STATISTICA esercizi svolti su: INTERPOLAZIONE PONDERATA, REGRESSIONE E CORRELAZIONE

1 INTERPOLAZIONE PONDERATA, REGRESSIONE E CORRELAZIONE

1.1 Esercizi

1. La seguente tabella riporta i dati relativi al numero Y di pezzi prodotti ed al numero X di addetti di 108 imprese di un certo settore economico:

Y / X	10	15	20	Totale
$0 \vdash\dashv 4$	12	12	0	24
$5 \vdash\dashv 11$	12	12	24	48
$12 \vdash\dashv 30$	0	36	0	36
Totale	24	60	24	108

   a) Stabilire se esiste indipendenza in media di Y da X ed in caso di risposta negativa valutare il grado di dipendenza in media utilizzando un indice adeguato;

   b) valutare il grado di correlazione lineare tra X e Y;

   c) calcolare i parametri della retta a minimi quadrati di Y in funzione di X;

   d) con riferimento alla retta ottenuta al punto precedente si calcoli da devianza spiegata e si scomponga opportunamente la devianza totale;

   e) si valuti la bontà di adattamento della retta individuata.

Svolgimento

a) Per stabilire se esiste indipendenza in media del carattere Y dal carattere X, è necessario calcolare le medie parziali di Y.

$$\begin{aligned} \bar{y}_1 = M_1(Y|X \in [0,4]) &= \frac{10 \cdot 12 + 15 \cdot 12 + 20 \cdot 0}{24} \\ &= \frac{300}{24} = 12.5. \end{aligned}$$

$$\begin{aligned} \bar{y}_2 = M_1(Y|X \in [5,11]) &= \frac{10 \cdot 12 + 15 \cdot 12 + 20 \cdot 24}{48} \\ &= \frac{780}{48} = 16.25. \end{aligned}$$

$$\begin{aligned} \bar{y}_3 = M_1(Y|X \in [12,30]) &= \frac{10 \cdot 0 + 15 \cdot 36 + 20 \cdot 0}{36} \\ &= \frac{540}{36} = 15. \end{aligned}$$

Calcoliamo ora anche la media totale del carattere Y:

$$\begin{aligned} \bar{y} = M_1(Y) &= \frac{10 \cdot 24 + 15 \cdot 60 + 20 \cdot 24}{108} \\ &= \frac{1620}{108} = 15. \end{aligned}$$

Poichè non si ha che

$$\bar{y}_1 = \bar{y}_2 = \bar{y}_3 = \bar{y}$$

possiamo concludere che non c'è indipendenza in media del carattere Y dal carattere X.

Calcoliamo la varianza di Y:

$$\begin{aligned} var(Y) = \sigma^2_{TOT} &= M_1(Y^2) - [M_1(Y)]^2 \\ &= \frac{100 \cdot 24 + 225 \cdot 60 + 400 \cdot 24}{108} - (15)^2 \\ &= \frac{25500}{108} - 225 \\ &= 236.\bar{1} - 225 = 11.\bar{1} \end{aligned}$$

e considerando i gruppi determinati dalle modalità del carattere X, calcoliamo la varianza **fra** i gruppi (fra le medie parziali):

$$\begin{aligned} \sigma^2_F &= \frac{1}{N} \sum_{i=1}^{r} (\bar{y}_i - \bar{y})^2 \cdot n_{i\cdot} \\ &= \frac{1}{108} \sum_{i=1}^{3} (\bar{y}_i - \bar{y})^2 \cdot n_{i\cdot} \\ &= \frac{1}{108} \cdot [(12.5 - 15)^2 \cdot 24 + (16.25 - 15)^2 \cdot 48 + (15 - 15)^2 \cdot 36] \\ &= 2.08\bar{3}. \end{aligned}$$

Possiamo a questo punto calcolare il rapporto di correlazione:

$$\eta^2_{(Y/X)} = \frac{\sigma^2_F}{\sigma^2_T} = \frac{2.08\bar{3}}{11.\bar{1}} = 0.1875$$

e concludere che la varianza fra i gruppi (fra le medie parziali) è il 18.75% della varianza totale.
Ricordando che l'indice $\eta^2_{(Y/X)}$ è sempre compreso tra 0 e 1, possiamo concludere che in questo caso, la dipendenza in media di Y da X è debole.

b) Per calcolare il coefficiente di correlazione lineare, è necessario calcolare lo scarto quadratico medio di Y:

$$\sigma(Y) = \sqrt{11.\bar{1}} = 3.\bar{3};$$

la media aritmetica di X:

$$M_1(X) = \bar{x} = \frac{1}{N}\sum_{i=1}^{r} x_i^c \cdot n_{i\cdot} = \frac{2\cdot 24 + 8\cdot 48 + 21\cdot 36}{108} = 11;$$

la varianza di X:

$$\begin{aligned}
\sigma^2(X) &= M_1(X^2) - [M_1(X)]^2 \\
&= \frac{1}{108}\sum_{i=1}^{3}(x_i^c)^2 \cdot n_{i\cdot} - (\bar{x})^2 \\
&= \frac{4\cdot 24 + 64\cdot 48 + 441\cdot 36}{108} - (11)^2 \\
&= \frac{1188}{108} - 121 \\
&= 55.\bar{3}
\end{aligned}$$

da cui si ottiene lo scarto quadratico medio di X:

$$\sigma_X = \sqrt{55.\bar{3}} = 7.4386.$$

Non ci rimane che calcolare la covarianza tra X e Y. É importante sottolineare che, avendo a disposizione una tabella a doppia entrata, il calcolo della covarianza tra X e Y deve tenere conto delle frequenze congiunte n_{ij}:

$$\begin{aligned}
cov(X,Y) &= \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c}(x_i - \bar{x})(y_j - \bar{y})n_{ij} \quad \text{(metodo diretto)} \\
&= \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c} x_i y_j n_{ij} - \bar{x}\bar{y} \qquad \text{(metodo indiretto).}
\end{aligned}$$

Per facilitare il calcolo, completiamo la seguente tabella nel seguente modo: nella cella (i,j) inseriamo il valore ottenuto moltiplicando la i-esima modalità di X per la j-esima modalità di Y per la frequenza congiunta corrispondente n_{ij}:

Y / X	10	15	20	
2	$2\cdot 10\cdot 12 =$ *240*	$2\cdot 15\cdot 12 =$ *360*	$2\cdot 20\cdot 0 =$ *0*	
8	$8\cdot 10\cdot 12 =$ *960*	$8\cdot 15\cdot 12 =$ *1440*	$8\cdot 20\cdot 24 =$ *3840*	
21	$21\cdot 10\cdot 0 =$ *0*	$21\cdot 15\cdot 36 =$ *11340*	$21\cdot 20\cdot 0 =$ *0*	
				$\overset{18180}{= \sum_{i=1}^{r}\sum_{j=1}^{c} x_i y_j n_{ij}}$

