Capitolo 3

Distribuzioni triple

3.1 Introduzione

In questo capitolo consideriamo il caso in cui il nostro dataset è costituito da tre variabili che denotiamo con X_1, X_2, X_3 oppure con Y, X_1, X_2 nel caso in cui si voglia studiare la dipendenza di una delle tre variabili (denotata con Y) rispetto alle altre due.

Una prima analisi esplorativa conduce a calcolare il vettore delle medie, la matrice di varianze e covarianze e la matrice di correlazione. Inoltre una comoda rappresentazione grafica è quella basata sulla *matrice di diagrammi a dispersione* in cui sono rappresentati i grafici a dispersione per tutte le coppie di variabili: (Y, X_1), (Y, X_2) e (X_1, X_2). Questo tipo di rappresentazione grafica mette in luce la dipendenza binaria e può essere utilizzato anche per distribuzioni aventi un numero di variabili superiore a 3.

Le distribuzioni triple possono essere anche rappresentate graficamente mediante istogrammi tridimensionali chiamati *stereogrammi* e/o grafici a dispersione nello spazio $\mathbb{R}^3$, quest'ultimo è un tipo di grafico che ovviamente non può essere costruito per distribuzioni aventi un numero di variabili maggiore di 3.

Esempio 3.1 Consideriamo il seguente dataset[1] dove, con riferimento ad un insieme di aziende, sono state rilevate le seguenti tre variabili:

$$\begin{aligned} Y &:= \text{reddito annuo in migliaia di euro} \\ X_1 &:= \text{superficie in ettari} \\ X_2 &:= \text{numero di bovini.} \end{aligned}$$

I dati sono riportati in Tabella 3.1. Il calcolo delle medie fornisce:

$$\bar{y} = 45.035 \qquad \bar{x}_1 = 13.95 \qquad \bar{x}_2 = 8.85\ ;$$

inoltre la matrice di varianze e covarianze e la matrice di correlazione sono rispettivamente

[1] fi ledati: *aziendeagricole.dat*

Y	X_1	X_2
49.58	6	18
42.87	22	0
65.07	18	14
31.5	8	6
30.47	12	1
46.48	10	9
42.35	17	6
45.45	11	12
44.42	16	7
39.25	23	2
52.68	7	17
55.78	12	15
49.58	24	7
36.15	16	0
41.32	9	12
58.36	11	16
39.25	22	2
38.22	11	6
50.61	16	12
41.32	8	15

Tabella 3.1: Dataset *aziendeagricole.dat* completo.

uguali a:

$$\begin{pmatrix} \sigma_{yy} & \sigma_{y1} & \sigma_{y2} \\ \sigma_{1y} & \sigma_{11} & \sigma_{12} \\ \sigma_{2y} & \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} 72.7523 & 0.3653 & 0.1363 \\ 0.3653 & 30.3475 & -19.7075 \\ 0.1363 & -19.7075 & 33.8275 \end{pmatrix}$$
$$\begin{pmatrix} r_{yy} & r_{y1} & r_{y2} \\ r_{1y} & r_{11} & r_{12} \\ r_{2y} & r_{21} & r_{22} \end{pmatrix} = \begin{pmatrix} 1.0000 & 0.0078 & 0.7083 \\ 0.0078 & 1.0000 & -0.6151 \\ 0.7083 & -0.6151 & 1.0000 \end{pmatrix} \tag{3.1}$$

dove ovviamente $\sigma_{ij} = \sigma_{ji}$ e $r_{ij} = r_{ji}$, per $i, j = y, 1, 2$.

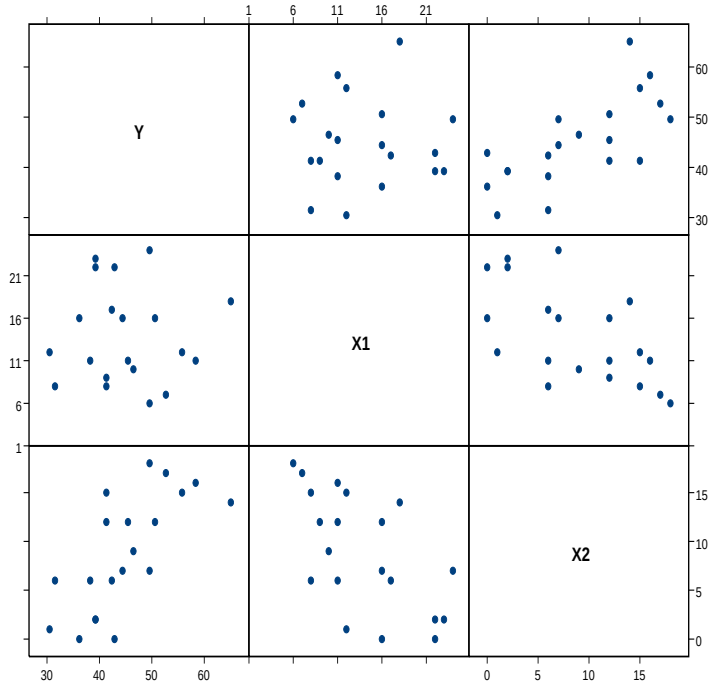
Alcune rappresentazioni grafiche per i i dati in Tabella 3.1 sono riportate in Figura 3.2.

♣

3.2 Regressione per distribuzioni triple

Consideriamo ora il modello di regressione in cui la variabile Y dipenda congiuntamente da due variabili X_1 e X_2:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \ . \tag{3.2}$$



Matrice di diagramma a dispersione

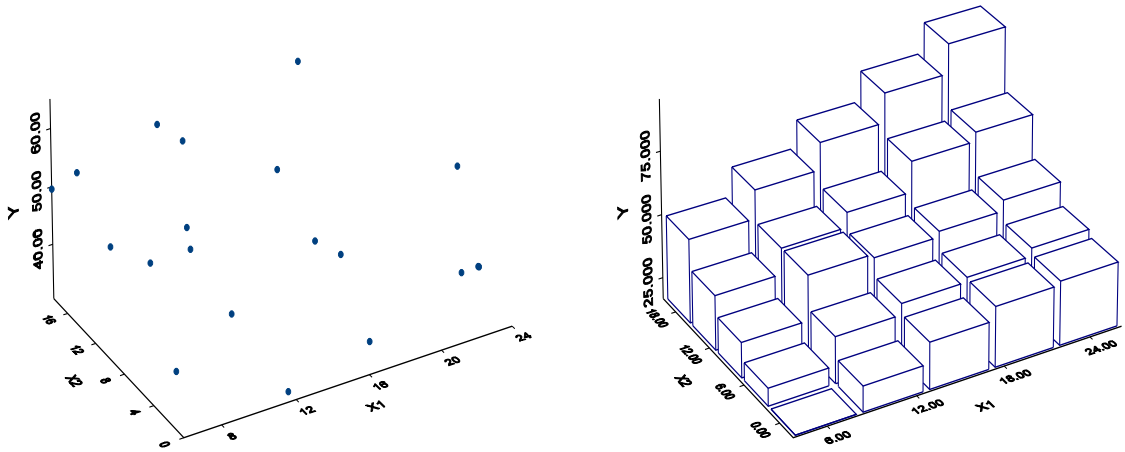


Diagramma a dispersione Stereogramma

Figura 3.1: Rappresentazioni grafiche per i dati in Tabella 3.1.

