Esercizi svolti

Insiemi di dati a coppie

Roma, Aprile 2020

Esercizio 1

I seguenti dati mostrano i punteggi QI (=Quoziente Intellettivo) di 10 madri e 10 figlie primogenite.

QI madre	135	127	124	120	115	112	104	96	94	85
QI figlia	121	131	112	115	99	118	106	89	92	90

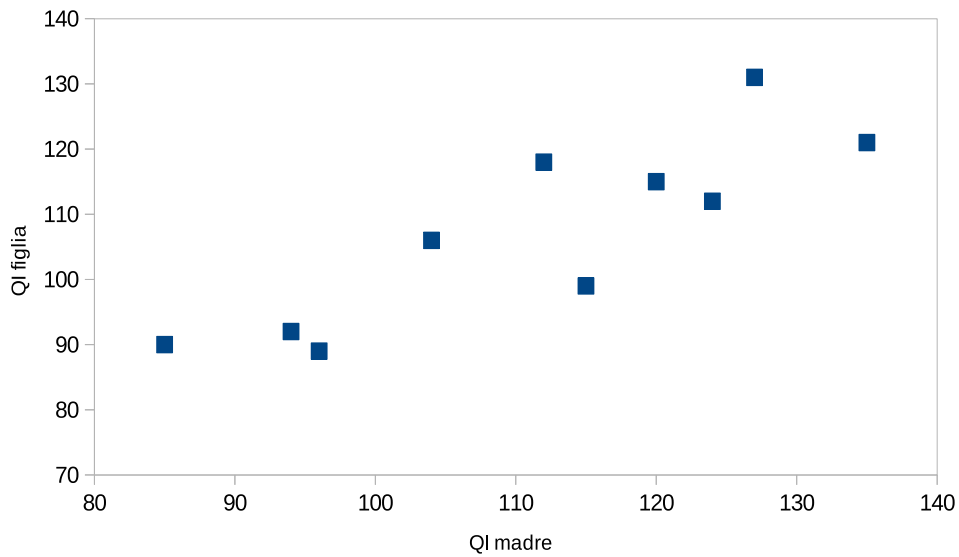
1. Rappresenta i dati in un diagramma di dispersione (scatter plot).
2. Guardando il grafico, pensi che ci sia una correlazione? Posiva o negativa?
3. Calcola il coefficiente ci correlazione campionaria r.
4. Quali conclusioni puoi trarre dalla relazione tra il QI delle madri e delle figlie?

Soluzione

1. Indico con x l'osservabile QI madre e con y l'osservabile QI figlia. Riordino i dati della tabella in modo che i valori di x siano crescenti.

QI(m)	85	94	96	104	112	115	120	124	127	135
QI(f)	90	92	89	106	118	99	115	112	131	121

Rappresento i dati in un diagramma di dispersione.



Soluzione

2. C'è una correlazione positiva tra i punteggi QI delle madri e delle figlie (a valori più grandi dell'osservabile x corrispondono mediamente valori più grandi della variabile y).
3. Il coefficiente di correlazione campionaria è dato da $r = \frac{1}{n-1}\frac{1}{s_x s_y}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ Per calcolarlo mi servono:
   - la media e la deviazione standard campionarie del punteggio QI delle madri, $\bar{x} = 111,2$, $s_x = 16,12$.
   - la media e la deviazione standard campionarie del punteggio QI delle figlie, $\bar{y} = 107,3$, $s_y = 14,47$.
   - la somma dei prodotti degli **scarti** dell'osservabile x, $(x_i - \bar{x})$, $i = 1,..,10$. e dell'osservabile y, $(y_i - \bar{y})$, $i = 1,...,n$, ovvero $S_{xy} = \sum_{i=1}^{10}(x_i - \bar{x})(y_i - \bar{y})$.

soluzione

tabella degli scarti

$(x_i - \bar{x})$	$(y_i - \bar{y})$
$85 - 111,2 = -26,2$	$90 - 107,3 = -17,3$
$94 - 111,2 = -17,2$	$-15,3$
$96 - 111,2 = -15,2$	$-18,3$
$104 - 111,2 = -7,2$	$-1,3$
$112 - 111,2 = 0,8$	$10,7$
$115 - 111,2 = 3,8$	$-8,3$
$120 - 111,2 = 8,8$	$7,7$
$124 - 111,2 = 12,8$	$4,7$
$127 - 111,2 = 15,8$	$23,7$
$135 - 111,2 = 23,8$	$13,7$

Calcolo la somma dei prodotti
$S_{xy} = \sum_{i=1}^{10}(x_i - \bar{x})(y_i - \bar{y}) = 1809,4$, quindi

$$r = \frac{1}{9}\,\frac{1}{(16,12 \times 14,47)}1809,4 = 0,86.$$

Soluzione

4. Poiché $r > 0$ c'è una correlazione positiva. Inoltre, poiché r è abbastanza vicino a 1, c'è una relazione lineare relativamente forte.

