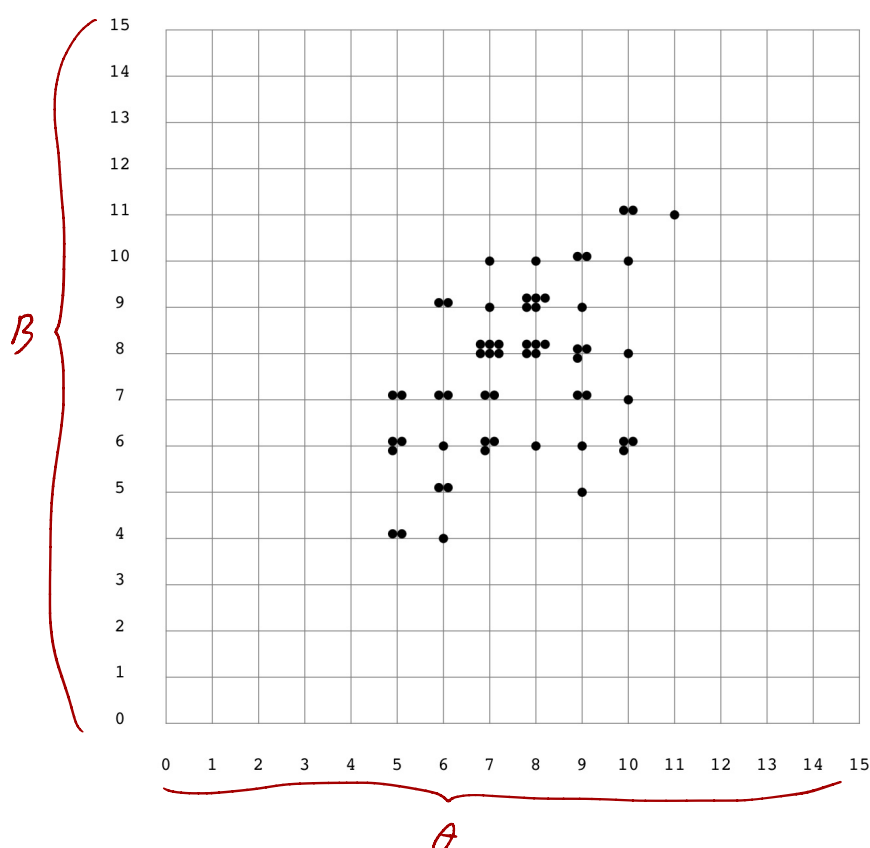


Regressione lineare

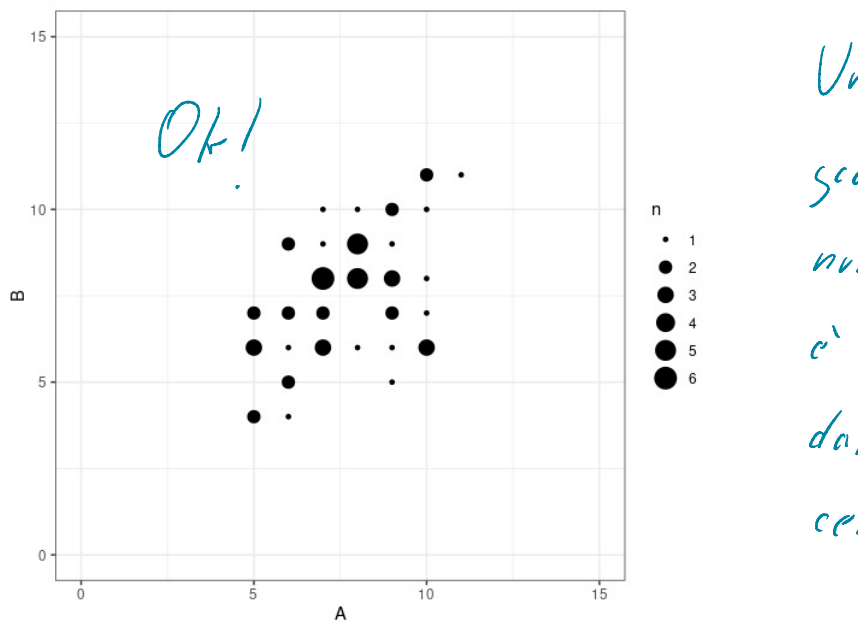
Date due variabili statistiche A e B , di tipo numerico, vogliamo quantificare in che misura i valori assunti da una siano correlati a quelli assunti dall'altra.