Possiamo calcolare quindi la covarianza tra X e Y:

$$\begin{aligned} cov(X,Y) &= \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c} x_i y_j n_{ij} - \bar{x}\bar{y} \\ &= \frac{1}{108}\cdot 18180 - 11\cdot 15 \\ &= 3.\bar{3} \end{aligned}$$

e il coefficiente di correlazione lineare tra X e Y:

$$r(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{3.\bar{3}}{7.4386\cdot 3.\bar{3}} = 0.1344.$$

Ricordando che il coefficiente di correlazione r è sempre compreso tra -1 e 1, possiamo affermare che tra i caratteri X e Y esiste una debole correlazione lineare positiva.

c) Calcoliamo ora i parametri della retta a minimi quadrati (retta di regressione). Ricordiamo che ciò significa determinare i parametri della retta interpolante fra i punti noti $(x_i^c; \bar{y}_i)$ $[i = 1,2,3]$ aventi come coordinate i valori centrali delle classi in cui è suddiviso X e le corrispondenti medie parziali di Y.
Tale interpolazione è però un'interpolazione ponderata: ciò significa che ciascun punto $(x_i^c; \bar{y}_i)$ va considerato avente frequenza pari alla numerosità del gruppo corrispondente $(n_{i\cdot})$.
Per maggiore chiarezza, esplicitiamo che in questo caso la nuvola di punti è costituita dai punti $(2; 12.5), (8; 16.25), (21; 15)$ rispettivamente con frequenze pari a $24, 48, 36$.

Impostiamo il sistema:

$$\begin{cases} \hat{\alpha}_1 = \dfrac{cod(\bar{Y}_i, X)}{dev(X)} \\ \\ \hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \cdot \bar{x}. \end{cases}$$

Calcoliamo per prima cosa la codevianza tra le medie parziali di Y e X:

$$\begin{aligned} cod(\bar{Y}_i, X) &= \sum_{i=1}^{3} (\bar{y}_i - \bar{y})(x_i - \bar{x}) \cdot n_{i\cdot} \\ &= (12.5-15)(2-11)24 + (16.25-15)(8-11)48 + (15-15)(21-11)36 \\ &= 540 - 180 + 0 = 360 \end{aligned}$$

e poi la devianza di X:

$$dev(X) = \sigma^2(X)\cdot N = 55.\bar{3}\cdot 108 = 5976.$$

Se ora sostituiamo nel sistema, otteniamo

$$\begin{cases} \hat{\alpha}_1 = \dfrac{360}{5976} \\ \\ \hat{\alpha}_0 = 15 - \hat{\alpha}_1 \cdot 11 \end{cases}$$

cioè

$$\begin{cases} \hat{\alpha}_1 = 0.0602 \\ \\ \hat{\alpha}_0 = 14.3378. \end{cases}$$

La retta di regressione ha perciò equazione:

$$\hat{Y} = 14.3378 + 0.0602 \cdot X.$$

Interpretiamo i parametri della retta di regressione:

- $\alpha_0 = 14.3378$ significa che (in teoria) un' impresa con 0 addetti ha una produzione media pari a 14.3378 pezzi. Notiamo che in questo caso il valore di α_0 è poco significativo (in quanto non ha senso valutare il numero di pezzi prodotti da un'impresa con 0 addetti);
- $\alpha_1 = 0.0602$ significa che all'aumentare di un addetto, il numero medio di pezzi prodotti aumenta di 0.0602 unità.

In figura (1) vediamo rappresentata graficamente la retta di regressione e la nuvola dei punti con le corrispondenti frequenze.

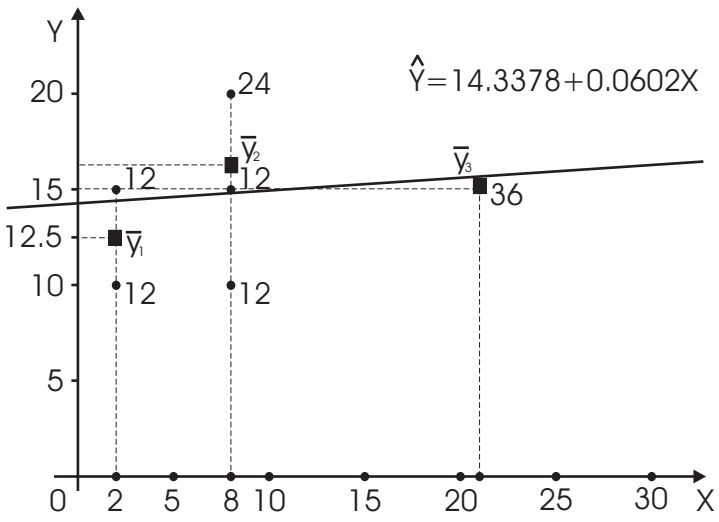

Fig. 1: *Grafico della retta di regressione* $\hat{Y} = 14.3378 + 0.0602 \cdot X$.

É importante notare che se a questo punto, si determinano i parametri della retta interpolante la nuvola di punti costituita dalle coppie $(x_i; y_j)$ $[i, j = 1, 2, 3]$,

tenendo ovviamente in considerazione le frequenze congiunte n_{ij}, si ottiene la stessa retta di regressione individuata precedentemente. Per verificarlo, è necessario seguire il seguente procedimento.

Per calcolare i parametri della retta interpolante la nuvola di punti costituita dalle coppie $(x_i; y_j)$

$$\hat{Y} = \hat{\alpha}_0 + \hat{\alpha}_1 \cdot X$$

è necessario impostare il seguente sistema:

$$\begin{cases} \hat{\alpha}_1 = \dfrac{cov(X,Y)}{var(X)} \\ \\ \hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \cdot \bar{x} \end{cases}$$

da cui, utilizzando le quantità precedentemente ricavate, otteniamo che

$$\begin{cases} \hat{\alpha}_1 = \dfrac{3.\bar{3}}{55.\bar{3}} \\ \\ \hat{\alpha}_0 = 15 - \hat{\alpha}_1 \cdot 11 \end{cases}$$

cioè

$$\begin{cases} \hat{\alpha}_1 = 0.0602 \\ \\ \hat{\alpha}_0 = 14.3378 \end{cases}$$

La retta interpolante la nuvola di punti costituita dalle coppie $(x_i; y_j)$ ha perciò equazione:

$$\hat{Y} = 14.3378 + 0.0602 \cdot X$$

e coincide con la retta di regressione già individuata.

d) Per calcolare la devianza spiegata e la devianza residua, sono necessari i valori $\hat{y}_i$, ovvero i valori previsti della retta di regressione in corrispondenza dei valori centrali delle classi di X:

$$\hat{y}_1 = 14.3378 + 0.0602 \cdot x_1 = 14.3378 + 0.0602 \cdot 2 = 14.4582$$

$$\hat{y}_2 = 14.3378 + 0.0602 \cdot x_2 = 14.3378 + 0.0602 \cdot 8 = 14.8194$$

$$\hat{y}_3 = 14.3378 + 0.0602 \cdot x_3 = 14.3378 + 0.0602 \cdot 21 = 15.602.$$

Per calcolare la devianza spiegata completiamo ora la seguente tabella.

