Esercitazione del 21.05.2013

Esercizi sulla regressione lineare.

Esercizio dal tema d'esame del 13.06.2011.

Si consideri il seguente campione di $n = 9$ osservazioni relative ai caratteri X ed Y:

X	7	17	8	36	23	19	17	2	5
Y	25	0	11	-26	-2	-18	8	26	35

1. Rappresentare la distribuzione congiunta dei due caratteri mediante un diagramma a dispersione. Sulla base del grafico il modello di regressione lineare $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, \ldots, n$ appare adeguato? Motivare la risposta.

2. Stimare i parametri $\beta_0,\ \beta_1$ col metodo dei minimi quadrati.

3. Costruire un intervallo di confidenza di livello 99% per β_1

4. Calcolare una misura della bontà di adattamento del modello.

5. Supponiamo di definire il carattere $Y^\star = 40 - Y$. Si consideri il modello $Y_i^\star = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, \ldots, n$. $\hat{\beta}_1^\star$, la nuova stima di minimi quadrati di β_1

i) rimane uguale a quella calcolata al punto 2:

ii) cambia di segno ma il valore assoluto resta lo stesso.

iii) cambiano sia il segno che il valore calcolato.

La stessa risposta vale anche per $\hat{\beta}_0^\star$?

Soluzione. 1. Il grafico di dispersione è rappresentato in Figura 1. Il grafico suggerisce una piena adeguatezza, confermata dal valore del coefficiente di correlazione lineare. Infatti, si ha:

$$\bar{x} = 14.89,\ \bar{y} = 6.56$$

$$\sum_i (x_i = \bar{x})(y_i - \bar{y}) = -1576.44,\ \sum_i (x_i = \bar{x})^2 = 910.89,\ \sum_i (y_i - \bar{y})^2 = 3328.22$$

da cui:

$$\rho = \frac{-1576.44}{\sqrt{910.89 \times 3328.22}} = -0.905$$

e, pertanto, si può concludere che esiste una forte correlazione lineare negativa.

2.

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{-1576.44}{910.89} = -1.73$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 6.56 + 1.73 \times 14.89 = 32.32$$

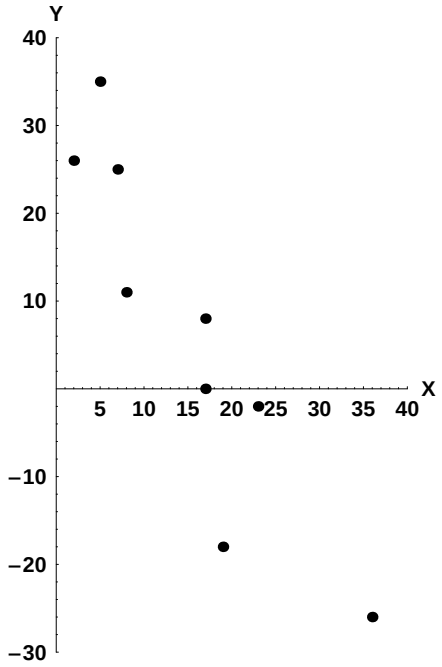

Figura 1: Diagramma di dispersione per l'Esercizio dal tema d'esame 13.06.2011.

3. L'intervallo di confidenza al livello del 99% per β_1 è dato da

$$\hat{\beta}_1 \pm t_{0.005}(7)\sqrt{\frac{\sum \hat{e}_i^2/(n-2)}{\sum_i (x_i - \bar{x})^2}}$$

ove $\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ sono i residui. Nel nostro caso: $t_{0.005}(7) = 3.50$ e $\sqrt{\sum \hat{e}_i^2/(n-2)} = 85.70$ e quindi

$$\text{IC}_{99\%}(\beta_1): \quad -1.73 \pm 3.50 \times \frac{\sqrt{85.70}}{\sqrt{910.89}}: \quad (-2.80,\ -0.66).$$

Siccome non contiene lo zero, possiamo rifiutare all'1% l'ipotesi nulla di $\beta_1 = 0$.

4. La misura è $R^2 = \rho^2 = 0.82$, che dice che l'82% della variabilità della variabile risposta (dipendente) è spiegata dal modello lineare.

Il grafico dei residui della regressione è rappresentato in Figura 2: lo si confronti con l'analogo dell'esercizio successivo, in Figura 6.