Esempio 1: È stato chiesto a 59 studenti di ripetere due volte una prova di abilità, il cui risultato è un punteggio fra 0 e 15. I punteggi delle due prove, A e B , sono riassunti in questo diagramma.



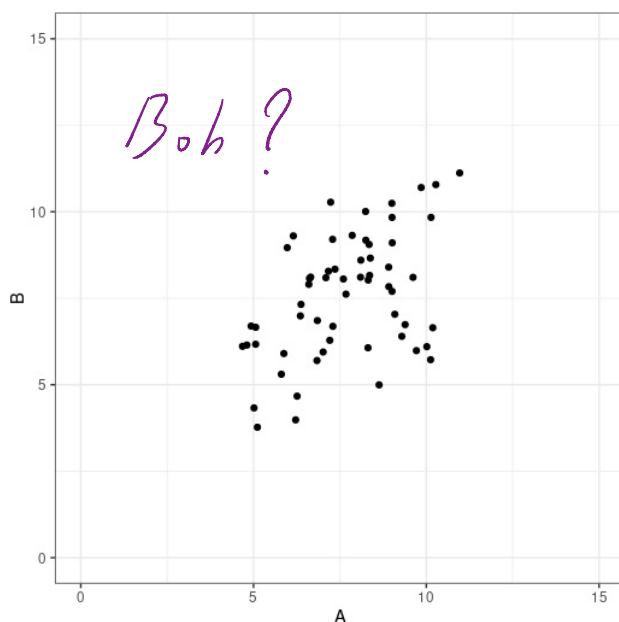
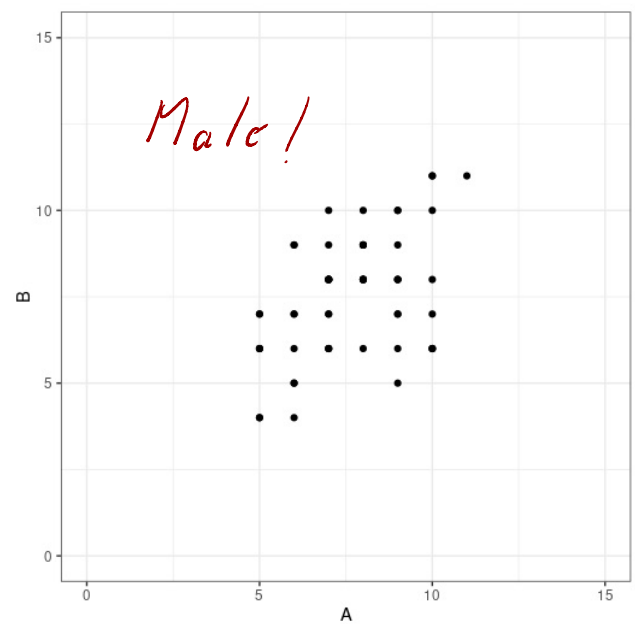
"Scatter plot" o
"diagramma di
dispersione"

Ciascun punto
rappresenta un
individuo, con la
sua coppia (A, B)



