

Analisi di regressione: approccio matriciale

Abbiamo rilevato i seguenti caratteri su $n = 25$ unità

Y	X_1	X_2	X_3
10.98	35.3	20	4
11.13	29.7	20	5
12.51	30.8	23	4
8.40	58.8	20	4
9.27	61.4	21	5
8.73	71.3	22	4
6.36	74.4	11	2
8.50	76.7	23	5
7.82	70.7	21	4
9.14	57.5	20	5
8.24	46.4	20	4
12.19	28.9	21	4
11.88	28.1	21	5
9.57	39.1	19	5
10.94	46.8	23	4
9.58	48.5	20	4
10.09	59.3	22	6
8.11	70.0	22	4
6.83	70.0	11	3
8.88	74.5	23	4
7.68	72.1	20	4
8.47	58.1	21	6
8.86	44.6	20	4
10.36	33.4	20	4
11.08	28.6	22	5

Y : libbre di vapore utilizzate in un mese

X_1 : temperatura media mensile in gradi F

X_2 : numero di giorni di operatività in un mese

X_3 : numero di riavviamenti (startup) in un mese

Problema: capire quali variabili e come influiscono sul consumo di vapore

Per capire quali variabili utilizzare nella regressione calcoliamo la **matrice di correlazione**

	Y	X_1	X_2	X_3
Y	1.000	-0.845	0.536	0.382
X_1		1.000	-0.210	-0.237
X_2			1.000	0.601
X_3				1.000

Tale matrice è simmetrica e all'incrocio della riga i e della colonna j c'è il coefficiente di correlazione tra la variabile della riga i e quella della colonna j .
Ad esempio

$$\rho(X_1, X_3) = -0.237$$

Si scelgono le variabili maggiormente correlate con la variabile da spiegare e meno correlate tra loro.

