Capitolo 2

La regressione lineare semplice

2.1 Introduzione

Siano assegnate $p+1$ variabili $X_1, X_2, \ldots, X_p, Y$ concernenti n osservazioni di p caratteri da una popolazione assegnata. I dati rilevati possono essere riportati come segue:

$$\begin{array}{cccccc} Y & & X_1 & X_2 & \cdots & X_p \\ \hline y_1 & & x_{11} & x_{21} & \cdots & x_{p1} \\ y_2 & & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & & \vdots & & \vdots & \vdots \\ y_h & & x_{1h} & x_{2h} & \cdots & x_{ph} \\ \vdots & & \vdots & & \vdots & \vdots \\ y_n & & x_{1n} & x_{2n} & \cdots & x_{pn} \end{array} .$$

Si supponga che la variabile Y "dipenda" congiuntamente dalle altre variabili, cioè, in altre parole, che i valori assunti da $X_1, X_2, \ldots, X_p$ influenzino (o sembrino influenzare) il valore della variabile Y. Il problema che qui intendiamo esaminare concerne lo studio di tale relazione. 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, almeno in un opportuno sottoinsieme di $\mathbb{R}^{p+1}$ mediante funzioni matematiche semplici (ad esempio dei polinomi) dipendenti da opportune variabili. Esaminando tale funzione possiamo cercare di comprendere meglio sia la relazione fra le variabili sia il contributo delle variabili, considerate separatamente o congiuntamente.

Facciamo rilevare che anche quando non esiste alcune 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. Nel seguito considereremo il caso in cui la relazione fra le variabili sia lineare in alcuni parametri incogniti. Tali parametri vengono *stimati*, sotto opportune ipotesi, in base ad un assegnato insieme di dati. In genere si considerano due tipi di variabili: le *variabili indipendenti* o *esplicative* o *predittori*, e le *variabili dipendenti* o *risposte*. Le prime, usualmente indicate con X, 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).

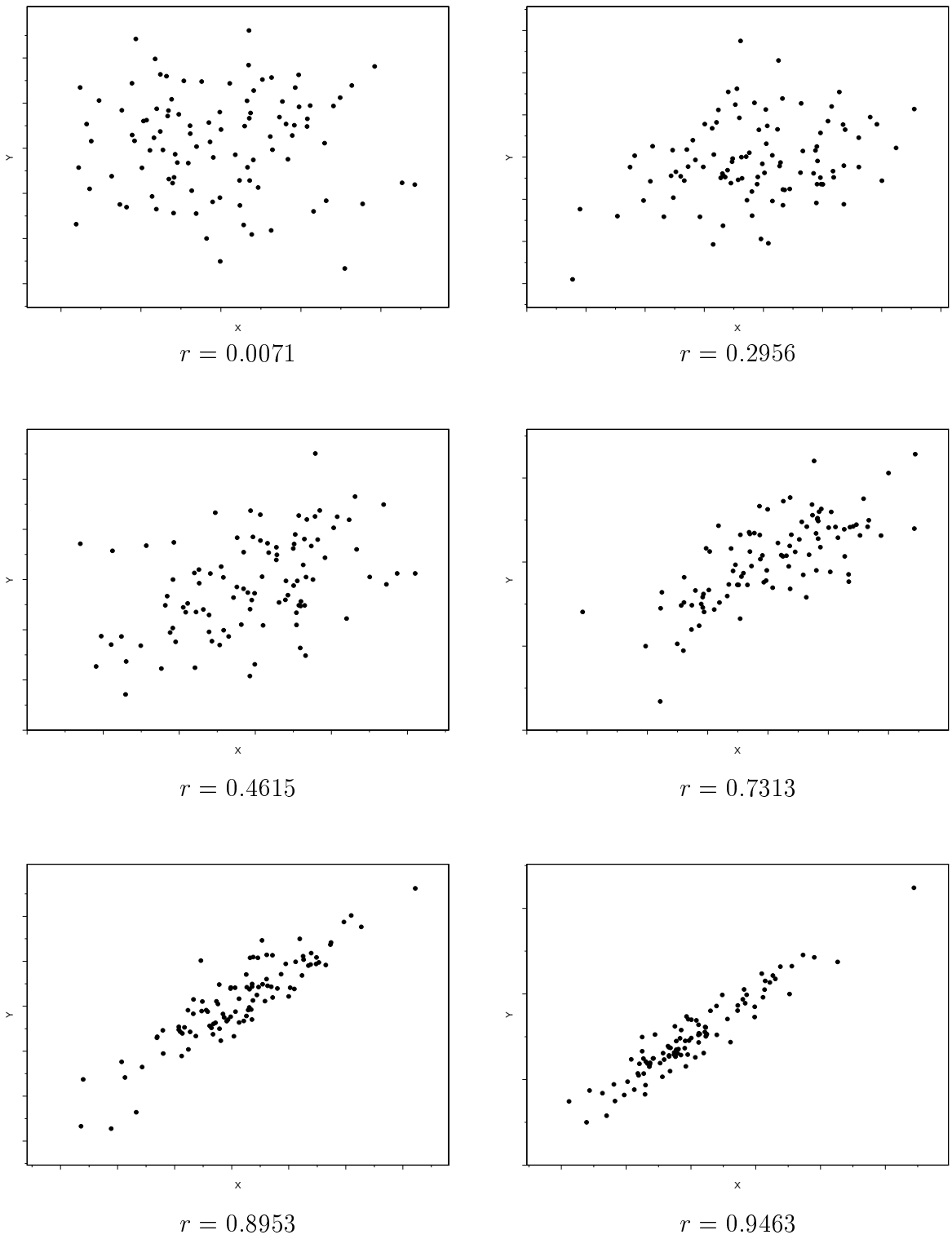
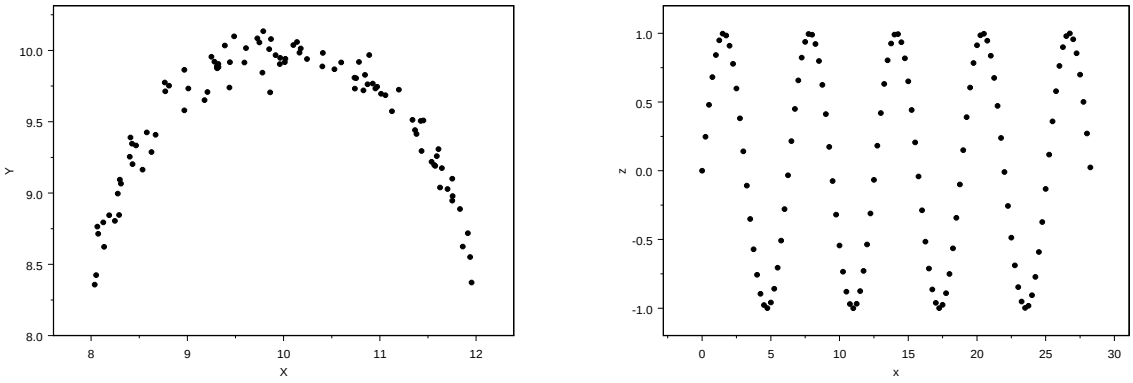


Figura 2.1: Esempi di diagramma a dispersione e corrispondenti valori del coefficiente di correlazione per alcune distribuzioni doppie.



$r = 0.0328$ $r = 0.0004$

Figura 2.2: Esempi di diagramma a dispersione e corrispondenti valori del coefficiente di correlazione per alcune distribuzioni doppie: si noti l'evidente relazione fra i valori di (X, Y).

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. Nella costruzione di un modello di regressione la fase esplorativa basata su tecniche grafiche riveste particolare importanza. Esplorare i dati significa descrivere in forma sintetica le informazioni raccolte al fine di evidenziare le strutture di relazione implicite che li percorrono e permettere così di proporre un modello interpretativo del fenomeno reale in esame. Le procedure esplorative sono euristiche in quanto hanno carattere intuitivo, analogico, previsionale, e danno risultati che dovranno essere in un secondo tempo controllati e convalidati per via rigorosa. Fra le tecniche esplorative, particolare importanza rivestono le rappresentazioni grafiche.[1] In Figura 2.1 riportiamo i diagrammi a dispersione di alcune distribuzioni doppie per valori crescenti del coefficiente di correlazione.

Ovviamente bisogna rilevare che l'assenza di correlazione lineare non indica assenza di relazione fra i dati, come evidenziato ad esempio in Figura 2.2. Il coefficiente di correlazione r misura infatti l'intensità della linearità della relazione fra X e Y; pertanto la relazione fra le due variabili potrebbe essere di altra natura ed in questo caso non verrebbe evidenziata da r direttamente sulle variabili X, Y.

Infine un altro aspetto grafico da considerare, al fine di evidenziare possibili relazioni, concerne la scelta della scala. In Figura 2.3 mostriamo due diagrammi a dispersione concernenti lo stesso insieme di dati, cambia unicamente il rapporto fra le scale dell'asse ordinate rispetto all'asse delle ascisse.

[1] L'approccio complementare a quello esplorativo `e detto *confermativo* e consiste nella verifi ca di alcune ipotesi formulate prima della rilevazione dei dati ed analizzate sulla base del metodo statistico-inferenziale. La scelta tra l'approccio confermativo e quello esplorativo nell'analisi dei dati dipende dagli obiettivi, dal grado di conoscenza sul modo di distribuirsi delle variabili, dalla stabilit`a delle distribuzioni nel tempo e nello spazio, dalla possibilit`a di controllare sperimentalmente l'osservazione dei fenomeni. Spesso si procede alternando l'esplorazione alla conferma di ipotesi sulle regolarit`a presenti nei dati.

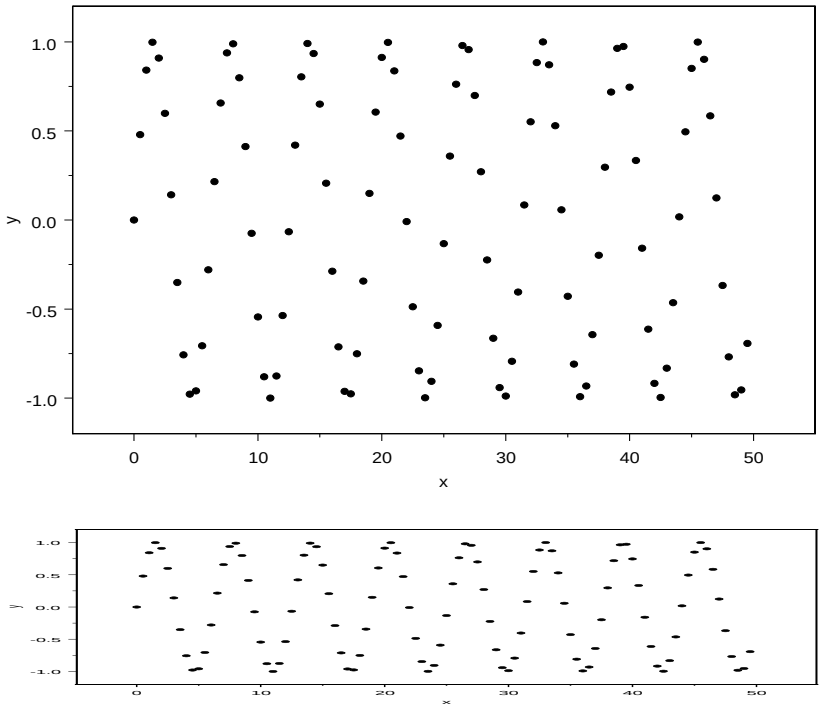


Figura 2.3: Esempio di diagrammi a dispersione per uno stesso insieme di dati.

2.2 Il modello lineare

Il caso più semplice concerne lo studio della relazione fra la variabile risposta Y e una variabile esplicativa X. Introduciamo il problema con un esempio, considerando la seguente tabella che riporta i valori della temperatura esterna media (in gradi centigradi) e del consumo di gas (in m^3) in un'area servita dalla *Norgas Distributing Company* rilevati con cadenza settimanale nel periodo Ottobre - Novembre 1996[2]:

Temperatura (in °C)	7.8	6.2	10.1	1.2	4.3	9.4	-2.0	3.2
Consumo Gas (in m^3)	16733	16596	14666	22620	20199	14848	24.086	21.748

Tabella 2.1: Dati: *norgas.dat*

I dati sono rappresentati graficamente in Figura 2.4. In questo caso si intende studiare come il consumo di gas risulta influenzato dai valori della temperatura esterna. Chiaramente quest'ultima è la variabile indipendente (denotata con X) ed il consumo di gas è la variabile dipendente (denotata con Y).

In generale si assume supporremo che il legame funzionale fra la variabile risposta Y ed il generico valore x della variabile X possa essere descritto da una relazione del tipo:

$$Y = \phi(x) + \varepsilon\,, \tag{2.1}$$

[2]Fonte: Tryfos P., *Methods for Business Analysis and Forecasting: Text and Cases*, John Wiley & Sons, New York, 1998.

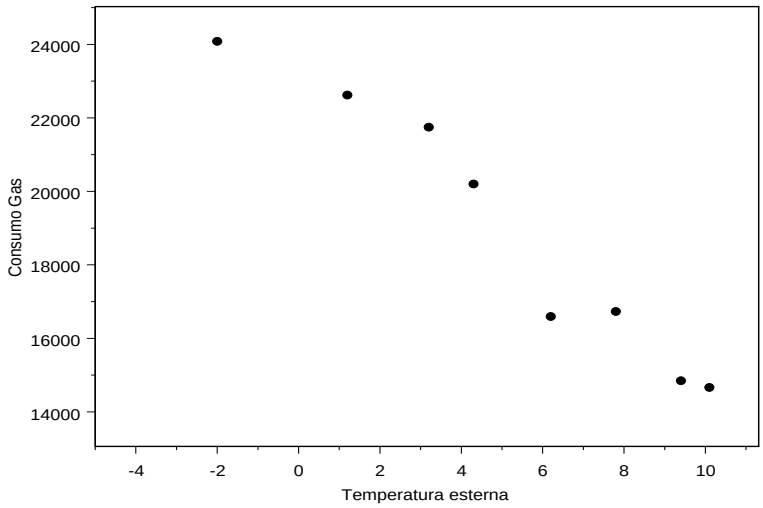


Figura 2.4: *Diagramma a dispersione Consumo Gas e Temperatura esterna.*

dove nel caso più semplice si assume che $\phi(x) = \beta_0 + \beta_1 x$, ottenendo quindi il modello:

$$Y = \beta_0 + \beta_1 x + \varepsilon, \tag{2.2}$$

che viene chiamato *modello di regressione lineare semplice*; più in generale – nel caso di p variabili esplicative $X_1, \ldots, X_p$ – si ipotizza una relazione del tipo:

$$Y = \beta_0 + \beta_1 x + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon = \phi(\mathbf{x}) + \varepsilon\,, \tag{2.3}$$

dove $\phi(\mathbf{x}) = \beta_0 + \beta_1 x + \beta_2 x_2 + \cdots + \beta_p x_p$, che viene chiamato *modello di regressione lineare multipla*. Si noti che il modello $\phi(x)$ nella (2.2) descrive una retta in $\mathbb{R}^2$; mentre il modello $\phi(\mathbf{x})$ nella (2.3) descrive un iperpiano in $\mathbb{R}^{p+1}$.