x_i	$\bar{y}_i$	$\hat{y}_i$	$\hat{y}_i - \bar{y}$	$n_{i\cdot}$	$(\hat{y}_i - \bar{y})^2 n_{i\cdot}$
2	12.5	14.4582	-0.5418	24	7.0451
8	16.5	14.8194	-0.1806	48	1.5656
21	15	15.602	0.602	36	13.0465
				108	*21.6572*

Si ha quindi che la devianza spiegata (dalla retta) è:

$$D_S = \sum_{i=1}^{3} (\hat{y}_i - \bar{y})^2 n_{i\cdot} = 21.6572.$$

Si calcola ora la devianza totale:

$$\begin{aligned} D_T &= \sigma_T^2(Y) \cdot N \\ &= 11.\bar{1} \cdot 108 = 1200. \end{aligned}$$

Per calcolare invece la devianza residua, si completa la seguente tabella in cui abbiamo inserito nella cella (i, j) la quantità $(y_j - \hat{y}_i)^2 n_{ij}$:

Y / X	10	15	20
2	$(10 - 14.4582)^2 \cdot 12 =$ *238.506*	$(15 - 14.4582)^2 \cdot 12 =$ *3.522*	$(20 - 14.4582)^2 \cdot 0 =$ *0*
8	$(10 - 14.8194)^2 \cdot 12 =$ *278.7192*	$(15 - 14.8194)^2 \cdot 12 =$ *0.3912*	$(20 - 14.8194)^2 \cdot 24 =$ *644.1264*
21	$(10 - 15.602)^2 \cdot 0 =$ *0*	$(15 - 15.602)^2 \cdot 36 =$ *13.0464*	$(20 - 15.602)^2 \cdot 0 =$ *0*

A titolo esemplificativo, riportiamo i calcoli effettuati per completare la cella centrale della prima colonna (corrispondente a $i = 2$ e $j = 1$).
Il valore contenuto nella cella (2,1) è stato calcolato nel seguente modo: individuato il valore centrale della seconda classe ($i = 2$) del carattere X, $x_2 = 8$, si è sottratto il valore previsto $\hat{y}_2$ dalla retta di regressione in corrispondenza di tale valore dall'effettivo primo ($j = 1$) valore assunto da Y, $y_1 = 10$:

$$y_1 - \hat{y}_2 = 10 - 14.8194 = -4.8194.$$

Il valore trovato è stato poi elevato al quadrato e moltiplicato per la frequenza n_{21}:

$$(y_1 - \hat{y}_2)^2 \cdot n_{21} = (-4.8194)^2 \cdot 12 = 278.7192.$$

I valori contenuti nelle altre celle sono stati calcolati in modo analogo.

Sommando tutti i valori contenuti nelle celle della precedente tabella, otteniamo la devianza residua:

$$D_R = \sum_{i=1}^{3} \sum_{j=1}^{3} (y_j - \hat{y}_i)^2 n_{ij} = 1178.3112.$$

Verifichiamo perciò la scomposizione:

$$\begin{array}{ccccc} 1178.3112 & + & 21.6572 & = & 1199.97\ (\cong 1200) \\ \text{DEVIANZA} & + & \text{DEVIANZA} & = & \text{DEVIANZA} \\ \text{RESIDUA} & & \text{SPIEGATA} & & \text{TOTALE} \end{array}$$

d) Per valutare la bontà di adattamento della retta di regressione, calcoliamo l'indice di determinazione delle medie parziali, rapportando la devianza spiegata alla devianza fra i gruppi:

$$I_d^{*2} = \frac{D_S}{D_F} = \frac{21.6572}{225} = 0.0962$$

dal momento che

$$D_F = \sigma_F^2(Y) \cdot 108 = 2.08\bar{3} \cdot 108 = 225.$$

Il valore di I_d^{*2} indica che la retta di regressione non rappresenta in maniera soddisfacente le medie parziali, visto che la varianza spiegata è pari al 9.62% della varianza fra le medie.

Passiamo a calcolare l'indice di determinazione

$$I_d^2 = \frac{D_S}{D_T} = \frac{21.6572}{1200} = 0.018.$$

Il valore di I_d^2 indica che la retta di regressione spiega solo l'1.8% della variabilità totale del carattere Y.

Entrambi i valori degli indici I_d^{*2} e I_d^2 ci permettono di concludere che la bontà di adattamento della retta di regressione alla situazione analizzata è bassissima.

2. Si consideri la seguente tabella che riporta la distribuzione bivariata delle variabili X e Y:

X / Y	-1	0	1	Totale
0	5	5	0	10
1	10	40	0	50
4	0	25	15	40
Totale	15	70	15	100

a) valutare il grado di dipendenza in media di Y da X;

b) calcolare i parametri della retta interpolante che spiega Y come funzione di X;

c) si calcolino opportuni indici dell'ordine di grandezza dei residui di interpolazione rispetto alla retta individuata al punto precedente;

d) analizzare con un opportuno indice quanta parte della devianza totale è spiegata dall'interpolante lineare;

e) valutare il grado di correlazione lineare tra X e Y e commentare.

Svolgimento

a) Per prima cosa, calcoliamo le medie parziali del carattere Y:

$$\begin{aligned} \bar{y}_1 = M_1(Y|X=-1) &= \frac{0\cdot 5 + 1\cdot 10 + 4\cdot 0}{15} \\ &= \frac{10}{15} = 0.\bar{6} \end{aligned}$$

$$\begin{aligned} \bar{y}_2 = M_1(Y|X=0) &= \frac{0\cdot 5 + 1\cdot 40 + 4\cdot 25}{70} \\ &= \frac{140}{70} = 2 \end{aligned}$$

$$\begin{aligned} \bar{y}_3 = M_1(Y|X=1) &= \frac{0\cdot 0 + 1\cdot 0 + 4\cdot 15}{15} \\ &= \frac{60}{15} = 4. \end{aligned}$$

Calcoliamo ora anche la media totale del carattere Y:

$$\begin{aligned} \bar{y} = M_1(Y) &= \frac{0\cdot 10 + 1\cdot 50 + 4\cdot 40}{100} \\ &= \frac{210}{100} = 2.1. \end{aligned}$$

Poichè non si ha che

$$\bar{y}_1 = \bar{y}_2 = \bar{y}_3 = \bar{y}$$

possiamo concludere che non c'è indipendenza in media del carattere Y dal carattere X.