Assegnato un campione di n terne (x_{1i}, x_{2i}, y_i), $i = 1, \ldots, n$, in base al metodo dei minimi quadrati si ottiene la seguente stima di $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$:

$$\mathbf{b} := \underset{(\beta_0,\beta_1,\beta_2)\in\mathbb{R}}{\arg\min} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2 \ .$$

Considerato il sistema di equazioni normali, analogo a quello (2.13), e risolvendo come in precedenza, si ottengono le seguenti stime (b_0, b_1, b_2) di $(\beta_0, \beta_1, \beta_2)$:

$$\begin{aligned} b_1 &= \frac{\sigma_{1y}\sigma_2^2 - \sigma_{2y}\sigma_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \\ b_2 &= \frac{\sigma_{2y}\sigma_1^2 - \sigma_{1y}\sigma_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \ , \\ b_0 &= \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 \ . \end{aligned} \tag{3.3}$$

Un'ulteriore formulazione di b_1, b_2, utile per sviluppi successivi, si ottiene moltiplicando e dividendo i termini a destra delle (3.3) per $\sigma_y/(\sigma_1^2\sigma_2^2)$:

$$b_1 = \frac{\dfrac{\sigma_{1y}\sigma_2^2}{\sigma_y\sigma_1^2\sigma_2^2} - \dfrac{\sigma_{2y}\sigma_{12}}{\sigma_y\sigma_1^2\sigma_2^2}}{\dfrac{\sigma_1^2\sigma_2^2}{\sigma_1^2\sigma_2^2} - \dfrac{\sigma_{12}^2}{\sigma_1^2\sigma_2^2}}\,\sigma_y = \frac{\dfrac{\sigma_{1y}}{\sigma_y\sigma_1} - \dfrac{\sigma_{2y}}{\sigma_y\sigma_2}\dfrac{\sigma_{12}}{\sigma_1\sigma_2}}{1 - r_{12}^2}\,\frac{\sigma_y}{\sigma_1} = \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}\frac{\sigma_y}{\sigma_1}$$

$$b_2 = \frac{\dfrac{\sigma_{2y}\sigma_2^2}{\sigma_y\sigma_1^2\sigma_2^2} - \dfrac{\sigma_{1y}\sigma_{12}}{\sigma_y\sigma_1^2\sigma_2^2}}{\dfrac{\sigma_1^2\sigma_2^2}{\sigma_1^2\sigma_2^2} - \dfrac{\sigma_{12}^2}{\sigma_1^2\sigma_2^2}}\,\sigma_y = \frac{\dfrac{\sigma_{2y}}{\sigma_y\sigma_2} - \dfrac{\sigma_{1y}}{\sigma_y\sigma_1}\dfrac{\sigma_{12}}{\sigma_1\sigma_2}}{1 - r_{12}^2}\,\frac{\sigma_y}{\sigma_2} = \frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}\frac{\sigma_y}{\sigma_2}$$

dove r_{1y}, r_{2y} e r_{12} sono i coefficienti di correlazione, rispettivamente di (X_1, Y), (X_2, Y) e (X_1, X_2).

Si ottiene pertanto la seguente stima della dipendenza di Y da X_1, X_2:

$$\hat{y} = f(x_1, x_2) = b_0 + b_1 x_1 + b_2 x_2 \tag{3.4}$$

che rappresenta l'equazione di un piano detto *piano regressione* e che passa per il punto di coordinate $(\bar{x}_1, \bar{x}_2, \bar{y})$.

Analogamente a quanto visto nel caso bivariato, si dimostra infine che sono verificate le condizioni:

$$\bar{e} = 0 \qquad \sigma_{1\varepsilon} = 0 \qquad \sigma_{2\varepsilon} = 0 \ ,$$

dove:

$$\sigma_{1\varepsilon} := \frac{1}{n}\sum_i (x_{1i} - \bar{x}_1)(\varepsilon_i - \bar{e}) \qquad\qquad \sigma_{2\varepsilon} := \frac{1}{n}\sum_i (x_{2i} - \bar{x}_2)(\varepsilon_i - \bar{e}) \ .$$

L'interpretazione dei coefficienti b_0, b_1, b_2 è analoga al caso precedente: il coefficiente b_1 misura la variazione media della variabile dipendente Y in corrispondenza di un incremento unitario della variabile X_1 mantenendo costante X_2; analogamente il coefficiente b_2 misura la variazione media della variabile dipendente Y in corrispondenza di un incremento unitario

della variabile X_2 mantenendo costante X_1; b_0 è il valore dell'ordinata all'origine (che potrebbe non avere alcun significato fisico o statistico). E' necessario sottolineare ancora che il modello ricavato ha validità nell'intervallo dei valori utilizzati per la stima dei parametri; nel caso in cui si voglia utilizzare il modello ricavato la stima di Y in corrispondenza di coppie di valori (x_1, x_2) esterni a tale intervallo bisogna procedere con estrema cautela.

3.3 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:

$$R^2 := \frac{\sigma_f^2}{\sigma_y^2} = \frac{\sum_{i=1}^n [f(x_{1i}, x_{2i}) - \bar{y}]^2}{\sum_{i=1}^n [y_i - \bar{y}]^2} \tag{3.5}$$

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 $i = 1, \ldots, n$, cioè se il piano è parallelo al piano X_1X_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 (3.5) sono tutti nulli, cioè quando $y_i = f(x_{1i}, x_{2i})$ per ogni $i = 1, \ldots, n$. 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 $R^2 = 0$ quanto 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 $i = 1, \ldots, n$.

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

$$R^2 = 1 - \frac{\begin{vmatrix} 1 & r_{y1} & r_{y2} \\ r_{1y} & 1 & r_{12} \\ r_{2y} & r_{21} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{12} \\ r_{21} & 1 \end{vmatrix}} = \frac{r_{y1}^2 + r_{y2}^2 - 2r_{12}r_{y2}r_{y1}}{1 - r_{12}^2} \ . \tag{3.6}$$

Ovviamente deve risultare $r_{12} \neq 1$; si tenga inoltre presente che la matrice di correlazione è simmetrica, cioè $r_{ij} = r_{ji}$ per $i, j = y, 1, 2$.

Esempio 3.1 (segue). Considerato il modello di regressione $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, vogliamo calcolare le stime b_0, b_1, b_2 in base al metodo dei minimi quadrati in base alle (3.3), che fornisce: $b_1 = 1.10, b_2 = 1.68$ e $b_0 = 14.74$, da cui la seguente equazione di regressione:

$$\hat{y} = 14.74 + 1.10x_1 + 1.68x_2 \, .$$

Infine dalla (3.6) si ottiene il coefficiente di determinazione $R^2 = 0.8179$. ♣

3.3.1 Miglioramento della bontà di adattamento nel passaggio da una a due variabili esplicative

Vogliamo ora calcolare l'aumento della varianza spiegata che si ha nel caso in cui si passa dal modello $f(x_1) = a_0 + a_1 x_1$ al modello $f(x_1, x_2) = b_0 + b_1 x_1 + b_2 x_2$, cioè si aggiunge una nuova variabile esplicativa al modello lineare semplice. In questo caso le devianze di di errore, indicate rispettivamente con $\mathrm{SS}_e^{(1)}$ e $\mathrm{SS}_e^{(12)}$, sono uguali a:

$$\mathrm{SS}_e^{(1)} = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_{1i})^2 \qquad\qquad \mathrm{SS}_e^{(12)} = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i})^2 \, .$$

che risultano uguale se e solo se $b_2 = 0$. In generale si ha $\mathrm{SS}_e^{(12)} \leq \mathrm{SS}_e^{(1)}$.

Proposizione 3.2 Siano $f(x_1) = a_0 + a_1 x_1$ e $f(x_1, x_2) = b_0 + b_1 x_1 + b_2 x_2$ i modelli di regressione ottenuti per uno stesso insieme di dati $\{(y_i, x_{1i}, x_{2i})\}_{i=1,\ldots,n}$ in base al metodo dei minimi quadrati. Si ha allora:

$$\sum_{i=1}^{n} (y_i - a_0 - a_1 x_{1i})^2 \geq \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i})^2$$

Dimostrazione. Poichè $f(x_1, x_2) = b_0 + b_1 x_1 + b_2 x_2$ è il modello ottenuto in base al metodo dei minimi quadrati, allora risulta

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i})^2 \leq \sum_{i=1}^{n} (y_i - \alpha_0 - \alpha_1 x_{1i} - \alpha_2 x_{2i})^2$$

per ogni scelta di $(\alpha_0, \alpha_1, \alpha_2) \in \mathbb{R}^3$. In particolare la relazione precedente vale per $\alpha_0 = a_0$, $\alpha_1 = a_1$ e $\alpha_2 = 0$. ■

Per valutare il miglioramento che si consegue nel passare dal modello più semplice a quello più complesso, possiamo considerare la differenza fra i due coefficienti di determinazione. Indicati con R^2_{y1} e R^2_{y12} rispettivamente i coefficienti di determinazione dei due modelli di regressione di Y su X_1 e di Y su X_1, X_2, dati rispettivamente in (2.26) e (3.6), possiamo considerare la quantità $R^2_{y12} - R^2_{y1}$, in particolare tenendo conto della (3.6), si scrive:

$$\begin{aligned} R^2_{y12} - R^2_{y1} &= \frac{r^2_{y1} + r^2_{y2} - 2r_{12}r_{y1}r_{y2}}{1 - r^2_{12}} - r^2_{y1} = \frac{r^2_{y1} + r^2_{y2} - 2r_{12}r_{y1}r_{y2} - r^2_{y1}(1 - r^2_{12})}{1 - r^2_{12}} \\ &= \frac{r^2_{y2} - 2r_{12}r_{y1}r_{y2} + r^2_{y1}r^2_{12}}{1 - r^2_{12}} = \frac{(r_{y2} - r_{y1}r_{12})^2}{1 - r^2_{12}} \geq 0 \, . \end{aligned} \tag{3.7}$$

che rappresenta la frazione di variabilità totale che viene spiegata passando dalla retta di regressione al piano di regressione. Si può operare anche con le varianze residue:

$$\frac{\mathrm{SS}_e^{(1)} - \mathrm{SS}_e^{(12)}}{\mathrm{SS}_e^{(1)}} = \frac{R^2_{y12} - R^2_{y1}}{1 - R^2_{y1}} = \frac{(r_{y2} - r_{y1}r_{12})^2}{(1 - r^2_{12})(1 - r^2_{y1})} \tag{3.8}$$

che indica la frazione della varianza residua della retta che viene spiegata nel passaggio dalla retta al piano di regressione.