Grafico di dispersione della variabile Y rispetto a X_1

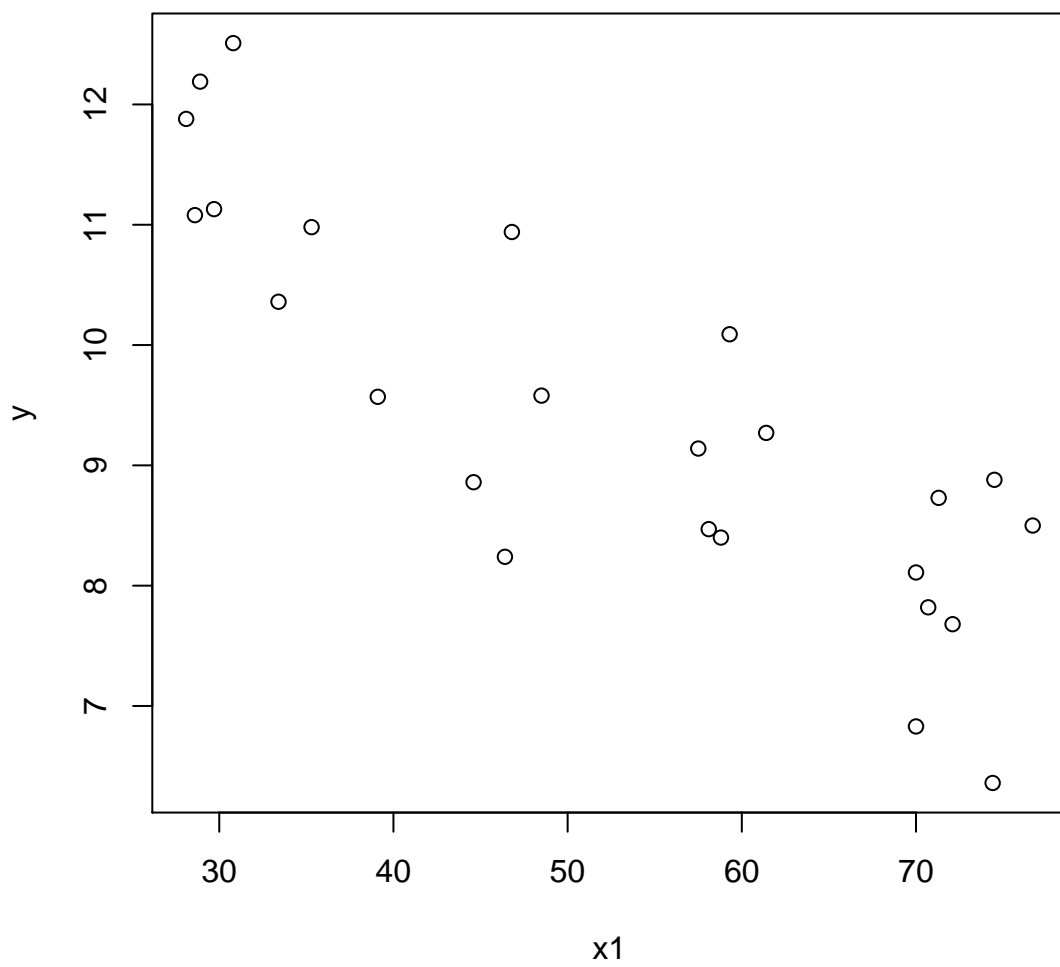


Grafico di dispersione della variabile Y rispetto a X_2

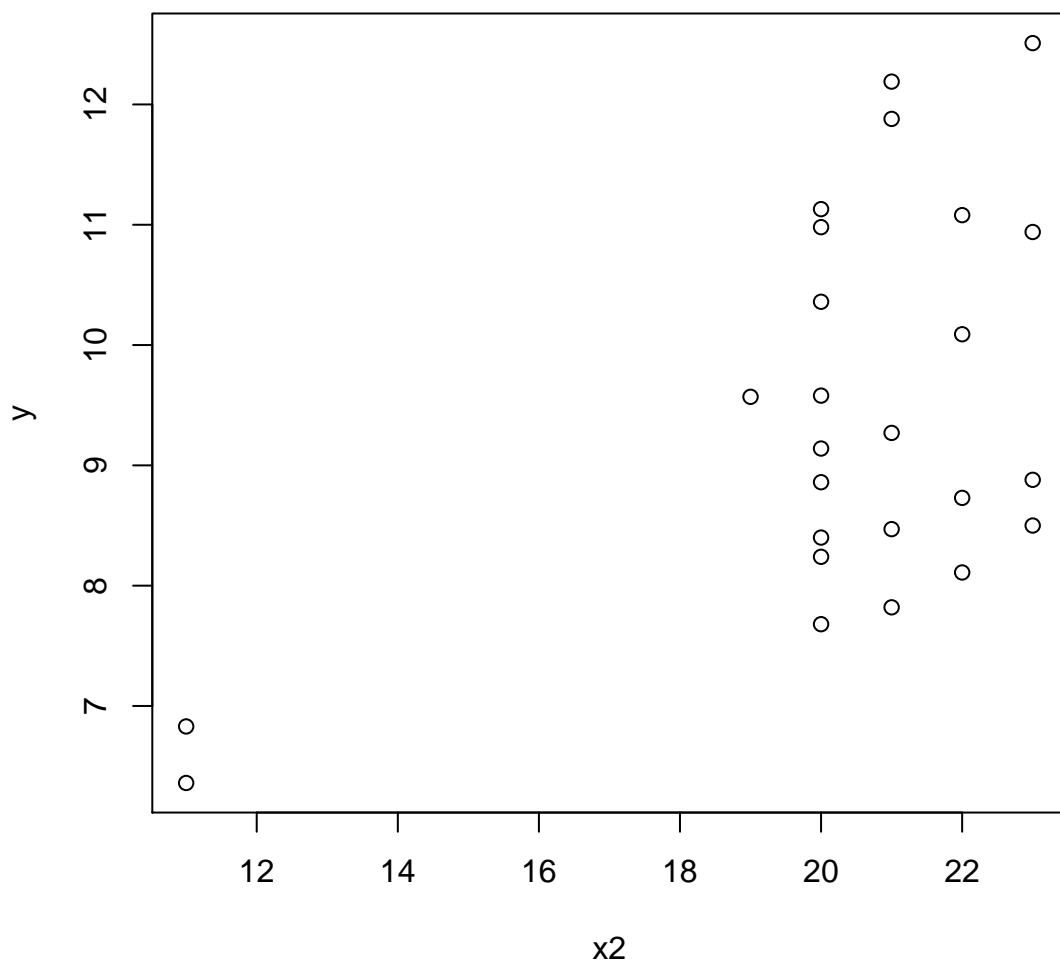
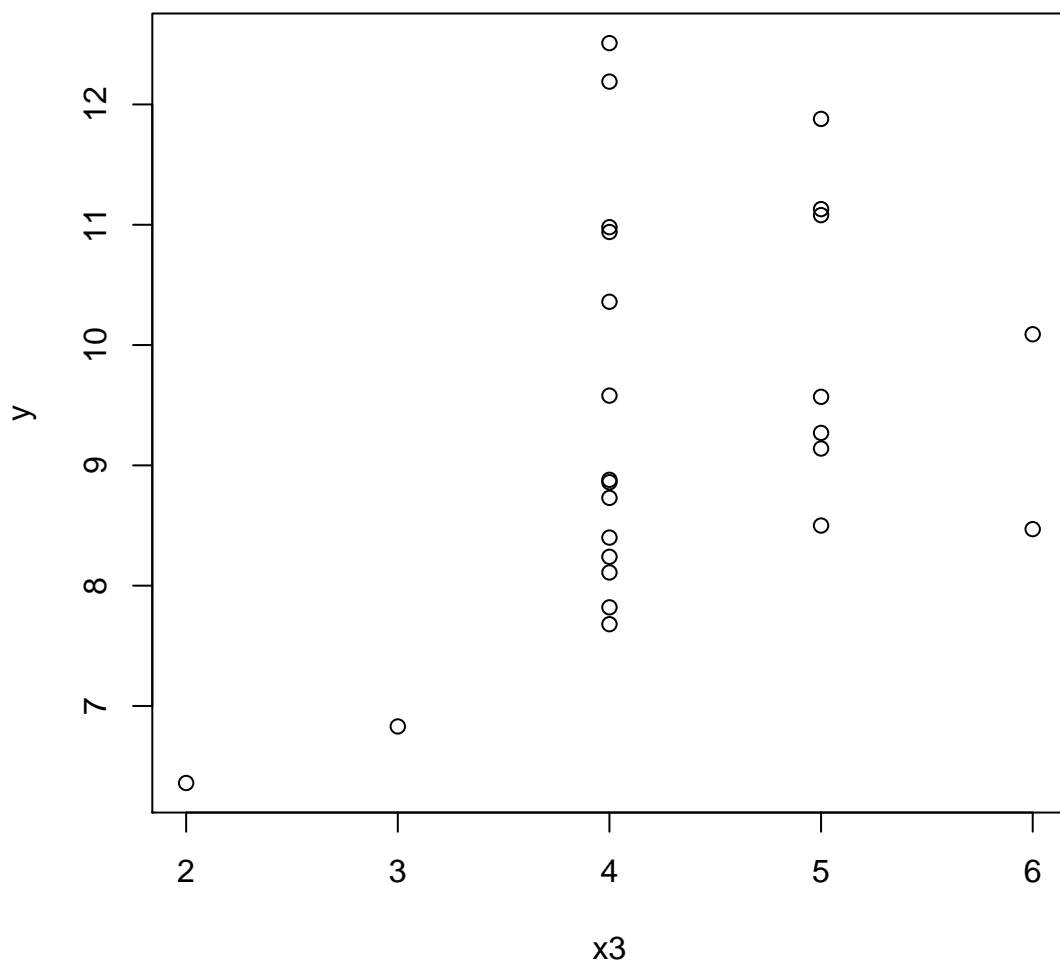


Grafico di dispersione della variabile Y rispetto a X_3



Cominciamo con modello lineare con una variabile esplicativa. Scegliamo la variabile X_1 per la quale abbiamo $\rho(Y, X_1) = -0.845$.