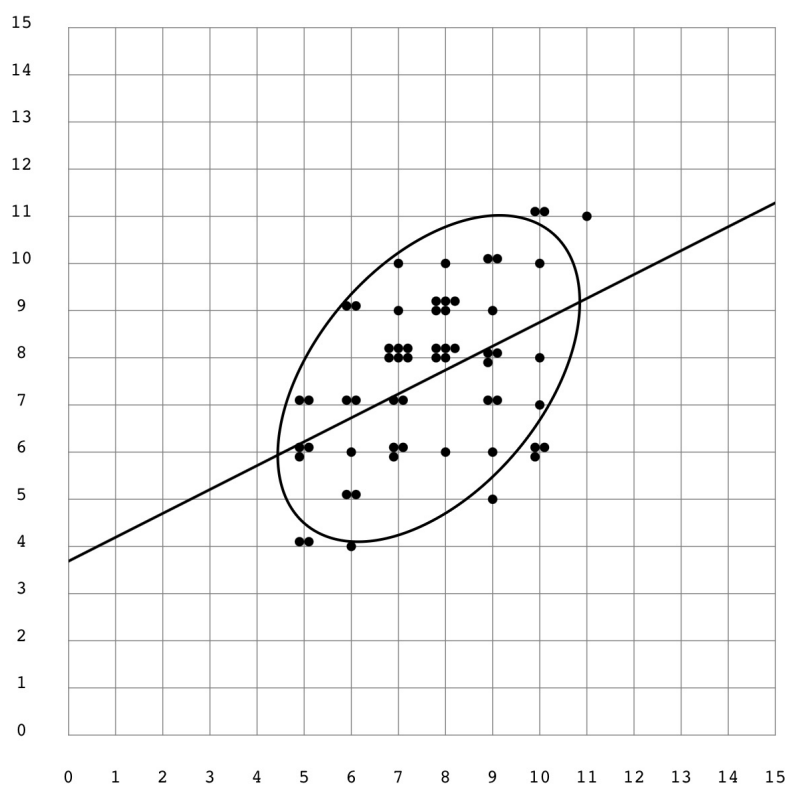
Un'altra forma di scatter plot: qui il numero di individui è rappresentato dalla dimensione dei cerchi

Rappresentando ogni coppia (A, B) con un punto si avrebbero delle sovrapposizioni che oscurano l'accumulo dei dati al centro della nuvola.



In questa versione i punti sono stati mossi di poco, in modo casuale, per rimuoverne le sovrapposizioni.

I punti — le coppie (A, B) — formano una nuvola grossolanamente ellittica. Questo avviene

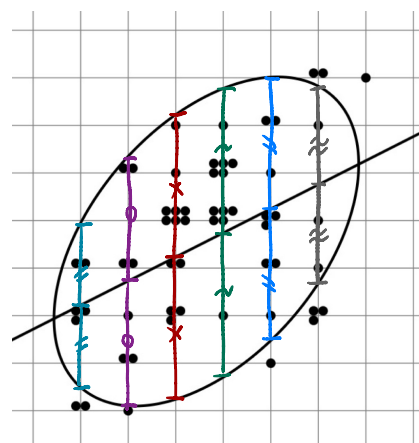


perché la distribuzione delle variabili A e B è approssimativamente congiuntamente gaussiana. L'ellisse è inclinato: questo indica una correlazione fra A e B , e, in particolare,

una correlazione positiva, perché, al crescere di A , anche B tende a crescere; la situazione opposta si chiamerebbe correlazione negativa.

Il diagramma mostra altresì la retta di regressione di B rispetto ad A . Questa retta divide a metà ogni sezione verticale della nuvola, vale a dire, per ogni fissato valore $A = A_0$, il corrispondente

$B_0 = \text{intercetta} + \text{pendenza} \cdot A_0$
 è una stima della media delle



coordinate B dei punti che hanno $A = A_0$.

