STATISTICA DESCRITTIVA - SCHEDA N. 5
REGRESSIONE LINEARE

Nella dcheda precedente abbiamo visto che il coefficiente di correlazione fra due variabili quantitative X e Y fornisce informazioni sull'esistenza o meno di un legame lineare fra le due variabili. Tale indice, però, non permette di individuare se è X che influisce su Y, oppure se è Y che influisce su X, oppure – ancora - se sia X che Y sono conseguenze di un fenomeno che influisce su tutte e due: solo la conoscenza del problema oggetto di studio può – in alcuni casi – permettere di fare alcune ipotesi di dipendenza di una variabile dall'altra.
Se si può ipotizzare l'esistenza di una dipendenza lineare ad esempio di Y da X, si può dire che le osservazioni della variabile Y si possono ottenere*, a meno di un errore* (o residuo), da una funzione lineare delle osservazioni della variabile X. Per ciascuna osservazione avremo quindi:

$$y_i = a\, x_i + b + \text{residuo}_i$$

La variabile Y viene detta **variabile risposta** (o **variabile dipendente**), la variabile X viene detta **variabile esplicativa** (o **variabile indipendente** o **regressore**).

Se indichiamo con ε_i l'errore di approssimazione di y_i tramite una funzione lineare delle x_i, allora la relazione precedente diventa:

$$y_i = a\, x_i + b + \varepsilon_i.$$

Naturalmente escludiamo il caso in cui X sia costante, perché in tal caso il modello perderebbe significato.

A fianco è riportato un esempio con 10 dati.
Indichiamo con $\hat{y}$ i valori della retta, per distinguerli dai dati della variabile risposta Y.
La retta indicata è:

$$\hat{y} = 22.9 - 1.52\, X$$

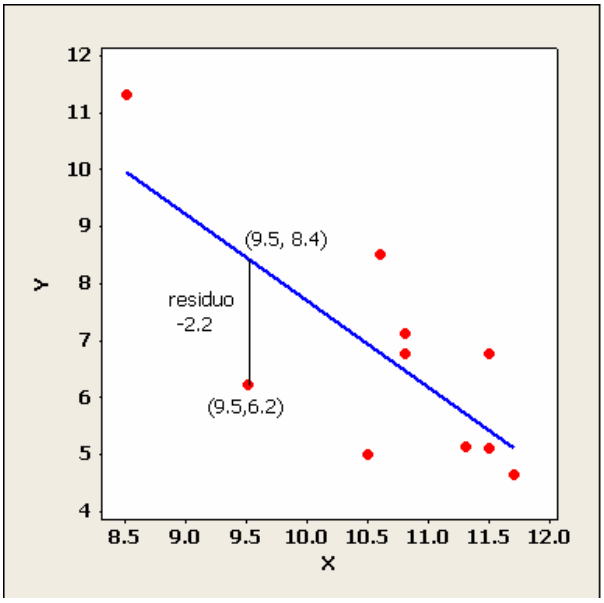
Nel grafico è anche evidenziato il residuo per il punto di coordinate (9.5, 6.2).

L'approssimazione lineare di Y tramite X per tale punto è:

$$\hat{y} = 22.9 - 1.52 \times 9.5 = 8.4$$

Il residuo risulta quindi:

$$6.2 - 8.4 = -2.2$$



Ma come scegliere i coefficienti a e b della retta?
Sicuramente vorremmo poter scegliere i parametri in modo da minimizzare i residui ε_i, ma in che senso? Inoltre devono essere "piccoli" sia i residui positivi, che quelli negativi. Quindi dobbiamo considerare una funzione dei residui che prescinda dal loro segno (il valore assoluto o il quadrato, ad esempio).

Si sceglie la coppia $(\hat{a}, \hat{b})$ che rende minima la somma dei residui al quadrato SS(a,b), ovvero la quantità

$$SS(a,b) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - a x_i - b\right)^2 .$$

Per risolvere questo problema è sufficiente calcolare le derivate parziali in a e b e porle uguali a zero. Si ottiene così un sistema a due equazioni e due incognite con un'unica soluzione.
Oppure si impone anche la condizione che la somma dei residui sia 0; usando questa condizione si ottiene che la somma dei quadrati dei residui è funzione di un solo coefficiente SS(a): più precisamente si ottiene una parabola con coefficiente di secondo grado positivo e minimo nel vertice.