Dalle relazioni precedenti si vede che la variabile aleatoria (risposta) Y può quindi essere riguardata come la somma di una componente *strutturale* o *deterministica* $\phi(x)$ (o $\phi(\mathbf{x})$, a seconda dei casi) e di una componente casuale di errore ε; le quantità $\beta_0, \beta_1, \ldots, \beta_p$ sono i *parametri* del modello. Infine per la variabile aleatoria ε si assumono le seguenti ipotesi:

$$\mathbb{E}(\varepsilon) = 0 \qquad \text{Var}(\varepsilon) = \sigma^2\,, \tag{2.4}$$

in particolare si assume che la distribuzione di ε è indipendente da x. L'ipotesi $\text{Var}(\varepsilon) = \sigma^2$ viene chiamata ipotesi di *omoschedasticità*; nel caso contrario in cui la varianza di ε dipende da x, si è in condizioni di *eteroschedasticità*. Pertanto, in base alle ipotesi (2.4), la media e la varianza di Y condizionata rispetto a x si scrivono:

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

L'equazione (2.2) costituisce il *modello* della relazione funzionale fra Y e X che si suppone essere valido; in un secondo momento ci si porrà il problema di valutare la validità di tale assunzione.

La componente casuale ε – che, come già osservato, viene chiamata *termine di errore* (che non vuol dire affatto "sbaglio") – comprende tutti i vari fattori che influenzano la variabile risposta Y ma che non sono osservati o non possono essere controllati. Ad esempio, nel controllo statistico di processi di produzione, la qualità del prodotto finale può essere influenzata da fattori non osservabili quali ad esempio le variazioni di temperatura/umidità dell'ambiente o differenze nel lavoro degli operai; nel caso di modelli basati sull'analisi di dati economici, variabili non osservate includono, ad esempio, fenomeni legati al mercato nero o quantità che sono inerentemente difficili da misurare come la produttività di un software. Pertanto la conoscenza delle grandezze in ingresso $\mathbf{x}$ non specifica univocamente i valori di risposta Y del sistema.

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 modelli del tipo:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2 X_3 + \varepsilon$$

sono modello lineari (rispetto ai parametri $\beta_0, \beta 1, \beta_2$).

Al contrario modelli del tipo:

$$Y = \beta_0 + \cos(\beta_1 + \beta_2 X) + \varepsilon$$
$$Y = \beta_1 \left[\left(\frac{\beta_3}{\beta_3 - \beta_4} \right) \mathrm{e}^{-\beta_4 X} + \left(\frac{\beta_1}{\beta_2} - \frac{\beta_3}{\beta_3 - \beta_4} \right) \mathrm{e}^{-\beta_3 X} \right]^{-1}$$

sono chiaramente non lineari.

Nota 2.2 Val la pena sottolineare che a volte l'esistenza di una dipendenza statistica fra variabili di ingresso e variabili di uscita viene (erroneamente) utilizzata per dimostrare una relazione di causa ed effetto fra tali variabili. A tal riguardo bisogna sottolineare che la *causalità* non può essere dimostrata unicamente in base all'analisi dei dati; al contrario essa deve essere assunta o dimostrata mediante argomenti esterni all'analisi statistica. In altre parole, una relazione di causa ed effetto fra variabili diverse può essere supportata ma mai dimostrata in base agli strumenti statistici. In generale si può assumere una dipendenza statistica fra variabili di ingresso e di uscita in una delle seguenti situazioni (o loro combinazione):

- le variabili di uscita dipendono causalmente da quelle osservate in ingresso;
- le variabili di ingresso dipendono causalmente da quelle di uscita (si pensi a sistemi di diagnostica in cui i sintomi sono gli ingressi e le cause le uscite);
- entrambe le variabili di ingresso e di uscita dipendono a loro volta da altri fattori non osservati;
- la correlazione ingresso-uscita non è causale.

Tuttavia, ciascuna di queste possibilità può essere specificata solo mediante argomenti esterni all'analisi statistica. In questo contesto, sono da criticare alcuni semplicistici approcci presenti nell'ambito del *data mining* e *knowledge discovery* che propongono l'uso di strumenti automatici per evidenziare associazioni significative (nel senso che ha un significato e quindi

un'interpretazione di tipo causale) fra variabili di ingresso e di uscita in insiemi di dati numerosi (*large data sets*). Una dipendenza significativa può essere estratta dai dati solo se la formulazione del problema riflette in qualche modo una conoscenza *a priori* circa il dominio specifico dell'applicazione.

2.3 La regressione lineare semplice

Il problema che ci poniamo adesso riguarda la determinazione pratica della relazione fra la variabile risposta Y ed una sola variabile esplicativa X, in accordo al modello lineare semplice:

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

In linea di principio, la determinazione di β_0, β_1 richiederebbe la conoscenza della distribuzione congiunta di (X, Y) sull'intera popolazione Ω cui si riferiscono le due variabili. Usualmente, invece, si dispone dei valori di (X, Y) rilevati su un campione casuale di n unità statistiche (x_i, y_i), $i = 1, \ldots, n$ di Ω, dove si assume che le x_i ($i = 1, \ldots, n$) non siano tutte uguali fra loro. Ne segue che i valori di β_0, β_1 possono essere solamente *stimati* sulla base dei valori rilevati nel campione a nostra disposizione; tali stime dipenderanno necessariamente dall'insieme delle coppie (x_i, y_i), $i = 1, \ldots, n$ dei valori rilevati per (X, Y).

In particolare, in base alla (2.2), il valore y_i costituisce il *dato empirico* rilevato per la variabile risposta Y in corrispondenza di x_i viene considerato come una realizzazione della variabile aleatoria:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad\qquad i = 1, \ldots, n, \tag{2.6}$$

dove $\beta_0 + \beta_1 x_i$ è la *componente strutturale* o *deterministica* e ε_i ($i = 1, \ldots, n$) sono variabili aleatorie per le quali si assumono le seguenti ipotesi:

$$\mathbb{E}(\varepsilon_i) = 0 \qquad 1 \le i \le n \tag{2.7}$$

$$\text{Var}(\varepsilon_i) = \sigma^2 \qquad 1 \le i \le n \tag{2.8}$$

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \qquad 1 \le i \ne j \le n \, . \tag{2.9}$$

Le prime due ipotesi sono congruenti con la (2.4), in particolare la (2.8) implica l'ipotesi di omoschedasticità; l'ipotesi (2.9) implica che le variabili $Y_1, Y_2, \ldots, Y_n$ siano incorrelate. 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 le ε_i seguono una distribuzione normale: $\varepsilon_i \sim N(0, \sigma^2)$ per ogni $i = 1, 2, \ldots, n$.

Le stime b_0, b_1 di β_0, β_1 vengono ottenute in base al metodo dei minimi quadrati, descritto successivamente, utilizzando le coppie di osservazioni (x_i, y_i), $i = 1, \ldots, n$. Si assume quindi che la relazione funzionale fra le coppie di valori osservati (x_i, y_i) possa essere scritta come segue:

$$y_i = b_0 + b_1 x_i + e_i = f(x) + e_i \tag{2.10}$$

per opportuni valori di b_0, b_1. Le quantità $e_i := y_i - f(x_i)$, per $i = 1, \ldots, n$ – che vengono chiamate *residui* – costituiscono gli scarti fra i valori osservati y_i ed i corrispondenti valori teorici $f(x_i) = b_0 + b_1 x_i$ ottenuti dal modello in corrispondenza di x_i. La funzione $f(x)$ costituisce la *stima della dipendenza $\phi(x)$ di Y da X sulla base dei dati empirici* (x_i, y_i), $i = 1, \ldots, n$. Il modello $f(x)$ viene anche denotato con $\hat{y}$ per indicare che $f(x)$ fornisce una stima della variabile Y.

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.

Il metodo dei minimi quadrati. Il metodo di stima dei minimi quadrati si applica soltanto ai parametri di interesse β_0, β_1 e spesso si utilizza in situazioni in cui risulta difficile poter assumere ipotesi ulteriori rispetto alle (2.7), (2.8) e (2.9).

Poniamo $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ e $\phi(\beta_0, \beta_1) = \beta_0 + \beta_1 x_i$ e scriveremo $f_i(b_0, b_1)$, $i = 1, \ldots, n$ se vogliamo evidenziare la dipendenza funzionale rispettivamente da (β_0, β_1) o da (b_0, b_1).

Il *metodo dei minimi quadrati* fornisce, quale stima di $\boldsymbol{\beta} = (\beta_0, \beta_1)'$, quel vettore $\mathbf{b} = (b_0, b_1)'$ che minimizza la cosidetta *funzione di errore*:

$$\mathcal{E}(\beta_0, \beta_1) := \sum_{i=1}^{n} (y_i - f_i(\beta_0, \beta_1))^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{2.11}$$

cioè

$$\mathbf{b} := \underset{(\beta_0,\beta_1)\in\mathbb{R}}{\arg\min}\, \mathcal{E}(\beta_0, \beta_1) = \underset{(\beta_0,\beta_1)\in\mathbb{R}}{\arg\min} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \ .$$

Ne segue che la stima dei minimi quadrati $\mathbf{b}$ è la soluzione del seguente sistema di equazioni:

$$\frac{\partial}{\partial b_j} \sum_{i=1}^{n} [y_i - f_i(b_0, b_1)]^2 = 0 \qquad j = 0, 1 \tag{2.12}$$

che vengono chiamate *equazioni normali*, cioè dal sistema di equazioni:

$$-2 \sum_{i=1}^{n} [y_i - f_i(b_0, b_1)] \frac{\partial f_i(b_0, b_1)}{\partial b_j} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) \frac{\partial (y_i - b_0 - b_1 x_i)}{\partial b_j} = 0 \qquad j = 0, 1 \ , \tag{2.13}$$

equivalente al sistema:

$$\begin{aligned} \sum_{i=1}^{n} (y_i - b_0 - b_1\, x_i) &= 0 \\ \sum_{i=1}^{n} x_i (y_i - b_0 - b_1\, x_i) &= 0 \ , \end{aligned} \tag{2.14}$$

da cui, risolvendo, si ottiene:

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\mathrm{S}_{xy}}{\mathrm{SS}_x} = \frac{\sigma_{xy}}{\sigma_x^2} \tag{2.15}$$

$$b_0 = \bar{y} - b_1\, \bar{x} \tag{2.16}$$

dove

$$\mathrm{S}_{xy} := \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \qquad \mathrm{SS}_x := \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad \text{e}$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \qquad \sigma_x^2 = \frac{\mathrm{SS}_x}{n} \qquad \sigma_{xy} = \frac{\mathrm{S}_{xy}}{n} \ .$$

Si noti che SS_x è la devianza di X e S_{xy} è la codevianza di X, Y.

La retta di equazione $f(x) = b_0 + b_1 x$ viene chiamata *retta di regressione* di $y_1, y_2, \ldots, y_n$ su $x_1, x_2, \ldots, x_n$. Si noti che b_1 misura la variazione della funzione f in corrispondenza di una variazione unitaria di x.

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 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Alcune conseguenze dell scelte effettuate riguardano la media dei residui e la covarianza $\sigma_{\varepsilon x}$ fra residui e e la variabile indipedente X. Considerati i residui:

$$e_i := y_i - (b_0 + b_1 x_i) \tag{2.17}$$

si ha infatti in base alla (2.16):

$$\begin{aligned} \bar{e} &= \frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n}(b_0 + b_1 x_i) \\ &= \bar{y} - b_0 - b_1 \bar{x} = 0 \, . \end{aligned} \tag{2.18}$$

Si ha inoltre, essendo $\bar{e} = 0$ e $\sum_i e_i = 0$ per le relazioni sopra ricavate:

$$\sigma_{xe} = \frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})(x_i - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n} e_i x_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) x_i = 0 \tag{2.19}$$

in quanto coincide con la seconda equazione normale (2.14).

Esempio 2.3 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'invecchiamento 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 2.2. Il grafico in Figura 2.5 riporta il diagramma a dispersione (scatter plot) di y_i rispetto a x_i.

Considerato un modello lineare del tipo:

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

in base al metodo dei minimi quadrati, in base alle (2.16) e (2.15) si ottengono le seguenti stime dei parametri:

$$b_0 = 2627.82 \qquad b_1 = -37.15 \ ,$$

e quindi il modello:

$$f(x) = 2627.82 - 37.15x \, . \tag{2.20}$$

Possiamo interpretare il valore $b_1 = -37.15$ come il decremento medio della forza d'urto in corrispondenza di un aumento di una settimana di invecchiamento 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 2.2 si riportano i valori stimati $f(x_i)$ della variabile risposta ed i residui e_i.

Nell'Esempio 2.4 si forniranno i dettagli per il calcolo pratico di b_0, b_1 e di un'altra quantità che verrà introdotta nel paragrafo successivo. ♣

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 2.2: Esempio 2.3: valori rilevati (x_i, y_i), valori stimati e residui $f(x_i), e_i$ in accordo al modello di regressione 2.20.

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?

Tutti questi aspetti devono essere indagati prima di utilizzare il modello ricavato; tale analisi può condurre a rigettare il modello oppure a cercare di migliorarlo. La risposta a queste domande concerne un importante aspetto della costruzione di modelli di regressione noto come *valutazione della bontà di adattamento* e che affronteremo più avanti.

2.4 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

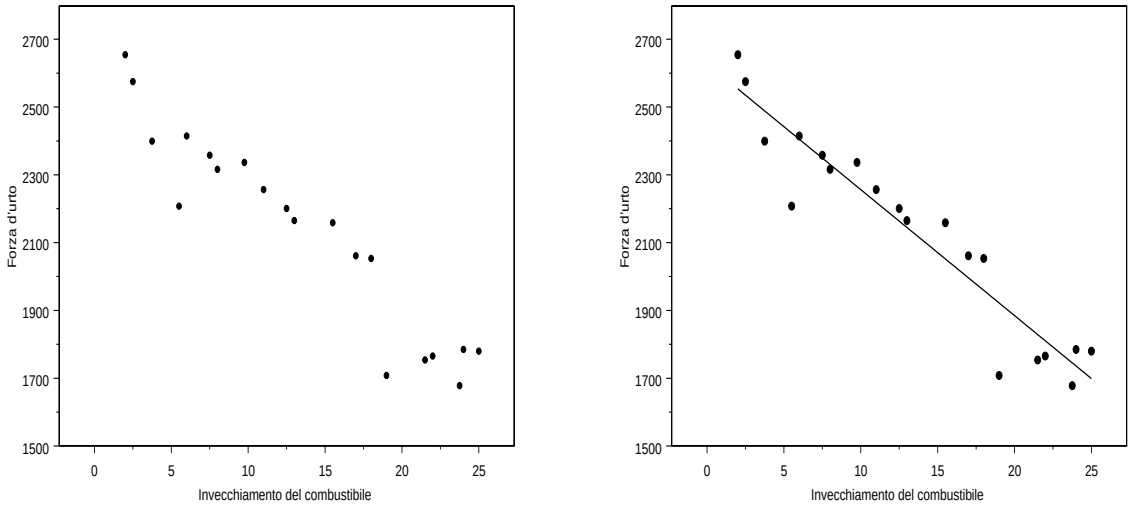


Figura 2.5: *Diagramma a dispersione forza d'urto-età del combustibile e retta di regressione.*

Y. Consideriamo la devianza della variabile risposta Y:

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

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:

$$\begin{aligned}\sum_{i=1}^{n} (y_i - f(x_i))(f(x_i) - \bar{y}) &= \sum_{i=1}^{n} e_i (b_0 - b_1 x_i - b_0 - b_1 \bar{x}) \\ &= b_1 \sum_{i=1}^{n} e_i (x_i - \bar{x}) = b_1 \sum_{i=1}^{n} (e_i - \bar{e})(x_i - \bar{x}) \\ &= b_1 \sigma_{xe} = 0\end{aligned}$$

in quanto $\bar{e} = 0$ e per la (2.19) si ha $\sigma_{ex} = 0$.

