati. In questo caso tali valori dovrebbero essere corretti o addirittura eliminati dalla distribuzione. Come detto in precedenza, la presenza di valori anomali ha effetti rilevanti sul modello di regressione, pertanto la loro rimozione porterebbe ad un migliore adattamento sul resto della distribuzione. Va comunque sottolineato che sono necessarie forti motivazioni "non statistiche" per la rimozione di valori anomali dall'insieme di dati.

A volte valori anomali, benchè rari, sono perfettamente plausibili. Pertanto l'eliminazione di tali punti per "migliorare l'adattamento del modello" può risultare pericolosa in quanto potrebbe dare la falsa impressione di un miglioramento nella precisione della stima o della previsione. Addirittura, un valore estremo potrebbe risultare molto importante in quanto evidenzia aspetti rilevanti della distribuzione o anche alcune inadeguatezze del modello, come ad esempio un cattivo adattamento al di fuori di un certo intervallo di valori del regressore.

2.8 Trasformazioni di variabili

Finora abbiamo considerato modelli di regressione del tipo $f(x) = \beta_0 + \beta_1 x$; tuttavia la dipendenza fra la risposta Y e la variabile X potrebbe non essere di tipo lineare, per considerazioni di tipo teorico o empirico. In alcuni casi, una funzione non lineare può essere ricondotta in una relazione lineare mediante opportune trasformazioni. Questi modelli non lineari vengono chiamati *intrinsecamente lineari*.

Consideriamo ad esempio una relazione del tipo:

$$Y = \beta_0 e^{\beta_1 x} \varepsilon .$$

Tale funzione è intrinsecamente lineare poichè può essere ricondotta ad una relazione lineare considerando una trasformazione logaritmica:

$$\ln Y = \ln \beta_0 + \beta_1 x + \ln \varepsilon .$$

Questa trasformazione richiede ovviamente che la trasformazione de termine di errore $\varepsilon' = \ln \varepsilon$ soddisfi le ipotesi richieste per i modelli di regressione: gli errori ε'_i si assumono indipendenti, identicamente distribuiti (i.i.d.) con media uguale a zero e varianza σ^2 costante; ulteriormente si può assumere che $\varepsilon' = \ln \varepsilon$ segua una distribuzione normale. L'analisi dei residui va quindi condotta su ε' .

Alcuni tipi di trasformazioni reciproche possono essere utili. Per esempio, il modello:

$$Y = \beta_0 + \beta_1 \frac{1}{x} + \varepsilon$$

possono essere linearizzate utilizzando la trasformazione $x' = 1/x$, da cui segue:

$$Y = \beta_0 + \beta_1 x' + \varepsilon .$$

Altri esempi di modello che possono essere linearizzati mediante trasformazioni reciproche sono:

$$\frac{1}{Y} = \beta_0 + \beta_1 x + \varepsilon$$

$$Y = \frac{x}{\beta_0 x - \beta_1 + \varepsilon} .$$

Quest'ultimo viene linearizzato ponendo $y' = 1/y$ e $x' = 1/x$ ottenendo:

$$Y' = \beta_0 - \beta_1 x' + \varepsilon x'$$

2.8.1 Trasformazioni stabilizzatrici della varianza

Nel modello di regressione si assume usualmente che gli errori ε_i presentino varianza costante su σ^2 . Può accadere che la risposta Y segua una distribuzione di probabilità la cui varianza è funzionalmente legata al valore atteso (Y). Ad esempio, se Y segue una distribuzione di Poisson, allora valore atteso e varianza coincidono; poichè il valor medio di Y è funzionalmente legato alla variabile X , allora la varianza di Y sarà proporzionale a X . In questi casi risultano utili trasformazioni stabilizzatrici della varianza. Se Y segue una distribuzione di Poisson, allora bisogna considerare il modello di regressione basato su $Y' = \sqrt{Y}$ rispetto a x .

Alcune trasformazioni stabilizzatrici della varianza di maggiore uso sono riassunte nella tabella qui di seguito:

Relazione di σ^2 a $\mathbb{E}(Y)$	Trasformazione
$\sigma^2 \propto \text{costante}$	$Y' = Y$
$\sigma^2 \propto \mathbb{E}(Y)$	$Y' = \sqrt{Y}$
$\sigma^2 \propto [\mathbb{E}(Y)]^2$	$Y' = \ln Y$
$\sigma^2 \propto [\mathbb{E}(Y)]^3$	$Y' = Y^{-1/2}$
$\sigma^2 \propto [\mathbb{E}(Y)]^4$	$Y' = Y^{-1}$

La forza di una trasformazione dipende dalla quantità di curvatura che essa apporta. Generalmente una trasformazione abbastanza moderata (come la radice quadrata) applicata su un intervallo di valori relativamente stretto (ad esempio y_{max}/y_{min} fra 2 e 3) ha poco effetto; al contrario una forte trasformazione su un ampio intervallo di valori, può avere un effetto molto fuorviante.

L'esperienza e considerazioni teoriche possono fornire criteri utili per la scelta della trasformazione. In molti casi, comunque, non vi sono ragioni a priori tali da far ritenere che la varianza di errore sia non costante; l'indicazione è quella di valutare opportunamente in base al diagramma a dispersione dei residui.

E' importante rilevare e correggere varianze di errori non costanti. Se questo problema non viene eliminato, benchè gli stimatori dei minimi quadrati risultano ancora non distorti, essi non avranno la proprietà di minima varianza. Ciò significa che i coefficienti di regressione avranno errori standard maggiori del necessario. L'effetto della trasformazione è usualmente quello di fornire stime più precise dei parametri del modello e di aumentare la sensitività dei test statistici.

Va rilevato che quando viene trasformata la variabile risposta, i valori stimati vengono ottenuti nella nuova scala; è quindi opportuno trasformare i valori ottenuti nella scala originaria.