Le soluzioni cercate sono

$$\hat{a} = \frac{Cov(X,Y)}{\sigma_x^2} \qquad \hat{b} = \bar{y} - \hat{a}\,\bar{x}$$

Sostituendo nell'equazione della retta si ottiene

$$\hat{Y} = \frac{Cov(X,Y)}{\sigma_x^2}(X - \bar{x}) + \bar{y}$$

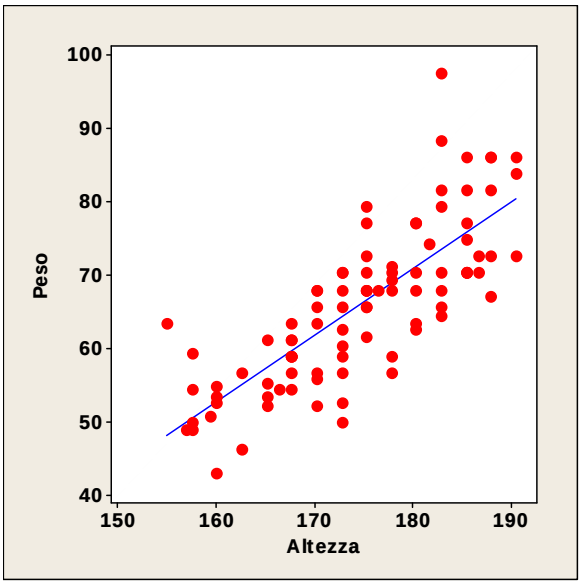
detta **retta di regressione di Y rispetto a X.**

Osserviamo che tale retta passa per il baricentro dei dati $(\bar{x}, \bar{y})$ e che il segno del suo coefficiente angolare è quello della covarianza fra le due variabili. L'errore che si commette approssimando Y con $\hat{y}$ è:

$$SS(\hat{a}, \hat{b}) = n\ \sigma_Y^2 (1-\rho^2(X,Y))$$

Quindi l'errore è tanto più piccolo quanto più è piccola la numerosità dei dati osservati e la varianza di Y e quanto più è grande il coefficiente di correlazione fra le variabili.

ESEMPIO: Riprendiamo l'esempio dei Pesi e delle Altezze (Scheda N. 3). Si può pensare l'altezza come variabile indipendente X (non si può intervenire facilmente per modificare l'altezza di una persona) e il peso come variabile risposta Y.



La retta di regressione risulta:

peso = - 92,9 + 0,909 altezza

La somma dei quadrati dei residui è:
4051,6

Il modello di regressione lineare può essere costruito anche quando si ha più di una variabile esplicativa:

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p + \text{residuo}$$

L'interpretazione geometrica è meno evidente; per esempio se abbiamo due variabili esplicative otteniamo l'equazione di un piano nello spazio.
La scelta dei coefficienti si effettua sempre minimizzando la somma dei quadrati dei residui; non riportiamo qui le espressioni delle soluzioni per i coefficienti, ma possiamo utilizzare Minitab per calcolarle.

Bontà del modello
Dati un insieme di osservazioni sperimentali è sempre possibile costruire un modello di regressione lineare, ma è necessario valutare la bontà dell'approssimazione.

Vediamo due strategie per verificare la bontà del modello.

a) Coefficiente R^2

Il modello è tanto migliore quanto più la variabile risposta, indicata con Y, e la sua approssimazione lineare tramite X, indicata con $\hat{Y}$, hanno una correlazione vicina a 1.

Il quadrato del coefficiente di correlazione fra Y e $\hat{Y}$, si indica con R^2:

$$R^2 = \rho^2(Y,\hat{Y})$$

e in genere viene fornito dai software statistici.