Calcoliamo la varianza di Y:

$$\begin{aligned} var(Y) = \sigma^2_{TOT} &= M_1(Y^2) - [M_1(Y)]^2 \\ &= \frac{0^2\cdot 10 + 1^2\cdot 50 + 4^2\cdot 40}{100} - (2.1)^2 \\ &= \frac{690}{100} - 4.41 \\ &= 6.9 - 4.41 = 2.49 \end{aligned}$$

e considerando i gruppi determinati dalle modalità del carattere X, calcoliamo la varianza **fra** i gruppi (fra le medie parziali):

$$\begin{aligned}\sigma_F^2 &= \frac{1}{N}\sum_{j=1}^{c}(\bar{y}_j - \bar{y})^2 \cdot n_{\cdot j} \\ &= \frac{1}{100}\sum_{j=1}^{3}(\bar{y}_j - \bar{y})^2 \cdot n_{\cdot j} \\ &= \frac{1}{100}\cdot[(0.\bar{6} - 2.1)^2 \cdot 15 + (2-2.1)^2 \cdot 70 + (4-2.1)^2 \cdot 15] \\ &= \frac{85.\bar{6}}{100} = 0.85\bar{6}.\end{aligned}$$

Si può a questo punto calcolare il rapporto di correlazione:

$$\eta^2_{(Y/X)} = \frac{\sigma_F^2}{\sigma_T^2} = \frac{0.85\bar{6}}{2.49} = 0.344$$

osservando che la varianza fra i gruppi (fra le medie parziali) rappresenta il 34.4% della varianza totale.
Ricordando che l'indice $\eta^2_{(Y/X)}$ è sempre compreso tra 0 e 1, possiamo concludere che esiste una bassa dipendenza in media di Y da X.

b) Si determinano ora i parametri $\hat{\alpha}_0$ e $\hat{\alpha}_1$ della retta interpolante

$$\hat{Y} = \hat{\alpha}_0 + \hat{\alpha}_1 \cdot X$$

con

$$\begin{cases} \hat{\alpha}_1 = \dfrac{cov(X, \bar{Y}_j)}{var(X)} \\ \\ \hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \cdot \bar{x} \end{cases}$$

dove si è indicato con $\bar{Y}_j$ il carattere che assume valori pari alle medie parziali di Y con frequenze pari alle numerosità dei gruppi.
Si calcola la media aritmetica di X:

$$M_1(X) = \bar{x} = \frac{-1 \cdot 15 + 0 \cdot 70 + 1 \cdot 15}{100} = 0$$

e la varianza di X:

$$\begin{aligned}\sigma^2(X) &= M_1(X^2) - [M_1(X)]^2 \\ &= \frac{1}{100}\sum_{j=1}^{3}(x_j)^2 \cdot n_{\cdot j} - (\bar{x})^2 \\ &= \frac{(-1)^2 \cdot 15 + 0^2 \cdot 70 + 1^2 \cdot 15}{100} - (0)^2\end{aligned}$$

$$\begin{aligned} &= \frac{30}{100} - 0 \\ &= 0.3 \end{aligned}$$

e si completa la tabella

x_j	$\bar{y}_j$	$n_{\cdot j}$	$x_j \bar{y}_j n_{\cdot j}$
-1	$0.\bar{6}$	15	$-9.\bar{9}$
0	2	70	0
1	4	15	60
		100	*50*

Calcolando quindi la covarianza tra X e le medie parziali di Y, si ha:

$$\begin{aligned} cov(X, \bar{Y}_j) &= \frac{1}{N} \sum_{j=1}^{3} x_j \bar{y}_j n_{\cdot j} - \bar{x}\bar{y} \\ &= \frac{1}{100} \cdot 50 - 2.1 \cdot 0 \\ &= 0.5. \end{aligned}$$

Sostituendo nel sistema, si ottiene

$$\begin{cases} \hat{\alpha}_1 = \dfrac{0.5}{0.3} \\ \\ \hat{\alpha}_0 = 2.1 - \hat{\alpha}_1 \cdot 0 \end{cases}$$

da cui

$$\begin{cases} \hat{\alpha}_1 = 1.\bar{6} \\ \\ \hat{\alpha}_0 = 2.1. \end{cases}$$

L'equazione della retta di regressione è pertanto:

$$\hat{Y} = 2.1 + 1.\bar{6} \cdot X.$$

In figura (2) è riportata la rappresentazione grafica della retta di regressione e la nuvola dei punti con la corrispondente frequenza.

Interpretiamo i parametri della retta di regressione:

- $\alpha_0 = 2.1$ significa che la retta di regressione prevede per la variabile Y, il valore medio 2.1, in corrispondenza del valore 0 della variabile X;

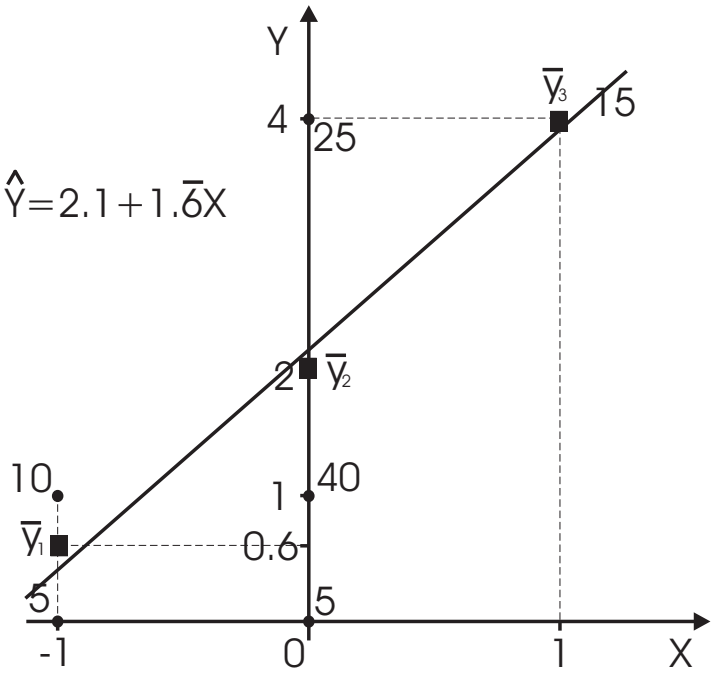

Fig. 2: *Grafico della retta di regressione* $\hat{Y} = 2.1 + 1.\bar{6} \cdot X$.

– $\alpha_1 = 1.\bar{6}$ significa che la retta prevede, attuando un incremento unitario della variabile X, un aumento medio del valore della variabile Y di $1.\bar{6}$.

c) Per valutare l'ordine di grandezza dei residui di interpolazione, completiamo la seguente tabella:

x_j	$\bar{y}_j$	$\hat{y}_j$	$n_{\cdot j}$	$\lvert\bar{y}_j - \hat{y}_j\rvert$	$\lvert\bar{y}_j - \hat{y}_j\rvert n_{\cdot j}$	$\lvert\bar{y}_j - \hat{y}_j\rvert^2$	$\lvert\bar{y}_j - \hat{y}_j\rvert^2 n_{\cdot j}$
-1	$0.\bar{6}$	$0.4\bar{3}$	15	$0.2\bar{3}$	3.5	0.054	0.81
0	2	2.1	70	0.1	7	0.01	0.7
1	4	$3.7\bar{6}$	15	$0.2\bar{3}$	3.5	0.054	0.81
			100		*14*		*2.32*

e calcoliamo la media aritmetica dei moduli dei residui:

$$\begin{aligned} A_1^* &= \frac{1}{N}\sum_{j=1}^{3} |\bar{y}_j - \hat{y}_j| n_{\cdot j} \\ &= \frac{1}{100} \cdot 14 = 0.14. \end{aligned}$$

Tale valore indica che mediamente i valori previsti dalla retta di regressione si discostano dalle medie parziali di 0.14.