3.3.2 Coefficienti di regressione totali (o grezzi) e parziali

Se nella (3.4) fissiamo il valore di $X_2 = x_2^\circ$, otteniamo una funzione della sola variabile x_1:

$$\hat{Y} = f(x_1, x_2^\circ) \quad = \quad b_0' + b_1 x_1 \tag{3.9}$$

dove $b_0' = b_0 + b_2 x_2^\circ$. La (3.4), che esprime come varia in media 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* $X_2 = x_2^\circ$. In questo caso scriveremo anche $b_{y1.2}$ anzichè b_1.

Si noti che facendo assumere a X_2 vari valori, si ottengono altrettante rette di regressione parziale di Y su X_1 che differiscono per il termine noto; esse pertanto sono parallele. Il termine $b_{y1.2}$ viene chiamato *coefficiente di regressione di* Y *rispetto a* X_1 *tenendo costante* X_2.

Vogliamo ora esaminare la relazione fra il coefficiente angolare della retta dei minimi quadrati con quelli del piano interpolatore. Per evitare confusione riscriviamo i parametri in maniera diversa:

$$\hat{y} = f(x_1, x_2) = b_0 + b_{y1.2}x_1 + b_{y2.1}x_2 \qquad \text{utilizzando le variabili} X_1, X_2 \tag{3.10}$$

$$\hat{y} = f(x_1) = b_0 + b_{y1}x_1 \qquad \text{utilizzando solo la variabile} X_1 \tag{3.11}$$

$$\hat{y} = f(x_2) = b_0 + b_{y2}x_2 \qquad \text{utilizzando solo la variabile} X_2\ .$$

Il parametro $b_{y1.2}$ del piano (3.10) indica il la variazione di $\hat{Y}$ che si ha in corrispondenza di un incremento unitario di X_1 nell'ipotesi in cui X_2 rimanga costante; il parametro b_{y1} nella retta (3.11) indica la variazione di $\hat{Y}$ in corrispondenza di un incremento unitario di X_1 (al lordo delle variazioni di X_2). Analoghi significati hanno $b_{y2.1}$ e b_{y2}. Bisogna considerare che quando si interpola Y mediante una retta del tipo $a+bx_1$ solitamente al variare di X_1 varia anche X_2 (nel caso in cui X_1 e X_2 siano correlati) e quindi le variazioni del valore Y dipendono sia da X_1 che da X_2. Per tale motivo, b_{y1} viene anche chiamato *coefficiente di regressione grezzo o totale*. Fra i coefficienti di regressione totali b_{y1}, b_{y2} e quelli parziali $b_{y1.2}, b_{y2.1}$ possono sussistere notevoli differenze.

I valori $b_{y1.2}$ e $b_{y2.1}$ coincidono numericamente rispettivamente con b_1 e b_2 ma hanno significato statistico un po' diverso. Infatti, facendo assumere a X_2 valori diversi, si ottengono altrettante rette di regressione di Y su X_1 (ed analogamente di X_1 su Y) che differiscono per l'ordinata all'origine (termine noto) ma che hanno sempre lo stesso valore di $b_{y1.2}$ (risp. $b_{1y.2}$), cioè lo stesso coefficiente angolare e pertanto sono parallele. Si intende per coefficiente di regressione parziale di Y su X_1 tenendo costante X_3 proprio questo valore comune.

Per ricavare la relazione fra i coefficienti di regressione parziali e quelli totali, dividiamo e

moltiplichiamo i coefficienti di regressione parziali per $\sigma_1^2\sigma_2^2$; dalle (3.3) otteniamo pertanto:

$$b_{y1.2} = \frac{\frac{\sigma_{1y}\sigma_2^2}{\sigma_1^2\sigma_2^2} - \frac{\sigma_{2y}\sigma_{12}}{\sigma_1^2\sigma_2^2}}{\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2\sigma_2^2} - \frac{\sigma_{12}^2}{\sigma_1^2\sigma_2^2}} = \frac{\frac{\sigma_{1y}}{\sigma_1^2} - \frac{\sigma_{2y}}{\sigma_2\sigma_y}\frac{\sigma_{12}}{\sigma_1\sigma_2}\frac{\sigma_y}{\sigma_1}}{\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2\sigma_2^2} - \frac{\sigma_{12}^2}{\sigma_1^2\sigma_2^2}} = \frac{b_{y1} - r_{y2}r_{12}\frac{\sigma_y}{\sigma_1}}{1 - r_{12}^2}$$

$$b_{y2.1} = \frac{\sigma_{2y}\sigma_1^2 - \sigma_{1y}\sigma_{12}}{\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2\sigma_2^2} - \frac{\sigma_{12}^2}{\sigma_1^2\sigma_2^2}} = \frac{\frac{\sigma_{2y}}{\sigma_1^2} - \frac{\sigma_{1y}}{\sigma_1\sigma_y}\frac{\sigma_{12}}{\sigma_1\sigma_2}\frac{\sigma_y}{\sigma_2}}{\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2\sigma_2^2} - \frac{\sigma_{12}^2}{\sigma_1^2\sigma_2^2}} = \frac{b_{y2} - r_{y1}r_{12}\frac{\sigma_y}{\sigma_2}}{1 - r_{12}^2} .$$

In particolare, se le variabili esplicative sono incorrelate, e quindi risulta $r_{12} = 0$, allora dalle relazioni precedenti segue che $b_{y1.2} = b_{y1}$ e $b_{y2.1} = b_{y2}$; negli altri casi ci possono essere rilevanti differenze fra i coefficienti di regressione parziale e totale.

Esempio 3.3 (Esempio 3.1 (segue)) Consideriamo ancora il dataset *aziendeagricole.dat* per cui si è ottenuto il modello di regressione $\hat{y} = 14.74 + 1.10x_1 + 1.68x_2$. In base alle (2.15), (2.16) costruiamo i modelli di regressione di Y su X_1 e di Y su X_2, ottenendo rispettivamente:

$$\hat{y} = 44.87 + 0.012x_1$$
$$\hat{y} = 35.84 + 1.039x_2$$

Si nota che i coefficienti angolari delle due rette di regressione sono abbastanza diversi dai valori dei coefficienti di regressione parziale. In particolare, dalla retta $\hat{y} = 44.87 + 0.012x_1$ sembrerebbe che la variabile X_1 (superficie aziendale) non abbia alcuna influenza sul reddito aziendale. La questione comincia a spiegarsi se si considera il coefficiente di correlazione fra X_1 e X_2 che risulta uguale a $r_{12} = -0.6151$: ciò vuol dire che le variabili X_1 e X_2 sono discordanti, cioè al crescere di una l'altra decresce e quindi hanno un effetto contrapposto su Y.

Il valore del coefficiente di regressione parziale $b_{y1.2}$ è positivo, ciò vuol dire che, tenendo costante la variabile X_2 (numero di bovini), il reddito aziendale Y dovrebbe aumentare con la superficie aziendale X_1. In effetti, se consideriamo le unità che presentano almeno due valori uguali di X_2, otteniamo i dati riportati in Tabella 3.2. si vede che Y cresce all'aumentare di X_1, in corrispondenza dello stesso valore di X_2. Analogamente, essendo $b_{y2.1} = 1.68$ positivo si ottengono i dati riportati in Tabella 3.3.