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

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

e ciò mostra che la relazione precedente mostra che devianza $\mathrm{SS}_y = \sum_{i=1}^n (y_i - \bar{y})^2$ è data dalla somma di due contributi: *i*) la *devianza di regressione* $\mathrm{SS}_f = \sum_{i=1}^n (f(x_i) - \bar{y})^2$; *ii*) la *devianza residua* o *di errore* $\mathrm{SS}_e = \sum_{i=1}^n (y_i - f(x_i))^2 = \sum_{i=1}^n e_i^2$, dovuta al fatto che non tutte le osservazioni giacciono sulla retta di regressione. Possiamo scriverla sinteticamente:

$$\mathrm{SS}_y = \mathrm{SS}_e + \mathrm{SS}_f \ . \tag{2.22}$$

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

$$\sigma_y^2 = \sigma_\varepsilon^2 + \sigma_f^2 \ . \tag{2.23}$$

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. Possiamo quindi considerare il rapporto:

$$R^2 := \frac{\mathrm{SS}_f}{\mathrm{SS}_y} = \frac{\sigma_f^2}{\sigma_y^2}$$

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. 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 uette 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.

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.

La relazione (2.22) può essere analizzata da un altro punto di vista. A ciascuna devianza (somma di quadrati) si associa un numero che concerne i suoi *gradi di libertà* che è dato dal numero di osservazioni n meno il numero di equazioni (vincoli) fra le variabili. Ad esempio la devianza di Y presenta $n - 1$ gradi di libertà in quanto le quantità $y_1 - \bar{y}, y_2 - \bar{y}, \ldots, y_n - \bar{y}$ soddisfano il vincolo $\sum_i (y_i - \bar{y}) = 0$.

Allo stesso modo la devianza di regressione $\sum_{i=1}^n (f(x_i) - \bar{y})^2$ presenta solo un grado di libertà in quanto essa può essere scritta come:

$$\sum_{i=1}^n (f(x_i) - \bar{y})^2 = \sum_{i=1}^n (b_0 + b_1 x_i - b_0 - b_1 \bar{x})^2 = b_1^2 \sum_{i=1}^n (x_i - \bar{x})^2$$

o, sinteticamente:

$$\mathrm{SS}_f = b_1^2 \mathrm{SS}_x \tag{2.24}$$

cioè risulta univocamente determinata una volta ottenuta la stima b_1 di β_1. Per sottrazione la devianza di errore SS_e presenta $n-2$ gradi di libertà. I due gradi di libertà che vengono sottratti a n riflettono il fatto che i residui possono essere calcolati dalla retta di regressione e questa richiede la stima b_0, b_1 dei due parametri β_0, β_1. La relazione (2.22) in termini di gradi di libertà viene pertanto scritta come:

$$n-1 \;\; = \;\; 1 + (n-2)\ . \tag{2.25}$$

Dalle (2.21) e (2.25) possiamo costruire la seguente tabella dell'*analisi della varianza*:

Sorgente di variazione	*Gradi di libertà*	*Devianza*
regressione	1	$\mathrm{SS}_f = \sum_{i=1}^n (f(x_i) - \bar{y})^2$
residui	$n-2$	$\mathrm{SS}_e = \sum_{i=1}^n (y_i - f(x_i))^2$
totale	$n-1$	$\mathrm{SS}_y = \sum_{i=1}^n (y_i - \bar{y})^2$

Calcolo pratico del coefficiente R^2. Per il calcolo del coefficiente di determinazione, tenendo conto che per la (2.24):

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

si ha:

$$R^2 = \frac{\sigma_f^2}{\sigma_y^2} = \frac{\sigma_{xy}^2}{\sigma_y^2 \sigma_x^2} = r^2\ . \tag{2.26}$$

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

Esempio 2.4 Riprendendo i dati riportati in Tabella 2.1, consideriamo un modello lineare del tipo $Y = \beta_0 + \beta_1 X + \varepsilon$, dove Y è il consumo di gas (in m^3) e X è la temperatura esterna (in $^\circ C$). Le stime b_1, b_0 di β_1, β_0, in base alle (2.15), (2.16), sono date da:

$$b_1 = \frac{\sigma_{xy}}{\sigma_x^2} \qquad\qquad b_0 = \bar{y} - b_1 \bar{x}\ ,$$

dove:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i \qquad \sigma_x^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2 \qquad\qquad \sigma_{xy} = \frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{x}\bar{y}\ .$$

Inoltre, il coefficiente di determinazione R^2, in base alla (2.26), è uguale al quadrato del coefficiente di correlazione; pertanto è necessario calcolare anche $\sigma_y^2 = \frac{1}{n}\sum_{i=1}^n y_i^2 - \bar{y}^2$.

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	7.8	16733	60.84	279993289	130517.4
2	6.2	16596	38.44	275427216	102895.2
3	10.1	14666	102.01	215091556	148126.6
4	1.2	22621	1.44	511709641	27145.2
5	4.3	20199	18.49	407999601	86855.7
6	9.4	14849	88.36	220492801	139580.6
7	-2.0	24086	4.00	580135396	-48172.0
8	3.2	21748	10.24	472975504	69593.6
somme	40.2	151498	323.82	2963825004	656542.3
medie	5.0	18937	40.48	370478126	82067.8

Tabella 2.3: Tabella per il calcolo delle stime b_0, b_1 e di R^2.

Pertanto per il calcolo delle stime b_0, b_1 e di R^2 bisogna ricavare le seguenti quantità:

$$\frac{1}{n}\sum_{i=1}^{n} x_i \qquad \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \frac{1}{n}\sum_{i=1}^{n} x_i^2 \qquad \frac{1}{n}\sum_{i=1}^{n} y_i^2 \qquad \frac{1}{n}\sum_{i=1}^{n} x_i y_i \ .$$

A tale scopo, si costruisce la Tabella 2.3 dei calcoli: da cui si ottiene:

$$\bar{x} = 5.0 \qquad \bar{y} = 18937$$
$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = 40.48 \qquad \frac{1}{n}\sum_{i=1}^{n} y_i^2 = 370478126 \qquad \frac{1}{n}\sum_{i=1}^{n} x_i y_i = 82067.8 \ .$$

In base alle relazioni precedenti, si ottiene quindi:

$$b_1 = -859.8 \qquad b_0 = 23257.7 \qquad R^2 = 0.9492 \ .$$

da cui la seguente equazione della retta di regressione:

$$\hat{Y} = 23257.7 - 859.8X \ .$$

In Figura 2.6 riportiamo il grafico a dispersione della distribuzione di (X, Y) e quello della corrispondente retta di regressione. Il coefficiente di determinazione è uguale a $R^2 = 0.9492$, che è un valore estremamente elevato; pertanto il modello proposto presenta un buon adattamento ai dati. Il modello ottenuto può essere utilizzato a fini predittivi: in corrispondenza di una temperatura esterna di $X = 8°$ si prevede un consumo medio di gas pari a $\hat{Y} = 23257.7 - 859.8 \cdot 8 = 16379.4m^3$. ♣

Esempio 2.5 (Continua Esempio 2.3) Per il modello ricavato in precedenza si ottiene un coefficiente di determinazione $R^2 = 0.9018$ e pertanto il modello presenta un buon adattamento ai dati. In corrispondenza di un combustibile avente un'età pari a $x = 10$ settimane, otteniamo una forza d'urto pari a $\hat{Y} = 2627.82 - 37.15 \cdot 10 = 2256.32$. ♣

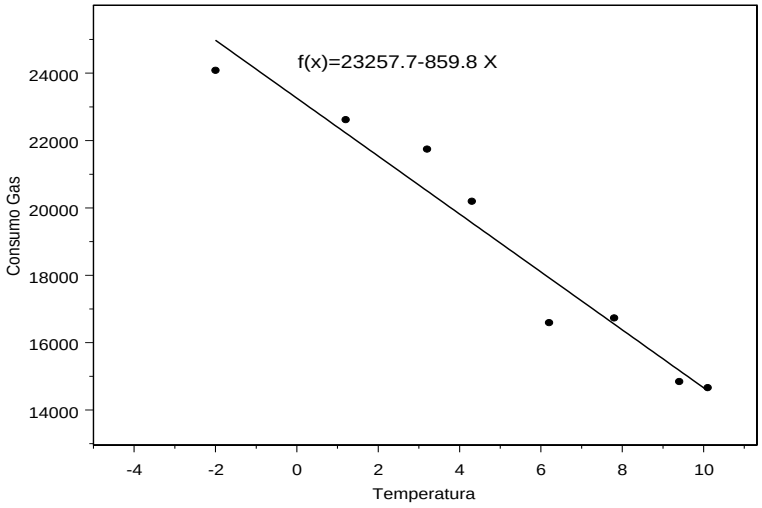


Figura 2.6: *Diagramma a dispersione temperatura-consumo di gas e relativa retta di regressione.*

Nota importante. 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). In particolare, 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. Inoltre il valore di R^2 cresce all'aumentare del numero di termini nel modello. Per esempio, se fra gli n punti osservati $x_1, \ldots, x_n$ non vi sono misure ripetute (cioè valori diversi della risposta Y in corrispondenza degli stessi valori della $\mathbf{x}$), 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.

2.5 Regressione di X su Y

Assegnate le variabili $X.Y$, oltre al modello di regressione di Y su X, nei casi in cui ciò abbia senso, consideriamo il modello di regressione di X su Y:

$$X = \alpha_0 + \alpha_1 x + \varepsilon$$

allora, procedendo come sopra, si perviene alla seguente stima della dipendenza di X su Y:

$$g(y) = a_0 + a_1 x$$

dove:

$$a_1 = \frac{\sigma_{xy}}{\sigma_y^2} \qquad \text{e} \qquad a_0 = \bar{x} - a_1 \bar{y}\ .$$

Si noti che il prodotto dei coefficienti di regressione b_1 (regressione di Y su X) e a_1 (regressione di X su Y) è uguale a :

$$b_1 \cdot a_1 = \frac{\sigma_{xy}}{\sigma_x^2}\frac{\sigma_{xy}}{\sigma_y^2} = r^2 \ ,$$

e quindi:

$$r = \sqrt{b_1 \cdot a_1} \tag{2.27}$$

cioè il coefficiente di correlazione lineare semplice fra due variabili Y e X può essere espresso come la media geometrica del coefficiente di regressione b_1 di Y su X e del coefficiente di regressione a_1 di X su Y.

Esercizio 2.1 Sia data la retta di regressione

$$y = 4 + 0.8x$$

relativa ad una variabile statistica doppia (X, Y).

1. Consideriamo le seguenti rette:

$$\begin{array}{lrcl} a. & x & = & 6 + 0.7y \\ b. & x & = & 6 + 1.25y \\ c. & x & = & 6 - 0.7y \\ d. & x & = & 6 + 2y \ . \end{array}$$

   Quale di queste quattro rette può essere quella di regressione di X su Y?

2. Assumendo la retta così individuata come relazione funzionale fra X e Y, calcolare i valori medi $\bar{x}$ e $\bar{y}$.

Il primo quesito viene affrontato considerando le relazioni che intercorrono fra il coefficiente b_1 della retta di regressione di Y su X ed il coefficiente a_1 della retta di regressione di X su Y relativamente ad una variabile statistica doppia (X, Y). Data una variabile statistica doppia (X, Y), le rette di regressione sono date da:

$$\begin{array}{cc} y = b_0 + b_1 x & x = a_0 + a_1 y \\ b_1 = \dfrac{\sigma_{xy}}{\sigma_x^2} & a_1 = \dfrac{\sigma_{xy}}{\sigma_y^2} \ . \end{array}$$

Pertanto segue che:

$$0 \leq b_1 a_1 = \frac{\sigma_{xy}}{\sigma_x^2 \sigma_y^2} \leq 1 \, .$$

Ciò porta ad escludere la retta c) perchè in questo caso risulterebbe:

$$b_1 a_1 = -0.7 \cdot 0.8 = -0.56 < 0 \ .$$

cioè il coefficiente di determinazione assumerebbe valore negativo, e ciò è assurdo. Per motivi analoghi si scarta anche la retta d). Infatti in questo caso si avrebbe:

$$b_1 a_1 = 0.8 \cdot 2 = 1.6 > 1 \ .$$

Un po' diverse le ragioni per cui si scarta l'equazione b). In questo caso, infatti, risulta:

$$b_1 a_1 = 0.8 \cdot 1.25 = 1$$

che è un valore ammissibile per il coefficiente di determinazione. Questo è però un valore particolare. In questo caso, infatti, $b_1 a_1 = 1$ significa che le due rette di regressione coincidono, per cui considerando l'equazione $x = 6 + 1,25y$ ed esplicitandola rispetto a y dovremmo trovare l'equazione che fornisce il testo. Ciò invece non accade, infatti da:

$$x = 6 + 1.25y$$

si passa a:

$$y = -6 + 0.8$$

che è evidentemente divera dall'equazione $y = 4 + 0.8x$ che fornisce il testo.

L'unica equazione possibile è la prima, infatti:

$$b_1 a_1 = 0.8 \cdot 0.7 = 0.56 \ .$$

Per quanto riguarda il secondo quesito, al fine di calcolare le due medie $\bar{x}$ e $\bar{y}$, partiamo dalle relazioni:

$$a_0 = \bar{x} - a_1 \bar{y} \qquad\qquad b_0 = \bar{y} - b_1 \bar{x}$$

In base ai valori noti possiamo scrivere:

$$6 = \bar{x} - 0.7\bar{y} \qquad\qquad 4 = \bar{y} - 0.8\bar{x}$$

ottenendo un sistema di due equazioni nelle due incognite $\bar{x}$ e $\bar{y}$. Risolvendo tale sistema (ad esempio col metodo di sostituzione), si perviene alla soluzione:

$$\bar{x} = 20 \qquad\qquad \bar{y} = 20 \ .$$

♣

2.6 Altre formulazioni del modello di regressione

Presentiamo ora un'altra formulazione del modello di regressione (2.2) che può risultare utile in alcuni casi. Aggiungendo e sottraendo la quantità $\beta_1 \bar{x}$ nel modello $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ si ottiene:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} + \beta_1 x_i + \varepsilon_i = (\beta_0 + \beta_1 \bar{x}) + \beta_1 (x_i - \bar{x}) + \varepsilon_i \\ &= \beta_0' + \beta_1 (x_i - \bar{x}) + \varepsilon_i \ , \end{aligned} \tag{2.28}$$

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

Le equazioni normali per il modello (2.28) sono le seguenti:

$$\begin{aligned} n b_0' &= \sum_{i=1}^{n} y_i \\ b_1' \sum_{i=1}^{n} (x_i - \bar{x})^2 &= \sum_{i=1}^{n} y_i (x_i - \bar{x}) \end{aligned} \tag{2.29}$$

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

$$b'_0 = \bar{y}$$
$$b'_1 = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\mathrm{S}_{xy}}{\mathrm{SS}_x}$$

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 (2.29) risulta più semplice da risolvere rispetto al precedente (2.14) in quanto ora ciascuna equazione presenta solo un'incognita. In base alla (2.28), il modello di regressione si scrive:

$$f(x) = \bar{y} + b'_1(x_i - \bar{x}) \tag{2.30}$$

e, benchè sia formalmente equivalente a quello precedente $f(x) = b_0 + b_1 x$, evidenzia immediatamente che esso ha validità nell'intervallo dei valori assegnati centrato in $\bar{x}$.