Supponiamo di voler spiegare la variabile Y come funzione della X_1 e che il legame sia lineare

$$Y = a + bX_1$$

Quindi vorremmo determinare a e b tali che

$$y_i^* = a + bx_{1,i}, \quad i = 1, \dots, 25$$

e

$$\sum_{i=1}^{25} (y_i - y_i^*)^2 = \min$$

Facendo i conti ricaviamo $b = \frac{\sigma_{xy}}{\sigma_x^2} = -0.08$ e $a = \bar{y} - b \cdot \bar{x} = 13.6$.

Possiamo riscrivere le 25 equazioni in un'unica equazione matriciale, ovvero

$$\mathbf{y}^* = \mathbf{X}\alpha$$

dove

$$\mathbf{y}^* = \begin{pmatrix} y_1^* \\ \vdots \\ y_{25}^* \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 35.3 \\ 1 & 29.7 \\ \vdots & \vdots \\ 1 & 28.6 \end{pmatrix}, \quad \alpha = \begin{pmatrix} a \\ b \end{pmatrix}$$

determiniamo $\alpha = (a, b)$ minimizzando

$$g(a, b) = (y - y^*)'(y - y^*) = (y - X\alpha)'(y - X\alpha),$$

dove $y = [10.98 \ 11.13 \ \dots \ 11.08]'$. La soluzione dell'equazione matriciale è

$$\alpha = (X'X)^{-1}X'y$$

In questo esempio otteniamo

$$\alpha = \begin{bmatrix} 13.62379 \\ -0.079848 \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix}$$

Si noti che la soluzione coincide con quella già nota. I conti in dettaglio sono:

$$X'X = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \quad X'y = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

$$(X'X)^{-1} = \frac{1}{n \sum (x_i - \bar{x})^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}$$

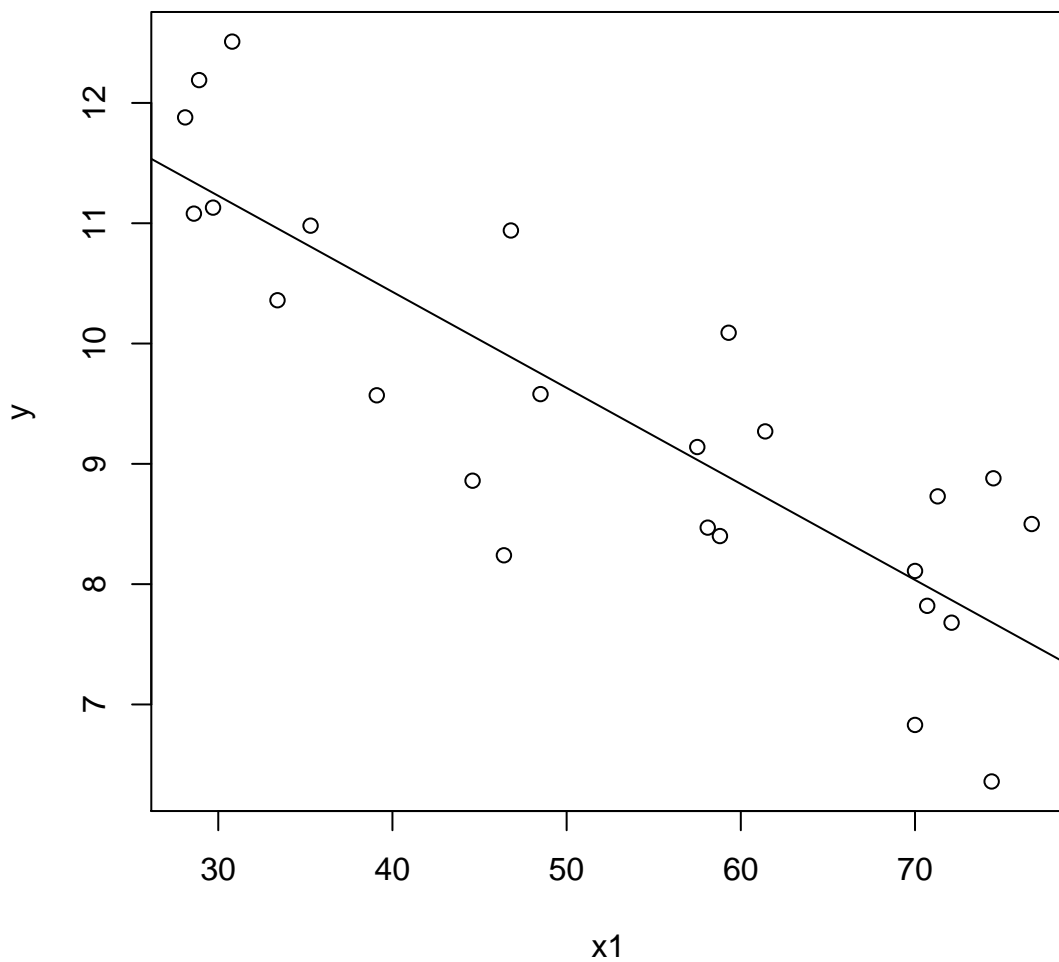
Nell'esempio abbiamo

$$X'X = \begin{bmatrix} 25 & 1315 \\ 1315 & 76323.42 \end{bmatrix} \quad X'y = \begin{bmatrix} 235.6 \\ 11821.4320 \end{bmatrix}$$

$$(X'X)^{-1} = \frac{1}{178860.5} \begin{bmatrix} 76323.42 & -1315 \\ -1315 & 25 \end{bmatrix}$$