I diagrammi delle due distribuzioni di Y su X_1 condizionatamente ai diversi valori di per X_2 e di di Y su X_2 condizionatamente ai diversi valori di per X_1 sono rappresentate graficamente in Figura

Nel complesso delle 20 aziende però le due variabili esplicative hanno un andamento contrapposto e, quindi, tenuto conto dei valori positivi di $b_{y1.2}$ e $b_{y2.1}$, anche il reddito è influenzato da due forze contrapposte: la prima tende a far aumentare Y e la seconda, dovuta alla contemporanea diminuzione di X_2, che tende a far diminuire Y. Consegue che all'aumento di X_1 si contrappone una quasi stazionarietà di Y, come si vede essendo $b_{y1} = 0.012$. Quest'esempio mostra che va usata molta cautela nell'interpretazione dei coefficienti totali di regressione. ♣

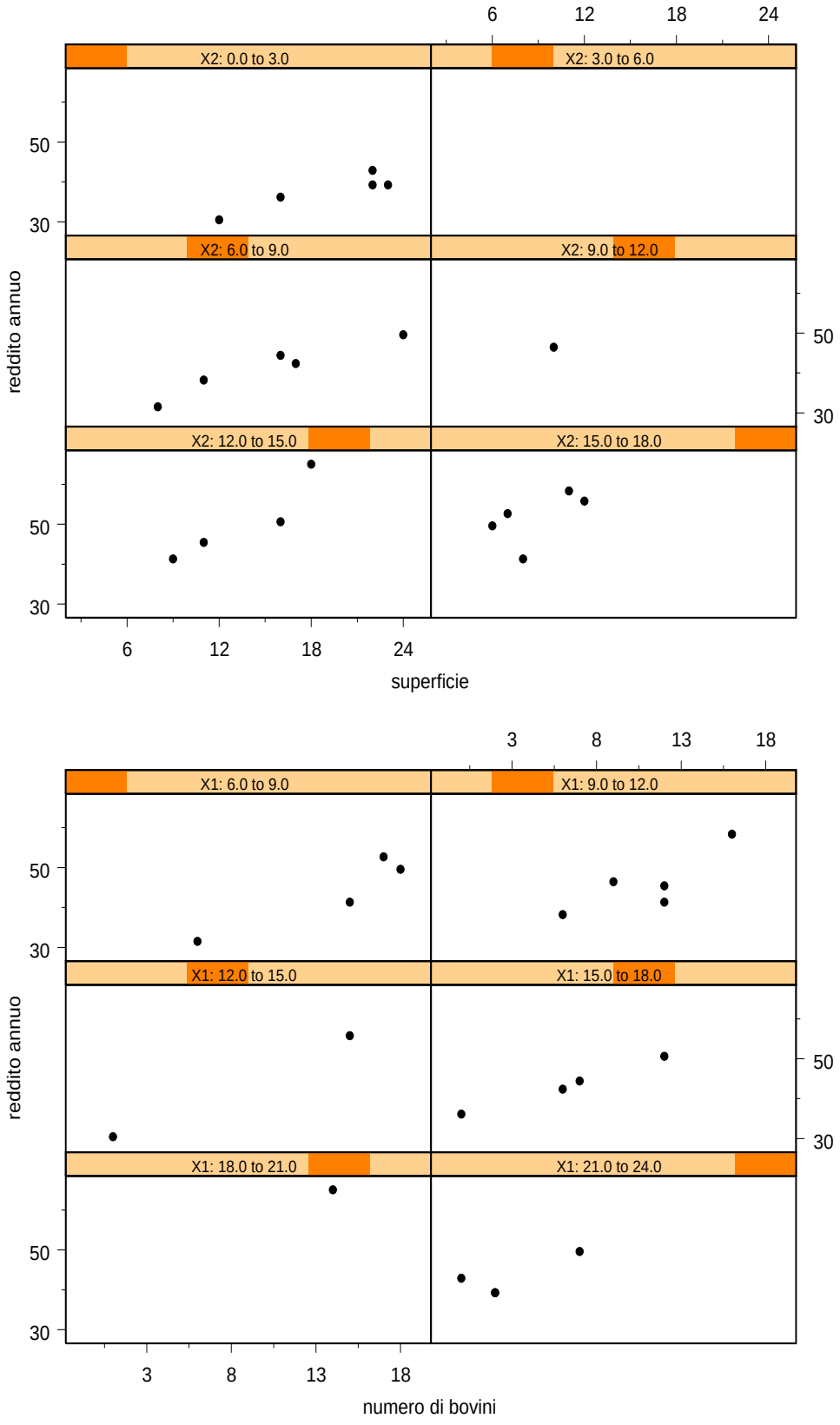


Figura 3.2: Rappresentazioni grafiche delle distribuzioni condizionate per i dati in Tabella 3.1 per X_2 costante (in alto) e X_1 costante (in basso) .

Y	X_1	X_2
31.5	8	6
38.22	11	6
42.35	17	6
44.42	16	7
49.58	24	7
41.32	9	12
45.45	11	12
50.61	16	12
41.32	8	15
55.78	12	15

Tabella 3.2: Distribuzione condizionata di Y su X_1 per X_2 (numero di bovini) costante.

Y	X_1	X_2
38.22	11	6
45.45	11	12
58.36	11	16
30.47	12	1
55.78	12	15
36.15	16	0
44.42	16	7
50.61	16	12
42.87	22	0
39.25	22	2

Tabella 3.3: Distribuzione condizionata di Y su X_2 per X_1 (superficie in ettari) costante.

3.4 Correlazione parziale

Da quanto visto in precedenza si evince che b_0, b_1, b_2 dipendono dalla correlazione di Y con X_1 e di Y con X_2, ma anche dalla correlazione fra X_1 e X_2. Soltanto se $\sigma_{12} = 0$, e quindi $r_{12} = 0$, allora $b_{y1.2}$ e $b_{y2.1}$ dipendono esclusivamente dalla correlazione di X_1 con Y e di X_2 con Y, anzi si ha i questo caso dalle (3.3):

$$b_{y1.2} = \frac{\sigma_y}{\sigma_1} r_{1y} = b_{y1} \ , \qquad b_{y2.1} = \frac{\sigma_y}{\sigma_2} r_{2y} = b_{y2} \ .$$

Nelle scienze economiche non è quasi mai possibile tenere fissa una variabile esplicativa (X_1 o X_2) e quindi la correlazione r_{1y} di Y e X_1, ad esempio, risente degli effetti che le variazioni di X_2 hanno su Y e X_1. Tuttavia è possibile pervenire lo stesso ad una valutazione del coefficiente di correlazione parziale di Y e X_1, al netto dell'influenza di X_2, che sarà indicato con $r_{y1.2}$; analogamente per il coefficiente di correlazione parziale di Y e X_2, al netto dell'influenza di X_1, che sarà indicato con $r_{y2.1}$.

Per comprendere il significato del coefficiente di correlazione parziale, preliminarmente ricordiamo che, in base alla relazione (2.27) il coefficiente di regressione lineare fra due variabili Y e X può essere visto come la media geometrica dei due coefficienti di regressione b_{yx} e b_{xy} rispettivamente della retta di regressione di Y su X e di quella di X su Y:

$$r_{xy} = \sqrt{b_{xy} \cdot b_{yx}}$$

Analogamente, nel caso di una distribuzione tripla, $r_{y1.2}$ può essere ricavato come media geometrica di $b_{y1.2}$ e $b_{1y.2}$. Supponiamo di aver determinato l'equazione di regressione di Y rispetto a X_1 e X_2 e quella di X_1 rispetto a Y e X_2:

$$\begin{aligned} \hat{y} &= f(x_1, x_2) = b_{y1.2}x_1 + b_{y2.1}x_2 + b_{y0} \\ \hat{x}_1 &= f(y, x_2) = b_{1y.2}y + b_{12.y}x_2 + b_{20} \ , \end{aligned}$$

dove $b_{y1.2}$ fornisce la variazione media di y rispetto ad una variazione unitaria di x_1 per x_2 fissato, e $b_{1y.2}$ fornisce la variazione media di x_1 per un aumento unitario di y, con x_1 costante.