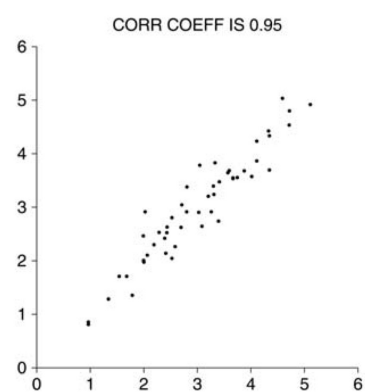
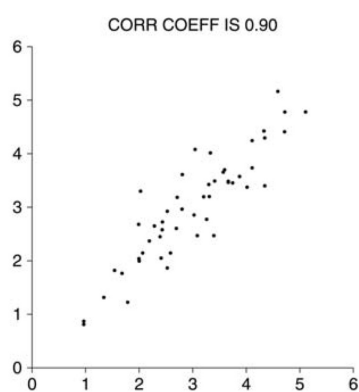
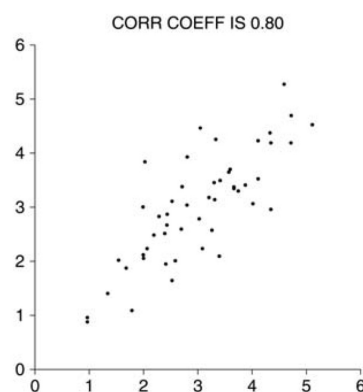
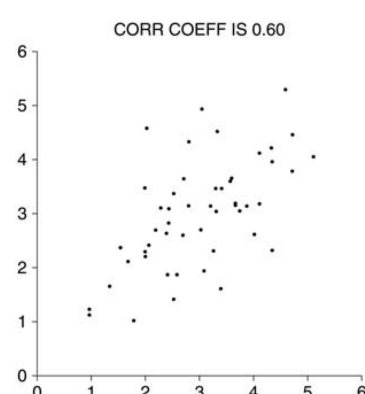
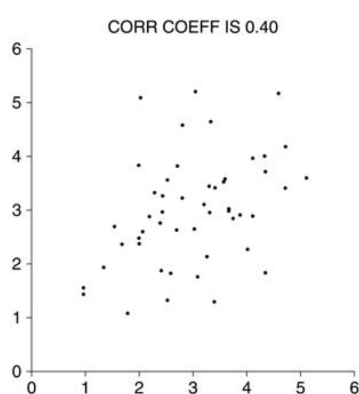
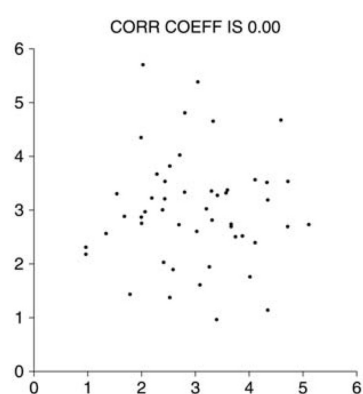
Vedremo come calcolare le seguenti quantità:

- I coefficienti della retta di regressione
- Il coefficiente di correlazione di Pearson r che misura la forza della correlazione.
- Un p -value che misura il grado di certezza con cui possiamo affermare che le variabili sono realmente correlate.

Abbiamo già descritto il significato della retta di regressione. Per quanto riguarda r , si tratta di un numero nell'intervallo $-1 \leq r \leq 1$, il cui segno indica il segno della correlazione. Il valore assoluto $|r|$, dice quanto la correlazione è forte, ossia in che misura è possibile prevedere B conoscendo A per mezzo della retta di regressione. A $|r| = 0$ corrispondono variabili totalmente scorrelate (in questo caso la retta di regressione è orizzontale: una costante pari alla media di B). A $|r| = 1$ corrisponde il caso in cui B è esattamente determinata da $B = \alpha + \beta \cdot A$ per opportuni coefficienti α e β .

Grossolanamente, r descrive la forma della nuvola di punti. Se $r=0$ gli assi dell'ellisse sono paralleli agli assi, per cui, a scegliere le giuste unità di misura, la nuvola appare circolare.

Se $r>0$ (risp. $r<0$) l'asse maggiore è disposto diagonalmente così $/$ (risp. così $\backslash$). Più $|r|$ si avvicina a 1, più l'ellisse è schiacciata, e, per $|r|=1$, questo ha la forma di un segmento di retta.



