**Corso integrato di informatica, statistica e analisi dei dati sperimentali**
**Esercitazione 4**

**Esercizio 1)** Regressione lineare. Il file "**altezza_peso.dat**" è un file ASCII che contiene i dati di altezza, peso e età di un numeroso gruppo di studenti, divisi per sesso. Importare i dati in un foglio Excel (utilizzare l'opzione *dati delimitati da spazi* per ripartire i dati nelle colonne). Il file **altezza_peso_soluz.xls** è un file Excel che può essere consultato e preso come modello

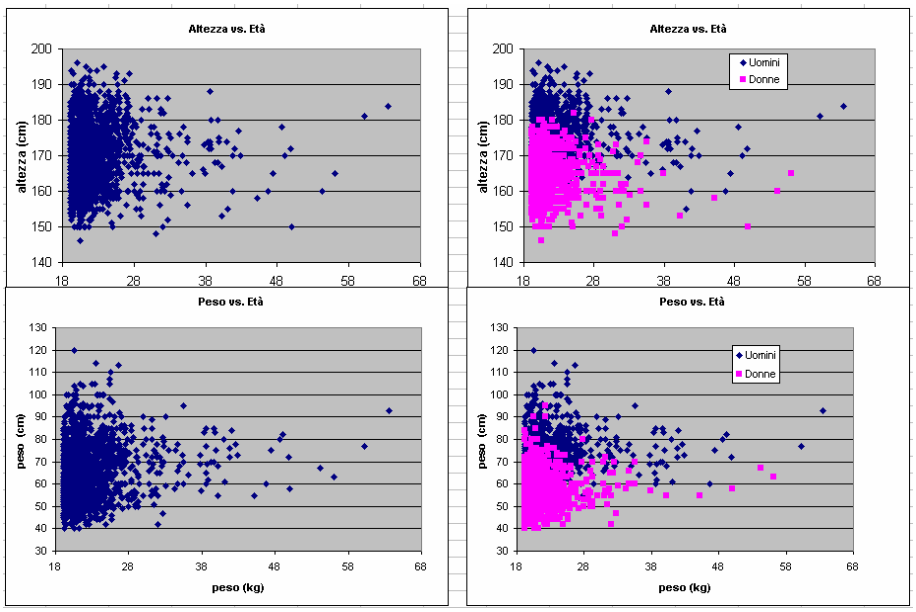
| SESSO | ETA' | ALTEZZA | PESO |
|---|---|---|---|
| *1=F 0=M* | *anni* | *cm* | *Kg* |
| 0 | 20.6 | 180 | 65 |
| 0 | 20.2 | 180 | 75 |
| 0 | 20.3 | 173 | 60 |
| 0 | 23.9 | 187 | 93 |
| 0 | 21.4 | 164 | 66 |
| 0 | 25 | 186 | 84 |
| 0 | 20.8 | 175 | 67 |
| 0 | 20.6 | 170 | 89 |
| 0 | 27.1 | 180 | 71 |
| 0 | 23.3 | 170 | 63 |

dare una breve descrizione statistica dei dati: valori minimi, massimi, media, var. etc…

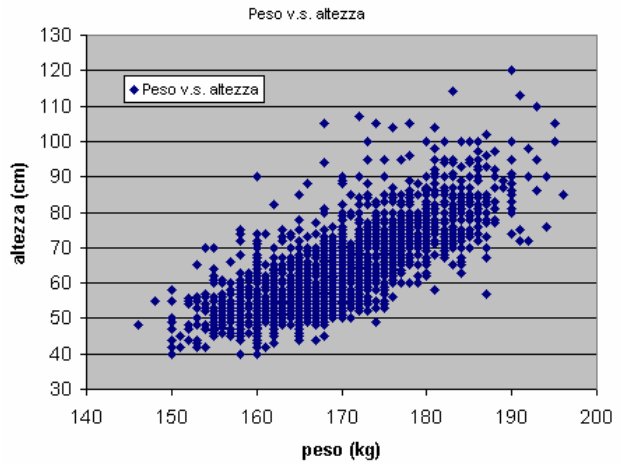
| Statistica | | | |
|---|---|---|---|
| SESSO | ETA' | ALTEZZA | PESO |
| N | | 2759 | |
| min | 19.1 | 146.0 | 40.0 |
| max | 63.5 | 196.0 | 120.0 |
| media | 21.85 | 169.01 | 63.38 |
| var | 12.60 | 67.87 | 133.69 |
| dev.st | 3.55 | 8.24 | 11.56 |

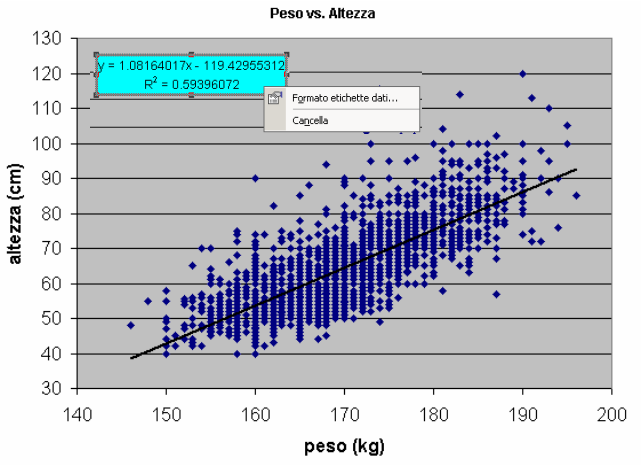
| Correlazione | |
|---|---|
| età-altezza | 0.088745 |
| età-peso | 0.194581 |
| altezza-peso | 0.770688 |

controllare i valori della correlazione tra i dati di età, altezza e peso. I dati di altezza e peso sono fortemente correlati. Concentriamosi su questi (volendo graficare i dati di peso o altezza in funzione dell'età)



Presentare i dati di peso in funzione dell'altezza in un grafico xy




Calcolare la retta di regressione riportando sul grafico la retta, l'equazione e il fattore $R^2$. In seguito sarà utile avere i parametri delle curve di regressione con precisione elevata, per aumentare il numero delle cifre significative utilizzare il tasto destro sul riquadro dei parametri per accedere al menu "formato etichette dei dati"

Utilizzare la macro "regressione" nelle opzioni di "analisi dati" per ottenere una tabella riassuntiva dei parametri della regressione lineare e il tracciato dei residui.