5. Sicuramente cambia di segno (da Y a $-Y$). Siccome i termini $(y_i - \bar{y})$ diventano

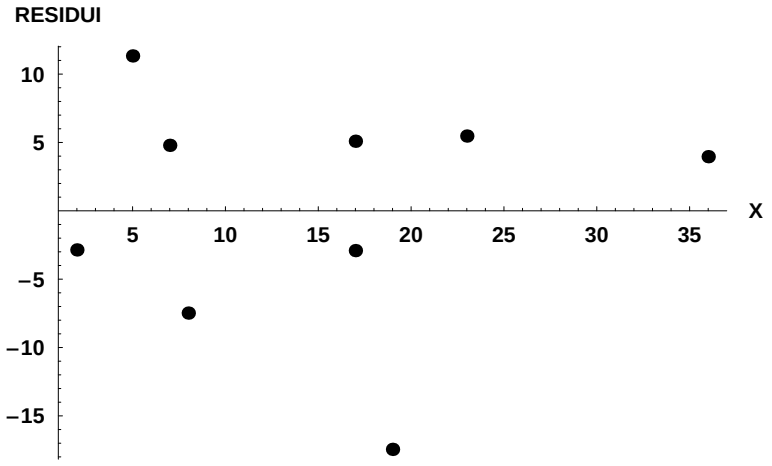

Figura 2: Residui per il modello di regressione lineare dell'Esercizio dal tema d'esame 13.06.2011.

$(40 - y_i) - (40 - \bar{y}) = -(y_i - \bar{y})$, la risposta esatta è la ii). Per la nuova intercetta si ha: $\hat{\beta}_0^\star = 40 - \bar{y} - (-\hat{\beta}_1)\bar{x} = 40 - 6.56 - 1.73 \times 14.89 = 7.68$, e cambia di valore ed il fatto che non cambi segno dipende solo dal valore 40 (intendo dire: se $Y^\star = 25 - Y$ cambia sia il segno che il valore, mentre la risposta per β_1 resta la medesima). La retta di regressione stimata è mostrata assieme ai dati nella Figura 3.

Esercizio dal tema d'esame del 16.01.2012.

Il diametro massimo di un tronco d'albero (Y, misurato in pollici) è influenzato, tra le altre cose, dalla piovosità della regione (X, misurata in pollici). I seguenti dati sono relativi ad un campione di 15 eucalipti.

Y	16.2	16.4	16.6	16.6	16.9	17	17.6	16.5	16.1	16.1	16.5	16.5	16.7	16.6	17.8
X	250	115	75	85	100	75	85	225	250	255	175	140	150	170	75

Per facilitare i conti si consideri che:

$$\sum_{i=1}^{15} x_i^2 = 396825 \quad \sum_{i=1}^{15} y_i^2 = 4173.35 \quad \sum_{i=1}^{15} x_i y_i = 36755.5.$$

1. Rappresentare la distribuzione dei due caratteri tramite un diagramma di dispersione.

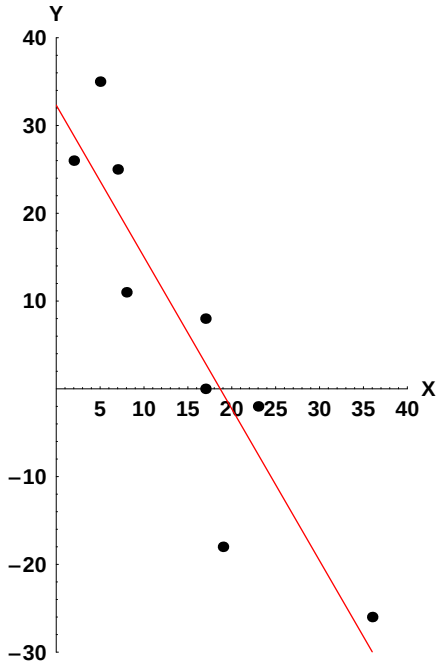

Figura 3: Diagramma di dispersione e retta di regressione stimata per l'Esercizio dal tema d'esame 13.06.2011.

2. Calcolare il coefficiente di correlazione lineare tra i due caratteri.

3. È noto che 1 pollice = 2.54 cm. La covarianza tra Y e X entrambi espressi in cm rimane la stessa? Perchè? Altrimenti è possibile ottenerne il valore senza trasformare i valori di tutte le osservazioni? Come?

4. (*aggiunto*) Stimare i parametri del modello di regressione lineare $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, \ldots, n$ e indicare una misura di buon adattamento del modello.