Osserviamo che se – come nei casi esaminati sopra – abbiamo una sola variabile esplicativa:

$$R^2 = \rho^2(Y,\hat{Y}) = \rho^2(Y,\hat{a}X + \hat{b}) = \rho^2(Y,X)$$

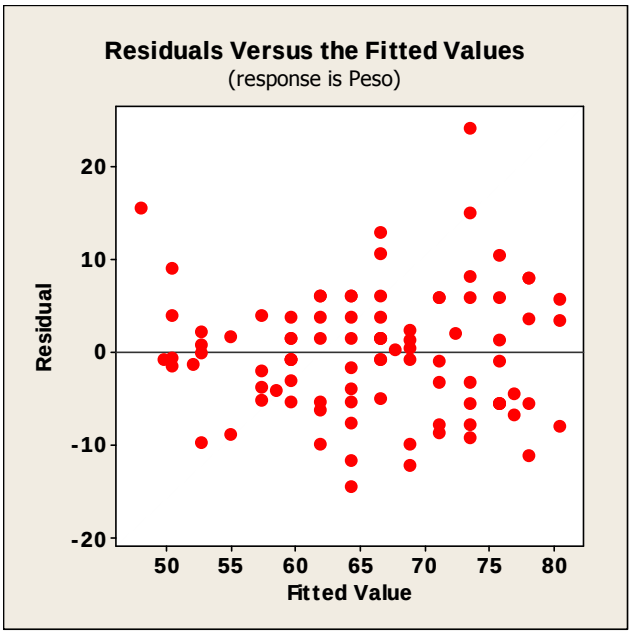
Abbiamo visto che se R^2 ha un valore alto allora ha un valore bsso la somma dei quadrati dei residui, ossia $SS(\hat{a},\hat{b}) = n\ \sigma_Y^2(1-\rho^2(X,Y))$..

Esiste un caso in cui $\rho^2(Y,X)$ è circa 0, ma i residui sono piccoli: pensateci.

b) Grafico dei residui

È interessante analizzare il grafico dei residui rispetto ai valori predetti, ovvero lo scatterplot dei punti $(\hat{a}x_i+\hat{b},\ y_i-\hat{a}x_i+\hat{b})$; se si ottiene una nuvola omogenea di punti il modello è corretto. Questo avviene nell'esempio precedente.

Osserviamo che l'"omogeneità" dei residui deve essere considerata rispetto alla retta Res=0.



Se la dipendenza della variabile risposta dalle variabili esplicative non è lineare (ad esempio, quadratica, esponenziale, logaritmica, ecc.) il grafico dei residui rispetto ai valori predetti enfatizzerà questa dipendenza non lineare.

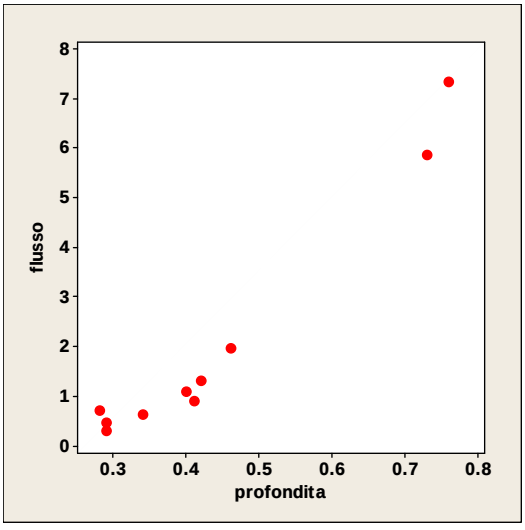
Vediamo questo fatto con un esempio.

ESEMPIO

Si vuole stabilire se esiste una dipendenza fra il flusso di un corso d'acqua (cioè la quantità di acqua che passa in un dato punto in un determinato intervallo di tempo) e la profondità del corso d'acqua. I dati sono rilevati in 10 stazioni.

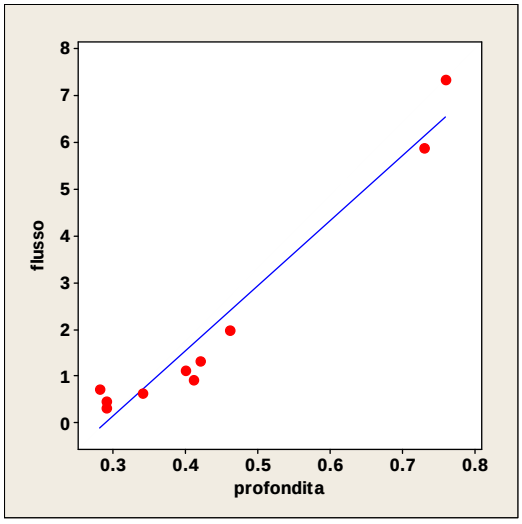
A fianco è riportato lo scatterplot dei dati osservati.
Provate a disegnare a matita una "buona" retta che approssimi il flusso rispetto alla profondità.