Grafico di dispersione della variabile Y rispetto a X_1 con sovrapposta la retta di equazione

$$y = 13.62 - 0.08x_1$$



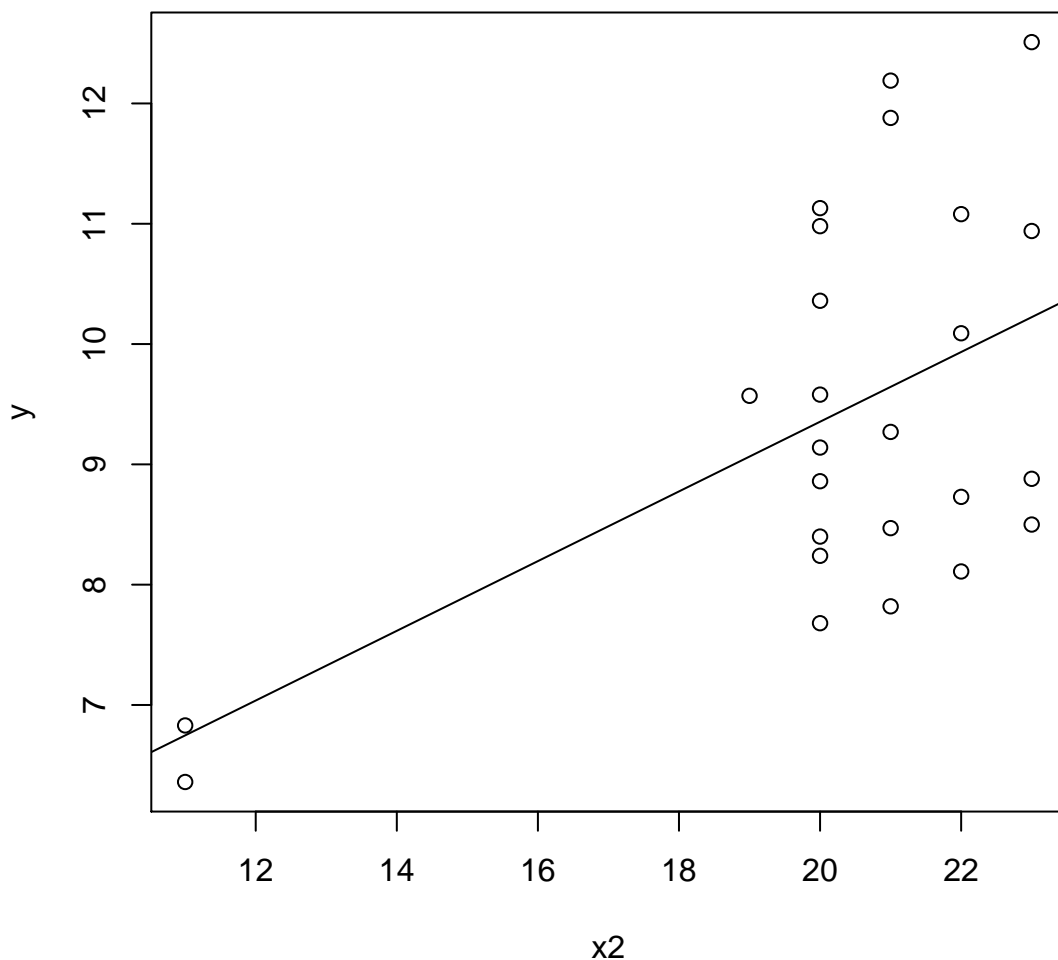
$$R^2 = 0.7144$$

Passiamo alla variabile X_2 per la quale $\rho(Y, X_2) = 0.54$. Supponiamo di voler spiegare la variabile Y in funzione di X_2 . In questo caso abbiamo

$$\alpha = \begin{bmatrix} 3.56055 \\ 0.28970 \end{bmatrix}$$

Grafico di dispersione della variabile Y rispetto a X_2 con sovrapposta la retta di equazione

$$y = 3.56 + 0.29x_2 \quad R^2 = 0.2874$$

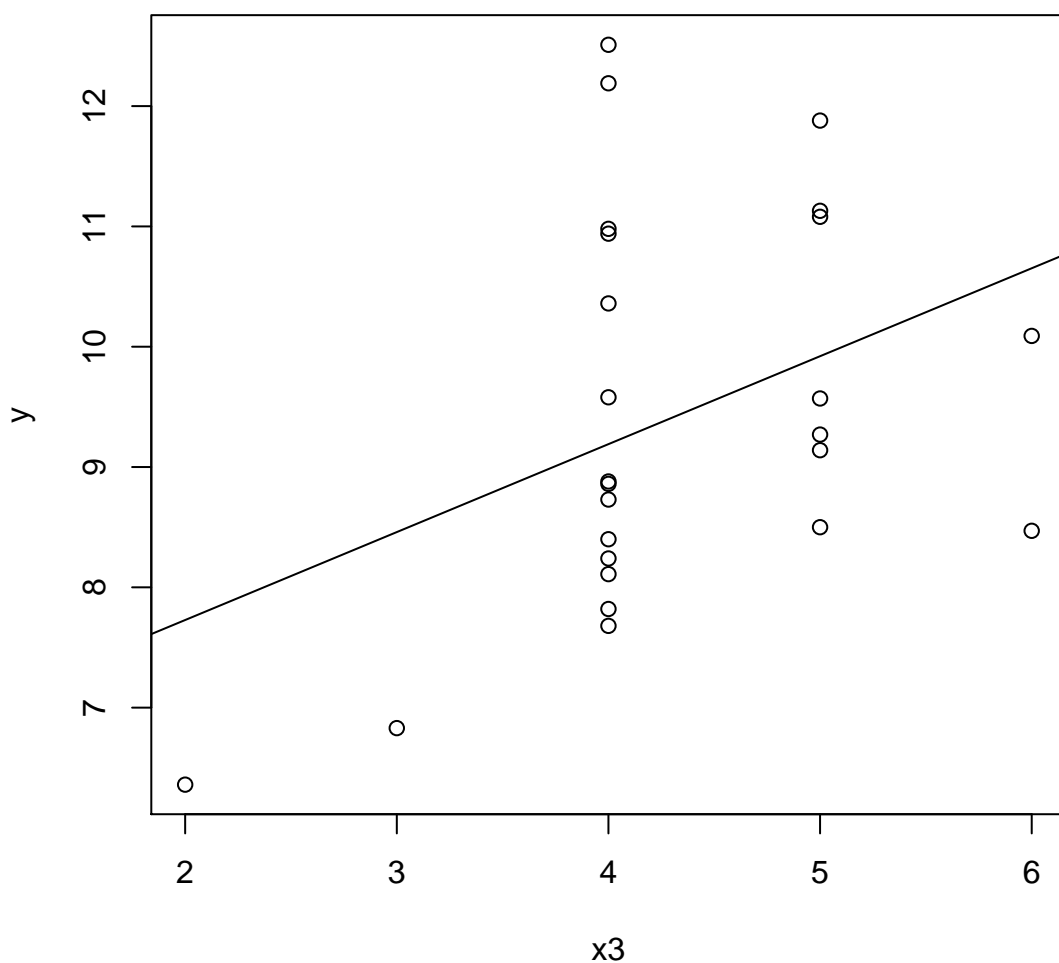


Infine consideriamo la variabile X_3 per la quale $\rho(Y, X_3) = 0.38$. In questo caso abbiamo