Si può quindi misurare la concordanza o la discordanza fra Y e X_1, per X_2 costante, con la media geometrica dei due coefficienti di regressione parziale $b_{y1.2}$ e $b_{1y.2}$ che viene chiamata *coefficiente di correlazione parziale fra X_1 e X_2 con X_3 costante*. Si ha pertanto:

$$r_{y1.2} \quad = \quad \sqrt{b_{y1.2} \cdot b_{1y.2}} \ ,$$

dove $b_{y1.2}$ è il coefficiente di regressione parziale di Y rispetto a X_1, fissato X_2 e $b_{1y.2}$ è il coefficiente di regressione parziale di X_1 rispetto a X_2, fissato X_2. Tenendo conto delle relazioni fra coefficiente di regressione e coefficienti di correlazione nella regressione multipla che, come visto nelle (3.3), sono dati rispettivamente da:

$$\begin{aligned} b_{y1.2} &= \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2} \frac{\sigma_y}{\sigma_1} \\ b_{1y.2} &= \frac{r_{1y} - r_{y2}r_{12}}{1 - r_{y2}^2} \frac{\sigma_1}{\sigma_y} \end{aligned}$$

da cui la media geometrica è data da (si tenga conto della simmetria della matrice di correlazione R, per cui $r_{1y} = r_{y1}$):

$$r_{y1.2} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{(1 - r_{y2}^2)(1 - r_{12}^2)}} \ . \tag{3.12}$$

In particolare, se Y e X_1 sono incorrelate con X_2, per cui $r_{y1} = r_{12} = 0$, allora risulta $r_{y1.2} = r_{y1}$. Si noti che il quadrato di $r_{y2.1}$ è uguale alla (3.8) e pertanto misura il grado di miglioramento inteso come riduzione relativa della varianza residua, che si ha nel passaggio dal modello ad una variabile esplicativa (retta di regressione) a due variabili esplicative (piano di regressione):

$$r_{y1.2}^2 = \frac{(r_{y1} - r_{y2}r_{12})^2}{(1 - r_{12}^2)(1 - r_{y2}^2)} = \frac{\mathrm{SS}_e^{(2)} - \mathrm{SS}_e^{(12)}}{\mathrm{SS}_e^{(2)}} = \frac{R_{y12}^2 - R_{y2}^2}{1 - R_{y2}^2} \ .$$

Analogamente si ottiene:

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

e quindi:

$$r_{y2.1}^2 = \frac{(r_{y2} - r_{y1}r_{12})^2}{(1 - r_{y1}^2)(1 - r_{12}^2)} = \frac{\mathrm{SS}_e^{(1)} - \mathrm{SS}_e^{(12)}}{\mathrm{SS}_e^{(1)}} = \frac{R_{y12}^2 - R_{y1}^2}{1 - R_{y1}^2} \ .$$

Esempio 3.4 (Esempio 3.1 (segue)) Calcoliamo i coefficienti di correlazione parziale per i dati in Tabella 3.1 a partire dalla matrice di correlazione (3.1):

$$r_{y1.2} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{(1-r_{y2}^2)(1-r_{12}^2)}} = \frac{0.0078 - 0.7083(-0.6151)}{\sqrt{(1-0.7083^2)(1-0.6151^2)}} = 0.7968$$
$$r_{y2.1} = \frac{r_{y2} - r_{y1}r_{12}}{\sqrt{(1-r_{y1}^2)(1-r_{12}^2)}} = \frac{0.7083 - 0.0078(-0.6151)}{\sqrt{(1-0.0078^2)(1-0.6151^2)}} = 0.9045\,,$$

che evidenziano nei due casi una forte correlazione parziale, come evidenziato anche nei grafici in Figura 3.2. Dal confronto con i valori della correlazione binaria, $r_{y1} = 0.0078$ e $r_{y2} = 0.7083$, si può desumere l'effetto nei due casi della terza variabile di cui abbiamo discusso in precedenza. ♣

Esempio 3.5 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 dati sono riportati nel file *manutenzione.dat*.

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;
5. la proporzione delle variazioni di Y che sono spiegate: solo da X_1; da X_1 e da X_2.

Il problema richiede la stima dei parametri del modello di regressione multipla $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ in base al metodo dei minimi quadrati. In base ai dati rilevati si ottengono le seguenti varianze e covarianze:

$$\sigma_y^2 = 231.38 \qquad \sigma_{y1} = 98.93$$
$$\sigma_1^2 = 45.46 \qquad \sigma_{y2} = 1317.21$$
$$\sigma_2^2 = 9431.28 \qquad \sigma_{12} = 539.70\,.$$

i	Y	X_1	X_2
1	16.68	7	170.69
2	11.50	3	67.06
3	12.03	3	103.63
4	14.88	4	24.38
5	13.75	6	45.72
6	18.11	7	100.58
7	8.00	2	33.53
8	17.83	7	64.01
9	79.24	30	445.01
10	21.50	5	184.40
11	40.33	16	209.70
12	21.00	10	65.53
13	13.50	4	77.72
14	19.75	6	140.82
15	24.00	9	136.55
16	29.00	10	236.52
17	15.35	6	60.96
18	19.00	7	40.23
19	9.50	3	10.97
20	35.10	17	234.70
21	17.90	10	42.67
22	52.32	26	246.89
23	18.75	9	137.16
24	19.83	8	193.55
25	10.75	4	45.72

Tabella 3.4: Dataset *manutenzione.dat*.

e successivamente gli scarti quadratici medi: $\sigma_y = 15.21$, $\sigma_1 = 6.74$ e $\sigma_2 = 97.11$. A questo punto possiamo ottenere le stime dei parametri del modello di regressione:

$$b_0 = 2.3412$$
$$b_1 = 1.6159$$
$$b_2 = 0.0144$$

e pertanto l'equazione di regressione multipla si scrive:

$$\hat{y} = 2.3412 + 1.6159x_1 + 0.0144x_2 \ . \tag{3.13}$$

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:

$$r_{y1.2} = \frac{r_{y1} - r_{y2} \, r_{12}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{y2}^2}} \ . \tag{3.14}$$

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

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

da cui, sostituendo nell'equazione (3.14), si ottiene:

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

In base alla (3.6) possiamo infine calcolare il coefficiente di correlazione multipla di Y su X_1 e X_2:

$$R^2 = 0.96 \, .$$

In particolare il miglioramento che si ottiene aggiungendo la variabile X_2, in base alla (3.7), è pari a:

$$R^2_{y12} - R^2_{y1} = 0.96 - 0.965^2 = 0.029 = 2.9\%$$

cioè la variabile X_2 apporta un aumento del 2.9% della variabilità spiegata. In un'ottica di compromesso fra semplicità del modello e bontà di adattamento, è pertanto preferibile il modello più semplice.

Per quanto riguarda l'analisi dei residui, in Figura 3.3 viene riportato il *box-plot* della distribuzione dei residui relativi al modello di regressione $\hat{y} = 2.3412 + 1.6159x_1 + 0.0144x_2$. In questo caso si ha: $x_{\min} = -5.788$, $x_{0.25} = -0.6626$, $x_{0.50} = 0.4363$, $x_{0.75} = 1.156$ e $x_{\max} = 7.42$. Essendo $DI = x_{0.75} - x_{0.25} = 1.156 - (-0.6626) = 1.8186$ Segue quindi:

$$\begin{aligned} t_i &= \max\{x_{\min}, x_{0.25} - 1.5DI\} = \max\{-5.788, -0.6626 - 1.5 \cdot 1.8186\} \\ &= \max\{-5.788, -3.3905\} = -3.3905 \\ t_s &= \min\{x_{\max}, x_{0.75} + 1.5DI\} = \min\{7.42, 1.156 + 1.5 \cdot 1.8186\} \\ &= \min\{7.42, 3.8839\} = 3.8839 \, . \end{aligned}$$