Vi sembra che il flusso sia ben approssimato *da una retta*?



Utilizzando Minitab possiamo calcolare
- la correlazione lineare: ρ(prof, flusso) = 0,973.
- la retta di regressione:

flusso = - 3,98 + 13,8 profondita



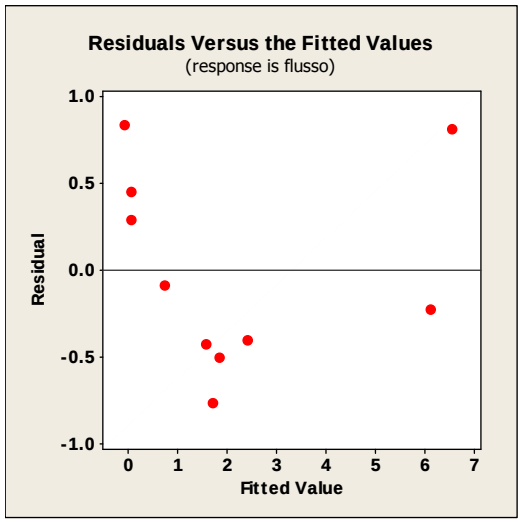
Osservando bene i dati, si può notare che la dipendenza fra il flusso e la profondità può essere anche quadratica. Cancellate la retta che avete disegnato sul primo grafico e provate a disegnare una "buona" parabola che approssimi il flusso rispetto alla profondità.

Per comprendere meglio se la dipendenza è lineare oppure no consideriamo i residui relativi a ciascuna stazione di rilevazione. Sul grafico con la retta:

flusso = - 3,98 + 13,8 profondita

indicate i residui di ciascun punto.

Nel grafico a fianco sono riportati:
- in ascissa i valori del flusso approssimati linearmente tramite la profondità
- in ordinata i corrispondenti residui.

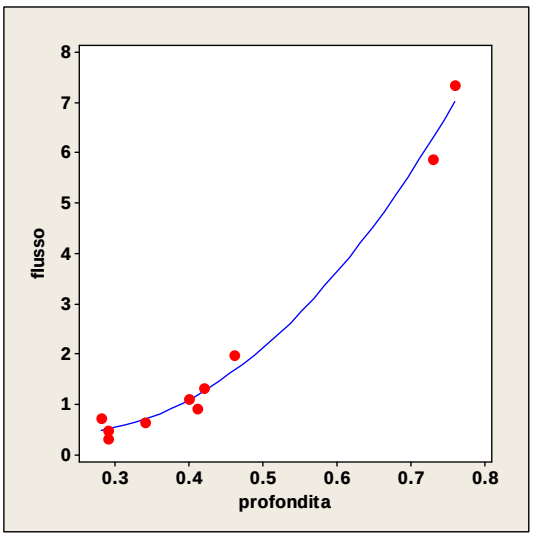


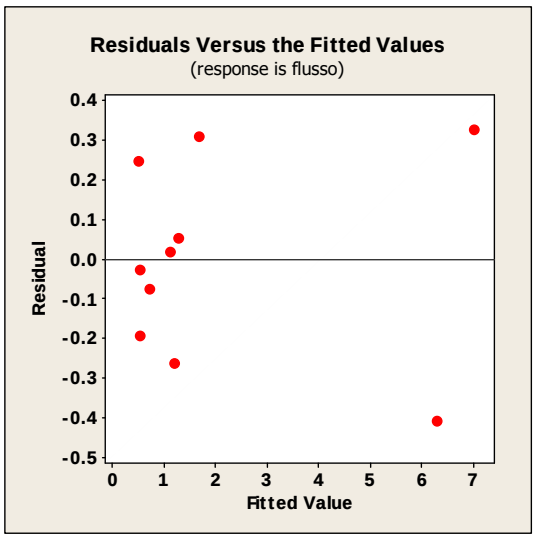
La dipendenza non lineare è sicuramente più evidente se si osserva il grafico dei residui piuttosto che se si osserva lo scatterplot del flusso e della profondità.

Supponiamo che esista una dipendenza quadratica e indichiamo con "profondita2" la variabile corrispondente al quadrato della profondità. Calcoliamo con l'ausilio Minitab® l'approssimazione del flusso tramite le variabili "profondita" e "profondita2"; si ha che:

Flusso=1.683 - 10.86 profondita + 23.54 profondita2

Il grafico dei residui diventa:





Questo tipo di modello:

$$Y = b_0 + b_1X + b_2X^2 + \square$$

viene detto *modello lineare polinomiale del secondo ordine*.