Possiamo anche calcolare la media quadratica dei residui:

$$A_2^* = \sqrt{\frac{1}{N}\sum_{j=1}^{3} |\bar{y}_j - \hat{y}_j|^2 n_{\cdot j}}$$

$$= \sqrt{\frac{1}{100} \cdot 2.32} = \sqrt{0.0232} = 0.152$$

e interpretare il valore ottenuto nel seguente modo: mediamente (in media quadratica) i valori previsti dalla retta di regressione si discostano dalle medie parziali di 0.152.

d) Per valutare quanta parte della varianza totale è spiegata dalla retta interpolante, bisogna calcolare l'indice di determinazione :

$$I_d^2 = \frac{\sigma_S^2}{\sigma_T^2}.$$

Calcoliamo perciò la varianza spiegata: per far ciò, completiamo la tabella seguente, ricordando che $\bar{y} = 2.1$.

x_j	$\hat{y}_j$	$n_{\cdot j}$	$(\hat{y}_j - \bar{y})^2$	$(\hat{y}_j - \bar{y})^2 n_{\cdot j}$
-1	$0.4\bar{3}$	15	$2.\bar{7}$	$41.\bar{6}$
0	2.1	70	0	0
1	$3.7\bar{6}$	15	$2.\bar{7}$	$41.\bar{6}$
		100		*83.3*

Abbiamo perciò che la varianza spiegata è

$$\sigma_S^2 = \frac{1}{100} \sum_{j=1}^{3} (\hat{y}_j - \bar{y})^2 n_{\cdot j} = \frac{83.\bar{3}}{100} = 0.8\bar{3}.$$

Ricordando che $var(Y) = \sigma_{TOT}^2 = 2.49$, ricaviamo l'indice di determinazione

$$I_d^2 = \frac{0.8\bar{3}}{2.49} = 0.335.$$

Tale valore indica che la retta di regressione spiega il 33.5% della variabilità totale di Y.
Se però calcoliamo l'indice I_d^{*2}, otteniamo

$$I_d^{*2} = \frac{\sigma_S^2}{\sigma_F^2} = \frac{0.8\bar{3}}{0.85\bar{6}} = 0.973.$$

Tale valore indica che la retta di regressione spiega il 97.3% della variabilità fra le medie parziali di Y.

Confrontando i valori dei due indici I_d^2 e I_d^{*2}, possiamo concludere che il modello $\hat{Y} = 2.1 + 1.\bar{6} \cdot X$ spiega bene la variabilità fra le medie parziali, ma non la variabilità totale, perchè la varianza nei gruppi è elevata.

d) Per valutare il grado di correlazione lineare tra X e Y, calcoliamo il coefficiente di correlazione lineare:

$$r(X,Y) = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)}.$$

Ricordando che

$$cov(X,Y) = cov(X,\bar{Y}_j) = 0.5,$$
$$\sigma^2(X) = 0.3,$$
$$\sigma^2(Y) = 2.49,$$

si ottiene

$$r(X,Y) = \frac{0.5}{\sqrt{0.3}\sqrt{2.49}} = 0.5785.$$

Dal valore ottenuto concludiamo che tra il carattere X e il carattere Y c'è una media correlazione lineare positiva.

3. La seguente tabella riporta i dati (in migliaia) relativi agli occupati con doppio lavoro classificati in base alle ore settimanali di lavoro impiegate nella attività principale (carattere X) e nelle attività secondarie (carattere Y):

X / Y	5–15	16–25	26–40	41–50	Totale
0–10	4	7	133	58	202
11–20	5	15	66	21	107
21–30	12	11	11	4	38
31–40	20	2	2	1	25
Totale	41	35	212	84	372

.

a) Esiste indipendenza distributiva? In caso di risposta negativa costruire la tabella delle frequenze congiunte in modo che i caratteri X e Y risultino indipendenti in distribuzione.

b) Calcolare e commentare le contingenze assolute.

c) Calcolare un indice che misuri il grado di connessione tra i due caratteri.

d) Esiste indipendenza in media di Y da X? In caso di risposta negativa si valuti il grado di dipendenza in media.

e) Calcolare i parametri della retta interpolante a minimi quadrati che si ritiene più opportuna dato il significato dei caratteri e tracciarne il grafico.

f) Dopo aver calcolato la varianza spiegata, scomporre opportunamente la varianza totale.

g) Valutare con un opportuno indice la bontà di adattamento della retta individuata al punto e).

h) Calcolare ed interpretare il coefficiente di correlazione lineare tra X e Y.

Svolgimento

a) Se esiste indipendenza distributiva tra X e Y, sappiamo che deve valere la relazione

$$n_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{N} \qquad \text{per } i = 1,2,3,4; \text{ per } j = 1,2,3,4.$$

Verifichiamo se vale tale relazione per $i = 1$ e $j = 1$:

$$\frac{n_{1\cdot} \cdot n_{\cdot 1}}{N} = \frac{202 \cdot 41}{372} = 22.26$$

e

$$n_{11} = 4.$$

Poichè $22.26 \neq 4$, possiamo concludere che non c'è indipendenza distributiva. Costruiamo perciò la tabella delle frequenze teoriche $\hat{n}_{ij}$ in caso di indipendenza distributiva.

X / Y	5–15	16–25	26–40	41–50	Totale
0–10	22.26	19	115.12	45.61	202
11–20	11.79	10.07	60.98	24.16	107
21–30	4.19	3.57	21.66	8.58	38
31–40	2.76	2.35	14.25	5.65	25
Totale	41	35	212	84	372

b) Costruiamo la tabella delle contingenze assolute $C_{ij} = n_{ij} - \hat{n}_{ij}$:

X / Y	5–15	16–25	26–40	41–50	Totale
0–10	-18.26	-12	17.88	12.39	0
11–20	-6.79	4.93	5.02	-3.16	0
21–30	7.81	7.43	-10.66	-4.58	0
31–40	17.24	-0.35	-12.25	4.65	0
Totale	0	0	0	0	0

Il valore delle contingenze assolute appena calcolate fornisce le seguenti informazioni:

- $C_{11} = -18.26$: la frequenza congiunta effettiva associata alle classi "$5-15$" del carattere X e "$0-10$" del carattere Y, risulta essere minore di quella teorica in ipotesi di indipendenza distributiva. Tra le classi "$5-15$" del carattere X e "$0-10$" del carattere Y vi è repulsione in quanto la frequenza congiunta che si è osservata è inferiore a quella che si sarebbe dovuta osservare se tra i due caratteri ci fosse stata indipendenza distributiva;

- $C_{12} = -12$: la frequenza congiunta effettiva associata alle classi "$16-25$" del carattere X e "$0-10$" del carattere Y, risulta essere minore di quella teorica in ipotesi di indipendenza distributiva. Tra le classi "$16-25$" del

carattere X e "$0-10$" del carattere Y vi è repulsione in quanto la frequenza congiunta che si è osservata è inferiore a quella che si sarebbe dovuta osservare se tra i due caratteri ci fosse stata indipendenza distributiva;