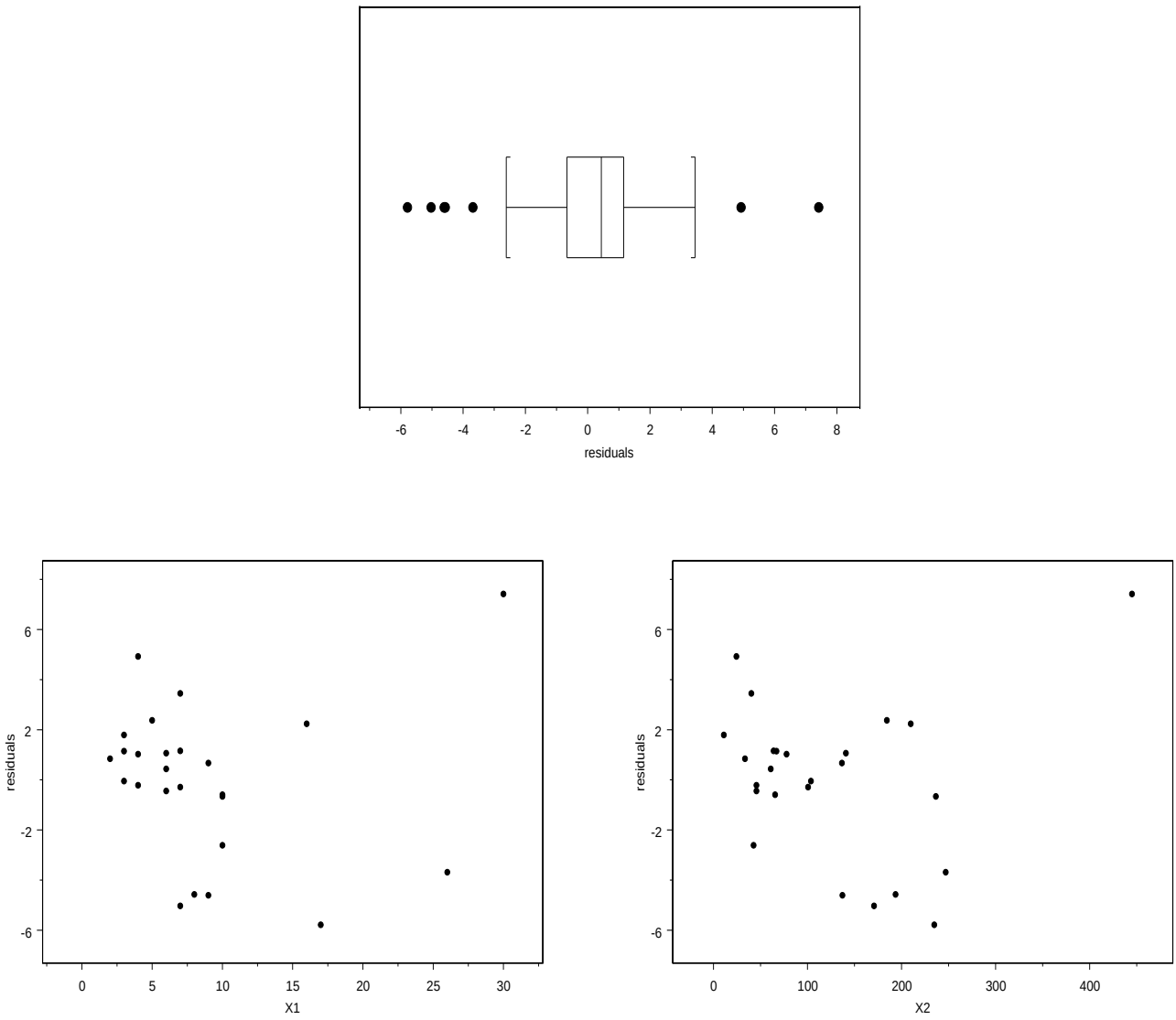


Figura 3.3: *Box plot della distribuzione dei residui e diagrammi dei residui rispetto a ciascun regressore ottenuti in base al modello di regressione* $\hat{y} = 2.3412 + 1.6159x_1 + 0.0144x_2$ *per i dati dell'Esempio 3.5.*

Pertanto i valori che risultano esterni all'intervallo $[t_i, t_s] = [-3.3905, 3.8839]$ sono da considerare anomali (o comunque richiedono ulteriori analisi). Si noti che, nel diagramma in Figura 3.3 i baffi si estendono fino ai valori -2.6142 e 3.4487 anzichè fino a quelli ricavati in precedenza. In effetti i valori della distribuzione dei residui effettivamente nell'intervallo $[-3.3905, 3.8839]$ sono quelli compresi fra -2.6142 e 3.4487.

I diagrammi dei residui rispetto a ciascuno dei due regressori X_1 e X_2 non evidenziano particolari andamenti, salvo alcuni valori anomali rilevati anche dal box-plot. ♣

3.4.1 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.

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:

$$r_{ij,p} := \frac{-R_{ij}}{\sqrt{R_{ii}R_{jj}}} \tag{3.15}$$

dove R_{ij} è il complemento algebrico o aggiunto[2] di r_{ij} nella matrice di correlazione **R**.

Ad esempio, nel caso di tre variabili, dalla (3.15) 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 (3.12).

[2] **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 **A** sopprimendo la i-esima riga e la j-esima colonna di **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:

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

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

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

Ritorniamo alla matrice dei dati dei vini di Bordeaux e calcoliamo $R_{12,34}$. Dalla (1.15) si ha preliminarmente:

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

$$R_{11} = (-1)^{1+1}\begin{vmatrix} r_{22} & r_{23} & r_{24} \\ r_{32} & r_{33} & r_{34} \\ r_{42} & r_{43} & r_{44} \end{vmatrix} = \begin{vmatrix} 1.0000 & 0.6459 & -0.4750 \\ 0.6459 & 1.0000 & -0.4011 \\ -0.4750 & -0.4011 & 1.0000 \end{vmatrix} = 0.4424$$

$$R_{22} = (-1)^{2+2}\begin{vmatrix} r_{11} & r_{13} & r_{14} \\ r_{31} & r_{33} & r_{34} \\ r_{41} & r_{43} & r_{44} \end{vmatrix} = \begin{vmatrix} 1.0000 & 0.8614 & -0.4118 \\ 0.8614 & 1.0000 & -0.4011 \\ -0.4118 & -0.4011 & 1.0000 \end{vmatrix} = 0.2121$$

e quindi:

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

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

$$\mathbf{R}^* = \begin{pmatrix} 1.0000 & 0.3870 & 0.7447 & -0.0188 \\ 0.3870 & 1.0000 & 0.0595 & -0.2774 \\ 0.7447 & 0.0595 & 1.0000 & -0.0797 \\ -0.0188 & -0.2774 & -0.0797 & 1.0000 \end{pmatrix} . \tag{3.16}$$

Dal confronto con la (1.15) si può notare l'influenza delle altre variabili nella correlazione bivariata.

3.5 Trasformazioni di variabili

Finora abbiamo considerato modelli di regressione del tipo $f(x) = \beta_0 + \beta_1 x$ (e più in generale si possono considerare $f(x) = \beta_0 + \beta_1 x + \cdots + \beta_p x_p$); 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:

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

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:

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

Esempio 3.6 Un'impresa vuole investigare l'utilizzo di generatori elettrici basati su energia eolica (mulini a vento). Per condurre tale studio, il settore *Ricerca e Sviluppo* provvede alla raccolta di $n = 25$ coppie di osservazioni relative alle variabili X =”velocità del vento” e Y =”tensione in uscita”[3] Il diagramma a dispersione della distribuzione dei valori rilevati è riportato in Figura 3.4. Il grafico precedente mostra che la relazione fra la tensione in uscita (Y, variabile dipendente) e la velocità del vento (X, variabile indipendente) potrebbe essere non lineare. Inizialmente, comunque, consideriamo un modello lineare semplice $Y = \beta_0 + \beta_1 X + \varepsilon$ ed otteniamo la seguente equazione di regressione:

$$\hat{y} = 0.1309 + 0.2411x$$

che presenta un coefficiente di determinazione $R^2 = 0.8745$ ed inoltre la varianza (corretta) dei residui (2.56) risulta uguale a $s_e^2 = 0.0557$. L'analisi dei residui, vedi Figura 3.5, mostra che il modello precedente è non adeguato in quanto si vede che la relazione ipotizzata non riesce a catturare tutta l'informazione presente nella variabile indipendente. Si noti che la curvatura rilevata nel diagramma a dispersione in Figura 3.4 viene evidenziata maggiormente nel diagramma dei residui.

Consideriamo successivamente un modello del tipo $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$. La stima dei minimi quadrati fornisce:

$$\hat{y} = -1.1559 + 0.7229x - 0.0381x^2$$

con un coefficiente di determinazione $R^2 = 0.9676$ e varianza dei residui pari a $s_e^2 = 0.0150$, vedi Figura 3.6

[3] fi le dati: *windmill.dat*.