Output di Minitab

Vediamo l'output di Minitab per quest'ultimo modello lineare; sono evidenziate in neretto le parti che siamo in grado di comprendere a questo livello di approfondimento delle nostre conoscenze statistiche.
I dati sono riportati a fianco.

Stazione	flusso	profondita	profondita2
1	0.636	0.34	0.1156
2	0.319	0.29	0.0841
3	0.734	0.28	0.0784
4	1.327	0.42	0.1764
5	0.487	0.29	0.0841
6	0.924	0.41	0.1681
7	7.350	0.76	0.5776
8	5.890	0.73	0.5329
9	1.979	0.46	0.2116
10	1.124	0.40	0.1600

The regression equation is
flusso = 1.68 - 10.9 profondita + 23.5 profondita2

Predictor	**Coef**	SE Coef	T	P
Constant	**1.683**	1.059	1.59	0.156
profondita	**-10.861**	4.517	-2.40	0.047
profondita2	**23.535**	4.274	5.51	0.001

S = 0.279418 **R-Sq = 99.0%** R-Sq(adj) = 98.7%

Analysis of Variance

Source	DF	**SS**	MS	F	P
Regression	2	54.105	27.053	346.50	0.000
Residual Error	7	**0.547**	0.078		
Total	9	54.652			

Source	DF	Seq SS
profondita	1	51.739
profondita2	1	2.367

Obs	**profondita**	**flusso**	**Fit**	SE Fit	**Residual**	St Resid
1	**0.340**	**0.6360**	**0.7107**	0.1029	**-0.0747**	-0.29
2	**0.290**	**0.3190**	**0.5123**	0.1477	**-0.1933**	-0.82
3	**0.280**	**0.7340**	**0.4868**	0.1634	**0.2472**	1.09
4	**0.420**	**1.3270**	**1.2727**	0.1344	**0.0543**	0.22
5	**0.290**	**0.4870**	**0.5123**	0.1477	**-0.0253**	-0.11
6	**0.410**	**0.9240**	**1.1860**	0.1281	**-0.2620**	-1.05
7	**0.760**	**7.3500**	**7.0223**	0.2138	**0.3277**	1.82
8	**0.730**	**5.8900**	**6.2961**	0.1830	**-0.4061**	-1.92
9	**0.460**	**1.9790**	**1.6667**	0.1575	**0.3123**	1.35
10	**0.400**	**1.1240**	**1.1040**	0.1218	**0.0200**	0.08

Negli output di Minitab, inizialmente è riportata l'approssimazione lineare.
Poi segue una prima tabella che riguarda i coefficienti dell'approssimazione lineare: nella colonna "Predictor" ci sono le variabili a cui si riferiscono i coefficienti o l'indicazione "Constant" per il primo coefficiente (l'intercetta); nella seconda colonna vengono riportati i valori dei coefficienti. Non prendiamo in considerazione le altre colonne.
Seguono poi alcune informazioni tra cui il coefficiente R^2 (indicato con R-sq).
Della seconda tabella "Analysis of variance" sappiamo solo leggere la somma dei quadrati dei residui di questo modello: si trova nella colonna "SS" in corrispondenza della riga "Residual Error"; è il valore che abbiamo indicato con SS($\hat{a}$, $\hat{b}$, $\hat{c}$) (in questo caso c'è anche il terzo coefficiente).
Nell'ultima tabella vengono riportati:

- nella colonna "Obs" il numero consecutivo delle unità sperimentali
- la prima variabile esplicativa, nel nostro caso "profondita".
- la variabile risposta, in questo caso "flusso"
- l'approssimazione lineare $\hat{Y}$ indicati con "Fit"
- i residui, cioè Y-$\hat{Y}$