$$\alpha = \begin{bmatrix} 6.26625 \\ 0.7310 \end{bmatrix}$$

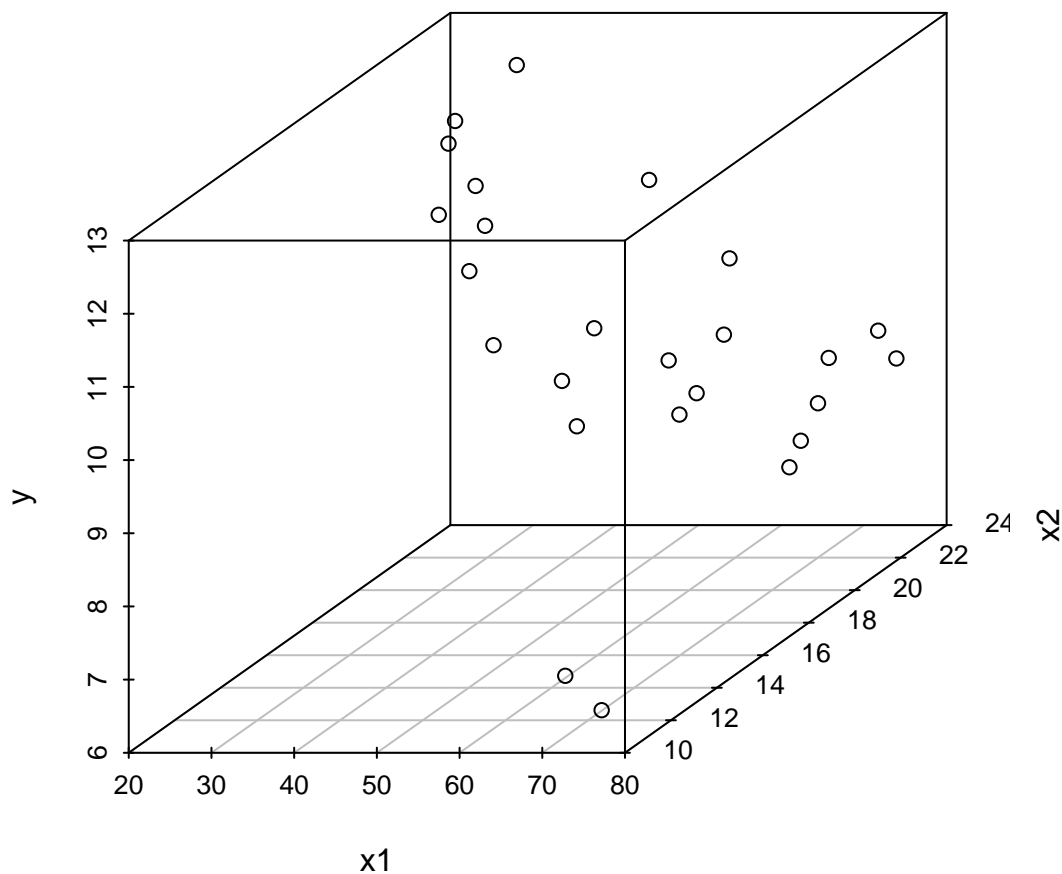
Grafico di dispersione della variabile Y rispetto a X_2 con sovrapposta la retta di equazione

$$y = 6.27 + 0.73x_2 \quad R^2 = 0.146$$



Vogliamo ora spiegare la Y come funzione di più variabili. Cominciamo con due variabili e scegliamo quella maggiormente correlata con Y , X_1 , e quella meno correlata con X_1 , cioè X_2 .