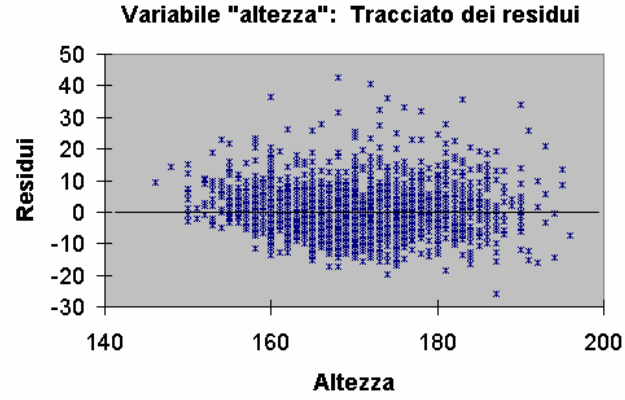
OUTPUT RIEPILOGO

| *Statistica della regressione* | |
|---|---|
| R multiplo | 0.770688 |
| R al quadr | 0.593961 |
| R al quadr | 0.593813 |
| Errore star | 7.369091 |
| Osservazio | 2759 |

ANALISI VARIANZA

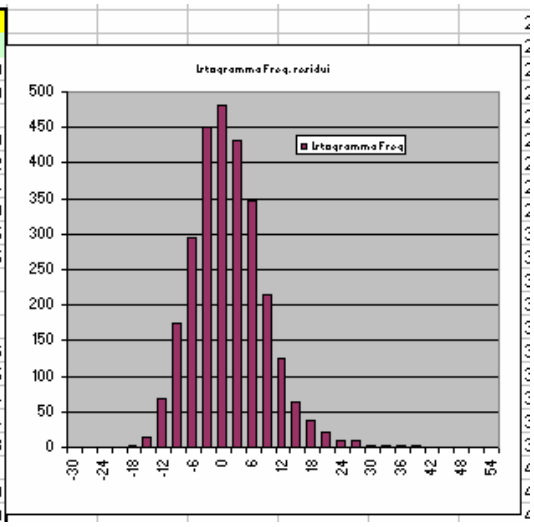
| | *gdl* | *SQ* | *MQ* | *F* | *Significatività F* |
|---|---|---|---|---|---|
| Regressior | 1 | 219005.1 | 219005.138 | 4032.98352 | 0.00E+00 |
| Residuo | 2757 | 149714.8 | 54.3035042 | | |
| Totale | 2758 | 368719.9 | | | |

| | *Coefficienti* | *Err. St.* | *Stat t* | *significatività* | *Inf. 95%* | *Sup. 95%* |
|---|---|---|---|---|---|---|
| Intercetta | -119.43 | 2.88 | -41.44 | 3.32E-292 | -125.08 | -113.78 |
| Variabile X | 1.08164 | 0.02 | 63.51 | 0 | 1.05 | 1.12 |



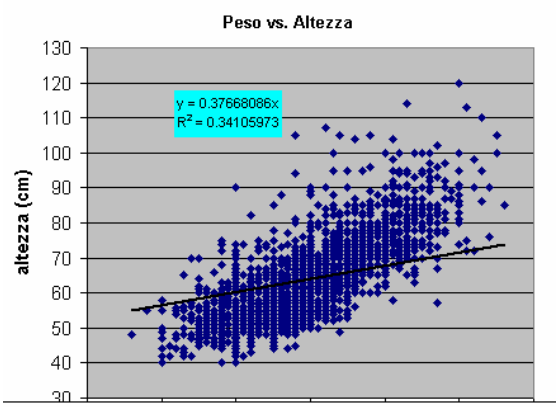

Notare che la distribuzione dei residui presenta una leggera asimmetria rispetto allo zero. Volendo si puó calcolare l'istogramma con la distribuzione dei residui

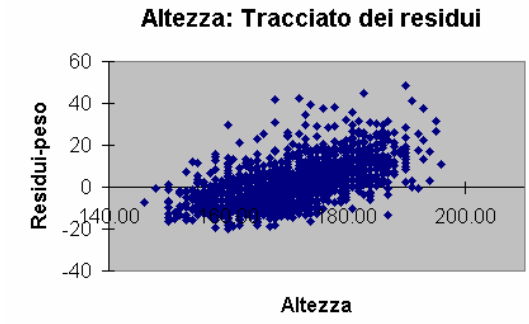
| Istogramma | |
|---|---|
| Residui | Freq. |
| -30 | 0 |
| -27 | 0 |
| -24 | 1 |
| -21 | 0 |
| -18 | 2 |
| -15 | 14 |
| -12 | 69 |
| -9 | 175 |
| -6 | 295 |
| -3 | 451 |
| 0 | 481 |
| 3 | 431 |
| 6 | 346 |
| 9 | 215 |
| 12 | 124 |
| 15 | 64 |
| 18 | 38 |
| 21 | 21 |
| 24 | 10 |
| 27 | 9 |