2.6.1 Un approccio matriciale

Introduciamo il vettore delle osservazioni $\mathbf{y}$, la matrice delle variabili indipendenti $\underset{\sim}{\mathbf{X}}$, il vettore dei parametri da stimare $\boldsymbol{\beta}$, il vettore degli errori $\boldsymbol{\varepsilon}$.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \underset{\sim}{\mathbf{X}} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{2.31}$$

Le equazioni (2.6) possono quindi essere scritte sinteticamente in forma di equazione matriciale:

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

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

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

Consideriamo le seguenti quantità:

$$\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

$$\underset{\sim}{\mathbf{X}}'\mathbf{y} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

Ciò significa che le equazioni normali (2.14) che qui riportiamo:

$$\sum_{i=1}^{n} (y_i - b_0 - b_1\, x_i) = 0$$
$$\sum_{i=1}^{n} x_i (y_i - b_0 - b_1\, x_i) = 0\ ,$$

possono essere scritte come:

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

dove

$$\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \tag{2.32}$$

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

$$\mathbf{b} = (\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}})^{-1}\underset{\sim}{\mathbf{X}}'\mathbf{y} \tag{2.33}$$

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 Modelli senza intercetta (caso $b_0 = 0$)

In alcuni casi, per ragioni strutturali, è più opportuno adottare un modello senza intercetta del tipo:

$$Y = \beta_1 x + \varepsilon\ . \tag{2.34}$$

Pertanto, assegnate n coppie di osservazioni (x_i, y_i), $i = 1, \ldots, n$, si ottiene la seguente funzione di errore:

$$\mathcal{E}(\beta_1) = \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2$$

che conduce ad un'unica equazione normale:

$$\sum_{i=1}^{n} x_i(y_i - b_1 x_i) = 0 \qquad \text{da cui} \quad b_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

ottenendo quindi il modello $\hat{y} = b_1 x$.

Bisogna osservare che spesso la relazione fra Y e X in prossimità dell'origine presenta un andamento abbastanza differente rispetto all'andamento che si può osservare nell'intervallo dei valori osservati della X. In questo caso, il diagramma a dispersione fornisce informazioni al fine di decidere se utilizzare un modello con intercetta ($b_0 \neq 0$) o senza intercetta ($b_0 = 0$). In alternativa si potrebbero costruire entrambi i modelli e poi valutare qual è il più opportuno. Va sottolineato che l'indice R^2 non è un buon criterio per tale confronto in quanto indica la proporzione di variabilità spiegata (intorno a $\bar{y}$) dal modello di regressione. Nel caso di modello senza intercetta, dalla (2.21), si ricava:

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

ottenendo quindi l'indice:

$$R_0^2 := \frac{\sum_{i=1}^{n} f(x_i)^2}{\sum_{i=1}^{n} y_i^2}$$

che indica la proporzione di variabilità spiegata dalla regressione intorno all'origine. In questi casi è meglio confrontare la devianza dei residui[3] dei due modelli e scegliere il modello per cui risulta inferiore.

2.7 Proprietà degli stimatori dei minimi quadrati

Come detto in precedenza, **b** è una stima di $\boldsymbol{\beta}$ in quanto dipende dalla particolare realizzazione (x_i, y_i), $i = 1, \ldots, n$ ottenuta; pertanto le stime b_0, b_1 ricavate rispettivamente nelle (2.15) e (2.16) e che qui riportiamo:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.35}$$

$$b_0 = \bar{y} - b_1 \bar{x} \tag{2.36}$$

sono da intendersi come realizzazioni di due variabili aleatorie denotate rispettivamente con B_1 e B_0, in quanto i valori empirici y_i, $(i = 1, \ldots, n)$ – come evidenziato in precedenza – sono da intendersi come realizzazioni delle v.a. $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. La (2.35) si scrive in questo caso:

$$B_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\mathrm{SS}_x} \tag{2.37}$$

e quindi calcolandone la speranza matematica, per la (2.6), si ottiene:

$$\begin{aligned} \mathbb{E}(B_1) &= \frac{\sum_{i=1}^{n}[(x_i - \bar{x})\mathbb{E}(Y_i)]}{\mathrm{SS}_x} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\mathrm{SS}_x} \\ &= \beta_0 \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\mathrm{SS}_x} + \beta_1 \frac{\sum_{i=1}^{n}(x_i - \bar{x})x_i}{\mathrm{SS}_x} = \beta_1 \end{aligned} \tag{2.38}$$

[3] Pi`u correttamente bisognerebbe considerare la stima corretta della varianza di errore che verr`a introdotta nella (2.46).

in quanto:

$$\sum_{i=1}^{n}(x_i-\bar{x})=0 \qquad \text{e} \qquad \frac{\sum_{i=1}^{n}(x_i-\bar{x})x_i}{\mathrm{SS}_x}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})x_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=1.$$

Pertanto per il modello in esame si ottiene $\mathbb{E}(B_1)=\beta_1$, e si dice che B_1 è uno *stimatore non distorto* di β_1. In maniera analoga, considerata la v.a. *media campionaria* $\bar{Y}$:

$$\bar{Y}:=\frac{Y_1+Y_2+\cdots+Y_n}{n}$$

osserviamo preliminarmente che, in base alla (2.6), si ha:

$$\begin{aligned}\mathbb{E}(\bar{Y})&=\mathbb{E}\left(\frac{Y_1+Y_2+\cdots+Y_n}{n}\right)=\left(\frac{\mathbb{E}(Y_1)+\mathbb{E}(Y_2)+\cdots+\mathbb{E}(Y_n)}{n}\right)=\mathbb{E}(Y)\\&=\beta_0+\beta_1\bar{x}\,.\end{aligned} \tag{2.39}$$

Pertanto segue dalla (2.36), mediante le (2.38) e (2.39):

$$\mathbb{E}(B_0)=\mathbb{E}(\bar{Y}-B_1\bar{x})=\mathbb{E}(\bar{Y})-\mathbb{E}(B_1)\bar{x}=\beta_0+\beta_1\bar{x}-\beta_1\bar{x}=\beta_0\,,$$

e quindi B_0 è uno *stimatore non distorto* di β_0.

Possiamo ricavare anche la varianza degli stimatori. Posto per semplicità:

$$c_i:=\frac{(x_i-\bar{x})}{\mathrm{SS}_x}=\frac{(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

dalla (2.15) segue:

$$\mathrm{Var}(B_1)=\mathrm{Var}\,(Y_i)=\sum_{i=1}^{n}c_i^2\mathrm{Var}(Y_i)$$

avendo assunto la non correlazione fra gli errori, cioè la (2.9), segue infatti che la varianza di una somma di v.a. è uguale alla somma delle rispettive varianze. Poichè, per la (2.8), si ha $\mathrm{Var}(Y_i)=\mathrm{Var}(e_i)=\sigma^2$, si ottiene la seguente relazione per la varianza di B_1:

$$\mathrm{Var}(B_1)=\sigma^2\sum_{i=1}^{n}c_i^2=\sigma^2\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{[\sum_{i=1}^{n}(x_i-\bar{x})^2]^2}=\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{\sigma^2}{\mathrm{SS}_x}\,. \tag{2.40}$$

Per quanto concerne la varianza di B_0 si ha:

$$\mathrm{Var}(B_0)=\mathrm{Var}(\bar{Y}-B_1\bar{x})=\mathrm{Var}(\bar{Y})+\bar{x}^2\mathrm{Var}(B_1)-2\bar{x}\mathrm{Cov}(\bar{Y},B_1)\,.$$

Tenendo conto che:

$$\text{Var}(\bar{Y}) = \text{Var}\left(\frac{Y_1 + \cdots + Y_n}{n}\right) = \frac{\sigma^2}{n} \tag{2.41}$$

$$\bar{x}^2\text{Var}(B_1) = \bar{x}^2\frac{\sigma^2}{\text{SS}_x}$$

$$\begin{aligned}\text{Cov}(\bar{Y}, B_1) &= \text{Cov}(\bar{Y}, \sum_{i=1}^{n} c_i Y_i) = \text{Cov}\left(\frac{1}{n}\sum_{j=1}^{n} Y_j, \sum_{i=1}^{n} c_i Y_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} c_i \text{Cov}(Y_i, Y_i) = \frac{1}{n}\sum_{i=1}^{n} c_i \text{Var}(Y_i) = \frac{1}{n}\sum_{i=1}^{n} c_i\sigma^2 \\ &= \frac{\sigma^2}{n}\sum_{i=1}^{n} c_i = 0 \end{aligned} \tag{2.42}$$

in quanto $\text{Cov}(Y_j, Y_i) = 0$ per $i \neq j$ per l'ipotesi di non correlazione fra gli errori e $\sum_i c_i = 0$. Segue pertanto:

$$\text{Var}(B_0) = \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{\text{SS}_x} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\text{SS}_x}\right) . \tag{2.43}$$

Un importante risultato concernente la qualità degli stimatori dei minimi quadrati B_0 e B_1 è noto come *teorema di Gauss-Markov*, che afferma che per il modello di regressione (2.2), sotto le ipotesi $\mathbb{E}(\varepsilon) = 0$ e $\text{Var}(\varepsilon) = \sigma^2$ e di non correlazione fra gli errori (2.9), gli stimatori dei minimi quadrati sono non distorti ed hanno varianza minima quando comparati con tutti gli altri stimatori che sono combinazioni lineari di Y_i. Spesso si dice che gli stimatori dei minimi quadrati sono i *migliori stimatori lineari non distorti*, dove "migliore" vuol dire "con minima varianza". Notiamo infine che risulta:

$$\text{Cov}(B_0, B_1) = \text{Cov}(\bar{Y} - B_1\bar{x}, B_1) = \text{Cov}(\bar{Y}, B_1) - \bar{x}\text{Var}(B_1) = -\bar{x}\frac{\sigma^2}{\text{SS}_x} \tag{2.44}$$

in quanto $\text{Cov}(B_1, B_1) = \text{Var}(B_1)$ e$\text{Cov}(\bar{Y}, B_1) = 0$ per la (2.42).

2.7.1 Varianza della risposta

Lo scopo principale del modello di regressione concerne la stima di $\mathbb{E}(Y|x)$ del valor medio della risposta Y in corrispondenza del valore x del regressore X. Ovviamente anche la risposta $\hat{y}_i$ può essere riguardata come una realizzazione della v.a. $\hat{Y}$, anch'essa dipendente dai valori empirici $y_1, \ldots, y_n$ che costituiscono la realizzazione del campione $Y_1, \ldots, Y_n$.

Per quanto riguarda la speranza di $\hat{Y}$, si ha:

$$\mathbb{E}(\hat{Y}) = \mathbb{E}(B_0 + B_1 x) = \mathbb{E}(B_0) + \mathbb{E}(B_1)x = \beta_0 + \beta_1 x .$$

Per quanto riguarda la varianza si ha, essendo $B_0 = \bar{Y} - B_1\bar{x}$:

$$\begin{aligned}\text{Var}(\hat{Y}) &= \text{Var}(B_0 + B_1 x) = \text{Var}(\bar{Y} + B_1(x - \bar{x})) \\ &= \text{Var}(\bar{Y}) + \text{Var}(B_1(x - \bar{x})) + \text{Cov}(\bar{Y}, B_1(x - \bar{x})) \\ &= \text{Var}(\bar{Y}) + (x - \bar{x})^2\text{Var}(B_1) + (x - \bar{x})\text{Cov}(\bar{Y}, B_1)\end{aligned}$$

ed essendo $\text{Var}(\bar{Y}) = \sigma^2/n$ per la (2.41), $\text{Var}(B_1) = \sigma^2/\text{SS}_x$ per la (2.40) e $\text{Cov}(\bar{Y}, B_1) = 0$ per la (2.42), segue:

$$\text{Var}(\hat{Y}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{\text{SS}_x}(x - \bar{x})^2 = \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\text{SS}_x} \right) . \tag{2.45}$$

Pertanto la varianza della stima assume valore minimo (pari a σ^2/n) per $x = \bar{x}$ e cresce all'aumentare della distanza del punto x da $\bar{x}$.

2.7.2 Stima di σ^2

Oltre alla stima di β_0 e β_1, è importante ottenere una stima della varianza della distribuzione degli errori, cioè della varianza di ε. Da un punto di vista ideale, questa stima non dipende dalla bontà del modello; tuttavia ciò si verifica solo quando vi sono parecchie osservazioni di Y per almeno un valore di x oppure quando sono note informazioni a priori sulla varianza di errore σ^2. Quando tale approccio non può essere utilizzato, la stima di σ^2 si ottiene come segue.

Consideriamo la somma dei quadrati dei residui:

$$\text{SS}_e := \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2 .$$

Come evidenziato dalla (2.25), la devianza dei residui presenta $n - 2$ gradi di libertà in quanto due gradi di libertà sono associati con le stime b_0 e b_1 nel modello di regressione. Si può dimostrare che $\mathbb{E}(\text{SS}_e) = (n - 2)\sigma^2$ e quindi una stima corretta s^2 di σ^2 è data da:

$$s^2 = \frac{\text{SS}_e}{n - 2} . \tag{2.46}$$

La quantità s (cioè la radice quadrata di s^2) viene chiamata *errore standard di regressione*. Si noti che la stima s dipende dal modello considerato.

2.7.3 Stima della varianza di B_0, B_1

Le quantità $\text{Var}(B_0)$ e $\text{Var}(B_1)$ non possono essere ricavate in quanto sono basate sulla varianza di errore σ^2 che non è nota. In tal caso si può dimostrare che tali quantità possono essere stimate utilizzando la stima s^2 di σ^2. La quantità:

$$\text{se}(b_1) := \sqrt{\frac{s^2}{\text{SS}_x}} \tag{2.47}$$

viene chiamata *errore standard* del coefficiente angolare intercetta e fornisce una misura dell'incertezza circa la stima b_1 di β_1; analogamente la quantità:

$$\text{se}(b_0) := \sqrt{s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\text{SS}_x} \right)} \tag{2.48}$$

viene chiamata *errore standard* dell'intercetta e fornisce una misura dell'incertezza circa la stima b_0 di β_0.