Esempi dal libro di Freedman, Pisani, Purves

Discuteremo in dettaglio il *p-value* più avanti. Molto sinteticamente, l'idea è che, in un campione casuale, potremmo osservare una correlazione che, in realtà, nella popolazione non c'è, per via di uno scherzo della fortuna. Il *p-value* misura la probabilità di osservare la correlazione apparentemente presente nei dati come mero frutto del caso. Quindi *minore* è il *p-value*, *più* è verosimile che la correlazione sia reale, ossia *maggiore* è la *significatività statistica* della correlazione. Un *p-value* grande, ossia un basso livello di significatività statistica, indica che la numerosità del campione — il numero di punti *(A, B)* rilevati — non è sufficiente per concludere che *A* e *B* sono linearmente correlate. Suglie tipiche con cui si confrontano i *p-value* nelle scienze della vita sono *<5%* per ritenere un risultato *significativo* e *<1%* per che sia *fortemente significativo*. Una triste macumba circonda queste soglie, per cui è bene non farne un feticcio: si tratta di semplici numeri scelti per convenzione.

Formule & calcoli

Chiamiamo $(A_1, B_1), (A_2, B_2), \dots, (A_n, B_n)$ i punti del nostro campione. Si comincia col calcolare media e deviazione standard di A e B : μ_A , σ_A , μ_B , σ_B . Poi si convertono i punti del campione in *unità standard*

$$A^* = \frac{A - \mu_A}{\sigma_A} \quad B^* = \frac{B - \mu_B}{\sigma_B}$$

Ottenendo $(A_1^*, B_1^*), (A_2^*, B_2^*), \dots, (A_n^*, B_n^*)$

Ora, il coefficiente di correlazione di Pearson r è la media dei prodotti $A_i^* B_i^*$.

$$r = \frac{A_1^* B_1^* + A_2^* B_2^* + \dots + A_n^* B_n^*}{n}$$

È la retta di regressione, in unità standard, è semplicemente

$$B^* = r A^*$$

Da cui, sostituendo A^* e B^* con le rispettive definizioni

$$\frac{B - \mu_B}{\sigma_B} = r \frac{A - \mu_A}{\sigma_A} \Rightarrow B = \underbrace{r \frac{\sigma_B}{\sigma_A}}_{\text{pendenza}} A + \underbrace{\mu_B - r \frac{\sigma_B}{\sigma_A} \mu_A}_{\text{intercetta}}$$

Non conviene memorizzare quest'ultima formula, ma la molto più semplice

Retta di regressione IN UNITA' STANDARD

$$B^* = r A^*$$

da cui è facile ricavare la retta di regressione che lega A e B come appena visto.

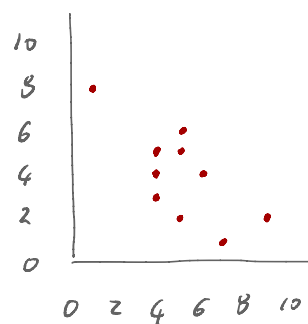
Esempio 2 Consideriamo

A_i	4	5	1	3	4	7	5	4	5	6
B_i	5	6	8	2	4	1	5	3	2	4

Otteniamo $M_A = 5.00$ $\sigma_A = 2.00$

$M_B = 4.00$ $\sigma_B = 2.00$

11/01/17 I dati sono finti...