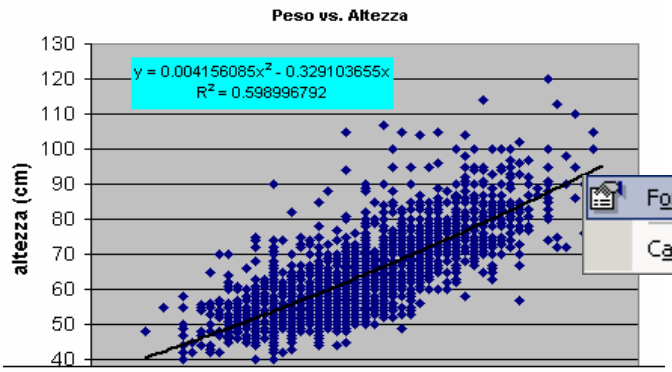


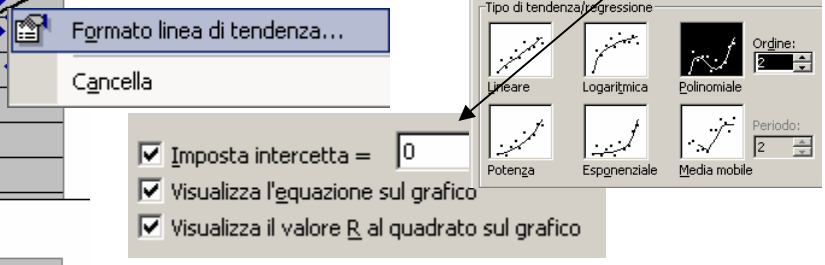
La regressione lineare ha un termine noto diverso da zero, fisicamente poco significativo dal momento che prevede un peso negativo ad altezza nulla! Possiamo pensare ad un modello più realistico in cui il termine noto sia nullo.

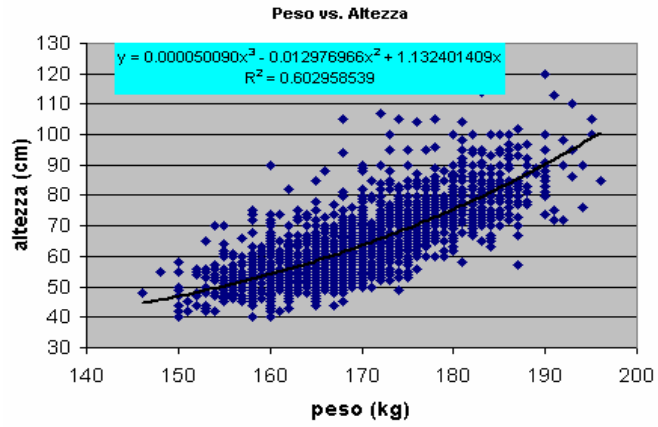
La regressione lineare passante per l'origine non riproduce bene il dato e il coefficiente $R^2$ diminuisce chiaramente. Il che suggerisce che il modello non sia corretto.



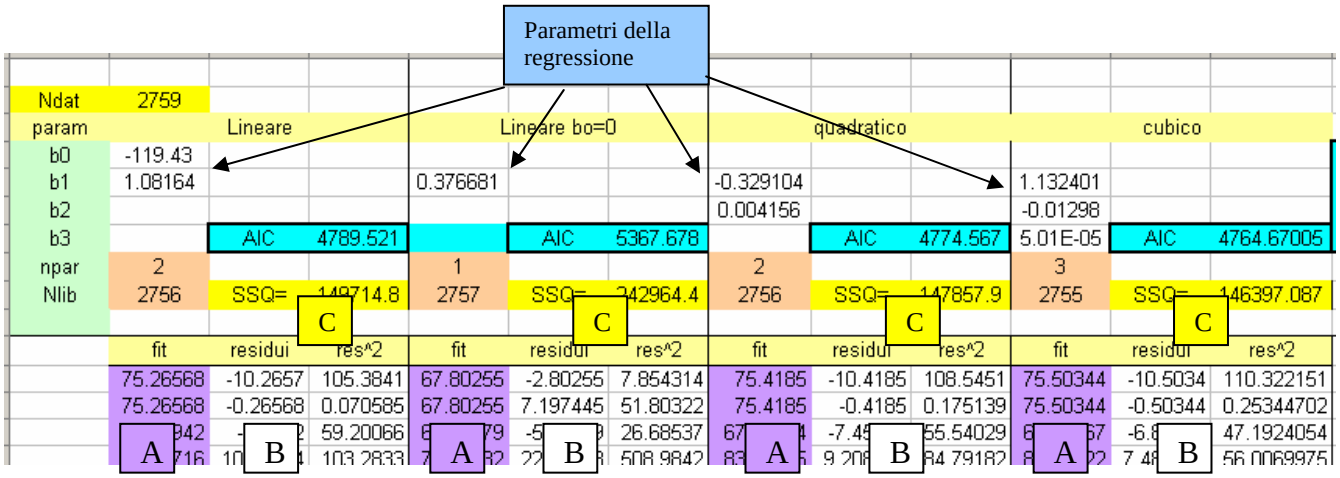




Si può pensare ad un modello più fisico con polinomi di secondo o terzo grado passanti per l'origine Calcolare le curve di regressione con funzioni polinomiali di secondo e terzo grado. I coefficienti $R^2$ aumentano aumentando l'ordine del polinomio. Ciò suggerisce un miglioramento dell'accordo.