Esercizio 2

Un'azienda vuole indagare il rapporto tra temperatura ambientale e numero di parti difettose prodotte . Per 22 giorni si registrano le temperature massime e il numero di difetti riscontrati.

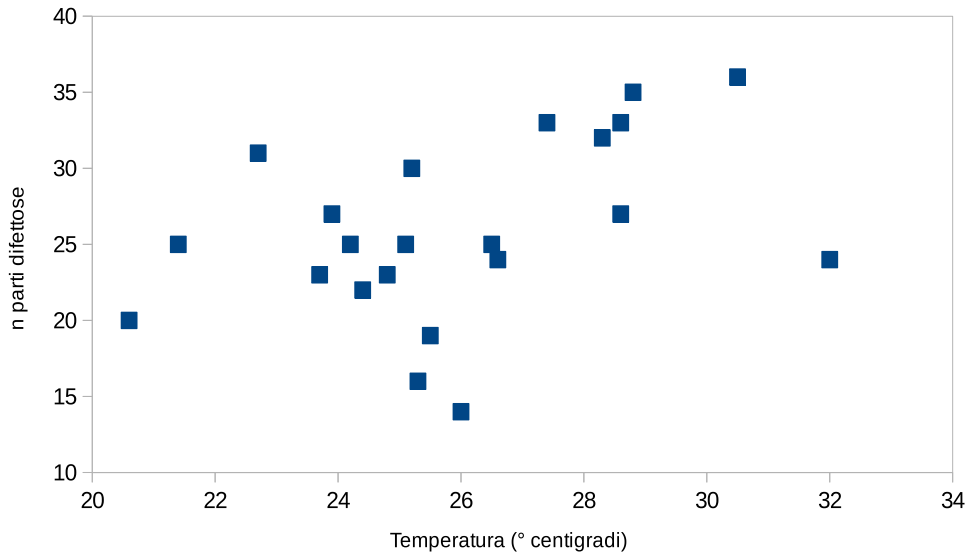
T (^{o}c)	n parti difettose	T (^{o}c)	n parti difettose
20,6	20	25,5	19
21,4	25	26	14
22,7	31	26,5	25
23,7	23	26,6	24
23,9	27	27,4	33
24,2	25	28,3	32
24,4	22	28,6	33
24,8	23	28,6	27
25,1	25	28,8	35
25,2	30	30,5	36
25,3	16	32	24

Esercizio 2

Disegna il diagramma di dispersione (scatter plot).

1. Guardando il grafico, pensi che ci sia una correlazione? Positiva o negativa?
2. Calcola il coefficiente di correlazione campionaria.
3. Quali conclusioni puoi trarre?

Scatter plot



Soluzione

1. Guardando il grafico, sembra esserci una debole correlazione positiva (l'orientazione della nuvola di punti suggerisce che all'aumentare della temperatura mediamente aumenta il numero di difetti).
2. Il coefficiente di correlazione campionaria è dato da $r = \frac{1}{n-1}\frac{1}{s_x s_y} S_{xy}$, dove $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$. $n = 22$. Calcolo la media e la deviazione standard campionarie delle temperature: $\bar{x} = 25,91$, $s_x = 2,81$. Calcolo la media e la deviazione standard campionarie del numero di difetti: $\bar{y} = 25,86$, $s_y = 5,91$. Calcolo $S_{xy} = 145,94$, quindi

$$r = \frac{1}{21}\frac{1}{2,81 \cdot 5,91}145,94 = 0,42$$

.

3. C'è una correlazione positiva tra le osservabili ($r > 0$), ma la relazione lineare è debole, perchè il valore di r è lontano da 1.