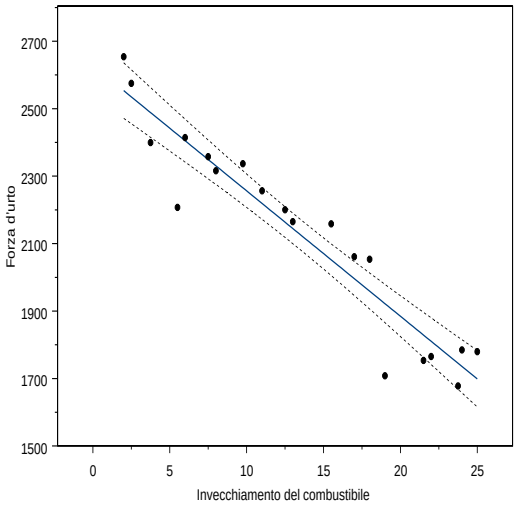


Figura 2.7: *Diagramma a dispersione forza d'urto-età del combustibile, retta di regressione e grafico di* $f(x) \pm \alpha \cdot \mathrm{se}(\hat{y})$ *per i dati dell'Esempio 2.3.*

Infine otteniamo anche una stima dello scarto quadratico medio di Y

$$\mathrm{se}(\hat{y}) = \sqrt{s^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\mathrm{SS}_x} \right)} \,. \tag{2.49}$$

Per comprendere meglio il significato della quantità sopra ricavata, con riferimento ai dati nell'Esempio 2.3, in Figura 2.7 viene fornito (in linea tratteggiata) il grafico di $f(x) \pm \alpha \cdot \mathrm{se}(\hat{y})$[4].

2.8 Alcuni rischi nella costruzione di modelli di regressione

I metodi di regressione trovano applicazione in numerosi ambiti; spesso, purtroppo, vi è un utilizzo errato che può portare a gravi errori di interpretazione o di predizione.

1. I modelli di regressione costituiscono delle interpolazioni lineari nel campo di variazione della variabile di regressione utilizzata per adattare il modello.

2. La disposizione dei valori di X gioca un ruolo importante nell'adattamento col metodo dei minimi quadrati. Mentre i punti hanno la stessa importanza nel determinare l'intercetta b_0, il coefficiente angolare b_1 – e quindi la pendenza della retta di regressione – è fortemente influenzata dai valori estremi di X. Si consideri ad esempio il diagramma in Figura 2.8. La pendenza della curva dipende pesantemente dalle osservazioni A e B che vengono chiamati *valori influenti*.

[4]Per valori opportuni di α tale quantit`a fornisce un intervallo di confi denza per la risposta in funzione di x.

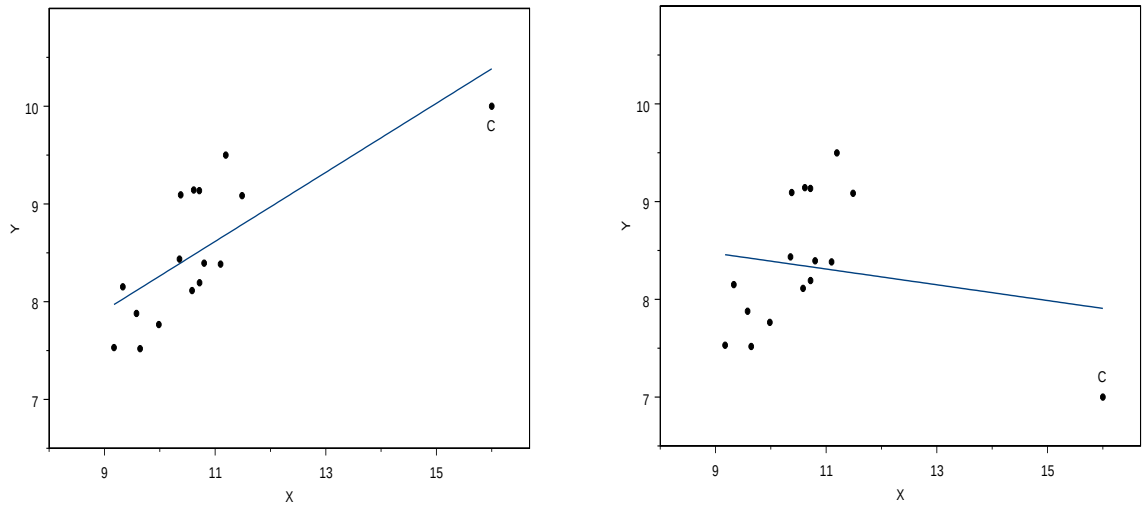


Figura 2.9: *Dipendenza del modello di regressione da valori remoti.*

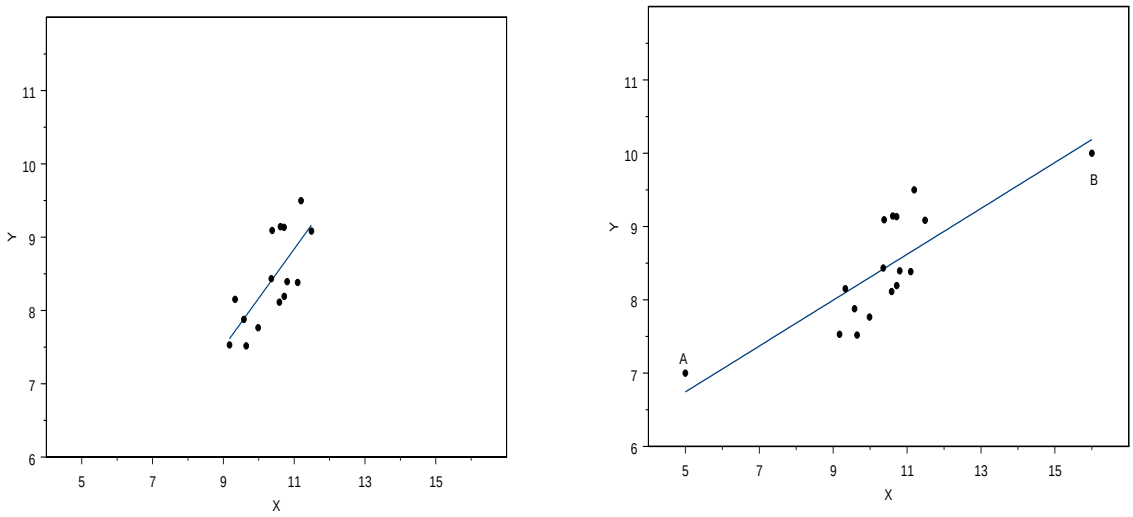


Figura 2.8: *Esempio di valori influenti A e B.*

Un caso simile è descritto in Figura 2.9 dove si può evidenziare come cambia il modello di regressione al variare della posizione del punto C. Situazioni come queste, che si verificano spesso nei casi pratici, richiedono azioni correttive come ad esempio ulteriori analisi e possibile eliminazione dei dati. In particolare, gli esempi precedenti mostrano come, in alcuni casi, piccole variazioni nei valori di poche osservazioni possono condurre a rilaventi cambiamenti nel modello di regressione.

3. L'adattamento dei minimi quadrati risente anche dei cosidetti *valori anomali*. Consideriamo l'esempio in Figura 2.10, in cui l'osservazione D sembra essere un valore anomalo

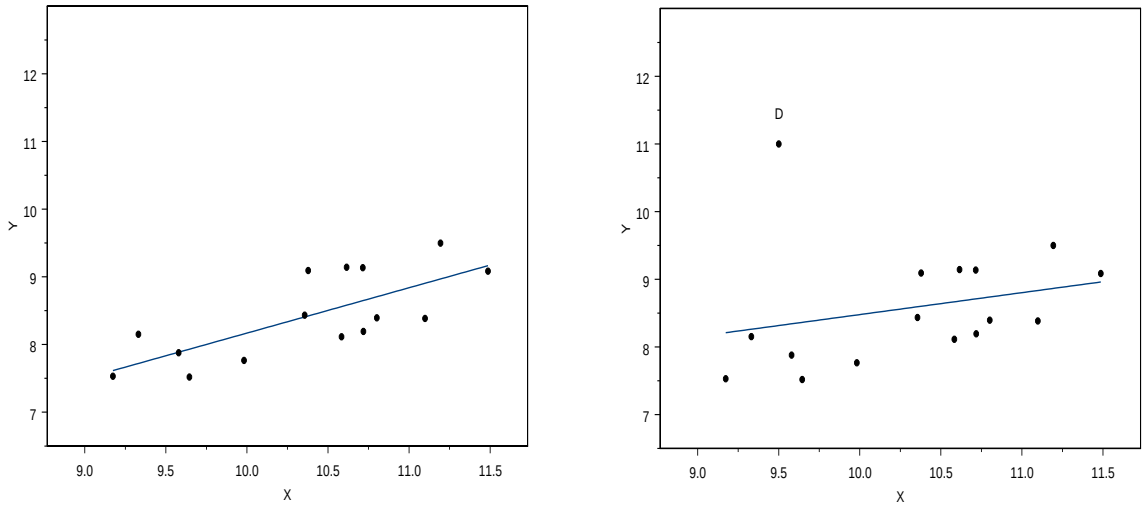


Figura 2.10: *Esempio di valore anomalo (punto D, figura a destra).*

in quanto ricade in una zona del piano abbastanza lontano dalle altre osservazioni. Se tale punto fosse un valore "errato", ciò modificherebbe il valore di b_0 così oltre ad alterare il valore di σ^2_ε. D'altra parte, il dato rilevato potrebbe non essere errato, al contrario potrebbe aiutare a comprendere meglio il fenomeno che si sta studiando.

4. Come evidenziato in precedenza, anche se l'analisi di regressione suggerisce una qualche relazione di dipendenza fra X e Y ciò non vuol dire che tale relazione abbia effettivamente senso: può essere assurda o anche fuorviante, come i due esempi seguenti mostrano.

   In Tabella 2.4 vengono riportate le osservazioni annuali dal 1924 al 1937 per seguenti variabili: Y:"numero di licenze rilasciate per apparecchiature radio nel Regno Unito (UK)" (dati in migliaia) e X:"numero di lettere del primo nome del Presidente degli Stati Uniti d'America"[5]: Il coefficiente di correlazione fra X e Y è pari a $r = 0.9553$; inoltre si ottiene il seguente modello di regressione:

   $$\hat{y} = -14251 + 2758x \qquad \text{con} \qquad R^2 = 0.9127\,.$$

   Pertanto la variabilità del numero di licenze per apparecchiature radio nel Regno Unito dal 1924 al 1937 può essere largamente spiegata dal numero di lettere del primo nome del Presidente degli Stati Uniti nello stesso periodo. Ciò è ovviamente assurdo!

   Il secondo esempio è riportato nella Tabella 2.5 in cui sono considerati i dati concernenti le variabili: X:"consumo di bevande alcoliche pro-capite" (in Cal/giorno.) e Y:"speranza di vita alla nascita" (in anni)[6] per alcune nazioni con riferimento all'anno 2000. Il coefficiente di correlazione fra X e Y è pari a $r = 0.7822$, l'analisi di regressione fornisce il

[5]Fonte: Montgomery D.C. & Peck E.A. (1992), *Introduction to Linear Regression Analysis*, John Wiley & Sons, New York.

[6]Fonte: Food and Agricoltural Organization of the United Nations, *www.fao.org*; The World Bank Group, *www.worldbank.org*

anno	*Nome Presidente USA*	X	Y
1924	Calvin	6	1350
1925	Calvin	6	1960
1926	Calvin	6	2270
1927	Calvin	6	2483
1928	Calvin	6	2730
1929	Calvin	6	3091
1930	Herbert	7	3647
1931	Herbert	7	4620
1932	Herbert	7	5497
1933	Herbert	7	6260
1934	Franklin	8	7012
1935	Franklin	8	7618
1936	Franklin	8	8131
1937	Franklin	8	8593

Tabella 2.4: *Esempio di relazione assurda fra le variabili.*

modello:

$$\hat{y} = 56.03 + 0.1318x \qquad \text{con} \qquad R^2 = 0.6118\,,$$

e, in base a tale modello, la speranza di vita alla nascita aumenta al crescere del consumo di bevande alcoliche. Ovviamente le cose non stanno così in quanto la correlazione trovata non ha nulla a che vedere con una relazione causa-effetto: la speranza di vita è ovviamente maggiore nei paesi più ricchi, per ragioni evidenti; in tali paesi vi è anche un maggiore consumo di carne e tabacco (ed infatti troveremmo una relazione analoga se avessimo considerato il consumo di carne o di tabacco).

2.9 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

nazione	X	Y
Bolivia	43.7	62.56
Bosnia and Herzegovina	156	73.33
Brazil	73.8	68.07
Bulgaria	174.2	71.55
Canada	134.2	78.93
Chile	70.4	75.86
Ethiopia	12.8	42.29
France	172.2	78.86
Ghana	21.1	56.96
Italy	144.5	78.67
Mexico	57.8	73.15
Poland	140.2	73.28
Spain	179.9	78.15
United Kingdom	176.4	77.33
United States	154.2	77.03

Tabella 2.5: *Esempio di relazione fuorviante fra le variabili.*

rilevare violazioni delle ipotesi del modello esaminando unicamente statistiche quali l'indice R^2, che è invece una misura "globale" delle proprietà del modello, e che quindi non assicura l'adeguatezza del modello ai dati. A tal scopo, consideriamo il seguente esempio.

Esempio 2.6 Consideriamo il seguente insiemi di dati[7]

X	Y_1	Y_2	Y_3	Z	W
10	8.04	9.14	7.46	8	6.58
8	6.95	8.14	6.77	8	5.76
13	7.58	8.74	12.74	8	7.71
9	8.81	8.77	7.11	8	8.84
11	8.33	9.26	7.81	8	8.47
14	9.96	8.1	8.84	8	7.04
6	7.24	6.13	6.08	8	5.25
4	4.26	3.1	5.39	8	5.56
12	10.84	9.13	8.15	8	7.91
7	4.82	7.26	6.42	8	6.89
5	5.68	4.74	5.73	19	12.5

Consideriamo i grafici a dispersione delle coppie di variabili: (X, Y_1), (X, Y_2), (X, Y_3) e (Z, W), riportati nella Figura 2.11 Considerati i modelli di regressione lineare per ciascuna coppia di variabili si ottiene:

[7] dataset: *anscombe.dat*

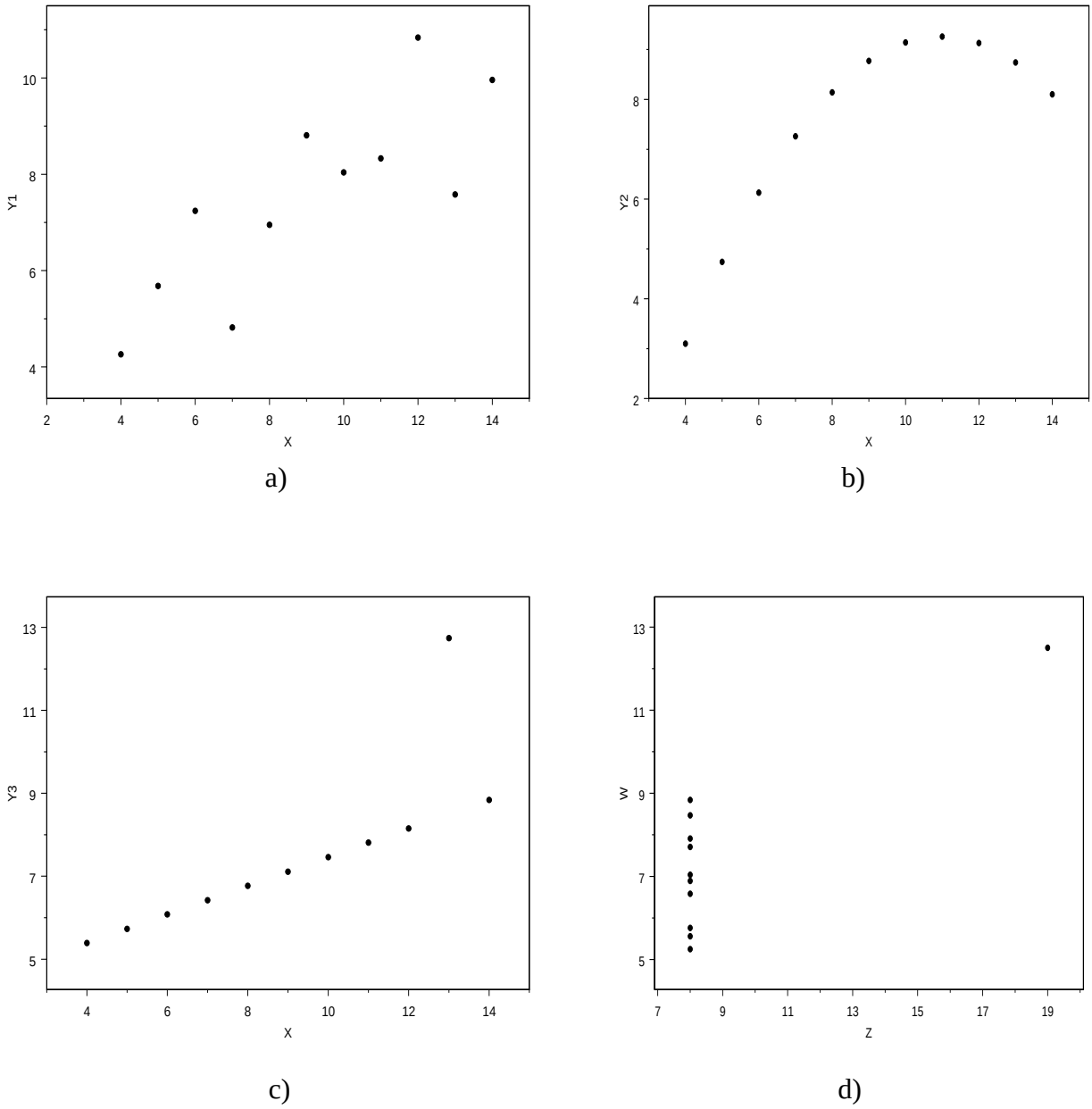


Figura 2.11: *Rappresentazioni grafiche dei delle coppie di variabili* (X, Y_1), (X, Y_2), (X, Y_3) *e* (Z, W) *del file dati anscombe.dat.*

dataset	*variabili*	*equazione retta di regressione*	R^2	s	$\mathrm{se}(b_0)$	$\mathrm{se}(b_1)$
a)	X, Y_1	$\hat{y_1} = 3.0001 + 0.5001x$	0.6665	1.237	1.1247	0.1179
b)	X, Y_2	$\hat{y_2} = 3.0009 + 0.5000x$	0.6662	1.237	1.1253	0.1180
c)	X, Y_3	$\hat{y_3} = 3.0025 + 0.4997x$	0.6663	1.236	1.1245	0.1179
d)	Z, W	$\hat{y} = 3.0017 + 0.4999z$	0.6665	1.236	1.1239	0.1178

e (Z, W), riportati nella Figura 2.11. Si vede che le quattro coppie di variabili conducono in pratica alla stessa equazione della retta di regressione ed agli stessi valori degli indici di valutazione di bontà dell'adattamento: il coefficiente di determinazione R^2, l'errore standard dei residui s e gli errori standard di b_0, b_1 rispettivamente $\mathrm{se}(b_0)$ e $\mathrm{se}(b_1)$.

I diagrammi a dispersione con la relativa retta di regresione sono riportati nella Figura 2.12 Per il primo data set (Figura 2.12 a) si vede che la retta di regressione costituisce un modello adeguato; nel secondo caso (Figura 2.12 b) si vede che la retta di regressione non è un modello adeguato, sarebbe più opportuno considerare una curva, per esempio di tipo quadratico; gli altri due casi mostrano rispettivamente l'effetto di valori anomali e di valori remoti: in entrambi i casi tutti i punti, tranne uno, sono allineati: in questi casi l'eliminazione di questo punto condurrebbe ad un'equazione della retta di regressione abbastanza diversa da quella ricavata. ♣

Poichè gli errori ε non sono quantità osservabili, la correttezza delle ipotesi 1-4 sopra riportate deve essere valutata indirettamente utilizzando i residui:

$$e_i := y_i - \hat{y}_i = y_i - f(x_i) \qquad i = 1, \ldots, n \tag{2.50}$$

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.

I residui presentano alcune importanti proprietà. Innanzitutto, come risulta dalla (2.18), hanno valor medio $\bar{e} = 0$ e la loro varianza viene stimata mediante:

$$s^2 := \frac{\mathrm{SS}_e}{n-2} = \frac{\sum_{i=1}^n (e_i - \bar{e})^2}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2} \,. \tag{2.51}$$

I residui non sono quantità indipendenti, infatti n residui presentano $n-2$ gradi di libertà a loro associati. Questa non indipendenza dei residui comporta alcune conseguenze nel loro utilizzo per la valutazione dell'adeguatezza del modello per valori piccoli di n.

A volte risulta utile lavorare con i *residui standardizzati*:

$$d_i := \frac{e_i}{\sqrt{s^2}} \qquad i = 1, \ldots, n \tag{2.52}$$

che presentano media uguale a zero e varianza approssimativamente unitaria, in quanto la (2.52) divide ciascun residuo per lo scarto quadratico medio della loro distribuzione.

Calcoliamo la varianza dei residui. Considerando e_i come realizzazione della variabile aleatoria $E_i = Y_i - \hat{Y}_i = Y_i - B_0 - B_1 x_i$, si ha

$$\begin{aligned} \mathrm{Var}(E_i) &= \mathrm{Var}(Y_i - \hat{Y}_i) = \mathrm{Var}(Y_i) + \mathrm{Var}(\hat{Y}_i) - 2\mathrm{Cov}(Y_i, \hat{Y}_i) \\ &= \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\mathrm{SS}_x} \right) - 2\mathrm{Cov}(Y_i, \hat{Y}_i) \,. \end{aligned} \tag{2.53}$$

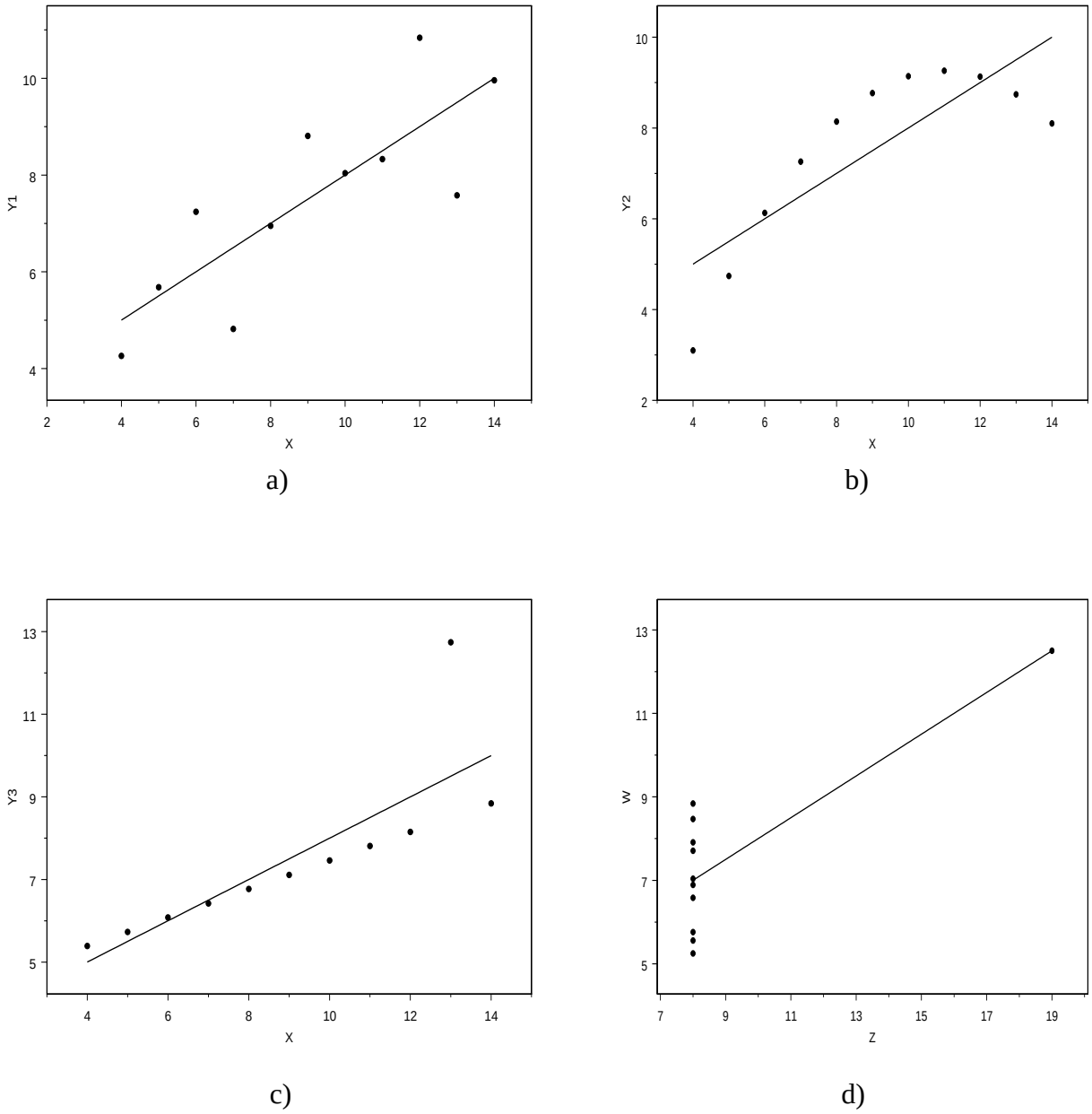


Figura 2.12: *Rappresentazioni grafiche dei delle coppie di variabili* (X, Y_1), (X, Y_2), (X, Y_3) *e* (Z, W) *e corrispondenti rette di regressione del file dati anscombe.dat.*

Tenendo conto che

$$\begin{aligned}\text{Cov}(Y_i, \hat{Y}_i) &= \text{Cov}(Y_i, B_0 + B_1 x_i) = \text{Cov}(\bar{Y} + B_1(x_i - \bar{x}), Y_i) \\ &= \text{Cov}(\bar{Y}, Y_i) + \text{Cov}(B_1(x_i - \bar{x}), Y_i) = \text{Cov}(\bar{Y}, Y_i) + \text{Cov}(B_1, Y_i)(x_i - \bar{x})\end{aligned}$$

e che:

$$\begin{aligned}\text{Cov}(\bar{Y}, Y_i) &= \text{Cov}\left(\frac{Y_1 + \cdots + Y_n}{n}, Y_i\right) = \frac{1}{n}\left(\text{Cov}(Y_1, Y_i) + \cdots + \text{Cov}(Y_n, Y_i)\right) \\ &= \frac{1}{n}\text{Cov}(Y_i, Y_i) = \frac{\text{Var}(Y_i)}{n} = \frac{\sigma^2}{n}\end{aligned}$$

in quanto per l'ipotesi di non correlazione fra gli errori si ha $\text{Cov}(Y_j, Y_i) = 0$ per $j \neq i$, e che:

$$\begin{aligned}\text{Cov}(B_1(x_i - \bar{x}), Y_i) = \text{Cov}(B_1, Y_i)(x_i - \bar{x}) &= \text{Cov}\left(\frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\text{SS}_x}, Y_i\right)(x_i - \bar{x}) \\ &= \frac{(x_i - \bar{x})^2}{\text{SS}_x}\end{aligned}$$

segue

$$\text{Cov}(Y_i, \hat{Y}_i) = \sigma^2\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\text{SS}_x}\right) . \tag{2.54}$$

Pertanto dalle (2.53) e (2.54), si ottiene infine:

$$\begin{aligned}\text{Var}(E_i) &= \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\text{SS}_x}\right) - 2\sigma^2\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\text{SS}_x}\right) \\ &= \sigma^2\left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\text{SS}_x}\right)\right] .\end{aligned}$$

Si definiscono *residui studentizzati* le quantità:

$$r_i = \frac{e_i}{\sqrt{s^2\left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\text{SS}_x}\right)\right]}} \qquad i = 1, \ldots, n . \tag{2.55}$$

Si noti che i residui studentizzati sono ottenuti dividendo il residuo e_i per la stima dello scarto quadratico medio della distribuzione di E_i, piuttosto che per la stima dello scarto di ε come nel caso dei residui standardizzati (2.52). Il termine:

$$\left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\text{SS}_x}\right)\right]$$

assume valore massimo (e quindi r_i assume valore minimo, prossimo a d_i) per $x_i = \bar{x}$; mentre r_i cresce all'aumentare della distanza $(x_i - \bar{x})^2$.

Nel caso di insiemi di dati piccoli, i residui studentizzati sono spesso più appropriati di quelli standardizzati poichè le differenze delle varianze residue possono risultare notevoli. Per valori grandi di n, le differenze fra i risultati derivanti dai due metodi saranno meno rilevanti tenendo conto che $\text{SS}_x = \sum_{i=1}^n (x_i - \bar{x})^2$.

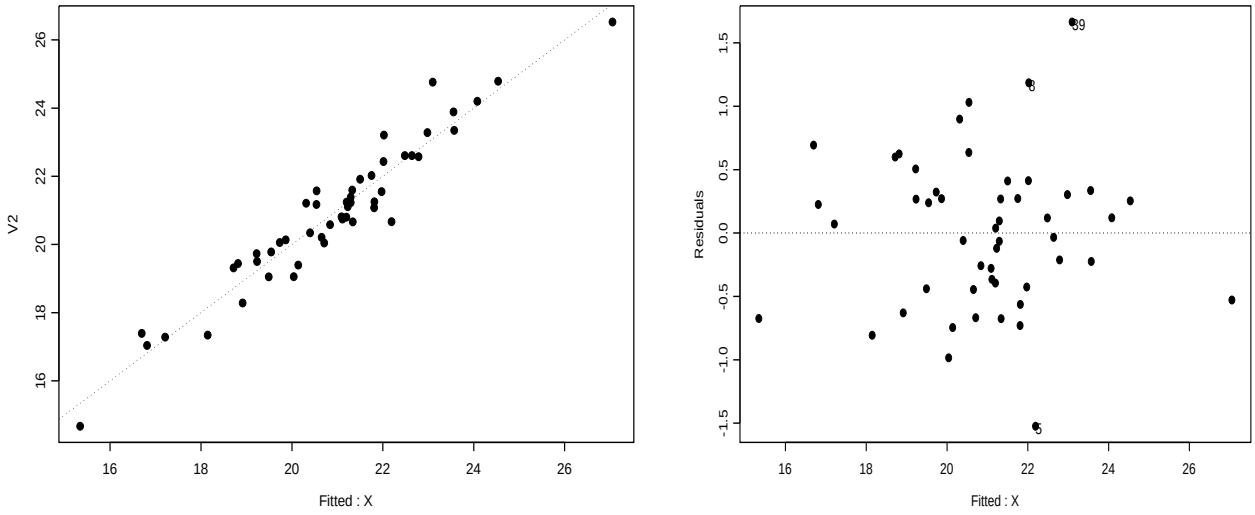


Figura 2.13: *Nube di punti, modello di regressione e diagramma dei corrispondenti residui (in condizioni di omoschedasticità)*

2.9.1 Diagrammi a dispersione

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:

1. diagramma a dispersione dei residui rispetto al modello adattato $\hat{y}$;
2. diagramma a dispersione dei residui rispetto a X;
3. grafici basati su quantili;
4. residui in sequenza temporale (ove nota).