- $C_{13} = 17.88$: la frequenza congiunta effettiva associata alle classi "$26-40$" del carattere X e "$0-10$" del carattere Y, risulta essere maggiore di quella teorica in ipotesi di indipendenza distributiva. Tra le classi "$26-40$" del carattere X e "$0-10$" del carattere Y vi è attrazione in quanto la frequenza congiunta che si è osservata è superiore a quella che si sarebbe dovuta osservare se tra i due caratteri ci fosse stata indipendenza distributiva;

- $C_{14} = 12.39$: la frequenza congiunta effettiva associata alle classi "$41-50$" del carattere X e "$0-10$" del carattere Y, risulta essere maggiore di quella teorica in ipotesi di indipendenza distributiva. Tra le classi "$41-50$" del carattere X e "$0-10$" del carattere Y vi è attrazione in quanto la frequenza congiunta che si è osservata è superiore a quella che si sarebbe dovuta osservare se tra i due caratteri ci fosse stata indipendenza distributiva;

- $C_{21} = -6.79$: la frequenza congiunta effettiva associata alle classi "$5-15$" del carattere X e "$11-20$" del carattere Y, risulta essere minore di quella teorica in ipotesi di indipendenza distributiva. Tra le classi "$5-15$" del carattere X e "$11-20$" del carattere Y vi è repulsione in quanto la frequenza congiunta che si è osservata è inferiore a quella che si sarebbe dovuta osservare se tra i due caratteri ci fosse stata indipendenza distributiva;

- $C_{22} = 4.93$: la frequenza congiunta effettiva associata alle classi "$16-25$" del carattere X e "$11-20$" del carattere Y, risulta essere maggiore di quella teorica in ipotesi di indipendenza distributiva. Tra le classi "$16-25$" del carattere X e "$11-20$" del carattere Y vi è attrazione in quanto la frequenza congiunta che si è osservata è superiore a quella che si sarebbe dovuta osservare se tra i due caratteri ci fosse stata indipendenza distributiva;

- $C_{23} = 5.02$: la frequenza congiunta effettiva associata alle classi "$26-40$" del carattere X e "$11-20$" del carattere Y, risulta essere maggiore di quella teorica in ipotesi di indipendenza distributiva. Tra le classi "$26-40$" del carattere X e "$11-20$" del carattere Y vi è attrazione in quanto la frequenza congiunta che si è osservata è superiore a quella che si sarebbe dovuta osservare se tra i due caratteri ci fosse stata indipendenza distributiva;

- $C_{24} = -3.16$: la frequenza congiunta effettiva associata alle classi "$41-50$" del carattere X e "$11-20$" del carattere Y, risulta essere minore di quella teorica in ipotesi di indipendenza distributiva. Tra le classi "$41-50$" del carattere X e "$11-20$" del carattere Y vi è repulsione in quanto la frequenza congiunta che si è osservata è inferiore a quella che si sarebbe dovuta

osservare se tra i due caratteri ci fosse stata indipendenza distributiva;

- $C_{31} = 7.81$: la frequenza congiunta effettiva associata alle classi "$5-15$" del carattere X e "$21-30$" del carattere Y, risulta essere maggiore di quella teorica in ipotesi di indipendenza distributiva. Tra le classi "$5-15$" del carattere X e "$21-30$" del carattere Y vi è attrazione in quanto la frequenza congiunta che si è osservata è superiore a quella che si sarebbe dovuta osservare se tra i due caratteri ci fosse stata indipendenza distributiva;

- $C_{32} = 7.43$: la frequenza congiunta effettiva associata alle classi "$16-25$" del carattere X e "$21-30$" del carattere Y, risulta essere maggiore di quella teorica in ipotesi di indipendenza distributiva. Tra le classi "$16-25$" del carattere X e "$21-30$" del carattere Y vi è attrazione in quanto la frequenza congiunta che si è osservata è superiore a quella che si sarebbe dovuta osservare se tra i due caratteri ci fosse stata indipendenza distributiva;

- $C_{33} = -10.66$: la frequenza congiunta effettiva associata alle classi "$26-40$" del carattere X e "$21-30$" del carattere Y, risulta essere minore di quella teorica in ipotesi di indipendenza distributiva. Tra le classi "$26-40$" del carattere X e "$21-30$" del carattere Y vi è repulsione in quanto la frequenza congiunta che si è osservata è inferiore a quella che si sarebbe dovuta osservare se tra i due caratteri ci fosse stata indipendenza distributiva;

- $C_{34} = -4.58$: la frequenza congiunta effettiva associata alle classi "$41-50$ del carattere X e "$21-30$" del carattere Y, risulta essere minore di quella teorica in ipotesi di indipendenza distributiva. Tra le classi "$41-50$" del carattere X e "$21-30$" del carattere Y vi è repulsione in quanto la frequenza congiunta che si è osservata è inferiore a quella che si sarebbe dovuta osservare se tra i due caratteri ci fosse stata indipendenza distributiva;

- $C_{41} = 17.24$: la frequenza congiunta effettiva associata alle classi "$5-15$" del carattere X e "$31-40$" del carattere Y, risulta essere maggiore di quella teorica in ipotesi di indipendenza distributiva. Tra le classi "$5-15$" del carattere X e "$31-40$" del carattere Y vi è attrazione in quanto la frequenza congiunta che si è osservata è superiore a quella che si sarebbe dovuta osservare se tra i due caratteri ci fosse stata indipendenza distributiva;

- $C_{42} = -0.35$: la frequenza congiunta effettiva associata alle classi "$16-25$" del carattere X e "$31-40$" del carattere Y, risulta essere minore di quella teorica in ipotesi di indipendenza distributiva. Tra le classi "$16-25$" del carattere X e "$31-40$" del carattere Y vi è repulsione in quanto la frequenza congiunta che si è osservata è inferiore a quella che si sarebbe dovuta osservare se tra i due caratteri ci fosse stata indipendenza distributiva;

- $C_{43} = -12.25$: la frequenza congiunta effettiva associata alle classi "$26-40$" del carattere X e "$31-40$" del carattere Y, risulta essere minore di quella teorica in ipotesi di indipendenza distributiva. Tra le classi "$26-40$" del carattere X e "$31-40$" del carattere Y vi è repulsione in quanto la frequenza congiunta che si è osservata è inferiore a quella che si sarebbe dovuta osservare se tra i due caratteri ci fosse stata indipendenza distributiva;

- $C_{44} = -4.65$: la frequenza congiunta effettiva associata alle classi "$41-50$" del carattere X e "$31-40$" del carattere Y, risulta essere minore di quella teorica in ipotesi di indipendenza distributiva. Tra le classi "$41-50$" del carattere X e "$31-40$" del carattere Y vi è repulsione in quanto la frequenza congiunta che si è osservata è inferiore a quella che si sarebbe dovuta osservare se tra i due caratteri ci fosse stata indipendenza distributiva.

c) Calcoliamo un indice che misuri il grado di connessione. Scegliamo l'indice di connessione quadratico di Pearson: troviamo quindi la media quadratica ponderata delle contingenze relative (ρ_{ij}), con pesi pari alle frequenze teoriche ($\hat{n}_{ij}$)

$$\begin{aligned} M_2(|\rho|) &= \sqrt{\frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c}\rho_{ij}^2\cdot\hat{n}_{ij}} \\ &= \sqrt{\frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}^2}\cdot\hat{n}_{ij}} \\ &= \sqrt{\frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}}} \\ &= \sqrt{\frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{C_{ij}^2}{\hat{n}_{ij}}} \end{aligned}$$

Per completare i conti, si completa la seguente tabella in cui inseriamo nella cella (i,j) la quantità $\dfrac{C_{ij}^2}{\hat{n}_{ij}}$:

X / Y	5–15	16–25	26–40	41–50
0–10	14.98	7.58	2.78	3.37
11–20	3.91	2.41	0.41	0.41
21–30	14.56	15.46	5.25	2.44
31–40	107.69	0.05	10.53	3.82

195.65

Si ha quindi

$$M_2(|\rho|) = \sqrt{\frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{C_{ij}^2}{\hat{n}_{ij}}} = \sqrt{\frac{1}{372}\cdot 195.65} = 0.53.$$

Il valore appena trovato informa che, in media quadratica, le frequenze effettive differiscono da quelle teoriche del 53% del valore di quest'ultime.