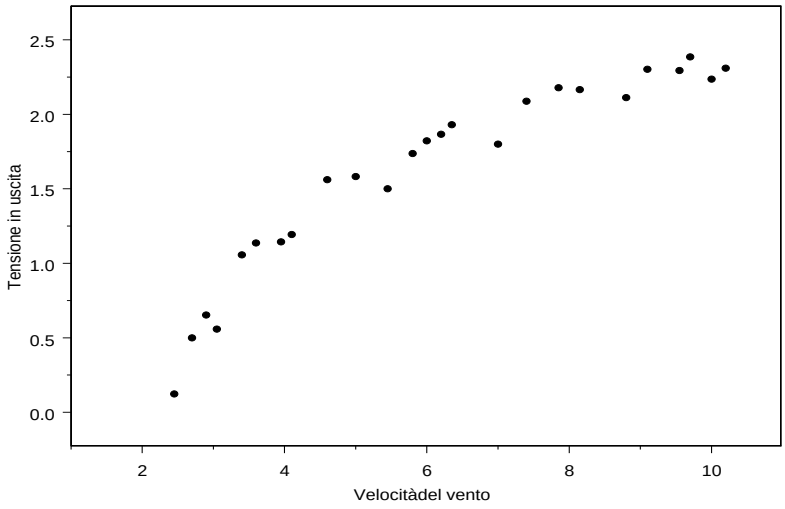


Figura 3.4: *windmill.dat: diagramma a dispersione.*

Comunque la Figura 3.4 suggerisce che al crescere della velocità del vento, la tensione tende ad un limite superiore (che in questo caso risulta pari a circa 2.5). Ciò è anche consistente con la teoria fisica sottostante. Poichè una relazione quadratica tende evidentemente a decrescere per valori di velocità del vento superiori ad una certa soglia (in corrispondenza del massimo della funzione di regressione stimata), una tale relazione risulta inaccettabile dal punto di vista fisico. Un modello più ragionevole potrebbe essere:

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

Posto $x' = 1/x$ la stima del modello precedente fornisce:

$$\hat{y} = 2.9789 - 6.9345 x'$$

per il quale si ha $R^2 = 0.9800$ e $s_e^2 = 0.0089$. In Figura 3.7 viene fornito il diagramma a dispersione di Y rispetto a $1/X$ e il corrisponte modello di regressione.

♣

3.5.1 Un esempio di linearizzazione: il modello di Cobb-Douglas

Consideriamo ad esempio il modello:

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \varepsilon \qquad \beta_0 > 0, X_1 > 0, X_2 > 0 \, , \tag{3.17}$$

che viene anche chiamato *modello di Cobb e Douglas*. Volendo applicare il metodo dei minimi quadrati, si tratta di minimizzare la seguente funzione di errore:

$$\mathcal{E}(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_0 x_{1i}^{\beta_1} x_{2i}^{\beta_2})^2 \tag{3.18}$$

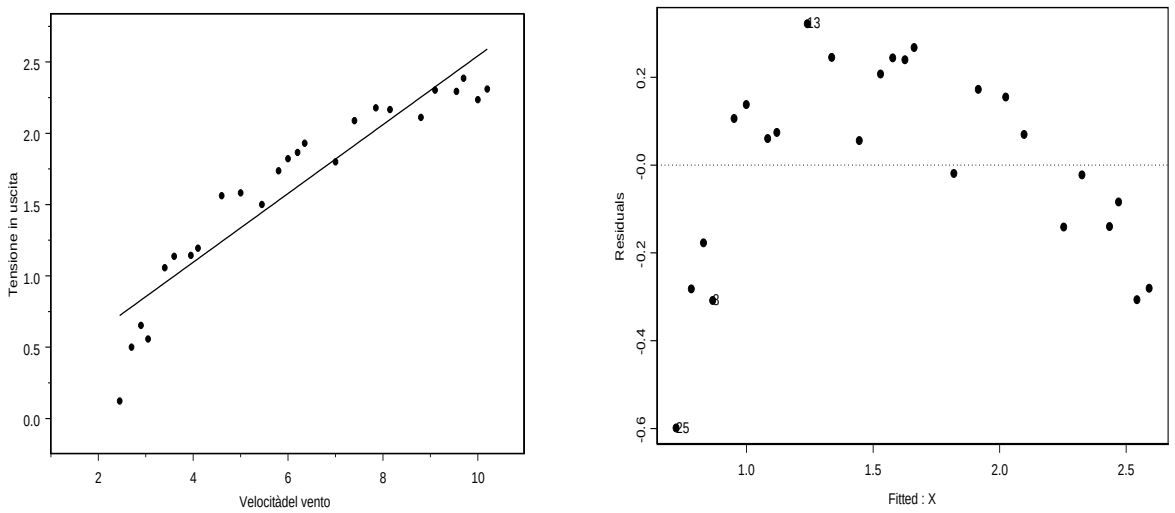


Figura 3.5: *Modello* $\hat{y} = 0.1309 + 0.2411x$ *e corrispondente diagramma* $(\hat{y}_i, e_i)$.

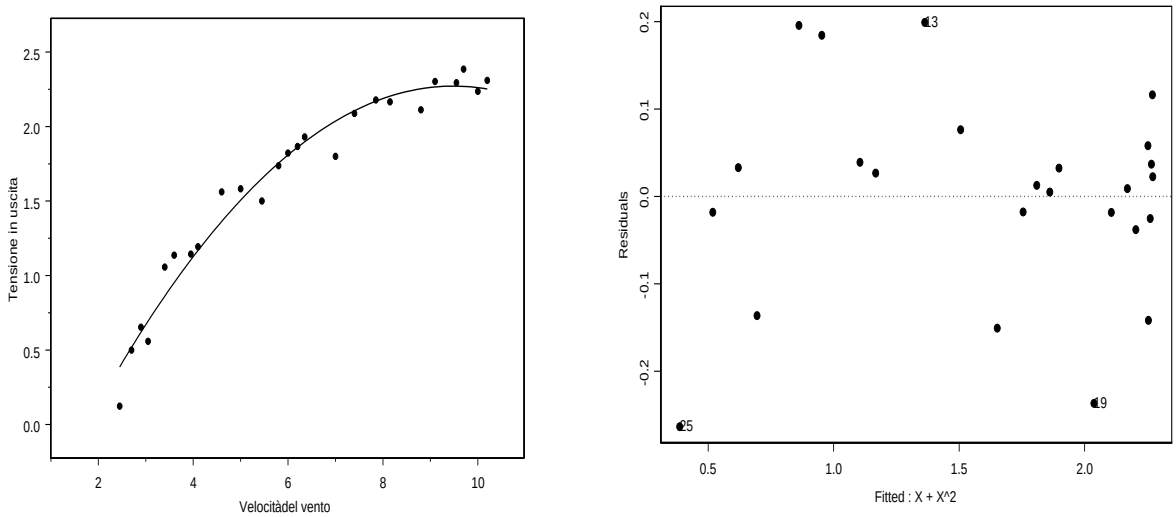


Figura 3.6: *Modello* $\hat{y} = -1.1559 + 0.7229x - 0.0381x^2$ *e corrispondente diagramma* $(\hat{y}_i, e_i)$.

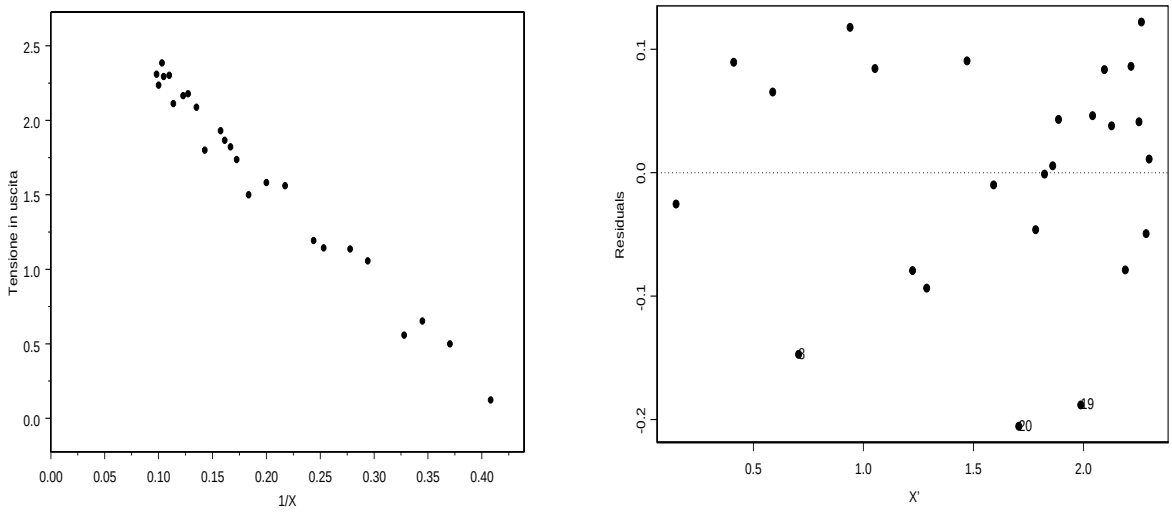


Figura 3.7: *Modello* $\hat{y} = 2.9789 - 6.9345/x$ *e corrispondente diagramma dei residui* $(\hat{y}_i, e_i)$.