Grafico di dispersione della variabile Y rispetto a X_1 e X_2



Cerchiamo il piano che passi il più vicino possibile ai punti del grafico

Il modello ipotizzato ora è

$$Y = a + bX_1 + cX_2$$

Vogliamo determinare a, b, c in modo tale

$$y_i^* = a + bx_{1,i} + cx_{2,i}, \quad i = 1, \dots, 25$$

Possiamo riscrivere le 25 equazioni in un'unica equazione matriciale, ovvero

$$\mathbf{y}^* = \mathbf{X}\alpha$$

dove

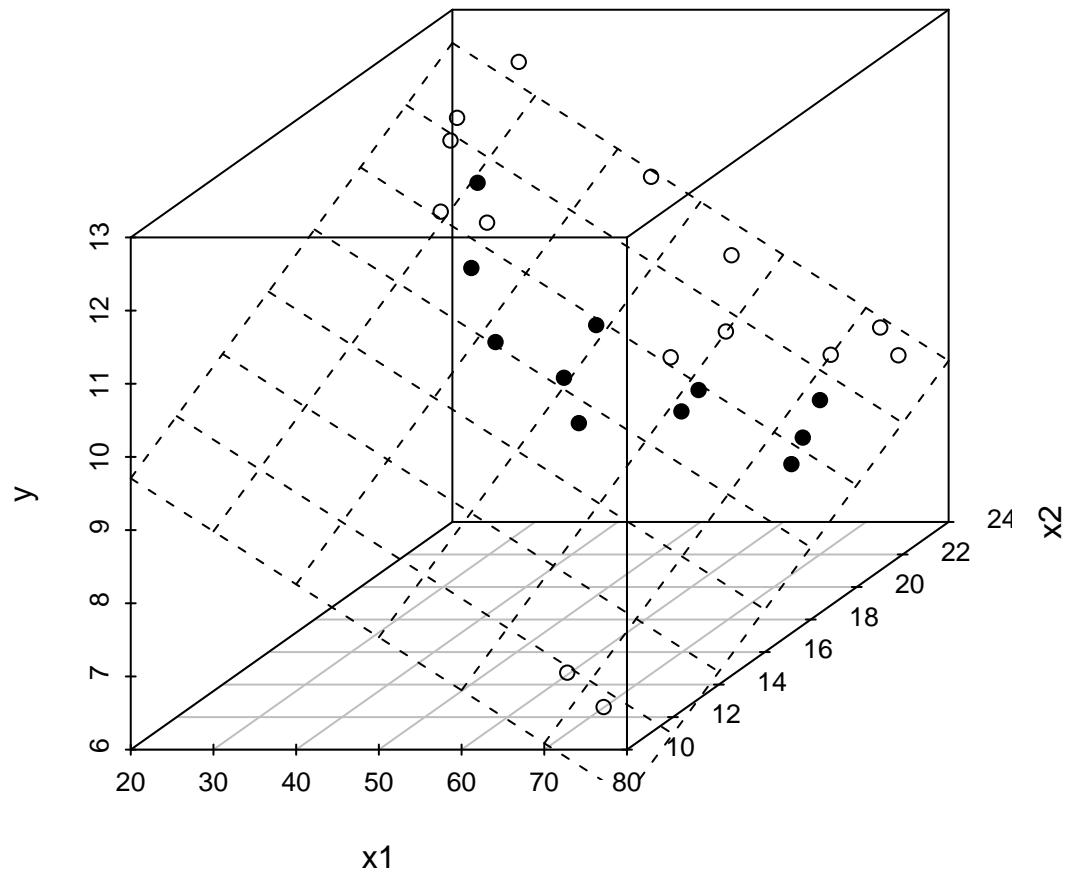
$$\mathbf{y}^* = \begin{pmatrix} y_1^* \\ \vdots \\ y_{25}^* \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 35.3 & 20 \\ 1 & 29.7 & 20 \\ \vdots & \vdots & \\ 1 & 28.6 & 22 \end{pmatrix}, \quad \alpha = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

La soluzione ottenuta col metodo dei minimi quadrati è ancora una volta

$$\alpha = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Fatti i conti otteniamo

$$\alpha = \begin{bmatrix} 9.1266 \\ -0.0724 \\ 0.2029 \end{bmatrix}$$



Il piano trovato ha equazione

$$Y^* = 9.1266 - 0.0724X_1 + 0.2029X_2$$

$$R^2 = 0.8491$$

L'indice R^2 è definito come

$$R^2 = 1 - \frac{\sum_i (y_i - y_i^*)^2}{\sum_i (y_i - \bar{y})^2}$$

Si osservi che nel caso della regressione multipla questo non è uguale a ρ^2 .

Le formule viste per il caso di due regressori si estendono al caso di $k > 2$ regressori.

Il valore dell'indice R^2 aumenta all'aumentare del numero delle variabili esplicative del modello. Occorre trovare un compromesso tra numero dei regressori e bontà di adattamento (principio di parsimonia).

Non seguendo questo principio potremmo incappare in problemi di *over fitting*, cioè modelli “molto buoni” ma inutilizzabili a fini previsivi.

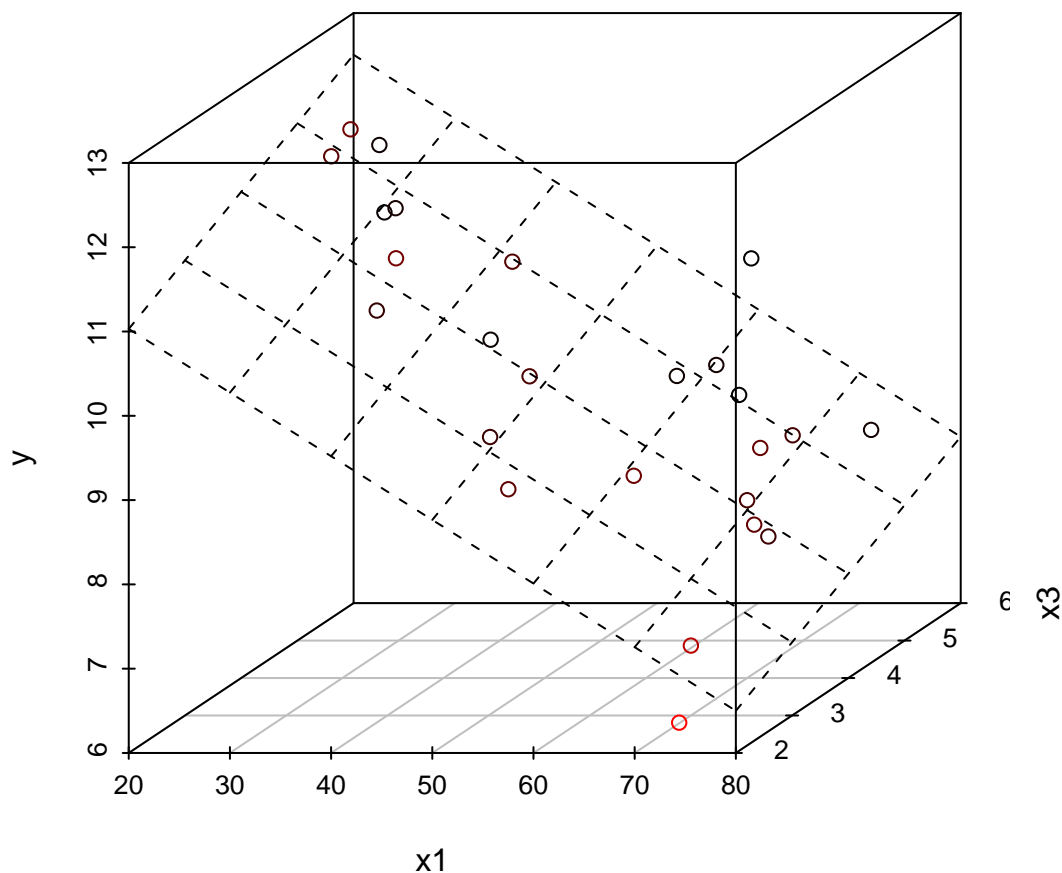
Ipotizziamo invece il modello

$$Y = a + bX_1 + cX_3$$

Il piano trovato ha equazione

$$Y^* = 11.80 - 0.075X_1 + 0.37X_3 \quad R^2 = 0.75$$

R^2 è più basso perché X_1 e X_3 sono maggiormente correlate. Il grafico di dispersione e il piano sono rappresentati in figura