Vogliamo innanzitutto verificare il miglioramento della regressione e poi quantificarlo




Per fare questo, per ognuno dei metodi di regressione utilizzati:

**A**: calcolare le curve teoriche utilizzando i parametri ottenuti dal raffinamento
**B**: calcolare i residui (*res = dati – fit*) (per la regressione lineare l'opzione "regressione" fornisce anche una tabella con andamento teorico e residui)
**C**: calcolare la somma dei quadrati dei residui (**SSQ**) calcolando i residui al quadrato e poi sommandoli
**D**: calcolare il numero di parametri, numero di osservazioni e numero di punti indipendenti (gradi di liberà) per i tre metodi.

I coefficienti AIC (Akaike information criteria)

$$AIC = N \log\left(\frac{SSQ_{res}}{N}\right) + 2n_p$$

Mostrano un effettivo miglioramento dell'accordo passando da regressione lineare, quadratica a cubica. L'AIC è decisamente più elevato per il modello lineare passante per l'origine.

Effettuare un test F per quantificare statisticamente il miglioramento dell'accordo:

$$F(n_{pc} - n_{ps}, N - n_{pc} - 1) = \left(\frac{SSQ_s - SSQ_c}{n_{pc} - n_{ps}}\right) \cdot \left(\frac{SSQ_c}{N - n_{pc} - 1}\right)^{-1}$$

$SSQ_c$ = somma dei quadrati dei residui per il modello più complesso (maggior numero di parametri)
$SSQ_s$ = somma dei quadrati dei residui per il modello più semplice (minor numero di parametri)
$n_{pc}$ = numero parametri per il modello più complesso
$n_{ps}$ = numero parametri per il modello più semplice
N = numero totale di dati

| Lin/lin b=0 | | test Cub/lin | | test Cub/quad | |
|---|---|---|---|---|---|
| F | 1716.571 | F | 62.43424 | F | 27.48985 |
| n1 | 1 | n1 | 1 | n1 | 1 |
| n2 | 2756 | n2 | 2755 | n2 | 2755 |
| stat F | 4.2E-292 | stat F | 3.95E-15 | stat F | 1.7E-07 |

DISRIB.F (x, n1, n2)

Si vede che la regressione che utilizza un polinomio di terzo grado è sicuramente migliore della regressione lineare o della regressione quadratica e che la regressione lineare con termine noto non nullo è sicuramente da preferire alla regressione passante per l'origine.

**Nota**: dal momento che passando da regressione lineare a regressione quadratica senza temine noto (passante per l'origine) in numero di parametri non cambia, quindi, diminuendo SSQ, il test F è sicuramente positivo.

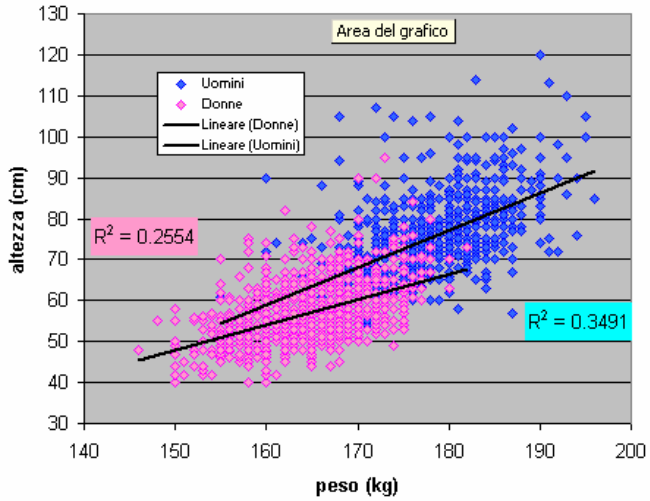
Volendo provare a vedere cosa succede non imponendo il passaggio per l'origine o utilizzando polinomi di grado più elevato.

**Ci sono differenze tra uomini e donne**?

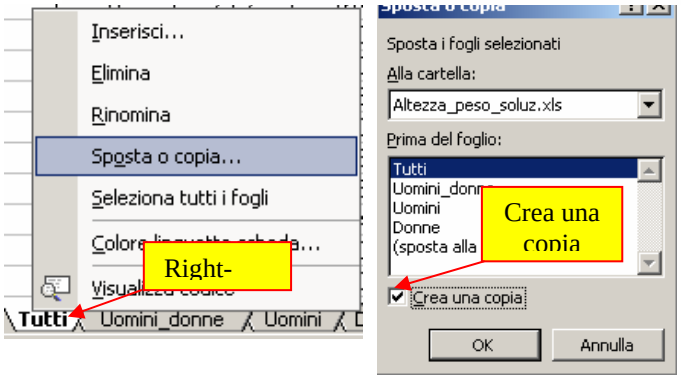
Copiare i dati in un nuovo foglio, separare i dati di uomini e donne e graficarli separatamente mostrando le curve di regressione lineare.


| | Uomini | | | Donne | | |
|---|---|---|---|---|---|---|
| | ETA' | ALTEZZA | PESO | ETA' | ALTEZZA | PESO |
| N | | 1079 | | | 1680 | |
| min | 19.1 | 155.0 | 48.0 | 19.1 | 146.0 | 40.0 |
| max | 63.5 | 196.0 | 120.0 | 56.1 | 182.0 | 95.0 |
| media | 22.77 | 176.25 | 73.66 | 21.26 | 164.36 | 56.78 |
| var | 19.67 | 40.38 | 96.36 | 7.18 | 30.31 | 46.14 |
| dev.st | 4.43 | 6.35 | 9.82 | 2.68 | 5.51 | 6.79 |
| | ETA' | ALTEZZA | PESO | ETA' | ALTEZZA | PESO |
| | *anni* | *cm* | *Kg* | *anni* | *cm* | *Kg* |
| | 20.6 | 180 | 65 | 20.9 | 163 | 49 |
| | 20.2 | 180 | 75 | 20.5 | 161 | 51 |
| | 20.3 | 173 | 60 | 19.1 | 167 | 58 |

Le differenze sono molto evidenti. Quanto sono significative?