5. Costruire l'intervallo di confidenza al livello del 95% per β_1.

6. Discutere il grafico dei residui della regressione.

Soluzione.

1. Il diagramma di dispersione è rappresentato in Figura 4.

2. Medie campionarie[1]: $\bar{x} = 148.3333,\ \bar{y} = 16.6733$. Quindi:

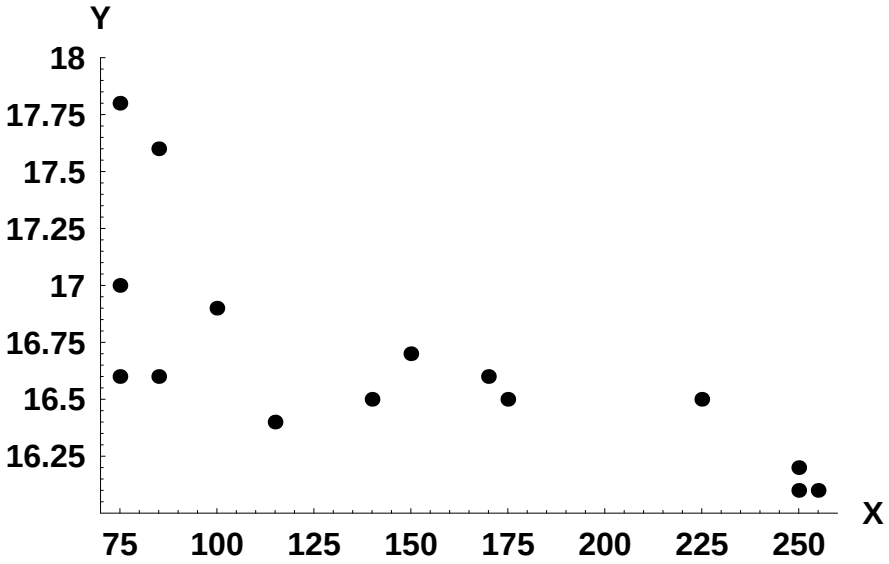

Figura 4: Diagramma di dispersione per l'Esercizio dal tema d'esame 16.01.2012.

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_i x_i^2 - n\bar{x}^2)(\sum_i y_i^2 - n\bar{y}^2)}} =$$

$$\frac{36755.5 - 15 \times 148.3333 \times 16.6733}{\sqrt{(396825 - 15 \times 148.3333^2)(4173.35 - 15 \times 16.6733^2)}} = \frac{-342.509}{\sqrt{66783.3 \times 3.3493}} = -0.72$$

3. L'indice ρ è un indice adimensionale (il cui valore è indipendente dalla scala) e dunque non varia passando da pollici a cm. La covarianza, invece, risulta riscalata del fattore $2.54^2 = 6.4516$, come si verifica facilmente sostituendo $2.54x_i$ e $2.54y_i$ nel numeratore della formula di ρ.

4.

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{-342.509}{66783.3} = -0.0051,$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 16.6733 + 0.0051 \times 148.3333 = 17.4298$$

La retta di regressione stimata è rappresentata assieme ai dati in Figura 5.

Nota per la studentessa che mi ha fatto una domanda sul valore 66783.3. Come si può vedere dai passaggi riportati, è il valore che si ottiene applicando la formula

[1]Uso più decimali di quanto fatto in aula. I conti, quindi, saranno un po' diversi come già anticipato durante l'esercitazione.