Infine ipotizziamo un modello del tipo

$$Y = a + bX_1 + cX_2 + dX_3$$

In questo caso non possiamo fare il grafico!!

L'equazione matriciale è

$$\mathbf{y}^* = \mathbf{X}\alpha$$

dove

$$\mathbf{y}^* = \begin{pmatrix} y_1^* \\ \vdots \\ y_{25}^* \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 35.3 & 20 & 4 \\ 1 & 29.7 & 20 & 5 \\ \vdots & \vdots & & \\ 1 & 28.6 & 22 & 5 \end{pmatrix}, \quad \alpha = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}$$

La soluzione ottenuta col metodo dei minimi quadrati è ancora una volta

$$\alpha = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Fatti i conti otteniamo

$$\alpha = \begin{pmatrix} 9.23 \\ -0.07 \\ 0.22 \\ -0.08 \end{pmatrix}$$

L'iperpiano che passa più vicino ai punti è

$$Y = 9.23 - 0.07X_1 + 0.22X_2 - 0.08X_3, \quad R^2 = 0.8501$$

R^2 è più alto ma non di molto rispetto al modello con solo X_1 e X_2

Esercizio Nella tabella sono riportati i dati ottenuti da un'esperimento per valutare quanto la resistenza all'abrasione di un tipo di gomma dipende dalla durezza della gomma e dalla sua resistenza alla tensione. Siano Y l'abrasione, misurata in grammi per ora. X_1 la durezza, misurata in gradi Shore e X_2 la resistenza misurata in chilogrammi per centimetro quadrato.

1. Si calcoli la matrice di correlazione e si dica quale delle variabili è più correlata a Y

2. Si determinino i coefficienti delle rette

$$Y = a + bX_1 \quad \text{e} \quad Y = c + dX_2$$

3. Si calcoli R^2 per le due rette

4. Si determinino i coefficienti del piano

$$Y = a + bX_1 + cX_2$$

e si calcoli R^2 . Si commenti il risultato

5. Si calcoli la corrosione nel caso in cui la durezza sia 80 gradi Shore e la resistenza sia 200 kg/cm²

Y	X_1	X_2
372	45	162
206	55	233
175	61	232
154	66	231
136	71	231
112	71	237
55	81	224
45	86	219
221	53	203
166	60	189
164	64	210
113	68	210
82	79	196
32	81	180
228	56	200
196	68	173
128	75	188
97	83	161
64	88	119
249	59	161
219	71	151
186	80	165
155	82	151
114	89	128
341	51	161
340	59	146
283	65	148
267	74	144
215	81	134
148	86	127

Abbiamo i seguenti valori

	Y	X_1	X_2
Var	8027	153	1382
Media	175.4	70.27	180.5

La matrice di correlazione è

	Y	X_1	X_2
Y	1.000	-0.738	-0.298
X_1	-0.738	1.000	-0.299
X_2	-0.298	-0.299	1.000

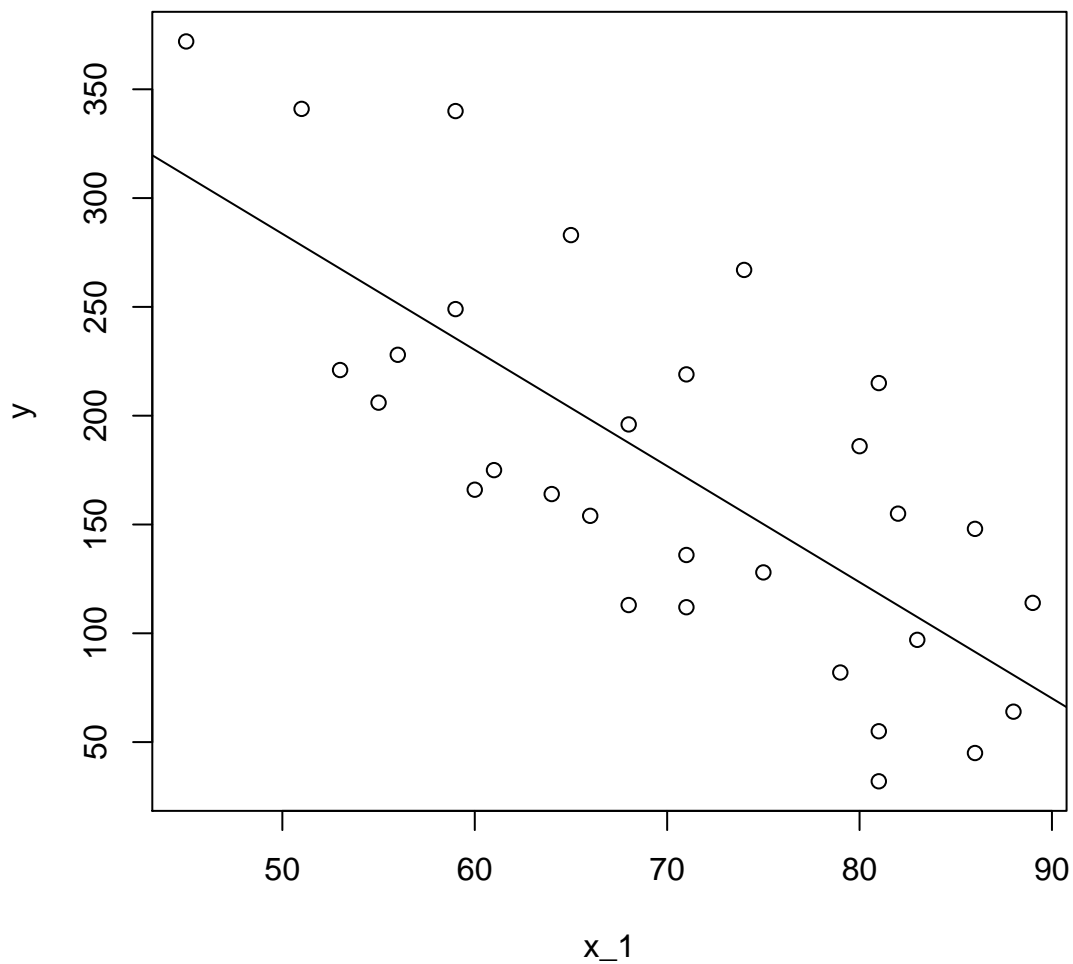
La variabile più correlata con Y risulta la variabile X_1 .