rispetto ai parametri precedenti. Il sistema di equazioni normali si scrive:

$$\frac{\partial \mathcal{E}(b_0, b_1, b_2)}{\partial b_0} = \sum_{i=1}^{n} (y_i - b_0 x_{1i}^{b_1} x_{2i}^{b_2}) \cdot x_{1i}^{b_1} x_{2i}^{b_2} = 0$$

$$\frac{\partial \mathcal{E}(b_0, b_1, b_2)}{\partial b_1} = \sum_{i=1}^{n} (y_i - b_0 x_{1i}^{b_1} x_{2i}^{b_2}) \cdot \beta_0 x_{1i}^{b_1} x_{2i}^{b_2} \ln x_{1i} = 0$$

$$\frac{\partial \mathcal{E}(b_0, b_1, b_2)}{\partial b_2} = \sum_{i=1}^{n} (y_i - b_0 x_{1i}^{b_1} x_{2i}^{b_2}) \cdot \beta_0 x_{1i}^{b_1} x_{2i}^{b_2} \ln x_{2i} = 0 \, .$$

Si tratta di un sistema non lineare che può essere risolto utilizzando tecniche di tipo numerico.

Il metodo dei minimi quadrati può invece essere applicato direttamente considerando i logaritmi di ambo i membri della (3.17):

$$\ln Y = \ln \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 \, . \tag{3.19}$$

In questo caso la corrispondente funzione di errore è

$$\mathcal{E}'(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{n} (\ln y_i - \ln \beta_0 - \beta_1 \ln x_{1i} - \beta_2 x_{2i})^2 \, .$$

Ovviamente le due funzioni di errore $\mathcal{E}(\beta_0, \beta_1, \beta_2)$ e $\mathcal{E}'(\beta_0, \beta_1, \beta_2)$, rispettivamente in (3.18) e (3.5.1) non assumono valore minimo in corrispodenza degli stessi valori (b_0, b_1, b_2), tuttavia utilizzando il modello (3.19) si ottiene una buona approssimazione della stima dei minimi quadrati del modello di Cobb-Douglas.

3.5.2 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
$\sigma^2 \propto$ 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 oportunamente 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.

Esempio 3.7 Un'azienda pubblica di produzione di energia elettrica desidera sviluppare un modello per l'analisi della relazione fra i picchi orari di domanda Y e l'energia totale utilizzata in un mese X. Tale problema assuma rilevanza in quanto mentre la maggior parte dei consumatori pagano direttamente per l'utilizzo di energia elettrica (in KWh), il sistema generatore deve essere dimensionato opportunamente per soddisfare i valori massimi di domanda richiesti. Vengono rilevati $n = 53$ coppie di osservazioni relative alle variabili X e Y nel mese di Agosto 1979[4]. I dati vengono riportati in Figura 3.8. Considerato un modello lineare semplice, si ottiene la seguente equazione della retta di regressione:

$$\hat{y} = -0.8313 + 0.00368x$$

con $R^2 = 0.7046$.

[4] fi le dati: *energy.dat*.

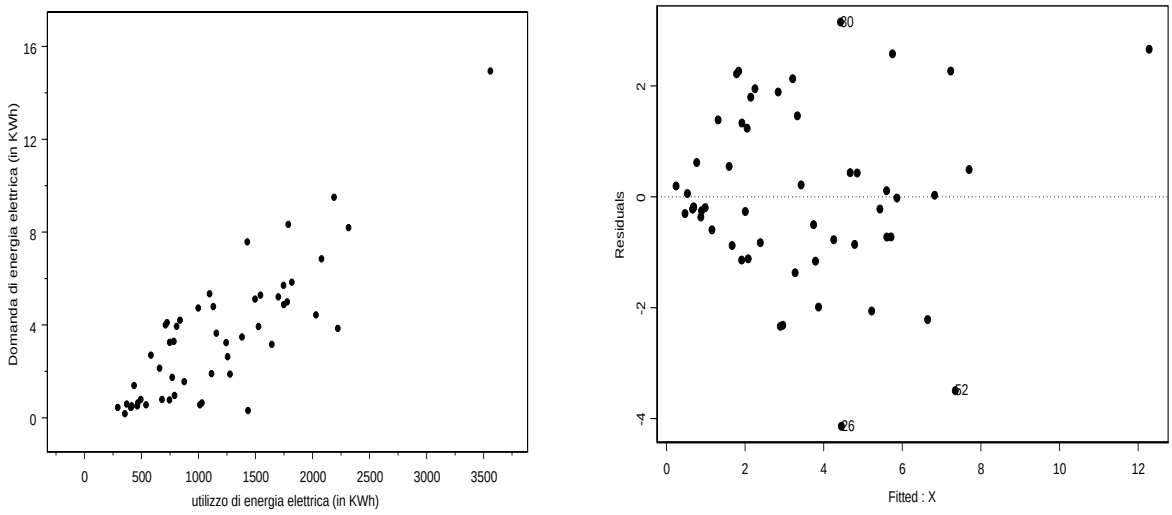


Figura 3.8: *Diagramma a dispersione del file dati energy.dat e corrispondente diagramma dei residui* $(\hat{y}_i, e_i)$ *ottenuto in base al modello* $\hat{y} = -0.8313 + 0.00368x$.

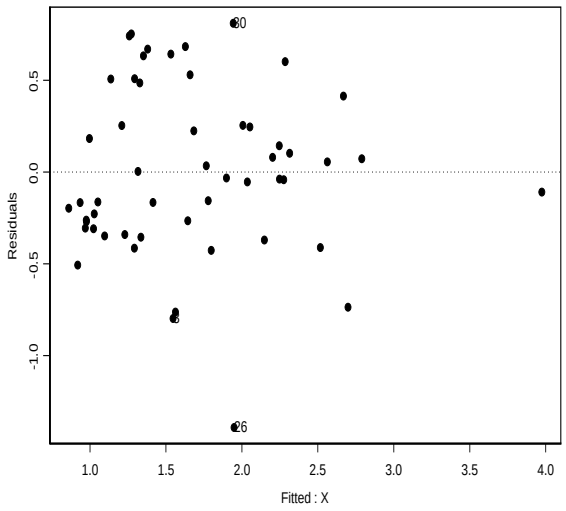


Figura 3.9: *Diagramma dei residui* $(\hat{y}_i, e_i)$ *ottenuto in base al modello* $\hat{y}' = 0.5822 + 0.0009529x$.

Il diagramma dei residui presenta una forma a imbuto che indica che la varianza di errore è crescente all'aumentare del consumo di energia. Il modello risulta pertanto inadeguato ed è necessario adottare una trasformazione stabilizzatrice della varianza. A tal fine, si noti che la variabile risposta Y può essere vista come il numero di Kilowatts utilizzati da un cliente in un'ora particolare. Il più semplice modello probabilistico che può essere adottato per dati di conteggio è la distribuzione di Poisson, e ciò suggerisce di scegliere, quale risposta, la trasformazione $Y' = \sqrt{Y}$ quale trasformazione stabilizzatrice della varianza. In base al metodo dei

minimi quadrati, si ottiene pertanto:

$$\hat{y}' = 0.5822 + 0.0009529x$$

con $R^2 = 0.6485$, leggermente inferiore al precedente modello. In Figura 3.9 viene fornita il grafico a dispersione dei residui $(\hat{y}'_i, e_i)$. L'analisi dei residui mostra che la varianza si mantiene stabile e pertanto si conclude che la trasformazione è adeguata. Si noti la presenza di un cliente il cui corrispondente residuo assume un valore alto ed un altro che presenta un valore grande di energia utilizzata. L'effetto di questi due punti dovrebbe essere ulteriormente valutato prima che il modello venga rilasciato per l'utilizzo sul campo. ♣