Esercizio 3

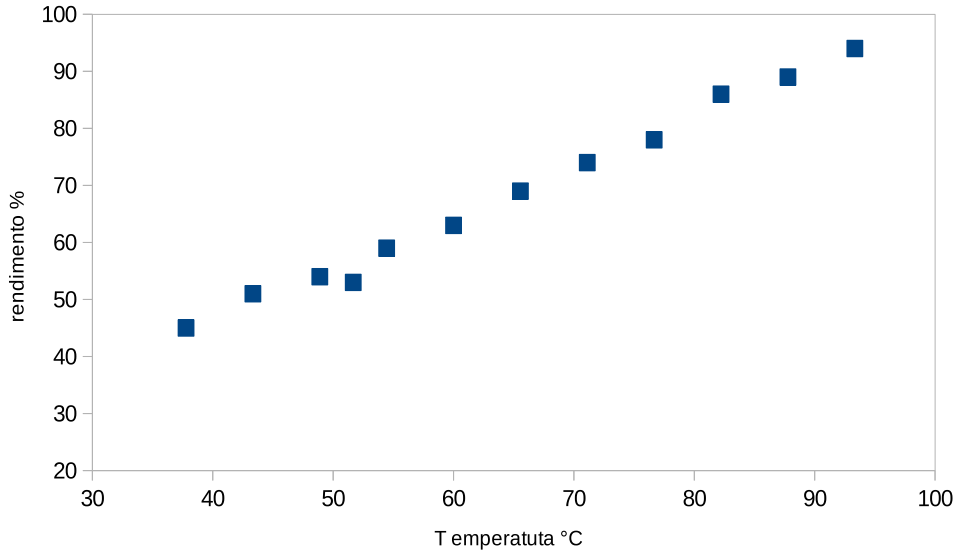
Le seguenti 12 coppie di dati mettono in relazione la resa percentuale di un esperimento di laboratorio (y) con la temperatura x, espressa in gradi Centigradi, a cui è stato condotto l'esperimento

x	38	43	49	52	54	60	66	71	77	82	88	93
y	45	51	54	53	59	63	69	74	78	86	89	94

1. Rappresenta questi dati in un diagramma di dispersione (scatter plot).
2. Pensi che un modello di regressione lineare semplice sia appropriato per descrivere la relazione tra resa percentuale e temperatura? Motiva la risposta.
3. Determina la retta di regressione dei minimi quadrati.
4. Prevedi il rendimento a una temperature di 75^o.

Soluzione

1. Scatter plot



2. Dal grafico si vede che i valori del campione bivariato si dispongono bene lungo una retta, quindi un modello di regressione semplice è appropriato. Se calcoliamo il coefficiente di correlazione troviamo $r = 0,996$, quindi effettivamente c'è una relazione lineare forte tra temperatura e rendimento.

Soluzione

3. La retta di regressione dei minimi quadrati ha equazione $y = a + bx$, con

$$b = r\frac{s_y}{s_x}, \qquad a = \bar{y} - b\bar{x},$$

dove $\bar{x}$, s_x sono la media e la varianza campionaria dell'osservabile x e $\bar{y}$, s_y sono la media e la varianza campionaria dell'osservabile y. Facendo i calcoli

$$\bar{x} = 16,40 \quad s_x = 18,02 \quad \bar{y} = 67,92 \quad s_y = 16,30,$$

quindi $b = 0,996 * 16,30/18,02 = 0,90$,
$a = 67,92 - 0,90 \cdot 16,40 = 9,89$. La retta di regressione dei minimi quadrati è quindi $y = 9,89 + 0,90x$.

4. Per prevedere il rendimento alla temperatura di 75^o usiamo la retta di regressione dei minimi quadrati. Il rendimento percentuale corrispondente a $x = 75$ è
$y = 9,89 + (0,90 \cdot 75) = 77,39$.

Esercizio 4

Si ritiene che più alcol c'è in circolo, più lento sia il tempo di reazione di una persona. Per verificare questa affermazione, 7 volontari assumono ciascuno una diversa quantità di alcol. La concentrazione di alcol nel sangue viene determinata come percentuale del peso corporeo. Viene quindi misurato il tempo di reazione a un certo stimolo. I dati ottenuti sono riportati nella seguente tabella, dove x è la concentrazione di alcol nel sangue (%) e y è il tempo di reazione in secondi.

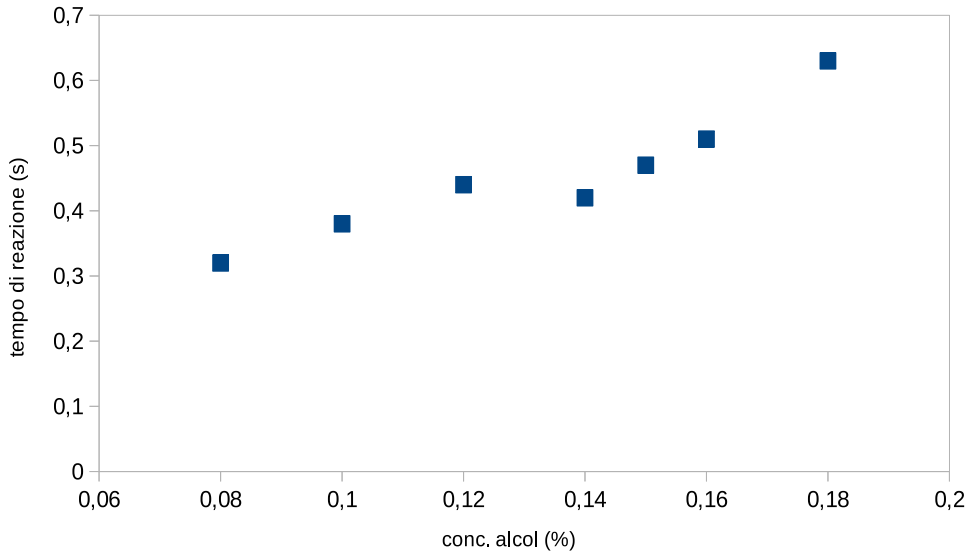
x	y
$0,08$	$0,32$
$0,10$	$0,38$
$0,12$	$0,44$
$0,14$	$0,42$
$0,15$	$0,47$
$0,16$	$0,51$
$0,18$	$0,63$