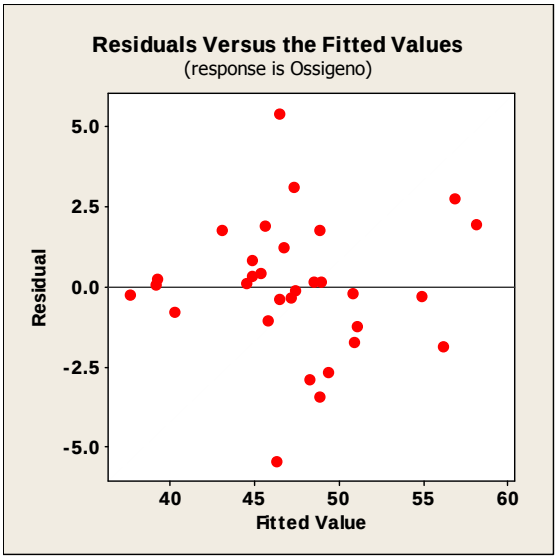
COMMENTO E POSSIBILI SVILUPPI

Perché si cotruiscono i modelli di regressione lineare?
Sia per fini descrittivi che previsionali.
Se si trova un buon modello per approssimare la variabile risposta Y è infatti possibile avere una valutazione dei valori di Y utilizzando le sole variabili esplicative in quei casi in cui raccogliere i dati per le variabili esplicative è più comodo o meno costoso che misurare la variabile risposta; vedremo un esempio a questo proposito.
Con metodologie statistiche che esulano da questo corso si può anche calcolare – per assegnati valori delle variabili esplicative - un intervallo in cui la risposta dovrebbe stare con una probabilità prefissata, per esempio del 95%.
Inoltre le metodologie statistiche dei modelli lineari permettono di capire quali variabili esplicative siano effettivamente utili per descrivere la variabile risposta e quali invece non strettamente necessarie. Anche questo tipo di procedure richiedono però conoscenze più approfondite di quelle imparate in queste schede.

ESEMPIO

Si vuole descrivere il consumo di ossigeno di atleti che praticano sport di fondo tramite le seguenti variabili esplicative: l'età, il peso, il tempo per effettuare un percorso di corsa, le pulsazioni cardiache da fermo, medie e massime durante la corsa. È evidente che tali variabili esplicative sono molto facilmente rilevabili su un campo, mentre per misurare il consumo di ossigeno occorrono strumenti adeguati.
L'output della regressione è il seguente:

```
The regression equation is
Ossigeno = 103 - 0.227 Eta - 0.0742 Peso - 2.63 TempoCorsa - 0.0215 PulsFermo
           - 0.370 PulsCorsa + 0.303 PulsMax

Predictor       Coef  SE Coef      T      P
Constant      102.93    12.40   8.30  0.000
Eta          -0.22697  0.09984  -2.27  0.032
Peso         -0.07418  0.05459  -1.36  0.187
TempoCorsa    -2.6287   0.3846  -6.84  0.000
PulsFermo    -0.02153  0.06605  -0.33  0.747
PulsCorsa     -0.3696   0.1199  -3.08  0.005
PulsMax        0.3032   0.1365   2.22  0.036

S = 2.31695   R-Sq = 84.9%   R-Sq(adj) = 81.1%

Analysis of Variance

Source           DF      SS      MS      F      P
Regression        6  722.54  120.42  22.43  0.000
Residual Error   24  128.84    5.37
Total            30  851.38
```



Il modello è buono, sia per quel che riguarda il grafico dei residui, ma anche per il coefficiente R^2.
Utilizzando strumenti statistici più approfonditi si deduce anche che il peso e le pulsazioni da fermo possono essere entrambe non necessarie per una buona approssimazione del consumo di ossigeno; in tal caso il modello diventa:

```
The regression equation is
Ossigeno = 98.1 - 0.198 Eta - 2.77 TempoCorsa - 0.348 PulsCorsa + 0.271 PulsMax
```

Calcolare quale consumo di ossigeno dovrebbe avere, secondo questo modello, un atleta con le seguenti rilevazioni:
`Eta: 44  TempoCorsa: 11.37  PulsCorsa: 178  PulsMax: 182`

ESERCIZI

1) A fianco sono riportati i risultati di due rilevazioni quantitative su 10 elementi.

Per questi dati si ha:

$\sum_{i=1}^{10} x_i = 431 \quad \sum_{i=1}^{10} x_i^2 = 18629$

$\sum_{i=1}^{10} y_i = 6514 \quad \sum_{i=1}^{10} y_i^2 = 4262260$

$\sum_{i=1}^{10} x_i y_i = 281627$

X	Y
46	654
44	672
40	613
41	630
45	679
40	577
45	718
43	642
41	612
46	717

a) Disegnare il grafico della distribuzione congiunta.

b) Calcolare media di X e di Y, la varianza di X e Cov(X,Y).

c) Calcolare la retta di regressione di Y rispetto a X e disegnarla sul sistema di assi dove è stata disegnata la distribuzione congiunta.