Come rilevato nel paragrafo 3.4, 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 2.13 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 2.14 forniamo un esempio di adattamento lineare in condizioni di eteroschedasticità; anche in questo caso si noti la presenza di tre valori anomali.

In Figura 2.14 forniamo un esempio di adattamento lineare in condizioni di eteroschedasticità; anche in questo caso si noti la presenza di tre valori anomali. Infine in Figura 2.15

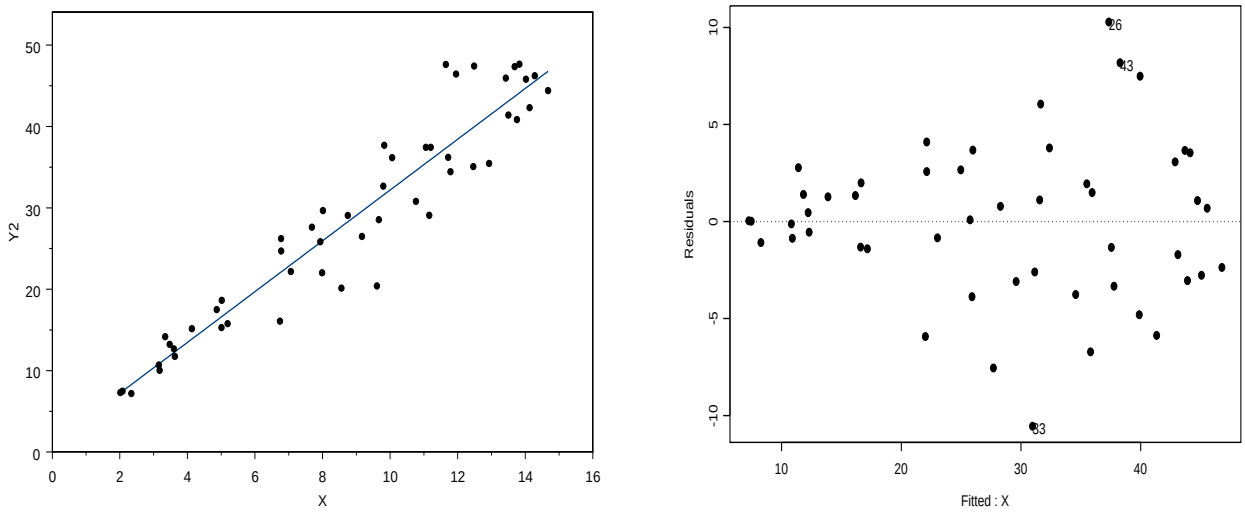


Figura 2.14: *Nube di punti, modello di regressione e diagramma dei corrispondenti residui (in condizioni di eteroschedasticità)*

forniamo un esempio di adattamento lineare nel vaso in cui la relazione fra X e Y è non lineare (si noti l'andamento dei residui in funzione di x).

Esempio 2.7 (Esempio 2.6 (segue)) In accordo ai rispettivi modelli di regressione prima ricavati, in Figura 2.16 riportiamo i diagrammi a dispersione dei residui $(\hat{y}_i, e_i)$.

2.9.2 Grafici basati su quantili

In molti casi, si assume che l'errore ε segua una distribuzione normale. Una semplice ispezione grafica per valutare l'ipotesi di normalità per la distribuzione dei residui è ovviamente basata su istogrammi.

Un altro semplice metodo grafico per una prima verifica di tale assunzione è basato sui quantili della distribuzione normale e viene chiamato *Q-Q plot* (*grafico Quantili-Quantili*.

Sia $w_1, w_2, \ldots, w_n$ una sequenza di valori osservati ed assumiamo che essi siano ordinati in ordine non decrescente: $w_1 \leq w_2 \leq \cdots \leq w_n$. Indichiamo con $F(w_1), F(w_2), \ldots, F(w_n)$ i valori della funzione di ripartizione definita come la frequenza dei valori $\leq w_i$ ($i = 1, \ldots, n$). Ovviamente risulta:

$$F(w_i) = \frac{i}{n} \qquad i = 1, \ldots, n$$

ed in particolare ciascun w_i può interpretarsi come il quantile di ordine i/n.

Nota 2.8 Ai fini del confronto con distribuzioni teoriche, si considera alcune correzioni di continuità, quale ad esempio: una definizione leggermente diversa:

$$F(w_i) = \frac{i - 1/2}{n} \qquad i = 1, \ldots, n\ .$$

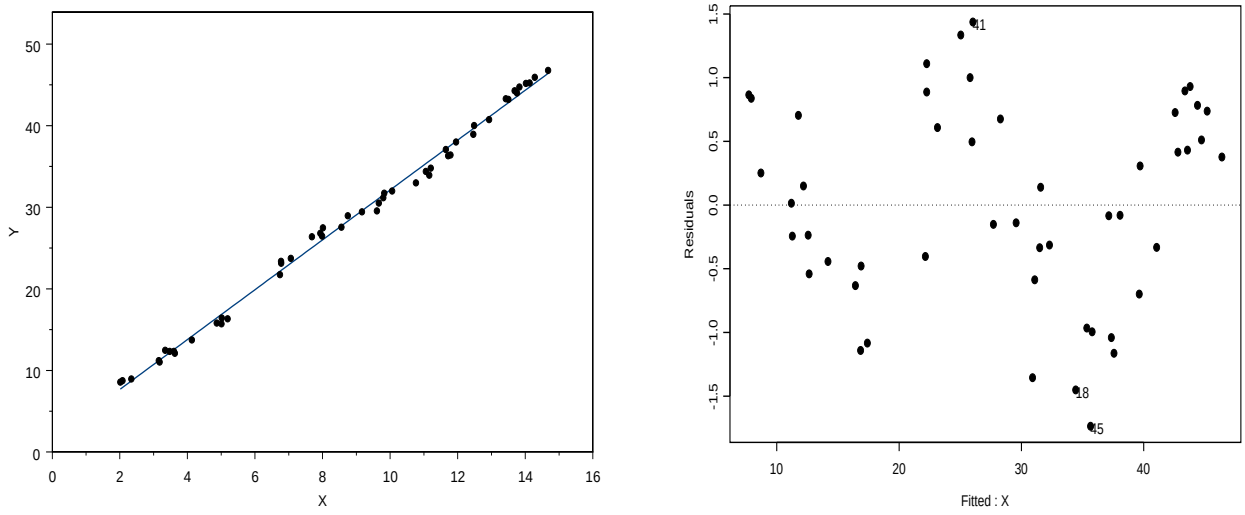


Figura 2.15: *Nube di punti, modello di regressione e diagramma dei corrispondenti residui (in condizioni di non linearità.)*

in particolare nel caso della distribuzione normale standard, per $n \leq 10$ si considera invece

$$F(w_i) = \frac{i - 3/8}{n + 1/4} \qquad i = 1, \ldots, n\,, \quad n \leq 10\,.$$

dovuta alla correzione . Ovviamente se n è molto elevato la correzione ha un'influenza trascurabile sul grafico.

Il diagramma *quantile-quantile*, o più brevemente *Q-Q plot*, consiste nella rappresentazione su uno stesso diagramma cartesiano dei quantili di una distribuzione A contro i corrispondenti quantili di un'altra distribuzione B. Usualmente si considerano i quantili di una distribuzione di dati rispetto ai quantili della distribuzione normale standard.

Tale grafico pone in luce le analogie e le differenze fra le due distribuzioni considerate in termini di forma, di ordine di grandezza e di variabilità. In Figura 2.17 vengono confrontati i diagrammi quantile-quantile per campioni aventi distribuzione normale standard e t-Student con due gradi di libertà. La linea nel grafico unisce il primo ed il terzo quartile ed è prossima alla bisettrice nel caso di distribuzioni normali standard. Le corrispondenti funzioni di densità e di ripartizione sono rappresentate in Figura 2.18 Si noti che la distribuzione t_2 è simmetrica ma presenta delle code più pesanti della distribuzione normale e quindi come ciò viene evidenziato dal *Q-Q plot*.

In Figura 2.19 forniamo un esempio di diagramma quantile-quantile per una distribuzione asimmetrica a destra.

Infine forniamo il diagramma quantile-quantile e l'istogramma per i residui dell'Esempio 2.3 in Figura 2.20. Si noti la leggera asimmetria a sinistra, rilevabile da entrambi i grafici.

Bisogna sottolineare che, comunque, campioni estratti da una distribuzione normale non conducono esattamente ad un andamento lineare, pertanto è necessaria un po' di esperienza per l'interpretazione di tali diagrammi. L'esperienza mostra che nel caso di piccoli campioni

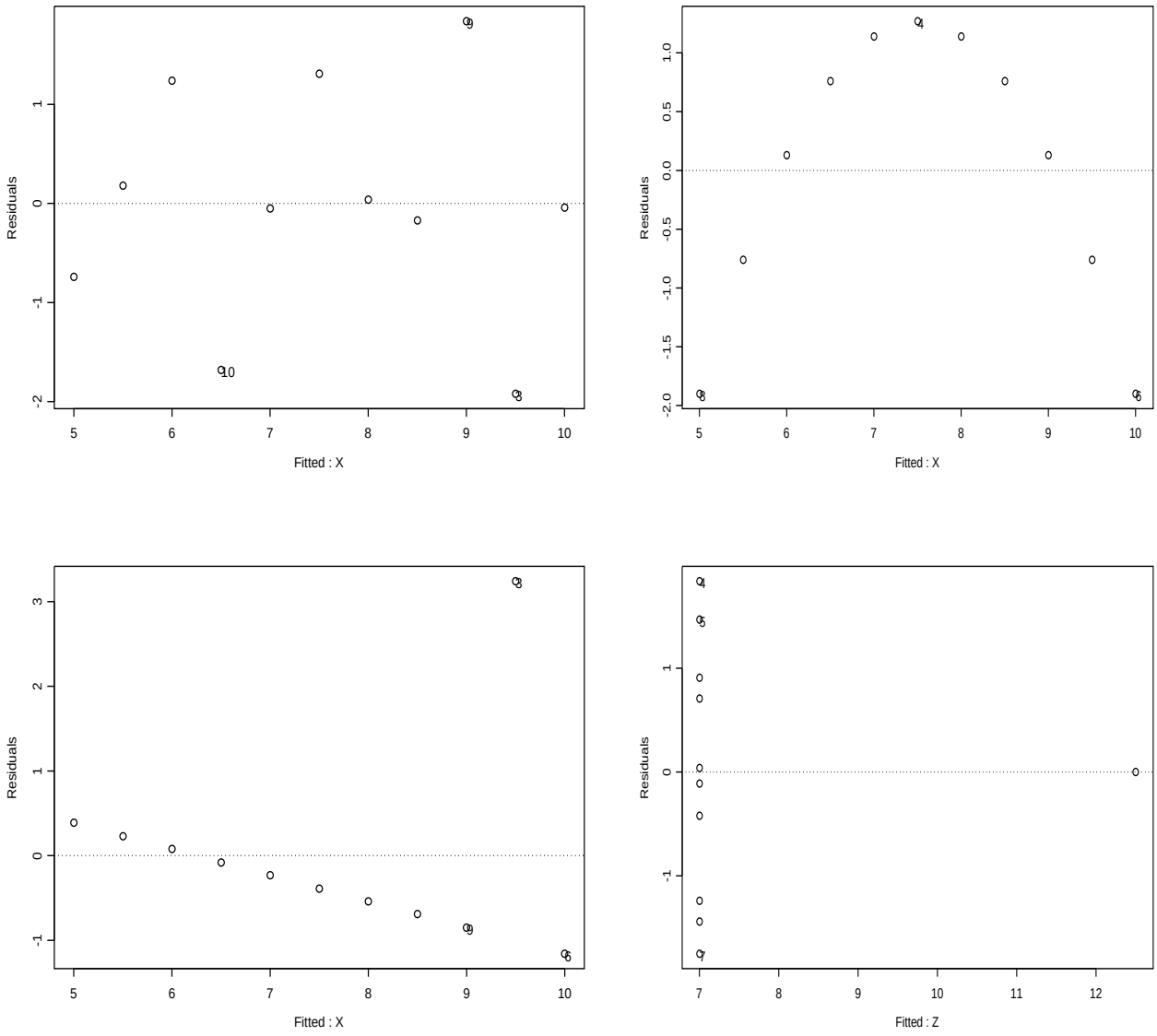


Figura 2.16: *Rappresentazioni grafiche dei residui delle rette di regressione per le coppie di variabili* (X, Y_1), (X, Y_2), (X, Y_3) *e* (Z, W) *del file dati anscombe.dat.*

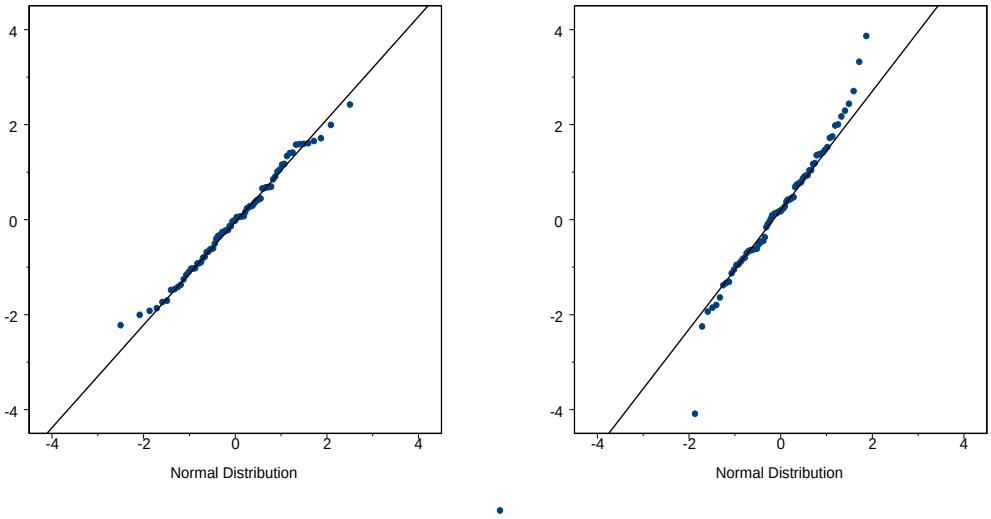


Figura 2.17: *Diagrammi quantile-quantile per distribuzione normale standard (a sinistra) e per distribuzione t_2 (t-Student con 2 gradi di libertà, grafico a destra).*