Esercizo 4

1. Disegna il diagramma di dispersione.
2. Pensi che un modello di regressione lineare semplice sia appropriato per descrivere la relazione tra tempo di reazione e concentrazione di alcol nel sangue? Motiva la risposta.
3. Determina la retta di regressione dei minimi quadrati.
4. Disegna il grafico dei residui. Sono disposti in modo casuale?

Soluzione

1. Scatter plot



2. Non è chiaro se un modello di regressione lineare semplice sia appropriato per descrivere la relazione tra concentrazione di alcol nel sangue e tempo di reazione. Anche se il coefficiente di correlazione è abbastanza vicino a 1 ($r = 0,94$), dalla figura si evince che i valori di y nella parte centrale tendono a scendere e poi a risalire, descrivendo una curva.
3. L'equazione della retta dei minimo quadrati è $y = 0,099 + 2,661x$.

Soluzione

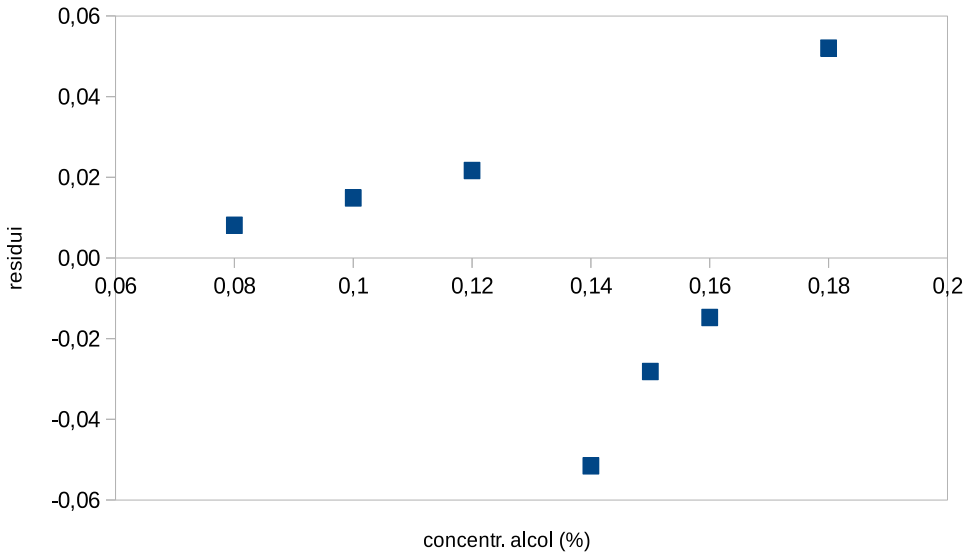
4. I residui sono dati dalla differenza tra valori y_i osservati e i valori teorici $Y_i = 0,099 + 2,661x_i$, per $i = 1, .., n$. Riportiamo i valori nella seguente tabella.

x	y	$Y = 0,099 + 2,661x$	$(y - Y)$
$0,08$	$0,32$	$0,31$	$0,01$
$0,10$	$0,38$	$0,37$	$0,01$
$0,12$	$0,44$	$0,42$	$0,02$
$0,14$	$0,42$	$0,47$	$-0,05$
$0,15$	$0,47$	$0,50$	$-0,03$
$0,16$	$0,51$	$0,52$	$-0,01$
$0,18$	$0,63$	$0,58$	$0,05$

Disegnamo il diagramma di dispersione per la coppia di variabili $(x, (y - Y))$ (grafico dei residui)

Soluzione

Grafico dei residui



I residui non sembrano disporsi in maniera casuale. Questo conferma l'ipotesi che la regressione lineare non sia un buon modello.