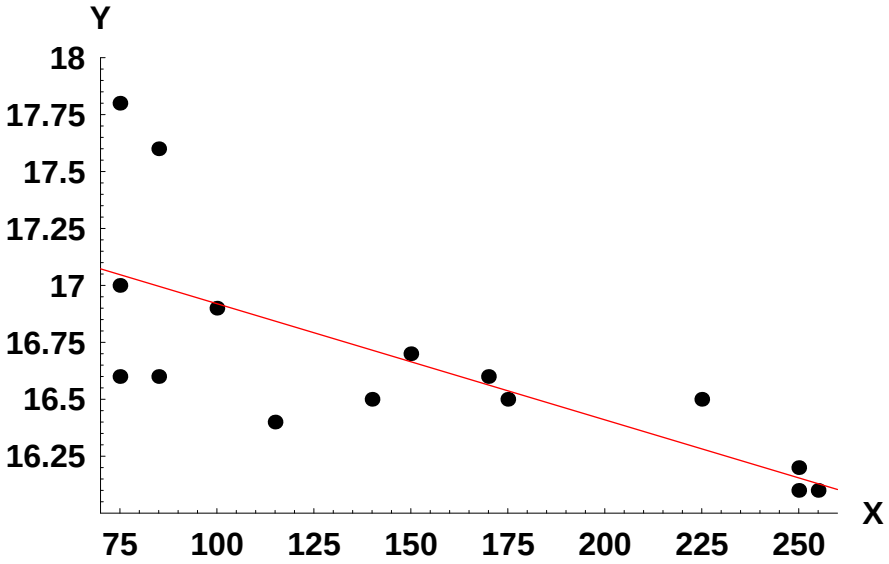

Figura 5: Diagramma di dispersione e retta di regressione stimata per l'Esercizio dal tema d'esame 16.01.2012.

$\sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - n\bar{x}$ col valore $\bar{x} = 148.3333$. Usando $\bar{x} = 148.3$ si ottiene 66931.6 da cui il valore per la varianza campionaria (distorta) $s_x^2 = 4462.11$ indicato in aula. Spero di aver risposto adeguatamente.

Una misura di buon adattamento è data da $R^2 = \rho^2 = -0.72^2 = 0.52$ che indica che solo il 50% circa della varianza totale delle y è spiegata dal modello lineare (**Nota**. A lezione la percentuale ottenuta era più alta per via dei conti un po' diversi ottenuti usando meno decimali.).

5. L'intervallo di confidenza al livello del 95% per β_1 è dato da

$$\hat{\beta}_1 \pm t_{0.025}(13)\sqrt{\frac{\sum \hat{e}_i^2/(n-2)}{\sum_i (x_i - \bar{x})^2}}$$

ove $\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ sono i residui. Nel nostro caso: $t_{0.025}(13) = 2.16$ e $\sqrt{\sum \hat{e}_i^2/(n-2)} = 0.350$ e quindi

$$\text{IC}_{95\%}(\beta_1): \quad -0.0051 \pm 2.16 \times \frac{0.350}{\sqrt{66783.3}}: \quad (-0.0080,\ -0.0022).$$

Siccome non contiene lo zero, possiamo rifiutare al 5% l'ipotesi nulla di $\beta_1 = 0$.

6. Il grafico dei residui è rappresentato in Figura 6 e mostra come questi si dispon-

gano con un andamento che appare crescente da sinistra verso destra, facendo così dubitare che siano soddisfatte le ipotesi di errori gaussiani indipendenti con media 0 e medesima varianza σ^2 (incognita).

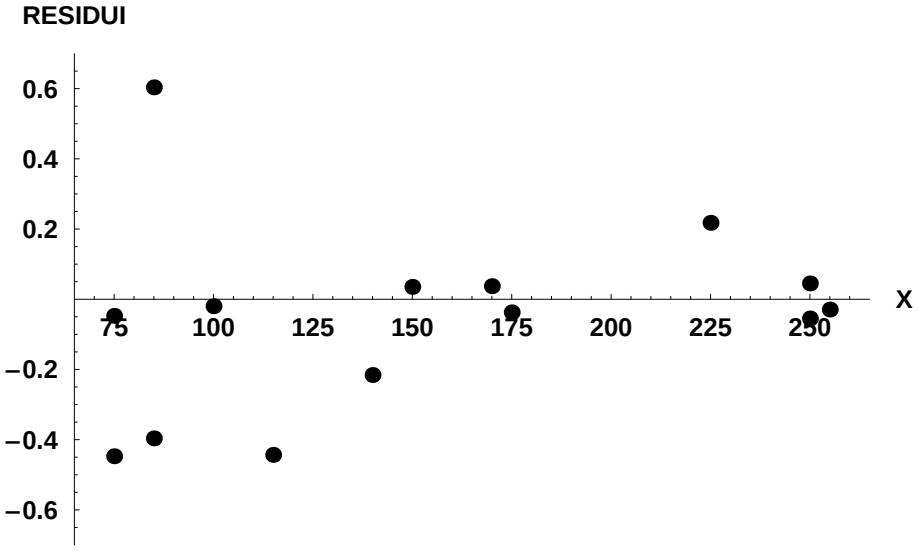

Figura 6: Residui per il modello di regressione lineare dell'Esercizio dal tema d'esame 16.01.2012.