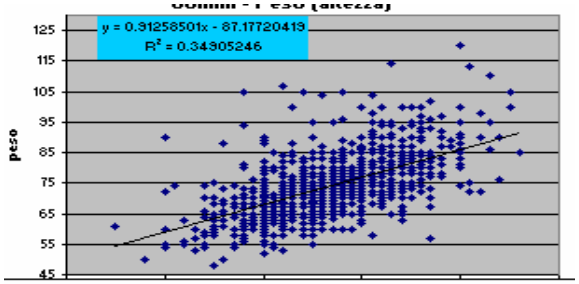

Per prima cosa effettuiamo sui due gruppi (uomini e donne) separatamente la stessa analisi fatta per i dati globali. Per non dover riscrivere tutto duplicare il foglio utilizzato per i dati globali usando la procedura descritta in figura. Fare due copie ed eliminare tutti i dati degli uomini da una e tutti i dati delle donne dall'altra. Con poche modifiche si dovrebbe riuscire ad ottenere un'analisi dei gruppi Uomini e Donne separatamente.

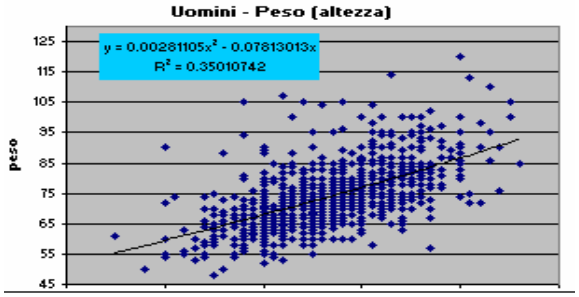


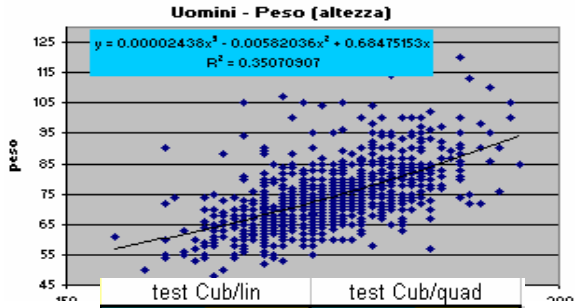

| | test Cub/lin | | test Cub/quad | |
|---|---|---|---|---|
| F | 2.738809 | F | 0.992184 | |
| n1 | 1 | n1 | 1 | |
| n2 | 1075 | n2 | 1075 | |
| stat F | 0.09823 | stat F | 0.319433 | |

| Ndat | 1079 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| param | Lineare | | | quadratico | | | cubico | | |
| b0 | -87.1772 | | | | | | | | |
| b1 | 0.912585 | | | -0.07813 | | | 0.684752 | | |
| b2 | | | | 0.002811 | | | -0.00582 | | |
| b3 | | AIC | 1942.983 | | AIC | 1942.223 | 2.44E-05 | AIC | 1943.79044 |
| npar | 2 | | | 2 | | | 3 | | |
| Nlib | 1076 | SSQ= | 67614.49 | 1076 | SSQ= | 67504.91 | 1075 | SSQ= | 67442.6667 |
| | fit | residui | res^2 | fit | residui | res^2 | fit | residui | res^2 |
| | 77.0881 | -12.0881 | 146.1221 | 77.0146 | -12.0146 | 144.3505 | 76.85977 | -11.8598 | 140.654178 |
| | 77.0881 | -2.0881 | 4.360152 | 77.0146 | -2.0146 | 4.058599 | 76.85977 | -1.85977 | 3.45874966 |
| | 70.7 | -10.7 | 114.4901 | 70.6154 | -10.6154 | 112.6868 | 70.4972 | -10.4972 | 110.191223 |

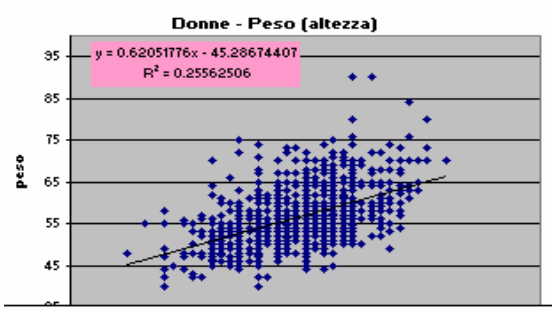
Il criterio AIC mostra che un modello cubico non è giustificato (AIC maggiore), mentre la regressione lineare e quadratica sono praticamente indistinguibili. Il test di F mostra una significatività del 9% utilizzando il modello cubico rispetto a quello lineare, quindi il rischio di sbagliare affermando che la regressione cubica è migliore di quella lineare è del 9%. Se si confronta la regressione cubica con la regressione quadratica il rischio di sbagliare affermando un miglioramento supera il 30%!