Da cui

A_i^*	-0.5	0.0	-2.0	2.0	-0.5	1.0	0.0	-0.5	0.0	0.5
B_i^*	0.5	1.0	2.0	-1.0	0.0	-1.5	0.5	-0.5	-1.0	0.0
$A_i^* B_i^*$	-0.25	0	-4	-2	0	-1.5	0	0.25	0	0

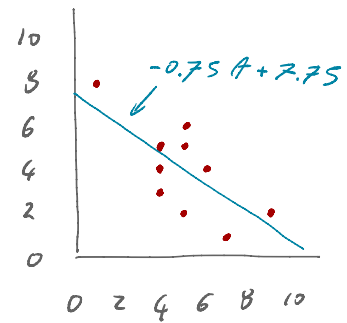
$r = \text{media di questi } \underbrace{\hspace{10em}} = -0.75$

Per cui la retta di regressione è $B^* = -0.75 A^*$

Che possiamo scrivere nella forma

$$\frac{B-4}{2} = -0.75 \frac{A-5}{2}$$

$$B = -0.75 A + 7.75$$



Esempio 3

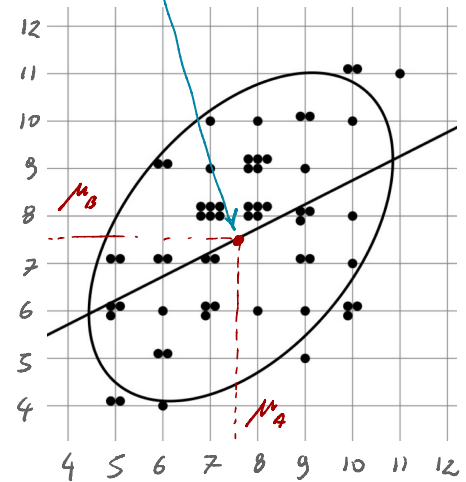
I risultati che abbiamo discusso di un test di abilità hanno

$$\mu_A = 7.644 \quad \sigma_A = 1.603$$

$$\mu_B = 7.560 \quad \sigma_B = 1.730 \quad r = 0.469$$

quindi la retta di regressione è

baricentro della distribuzione



$$\frac{B-7.560}{1.730} = 0.469 \frac{A-7.644}{1.603} \Rightarrow B = 0.506 A + 3.691$$

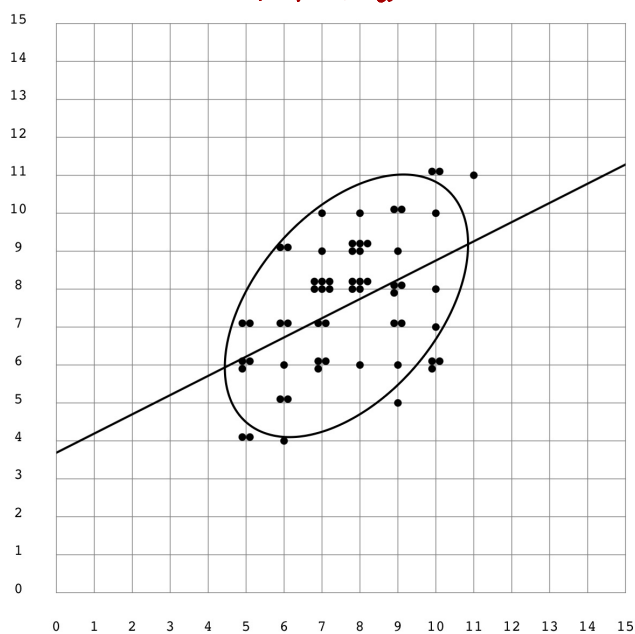
Non stupisce che ci sia correlazione fra due ripetizioni della medesima prova: questo avviene, in generale, con i test di abilità.

Osservazione La retta di regressione passa sempre per il baricentro (μ_A, μ_B) dei punti del campione, quindi la si può disegnare tracciando la retta per questo punto che ha pendenza $r \cdot \frac{\sigma_B}{\sigma_A}$.