Per valutare il grado di connessione tra i caratteri X e Y è necessario calcolare un indice di connessione normalizzato. Calcoliamo perciò l'indice

$$C = \frac{M_2(|\rho|)}{(k-1)^{\frac{1}{2}}}$$

dove k è il minimo tra il numero di modalità del carattere X e il numero di modalità del carattere Y.
Nel nostro caso, $k = 4$, quindi:

$$C = \frac{0.53}{(4-1)^{\frac{1}{2}}} = \frac{0.53}{\sqrt{3}} = 0.30.$$

Il valore ottenuto ci informa che l'indice quadratico di connessione di Pearson ($M_2(|\rho|)$) è pari al 30% del suo massimo valore (che corrisponde al caso di massima connessione).
Possiamo pertanto affermare che tra i due caratteri X e Y vi è un basso grado di connessione.

d) Per valutare se c'è indipendenza in media di Y da X, calcoliamo le medie parziali di Y, utilizzando i valori centrali delle classi:

$$\bar{y}_1 = M_1(Y|X \in [5,15]) = \frac{5\cdot 4 + 15.5\cdot 5 + 25.5\cdot 12 + 35.5\cdot 20}{41} = 27.16$$

$$\bar{y}_2 = M_1(Y|X \in [16,25]) = \frac{5\cdot 7 + 15.5\cdot 15 + 25.5\cdot 11 + 35.5\cdot 2}{35} = 17.69$$

$$\bar{y}_3 = M_1(Y|X \in [26,40]) = \frac{5\cdot 133 + 15.5\cdot 66 + 25.5\cdot 11 + 35.5\cdot 2}{212} = 9.62$$

$$\bar{y}_4 = M_1(Y|X \in [41,50]) = \frac{5\cdot 58 + 15.5\cdot 21 + 25.5\cdot 4 + 35.5\cdot 1}{84} = 8.96.$$

Calcoliamo ora anche la media totale del carattere Y:

$$\bar{y} = M_1(Y) = \frac{5\cdot 202 + 15.5\cdot 107 + 25.5\cdot 38 + 35.5\cdot 25}{372} = 12.16.$$

Poichè non si ha che

$$\bar{y}_1 = \bar{y}_2 = \bar{y}_3 = \bar{y}_4 = \bar{y}$$

possiamo concludere che non c'è indipendenza in media del carattere Y dal carattere X.

Calcoliamo la varianza di Y:

$$\begin{aligned} var(Y) = \sigma^2_{TOT} &= M_1(Y^2) - [M_1(Y)]^2 \\ &= \frac{5^2 \cdot 202 + 15.5^2 \cdot 107 + 25.5^2 \cdot 38 + 35.5 \cdot 25}{372} - (12.16)^2 \\ &= \frac{86972.5}{372} - 147.85 \\ &= 233.80 - 147.85 = 85.93. \end{aligned}$$

e la varianza **fra** i gruppi (fra le medie parziali):

$$\begin{aligned} \sigma^2_F &= \frac{1}{N}\sum_{j=1}^{c}(\bar{y}_j - \bar{y})^2 \cdot n_{\cdot j} \\ &= \frac{1}{372}\sum_{j=1}^{4}(\bar{y}_j - \bar{y})^2 \cdot n_{\cdot j} \\ &= \frac{1}{372} \cdot [(27.16 - 12.16)^2 \cdot 41 + (17.69 - 12.16)^2 \cdot 35 \\ &\qquad +(9.62 - 12.16)^2 \cdot 212 + (8.96 - 12.16)^2 \cdot 84] \\ &= \frac{12551.28}{372} = 33.74. \end{aligned}$$

Possiamo a questo punto calcolare il rapporto di correlazione:

$$\eta^2_{(Y/X)} = \frac{\sigma^2_F}{\sigma^2_T} = \frac{33.74}{85.93} = 0.392$$

ed osservare che la varianza fra i gruppi (fra le medie parziali) è il 39.2% della varianza totale.
Ricordando che l'indice $\eta^2_{(Y/X)}$ è sempre compreso tra 0 e 1, possiamo concludere che esiste una bassa dipendenza in media di Y da X.

e) Determiniamo i parametri della retta di regressione

$$\hat{Y} = \alpha_0 + \alpha_1 \cdot X$$

in modo da ricavare il numero di ore destinate alle attività secondarie in funzione delle ore dedicate all'attività principale.

Calcoliamo la media aritmetica di X:

$$M_1(X) = \bar{x} = \frac{1}{N}\sum_{j=1}^{c} x^c_j \cdot n_{\cdot j}$$

$$\begin{aligned} &= \frac{10 \cdot 41 + 20.5 \cdot 35 + 33 \cdot 212 + 45.5 \cdot 84}{372} \\ &= 32.11 \end{aligned}$$

e la varianza di X:

$$\begin{aligned} \sigma^2(X) &= M_1(X^2) - [M_1(X)]^2 \\ &= \frac{1}{372}\sum_{j=1}^{4}(x_j^c)^2 \cdot n_{\cdot j} - (\bar{x})^2 \\ &= \frac{(10)^2 \cdot 41 + (20.5)^2 \cdot 35 + (33)^2 \cdot 212 + (45.5)^2 \cdot 84}{372} - (32.11)^2 \\ &= \frac{423577.75}{372} - 1031.0521 \\ &= 107.5 \end{aligned}$$

Avendo già calcolato le medie parziali di Y e completiamo la seguente tabella.

x_j^c	$\bar{y}_j$	$n_{\cdot j}$	$x_j \bar{y}_j n_{\cdot j}$
10	27.16	41	11135.6
20.5	17.69	35	12692.575
33	9.62	212	67301.52
45.5	8.96	84	34245.12
			125374.815

e calcoliamo la covarianza tra X e le medie parziali di Y (che sappiamo coincidere con $cov(X, Y)$):

$$\begin{aligned} cov(X, \bar{Y_j}) = cov(X, Y) &= \frac{1}{N}\sum_{j=1}^{4} x_j \bar{y}_j n_{\cdot j} - \bar{x}\bar{y} \\ &= \frac{1}{372} \cdot 125374.815 - (32.11 \cdot 12.16) \\ &= -53.42. \end{aligned}$$