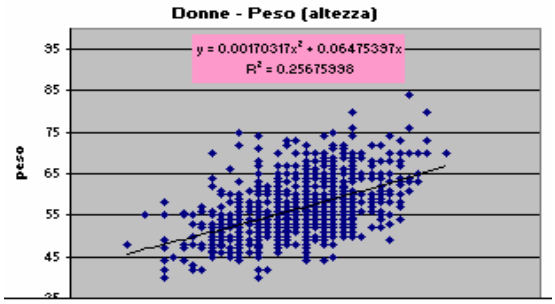


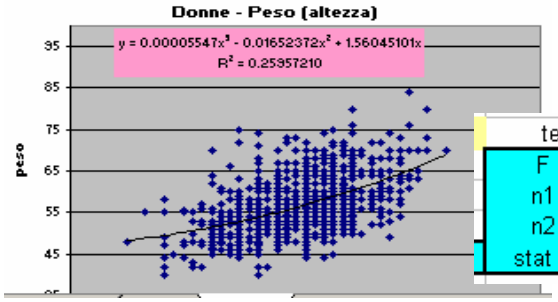




| test Cub/lin | | test Cub/quad | |
|---|---|---|---|
| F | 2.738809 | F | 0.992184 |
| n1 | 1 | n1 | 1 |
| n2 | 1075 | n2 | 1075 |
| stat F | 0.09823 | stat F | 0.319433 |




| test Cub/lin | | test Cub/quad | |
|---|---|---|---|
| F | 5.730405 | F | 4.082664 |
| n1 | 1 | n1 | 1 |
| n2 | 1075 | n2 | 1075 |
| stat F | 0.016844 | stat F | 0.043573 |

| Ndat | 1079 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| param | | Lineare | | | quadratico | | | cubico | |
| b0 | -45.2867 | | | | | | | | |
| b1 | 0.620518 | | | 0.064754 | | | 1.560451 | | |
| b2 | | | | 0.001703 | | | -0.01652 | | |
| b3 | | AIC | 1637.195 | | AIC | 1636.48 | 5.55E-05 | AIC | 1636.70411 |
| npar | 2 | | | 2 | | | 3 | | |
| Nlib | 1076 | SSQ= | 35207.96 | 1076 | SSQ= | 35154.28 | 1075 | SSQ= | 35021.2801 |
| | | | | | | | | | |
| | fit | residui | res^2 | fit | residui | res^2 | fit | residui | res^2 |

Anche per le donne il miglioramento dell'accordo usando una regressione con polinomi di terzo grado non è molto alto. È' possibile che il miglioramento che si era osservato passando da regressione lineare

a regressione cubica senza considerare le differenza uomini/donne fosse fittizio? Per verificare questa ipotesi vediamo se è preferibile un fit cubico sul gruppo intero dei dati o due fit lineari indipendenti su uomini e donne.

Per questo prepariamo una tabella in cui riassumiamo i valori ottenuti per SSQ nei diversi casi: lineare, quadratico e cubico sia per i dati globali, sia separatamente uomini e donne

| | lin | | quad | | cub | |
|---|---|---|---|---|---|---|
| tutti | SSQ= | 149714.76 | SSQ= | 147857.86 | SSQ= | 146397.09 |
| uomini | SSQ= | 67614.492 | SSQ= | 67504.914 | SSQ= | 67442.667 |
| donne | SSQ= | 35207.965 | SSQ= | 35154.285 | SSQ= | 35021.28 |

Per le regressioni effettuate separatamente sui due gruppi $SSQ_T$ è la somma degli SSQ calcolati per Uomini e Donne separatamente e va usato nel calcolo dell'AIC o del test F. Il numero di parametri è la somma dei parametri utilizzati per uomini e donne separatamente. Si vede che utilizzare due modelli lineari indipendenti separatamente per uomini e donne migliora l'accordo rispetto all'uso di un modello cubico che non tenga conto delle differenze tra i gruppi.

**Dal foglio** *Tutti*

**Dai fogli** *Uomini* **e** *Donne*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Ndat | 2759 | | | | | | | |
| param | | cub tutti | | | lin Uomini + lin donne | | test insieme/separati | |
| npar | 3 | AIC | 4764.6701 | 4 | AIC | 4343.3168 | F | 1167.1043 |
| Nlib | 2755 | SSQ= | 146397.09 | 2754 | SSQ_U= | 67614 | n1 | 1 |
| | | | | | SSQ_D= | 35208 | n2 | 2754 |
| | | | | | SSQ_T= | 102822 | stat F | 1.42E-213 |

**Esercizio 2**) **Propagazione degli errori** Volendo calcolare la densità di un blocchetto di rame di forma cilindrica sono state effettuate una serie di misure del diametro D (mm), dell'altezza H(cm) e del peso W (g). Le misure di D e H sono effettuate con uno strumento (calibro) avente una sensibilità di 0.05 mm mentre il peso è misurato con una bilancia di precisione avente una sensibilità di 0.02 g. I dati sono registrati nel file ASCII **cilindro1.dat**. Il file Excel **densita.xls** (cartella **misura1**) è utile come linea guida dell'esercizio mostra.

Importare i dati in un foglio elettronico e, per ognuna delle misure, calcolare il volume e la densità, ricordando che il volume del cilindro è:

$$V = \pi r^2 \cdot H$$

Mentre la densità è:

$$\rho = \frac{W}{V} = \frac{W}{\pi\, r^2\, H} = \frac{4\,W}{\pi\, D^2\, H}$$

A) Utilizzare un foglio elettronico per calcolare la densità per ognuna delle misure effettuate facendo attenzione alle unità di misura utilizzate: D [mm], H [cm], W [g] e ρ [g $cm^{-3}$]

| misure | | | Volume | | Densità | |
|---|---|---|---|---|---|---|
| D | H | M | V | | ρ | |
| mm | cm | g | cm^3 | | g/cm^3 | |
| 0.05 | 0.005 | 0.02 | V | err | | err. |
| 4.98348 | 10.00234 | 1.753811 | 1.951 | | 0.899 | |
| 4.944095 | 9.98986 | 1.768074 | 1.918 | | 0.922 | |
| 5.013508 | 10.00072 | 1.754515 | 1.974 | | 0.889 | |
| 5.055875 | 9.997964 | 1.757555 | 2.007 | | 0.876 | |
| 5.053845 | 9.99654 | 1.771167 | 2.005 | | 0.883 | |
| 5.064185 | 10.00055 | 1.752667 | 2.014 | | 0.870 | |
| 4.932029 | 9.987474 | 1.769578 | 1.908 | | 0.927 | |
| 4.987039 | 10.00539 | 1.769365 | 1.954 | | 0.905 | |
| 5.050865 | 9.998944 | 1.754204 | 2.003 | | 0.876 | |