Regressione lineare e outlier

In statistica un **outlier** è un dato che pare anomalo se confrontato con la distribuzione degli altri. Questa definizione è imprecisa perché **non c'è una definizione precisa**. Possiamo introdurre artificialmente un outlier tra i dati del nostro test di abilità aggiungendo il singolo punto **(15, 0)** ai 59 presenti

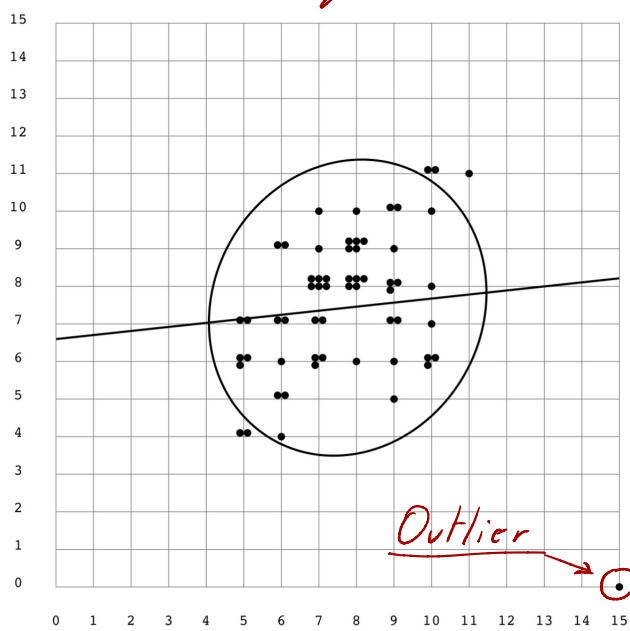
Prima



$$r = 0.469$$

$$p\text{-value} = 0.018 \%$$

Dopo



$$r = 0.101$$

$$p\text{-value} = 44 \%$$

Come si vede, un singolo outlier confonde completamente il calcolo: il coefficiente di correlazione e la retta di regressione **non** sono **statistiche robuste**. Allo stesso modo la media non è robusta,

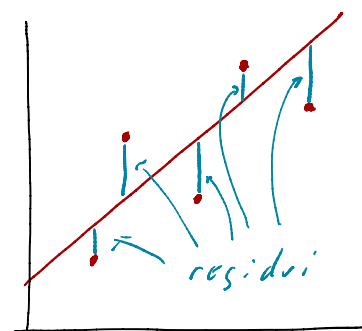
perché un singolo outlier può avere un peso arbitrariamente grande sulla media, mentre, per esempio, questo non accade per la mediana.

Perché il metodo che abbiamo visto sia giustificato, conviene che i residui

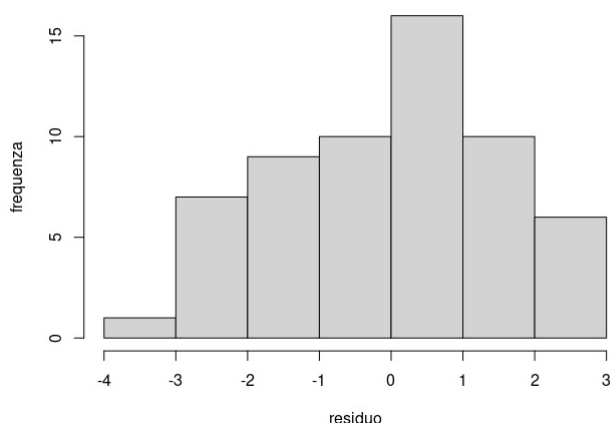
$$i\text{-esimo residuo} = B_i - (\alpha + \beta A_i)$$

retta di regressione

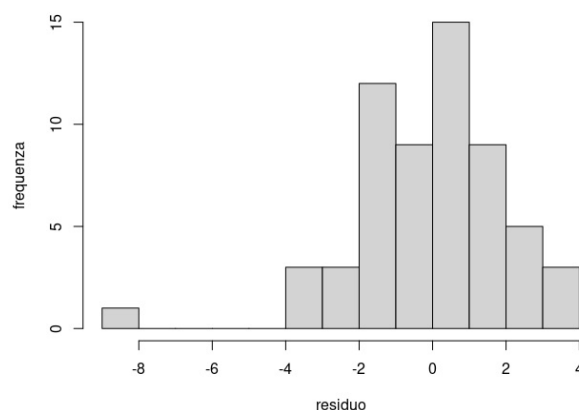
siano distribuiti in maniera (approssimativamente) gaussiana. Questa condizione si può verificare a posteriori, per esempio calcolando i residui e tracciandone un istogramma.