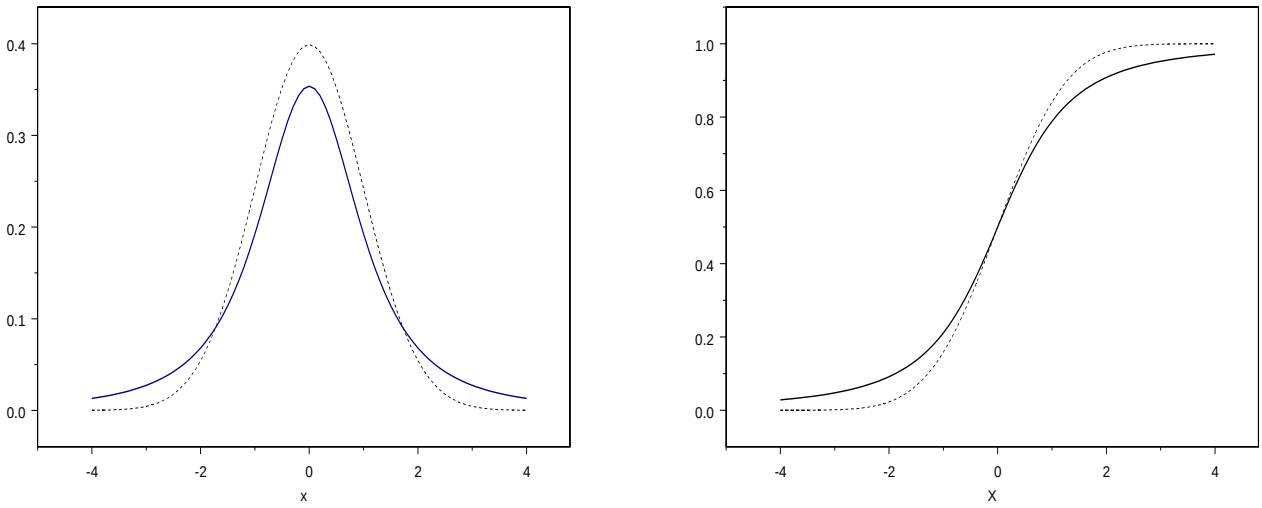


Figura 2.18: *Funzione di densità (a sinistra) e di ripartizione (destra) per le distribuzioni $N(0,1)$ (linea tratteggiata) e per t_2 (linea continua).*

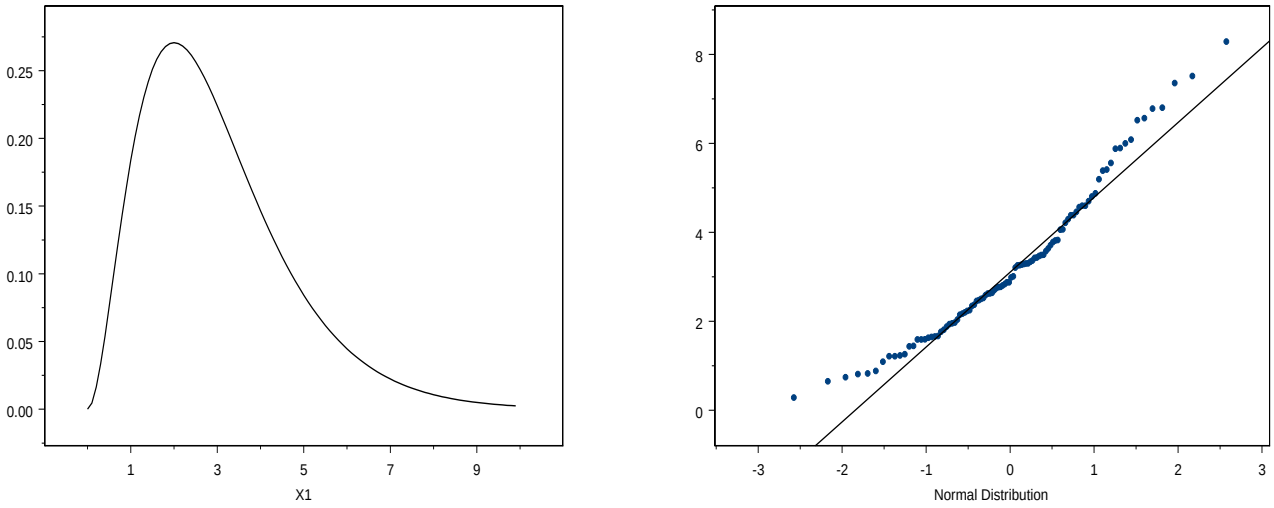


Figura 2.19: *Funzione di densità (a sinistra) e grafico Quantile-Quantile per distribuzione asimmetrica positiva.*

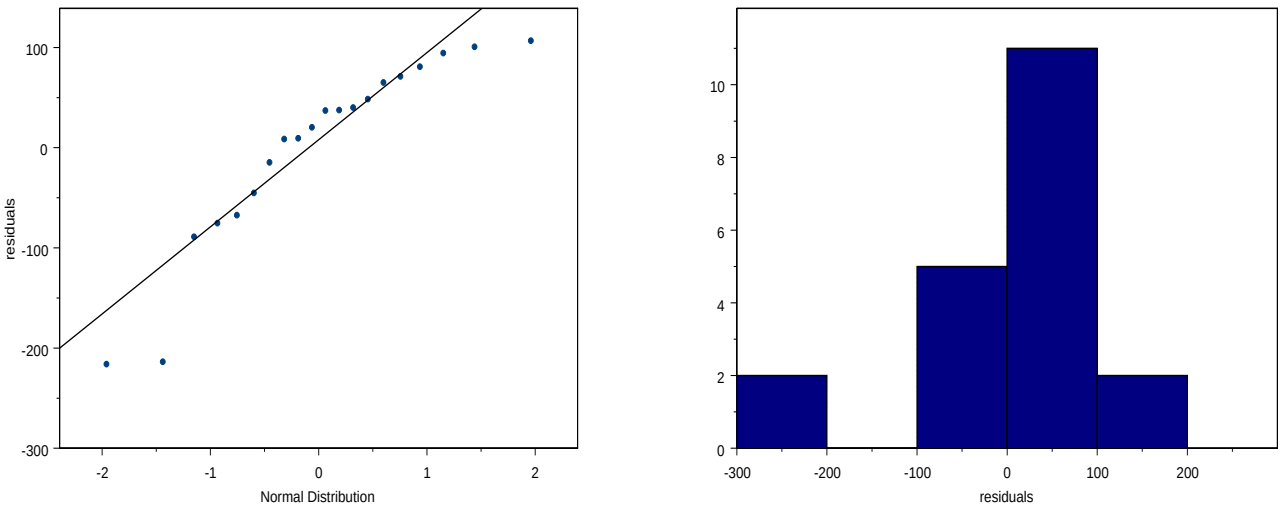


Figura 2.20: *Diagramma Quantile-Quantile e istogramma per la distribuzione dei residui dell'Esempio 2.3.*

($n \leq 16$) estratti con distribuzione normale, spesso si ottengono diagrammi che si allontanano dalla linearità; comportamenti migliori si ottengono in corrispondenza di campioni più ampi ($n \geq 32$). Usualmente si richiede $n \geq 20$ per ottenere un diagramma che sia abbastanza facile da interpretare.

2.10 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 rilevanti sul modello di regressione. Possibili fonti di valori anomali sono:

- *Errori grossolani di misurazione o di registrazione.* Anche se un gruppi di dati può contenere soltanto uno o due valori anomali, il campione potrebbe contenere molte più osservazioni "difettose" che non destano sospetti. Identificando quindi le cause di tali valori anomali, ed in seguito eliminandole nel modo più adeguato, si può ridurre considerevolmente la proporzione di errori nei dati nei campioni futuri. Ciò è particolarmente importante nella pratica del controllo di qualità, dove è consuetudine raccogliere campioni ripetitivi.

- *Ipotesi sbagliata sulla distribuzione.* Un errore comune è quello di ipotizzare che i dati siano distribuiti normalmente quando invece provengono da una distribuzione completamente diversa. Tali dati possono di frequente contenere osservazioni che sembrano valori anomali. Utilizzando il modello statistico giusto di solito si vede che queste osservazioni non sono veri valori anomali, ma semplicemente osservazioni non verosimili per la distribuzione normale. In tali situazioni l'indagine sui valori anomali porta a trovare un modello statistico più corretto e quindi ad inferenze statistiche più appropriate.

- *I dati contengono una struttura più complessa di quella utilizzata.* Per esempio, campioni che si ritengono prelevati casualmente durante la giornata possono essere effettivamente provenire da due distribuzioni diverse, ad esempio, quelli rilevati al mattino e quelli rilevati nel pomeriggio.

- A volte un'*osservazione inusuale* indica semplicemente che tale valore è possibile. Un'attenta indagine sulle condizioni che hanno portato a questa osservazione può essere utile ed importante. Nel caso del controllo di qualità, ad esempio, essa infatti potrebbe rivelarsi di grande aiuto nel migliorare sensibilmente il processo o nel portare alla produzione di prodotti alternativi o di qualità superiore.

In presenza di valori anomali, è pertanto importante valutare se vi sono delle ragioni che possono spiegare tali valori. Come mostrato in precedenza, la presenza di valori anomali ha infatti 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.

Come rilevato sopra, 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.

Come contrassegnare i valori anomali Un modo di contrassegnare in dati anomali è basato sull'utilizzo degli scarti standardizzati. Nell'ipotesi in cui la componente di errire segue una distribuzione normale, cioè $\varepsilon_i \sim N(0, \sigma^2)$, una regola è quella di considerare e_i come dato anomalo se, in valore assoluto, assume un valore maggiore o uguale a $k \cdot s_e$ dove s_e^2 è la stima corretta della varianza di errore σ^2, data da:

$$s^2 := \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} \tag{2.56}$$

e k è un valore usualmente pari a 2 o 3. Questo criterio è molto semplice da costruire ed è implementato in molti software statistici; tuttavia presenta qualche inconveniente.

Il diagramma a scatola. Il *diagramma a scatola* o *box-plot* costituisce un interessante rappresentazione grafica che sintetizza la forma di una distribuzione univariata e consente anche di contrassegnare i valori anomali. Esso si basa sulla mediana ($x_{0.50}$), sul primo e terzo quartile (rispettivamente $x_{0.25}$ e $x_{0.50}$) e sulla differenza interquartilica. Il *boxplot* consente di porre in chiara luce l'ordine di grandezza del fenomeno, la sua dispersione, la simmetria o asimmetria, la lunghezza delle "code" della distribuzione e l'eventuale presenza di valori anomali. Il grafico può essere costruito in senso orizzontale o verticale. Nel primo caso, dopo aver scelto una scala adeguata a rappresentare l'insieme dei valori, si posiziona sull'asse orizzontale la mediana e la si indica con un segmento verticale. A sinistra della medesima si colloca (alla distanza opportuna in base alla scala adottata) il valore del primo quartile ed alla destra della mediana si posiziona il valore del terzo quartile, segnando un segmento verticale in corrispondenza di ciascuno di tali quartili. Si uniscono gli estremi di tali segmenti, formando una scatola rettangolare, come descritto in Figura 2.21.

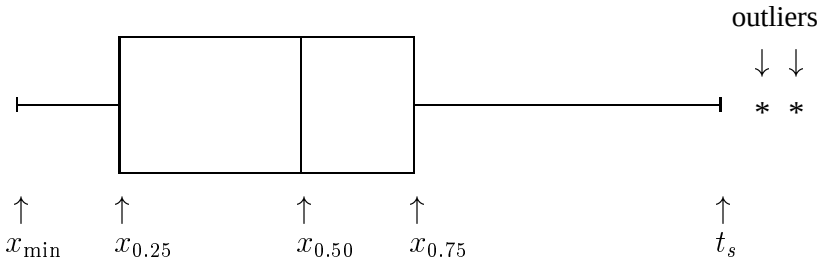


Figura 2.21: *Esempio di diagramma a scatola (box-plot).*

Posto $DI = x_{0.75} - x_{0.25}$ (distanza interquartilica), si individuano quindi:

1. il "punto di troncamento inferiore", denotato con t_i, definito come segue:

$$t_i := \max\{x_{\min}, x_{0.25} - 1.5DI\}\,;$$

2. il ”punto di troncamento superiore”, denotato con t_s, definito come segue:

$$t_s := \min\{x_{\max}, x_{0.75} + 1.5DI\} .$$

Si tracciano quindi a sinistra ed a destra della scatola due segmenti orizzontali che la uniscono rispettivamente ai punti t_i e t_s.

Gli eventuali valori esterni rispetto ai punti di troncamento vengono considerati come possibili valori anomali e sono indicati con asterischi (o altro simbolo) sulla retta in prosecuzione del rispettivo segmento. Bisogna sottolineare che questo criterio di definizione di valore anomalo è in qualche modo arbitrario, comunque l'esperienza effettuata su molti insiemi di dati indica che esso è appropriato per identificare i valori che richiedono speciale attenzione.

Il *boxplot* consente di percepire in via immediata le caratteristiche salienti dei valori assunti da un fenomeno quantitativo nell'insieme delle unità statistiche assegnate:

- la posizione della mediana, cioè $x_{0.50}$;
- la lunghezza della scatola che individua la distanza interquartilica $DI = x_{0.75} - x_{0.50}$;
- i segmenti esterni alla scatola che individuano la lunghezza delle ”code” (ad esclusione degli *outliers*): questi segmenti sono a volte chiamati ”baffi” (*whiskers*) per cui il box-plot viene anche chiamato *box-and-whisker plot*.

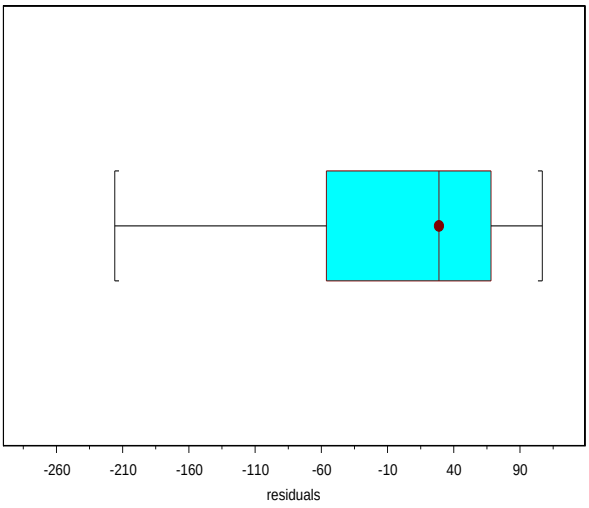


Figura 2.22: *Box plot della distribuzione dei residui per l'Esempio 2.3.*

In Figura 2.22 viene riportato il *box-plot* della distribuzione dei residui ottenuti in base al modello di regressione ricavato nell'Esempio 2.3. In questo caso si ha: $x_{\min} = -215.98$, $x_{0.25} = -50.68$, $x_{0.50} = 28.74$, $x_{0.75} = 66.61$ e $x_{\max} = 117.29$. Essendo $DI = x_{0.75} - x_{0.25} =$

$66.61 - (-50.68) = 106.76$. Segue quindi:

$$\begin{aligned} t_i &= \max\{x_{\min}, x_{0.25} - 1.5DI\} = \max\{-215.98, -50.68 - 1.5 \cdot 117.29\} \\ &= \max\{-215.98, -223.62\} = -215.98 \\ t_s &= \min\{x_{\max}, x_{0.75} + 1.5DI\} = \min\{106.76, 66.61 + 1.5 \cdot 117.29\} \\ &= \min\{106.76, 242.55\} = 106.76\,. \end{aligned}$$

In questo caso, il box-plot non evidenzia valori anomali.