A questo punto possiamo risolvere il sistema

$$\begin{cases} \hat{\alpha}_1 = \dfrac{cov(X, Y)}{var(X)} \\ \\ \hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \cdot \bar{x} \end{cases}$$

sostituendo i valori:

$$\begin{cases} \hat{\alpha}_1 = \dfrac{-53.42}{107.5} \\ \\ \hat{\alpha}_0 = 12.16 - \hat{\alpha}_1 \cdot 32.11 \end{cases}$$

e otteniamo

$$\begin{cases} \hat{\alpha}_1 = -0.50 \\ \\ \hat{\alpha}_0 = 28.22. \end{cases}$$

L'equazione della retta di regressione è pertanto

$$\hat{Y} = 28.22 - 0.5 \cdot X.$$

Interpretiamo i parametri della retta di regressione:

- $\alpha_0 = 28.22$ significa che la retta di regressione prevede per la variabile Y, il valore medio 28.22, in corrispondenza del valore 0 per la variabile X;
- $\alpha_1 = 0.5$ significa che all'incremento unitario della variabile X, il valore medio della variabile Y aumenta di 0.5.

Il grafico della retta è riportato in figura (3).

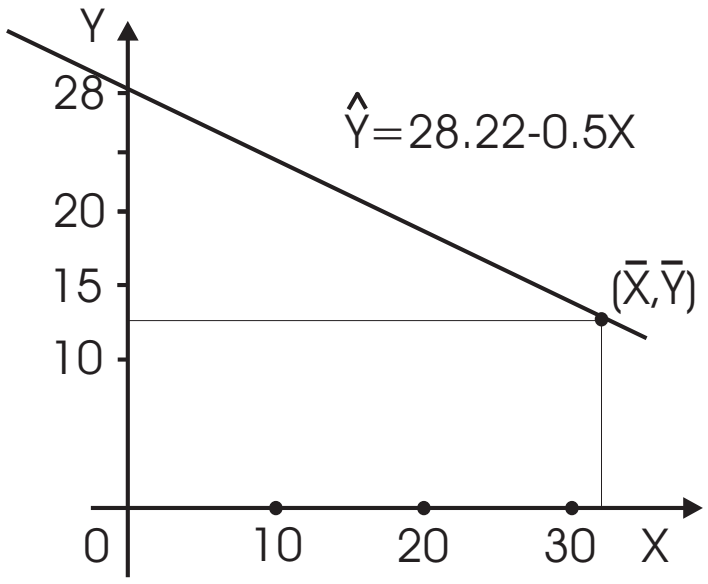

Fig. 3: *Grafico della retta di regressione* $\hat{Y} = 28.22 - 0.5 \cdot X$.

f) Per calcolare la devianza spiegata e la devianza residua, sono necessari i valori $\hat{y}_j$, ovvero i valori previsti della retta di regressione in corrispondenza dei valori centrali delle classi di X: calcoliamoli.

$$\hat{y}_1 = 28.22 - 0.5 \cdot x_1^c = 28.22 - 0.5 \cdot 10 = 23.22$$

$$\hat{y}_2 = 28.22 - 0.5 \cdot x_2^c = 28.22 - 0.5 \cdot 20.5 = 17.97$$

$$\hat{y}_3 = 28.22 - 0.5 \cdot x_3^c = 28.22 - 0.5 \cdot 33 = 11.72$$

$$\hat{y}_4 = 28.22 - 0.5 \cdot x_4^c = 28.22 - 0.5 \cdot 45.5 = 5.47.$$

Per calcolare la varianza spiegata, completiamo la seguente tabella.

x_j^c	$\hat{y}_j$	$n_{\cdot j}$	$(\hat{y}_j - \bar{y})^2 n_{\cdot j}$
10	23.22	41	5015.27
20.5	17.97	35	1181.46
33	11.72	212	41.04
45.5	5.47	84	3759.51
			9997.28

Quindi la varianza spiegata è:

$$\sigma_S^2 = \frac{1}{372} \cdot \sum_{j=1}^{4} (\hat{y}_j - \bar{y})^2 n_{\cdot j} = 26.87.$$

Calcoliamo ora la varianza residua, completando la seguente tabella in cui andiamo a calcolare nella cella (i, j) la quantità $(y_i - \hat{y}_j)^2 n_{ij}$.

x_j^c / y_i^c	10	20.5	33	45.5
5	$(5 - 23.22)^2 \cdot 4$ $= 1327.87$	$(5 - 17.97)^2 \cdot 7$ $= 1177.55$	$(5 - 11.72)^2 \cdot 133$ $= 6006.07$	$(5 - 5.47)^2 \cdot 58$ $= 12.81$
15.5	$(15.5 - 23.22)^2 \cdot 5$ $= 297.99$	$(15.5 - 17.97)^2 \cdot 15$ $= 91.51$	$(15.5 - 11.72)^2 \cdot 66$ $= 943.03$	$(15.5 - 5.47)^2 \cdot 21$ $= 2112.62$
25.5	$(25.5 - 23.22)^2 \cdot 12$ $= 62.38$	$(25.5 - 17.97)^2 \cdot 11$ $= 623.71$	$(25.5 - 11.72)^2 \cdot 11$ $= 2088.77$	$(25.5 - 5.47)^2 \cdot 4$ $= 1604.8$
35.5	$(35.5 - 23.22)^2 \cdot 20$ $= 3015.97$	$(35.5 - 17.97)^2 \cdot 2$ $= 614.6$	$(35.5 - 11.72)^2 \cdot 2$ $= 1130.98$	$(35.5 - 5.47)^2 \cdot 1$ $= 901.8$

Facendo la media di tutti i valori, otteniamo la varianza residua:

$$\sigma_R^2 = \frac{1}{372} \sum_{i=1}^{4} \sum_{j=1}^{4} (y_i - \hat{y}_j)^2 n_{ij} = \frac{22012.46}{372} = 59.17.$$

É quindi verificata la scomposizione:

$$\begin{array}{ccccc} 59.17 & + & 26.87 & = & 86.04\ (\cong 85.93) \\ \text{DEVIANZA} & + & \text{DEVIANZA} & = & \text{DEVIANZA} \\ \text{RESIDUA} & & \text{SPIEGATA} & & \text{TOTALE} \end{array}$$

g) Valutiamo la bontà di adattamento della retta di regressione, calcolando l'indice di determinazione:

$$I_d^2 = \frac{D_S}{D_T} = \frac{26.87}{85.93} = 0.31.$$

Il 31% della variabilità totale del carattere Y è spiegato dalla retta di regressione: abbiamo quindi una scarsa bontà di adattamento.

h) Calcoliamo il coefficiente di correlazione lineare:

$$r(X,Y))\frac{cov(X,Y)}{\sqrt{var(X)}\sqrt{var(Y)}} = \frac{-53.42}{\sqrt{107.5}\sqrt{85.93}} = -0.55.$$

Dal valore del coefficiente di correlazione lineare, possiamo dedurre che esiste una discreta correlazione lineare negativa tra i due caratteri.