Qui sotto sono riportati i risultati della regressione effettuata con il software Minitab da cui sono stati cancellati (e indicati con xxx) alcuni valori.

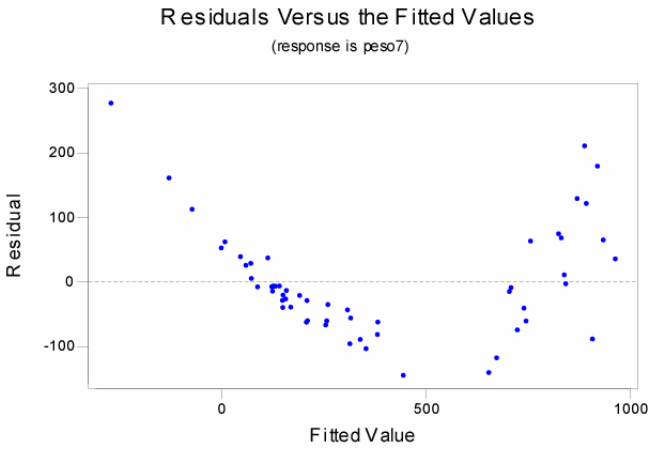
```
Predictor        Coef     SE Coef          T        P
Constant          xxx       142.5      -0.42      xxx
X                 xxx       3.302       5.00    0.001

S = 24.01       R-Sq = xxx%     R-Sq(adj) = 72.7%

Analysis of Variance
Source            DF          SS          MS         F        P
Regression         1       14427       14427     25.02    0.001
Residual Error     8        4614         577
Total              9       19040

Obs         C1         C4         Fit      SE Fit    Residual    St Resid
  1       46.0     654.00       699.2       12.22      -45.29       -2.19
  2       44.0     672.00         xxx        8.15         xxx        0.25
  3       40.0     613.00      600.21       12.75       12.79        0.63
  4       41.0     630.00      616.72       10.28       13.28        0.61
  5       45.0     679.00      682.78        9.85       -3.78       -0.17
  6       40.0     577.00      600.21       12.75      -23.21       -1.14
  7       45.0     718.00      682.78        9.85       35.22        1.61
  8       43.0     642.00      649.75        7.60       -7.75       -0.34
  9       41.0     612.00      616.72       10.28       -4.72       -0.22
 10       46.0     717.00      699.29       12.22       17.71        0.86
```

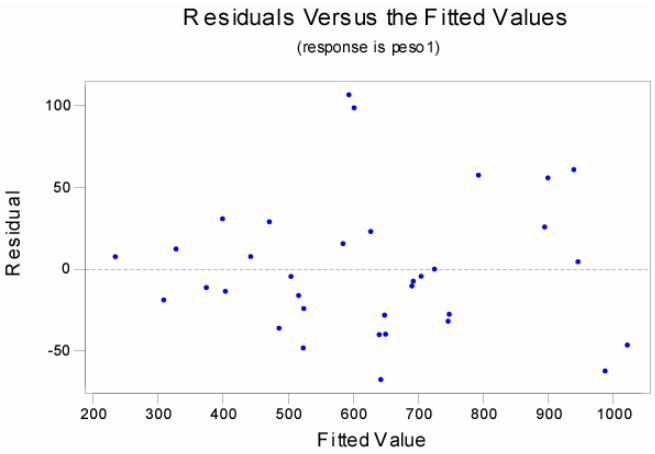
d) Calcolare l'indice R^2

e) Calcolare il valore di Y approssimato e il residuo per la seconda unità sperimentale.

f) Disegnare il grafico dei residui e valutare la bontà del modello.

2) Si considerano, per due specie di pesci, il peso, la lunghezza, la larghezza e l'altezza e si vuole stabilire se il peso è esprimibile come funzione lineare delle altre variabili. Qui sotto sono riportati i grafici dei residui della regressione lineare per le due specie.





Commentare i due grafici e stabilire - tramite essi - se il peso delle due specie di pesci è esprimibile come funzione lineare della lunghezza, larghezza e altezza.