Prima



Dopo



In conclusione, è importante, prima di applicare la regressione lineare, controllare su uno scatter plot che non ci siano evidenti outlier. È buona norma verificare altresì la distribuzione dei residui.

Il p-value

I p-value compaiono, in generale, nel contesto dei *test di ipotesi*. Nel caso della regressione lineare, noi vogliamo testare l'ipotesi seguente:

IPOTESI "ALTERNATIVA": A e B sono correlate linearmente.

Per farlo, raccogliamo un campione di n individui; calcoliamo il coefficiente $r = r_0$ e la retta di regressione, e ci convinciamo che i residui sono distribuiti in modo compatibile con una gaussiana. Ora possiamo chiederci: se A e B fossero scorrelate, quanto sarebbe probabile che un campione di n individui

abbia $|r| \geq |r_0|$? Detto in altri termini: se A e B fossero scorrelate, quanti dei campioni di n individui, in proporzione a tutti quelli possibili, sarebbero correlati al pari del nostro, o più?

Questo è il *p-value* e misura il grado di significatività statistica con cui possiamo rigettare la *ipotesi nulla*: A e B non sono correlate. Più

basso è il p-value, meno è verosimile che la correlazione osservata sia uno scherzo del caso, più,

quindi, è ragionevole che sia reale.

Calcolare questo p-value può essere complicato, per via della necessità di caratterizzare l'insieme di tutti i campioni di n individui possibili. Infatti, a seconda di quali ipotesi si fanno sulla popolazione, si otterranno diversi insiemi di campioni possibili, quindi p-value diversi. La regola seguente è basata sull'ipotesi che i regimi siano indipendenti fra loro e distribuiti in maniera gaussiana.

Se valesse l'ipotesi nulla, allora

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

sarebbe distribuito secondo una t di Student con $n-2$ gradi di libertà

Esempio 4

Consideriamo la situazione dell'esempio 2

A_i	4	5	1	3	4	7	5	4	5	6
B_i	5	6	8	2	4	1	5	3	2	4

$n = 10$

abbiamo già calcolato $r = -0.75$, da cui

$$t = -0.75 \sqrt{\frac{10-2}{1-0.75^2}} \approx -3.40$$

Il p-value con cui possiamo rigettare l'ipotesi nulla — l'ipotesi di non correlazione — è la probabilità che una t di Student con $8 = 10 - 2$ gradi di libertà assuma un valore assoluto maggiore o uguale a **3.40**.

Student's curve, with degrees of freedom shown at the left of the table

The shaded area is shown along the top of the table

t is shown in the body of the table

Degrees of freedom	25%	10%	5%	p-val. 0.05 2.5%	1%	p-val 0.01 0.5%
1	1.00	3.08	6.31	12.71	31.82	63.66
2	0.82	1.89	2.92	4.30	6.96	9.92
3	0.76	1.64	2.35	3.18	4.54	5.84
4	0.74	1.53	2.13	2.78	3.75	4.60
5	0.73	1.48	2.02	2.57	3.36	4.03
6	0.72	1.44	1.94	2.45	3.14	3.71
7	0.71	1.41	1.89	2.36	3.00	3.50
8	0.71	1.40	1.86	2.31	2.90	3.36
9	0.70	1.38	1.83	2.26	2.82	3.25
10	0.70	1.37	1.81	2.23	2.76	3.17

Vediamo che la t di Student con 8 gradi di libertà vale più di **3.36** < 3.40 nello 0.5% dei casi. Simmetricamente varrà meno di **-3.36** in un altro 0.5% dei casi. In totale, questa distribuzione supera **3.36**, in valore assoluto, nello 1% dei casi, per cui supera **3.40** in meno casi: possiamo concludere che il p-value è ≤ 0.01 .