B) Volume e Densità sono due grandezze derivate. Il calcolo degli errori, per ogni misura, richiede l'applicazione delle formule di propagazione degli errori statistici assumendo che gli errori sulle grandezze misurate (D, H e W) siano casuali e distribuiti "normalmente" e che le grandezze misurate siano indipendenti.

Il volume è calcolato da misure di D e H: $V = \pi H \frac{D^2}{4}$

nelle ipotesi fatte possiamo stimare l'errore sul volume utilizzando la formula per la propagazione degli errori statistici:

$$\sigma_f^2 = \sum_i \left(\frac{\partial f}{\partial p_i}\right)^2 \sigma_{p_i}^2$$

che, nel nostro caso diventa:

$$\epsilon_V = \sqrt{\left|\frac{\partial V}{\partial H}\right|^2 \epsilon_H^2 + \left|\frac{\partial V}{\partial D}\right|^2 \epsilon_D^2}$$

con

$$\frac{\partial V}{\partial D} = \frac{\pi D H}{2} = \frac{2V}{D} \quad \text{e} \quad \frac{\partial \mathrm{V}}{\partial \mathrm{H}} = \frac{\pi \mathrm{D}^2}{4} = \frac{\mathrm{V}}{\mathrm{H}}$$

e $\varepsilon_D$ e $\varepsilon_H$ sono gli errori di misura. Per ora assumiamo che questo sia ½ dell'errore di lettura. Nel calcolo fare attenzione alle unità di misura utilizzate: H [cm] e D[mm]

| Volume | | Errori parziali | |
|---|---|---|---|
| V | | errore dovuto | errore dovuto |
| cm^3 | | al Diametro | all'altezza |
| V | err | | |
| 1.951 | 0.010 | 0.010 | 0.0005 |
| 1.918 | 0.010 | 0.010 | 0.0005 |
| 1.974 | 0.010 | 0.010 | 0.0005 |
| 2.007 | 0.010 | 0.010 | 0.0005 |
| 2.005 | 0.010 | 0.010 | 0.0005 |
| 2.014 | 0.010 | 0.010 | 0.0005 |
| 1.908 | 0.010 | 0.010 | 0.0005 |

Si può notare come l'errore sulla misura del diametro del cilindro fornisca il contributo dominante all'errore sul volume

| Densità | | errori parziali | | |
|---|---|---|---|---|
| ρ | | errore dovuto | errore dovuto | errore dovuto |
| g/cm^3 | | al Diametro | all'altezza | al peso |
| | err | | | |
| 8.980 | 0.090 | 0.090 | 0.002 | 0.005 |
| 9.149 | 0.093 | 0.093 | 0.002 | 0.005 |
| 8.890 | 0.089 | 0.089 | 0.002 | 0.005 |
| 8.708 | 0.086 | 0.086 | 0.002 | 0.005 |
| 8.754 | 0.087 | 0.087 | 0.002 | 0.005 |
| 8.699 | 0.086 | 0.086 | 0.002 | 0.005 |
| 9.175 | 0.093 | 0.093 | 0.002 | 0.005 |
| 8.941 | 0.090 | 0.090 | 0.002 | 0.005 |

Per l'errore sulla densità, la propagazione degli errori è:

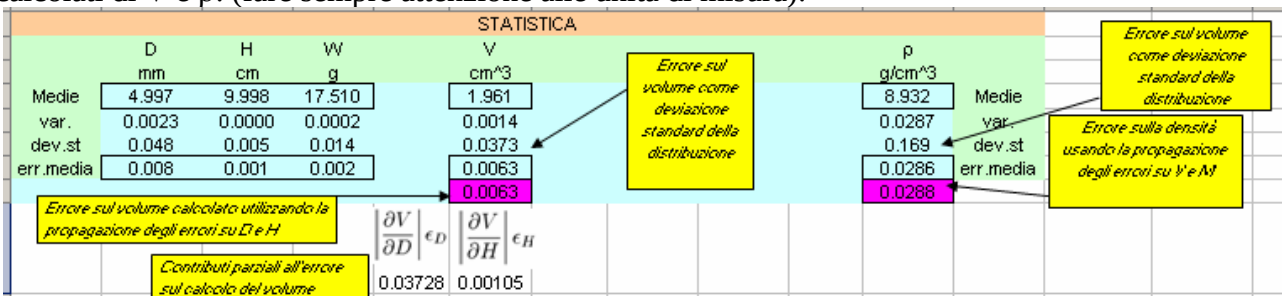
$$\epsilon_\rho^2 = \left|\frac{\partial \rho}{\partial D}\right|^2 \epsilon_D^2 + \left|\frac{\partial \rho}{\partial H}\right|^2 \epsilon_H^2 + \left|\frac{\partial \rho}{\partial W}\right|^2 \epsilon_W^2$$

dove:

$$\left|\frac{\partial \rho}{\partial D}\right| = \frac{8\,W}{\pi\,D^3\,H} = \frac{2\rho}{D}, \qquad \left|\frac{\partial \rho}{\partial H}\right| = \frac{4\,W}{\pi\,D^2\,H^2} = \frac{\rho}{H}, \qquad \left|\frac{\partial \rho}{\partial W}\right| = \frac{4}{\pi\,D^2\,H} = \frac{\rho}{W}$$

Si può usare una formula più semplice utilizzando direttamente l'errore su V anzichè separatamente su D e H. Provare!