La retta di regressione $Y = a + bX_1$

	Stima	R^2
a	550.4151	0.54
b	-5.3366	

In questo caso R^2 non è molto alto.

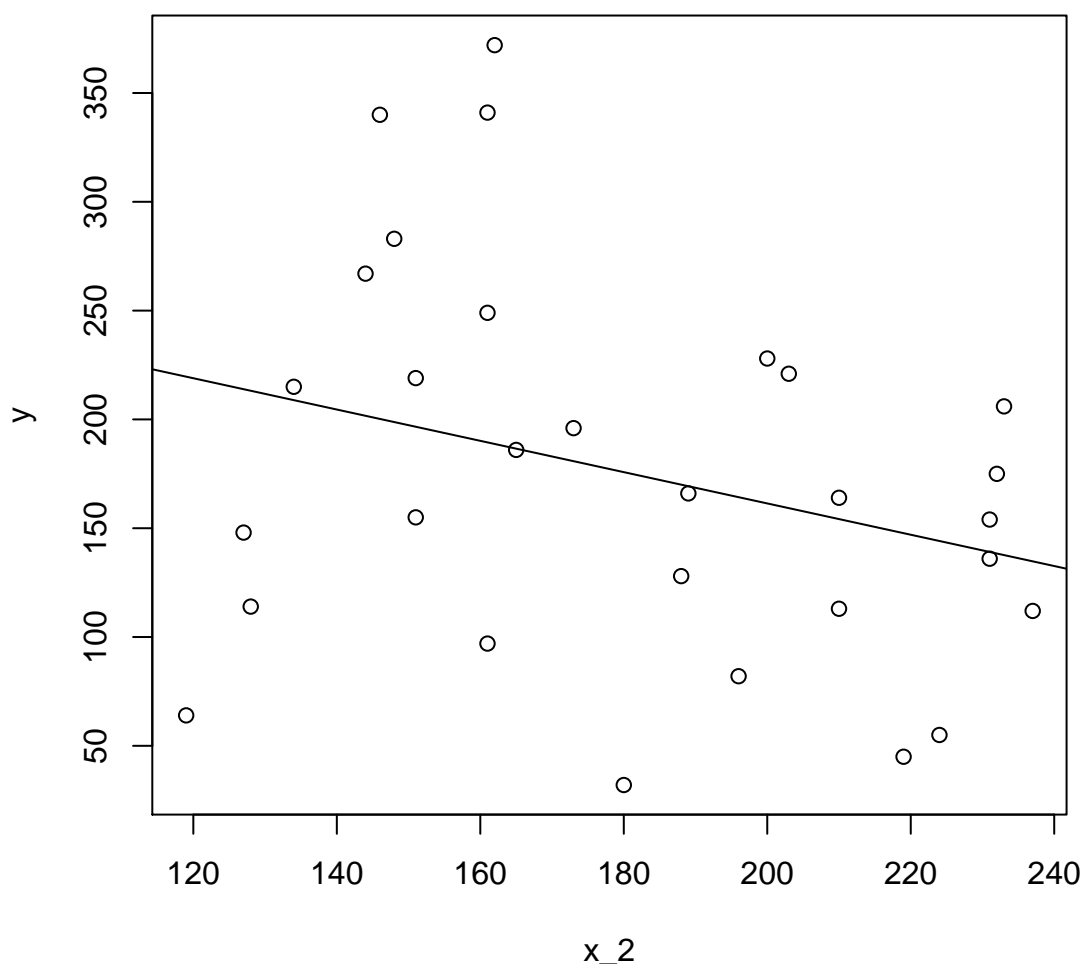
Il grafico di dispersione e la retta sono rappresentati in figura



Passiamo al modello $Y = c + dX_2$. Abbiamo

	Stima	R^2
c	305.2248	0.09
d	-0.7192	

Il grafico di dispersione e la retta sono rappresentati in figura



La retta spiega molto poco della variabilità di Y .

Consideriamo il modello

$$Y = a + bX_1 + cX_2$$

Abbiamo

	Stima	R^2	$\sum (y_i - y_i^*)^2$	$\sum (y_i - \bar{y})^2$
a	885.1611	0.84	35949.74	225011.4
b	-6.5708			
c	-1.3743			

R^2 è ottenuto come

$$R^2 = 1 - \frac{\sum_i (y_i - y_i^*)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{35949.74}{225011.4} = 0.84$$

Si nota come il valore dell'indice R^2 si incrementa notevolmente con le due variabili, rispetto a entrambi i modelli con una sola variabile esplicativa.

Il modello può essere usato a scopi previsivi: nel caso in cui $x_1 = 80$ e $x_2 = 200$ abbiamo

$$y = 885.16 - 6.57x_1 - 1.37x_2 = 885.16 - 6.57 \cdot 80 - 1.37 \cdot 200$$

Nel grafico è rappresentata la nuvola di punti e il piano ottenuto col metodo dei minimi quadrati