**C)** calcolare i valori medi, varianza e deviazione standard per i valori misurati D, H e W e per i valori calcolati di V e ρ. (fare sempre attenzione alle unità di misura):

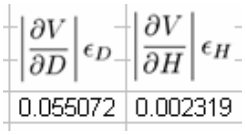



L'errore standard della media è, per definizione: $\epsilon_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$

dove N è il numero di misure effettuate. In base alle deviazioni standard delle distribuzioni D, H, W (misurate) e W e ρ (calcolate) si possono quindi stimare gli errori sulle medie, che risultano migliori rispetto a quelli ottenuti da una singola misura.
Osservando la deviazione standard dalle misure di D, H e M si vede come queste siano in ottimo accordo con quelle attese in base alla sensibilità degli strumenti.

Per calcolare l'errore sui valori medi calcolati su Volume e densità possiamo applicare la propagazione degli errori utilizzando gli errori calcolati dalle distribuzioni di D, H e W.
Il calcolo degli errori effettuato su V e ρ utilizzando le regole di propagazione degli errori o dalla deviazione standard della distribuzione sono del tutto simili.

Nota: come prima sulle singole misure il contributo all'errore sul volume dovuto agli errori su D ed H si nota che l'errore su D ha un effetto molto più grande che non l'errore su H:
Lo stesso vale per l'errore sul calcolo della densità.




Nella tabella finale riportare il risultato è riportato con il numero corretto di cifre significative accordate con l'errore:

| Risultato | |
|---|---|
| ρ= | 8.932 |
| err.= | 0.029 |

Generalmente l'errore si riporta arrotondato alla prima cifra significativa diversa da 0. Se questa e 1 o 2

si riportano due cifre significative. Nelle indicazioni del NIST (national institute for standards) è possibile utilizzare sempre due significative per gli errori. Questo semplifica la produzione automatica di tabelle, ad esempio con fogli elettronici.

**Esercizio 3) Propagazione degli errori: Variabili correlate** Per calcolare la densità del rame sono stati misurati il volume e il peso di una serie di cilindretti nominalmente eguali. Il file **cilindro2.dat** contiene una serie di misure del diametro D (mm), dell'altezza H(cm) e del peso W (g) per i cilindretti. Le misure di D e H sono effettuate con uno strumento (calibro) avente una sensibilità di 0.05 mm mentre il peso è misurato con una bilancia di precisione avente una sensibilità di 0.02 g.

Come sopra utilizzare un foglio elettronico per calcolare la densità per ognuna delle misure effettuate facendo attenzione alle unità di misura utilizzate: D [mm], H [cm], W [g] e ρ [g $cm^{-3}$]

Utilizzare il modello precedente per calcolare i parametri statistici delle distribuzione.

C) calcolare i valori medi, varianza e deviazione standard per i valori misurati D, H e W e per i valori calcolati di V e ρ. (fare sempre attenzione alle unità di misura)

| | STATISTICA | | | | |
|---|---|---|---|---|---|
| | D | H | M | V | ρ |
| | mm | cm | g | cm^3 | g/cm^3 |
| Medie | 4.996 | 10.000 | 17.483 | 1.960 | 8.902 |
| var. | 0.0049 | 0.0001 | 0.2353 | 0.0030 | 0.0095 |
| dev.st | 0.070 | 0.012 | 0.485 | 0.055 | 0.097 |

Dev.st(…)

Le deviazioni standard delle distribuzioni forniscono una determinazione sperimentale della dispersione delle misure e, quindi, dell'errore associato alle diverse grandezze. Osservando la deviazione standard dalle misure di D, H e M si vede come queste siano sensibilmente maggiori di quelle attese in base alla sensibilità degli strumenti. Questo effetto può essere dovuto a reali differenze tra i vari cilindretti utilizzati per la misura. Assumendo che gli effetti fisici (dimensioni e densità) e gli effetti dell'incertezza della misura siano indipendenti, questi si sommano in quadratura. Possiamo quindi stimare il contributo alla varianza dovuto alle diversità fisiche tra i vari cilindretti: $\sigma^2_{tot} = \sigma^2_{mis} + \sigma^2_{real}$ e quindi: $\sigma^2_{real} = \sigma^2_{tot} - \sigma^2_{mis}$

| D | H | M |
|---|---|---|
| mm | cm | g |
| 4.996 | 10.000 | 17.451 |
| 0.0049 | 0.0001 | 0.2437 |
| 0.070 | 0.012 | 0.494 |
| dev.st fisica | | |
| 0.066 | 0.012 | 0.494 |

Vediamo che, in questo caso, gran parte dell'errore osservato sulle grandezze misurate sia dovuto a effetti fisici (differenze reali tra i cilindretti) piuttosto che a errori di lettura (che appaiono praticamente trascurabili).

D) Per calcolare l'errore sui valori medi calcolati su Volume e densità possiamo applicare la propagazione degli errori utilizzando gli errori calcolati dalle distribuzioni di D, H e W che sono più realistici degli errori di lettura.
Se come per il volume si calcola l'errore sulla densità applicando la formula per la propagazione dell'errore statistico si vede che la stima dell'errore è molto diversa dall'errore calcolato utilizzando la distribuzione delle densità (deviazione standard)

| ρ | | |
|---|---|---|
| g/cm^3 | | |
| 8.902 | Medie | |
| 0.0093 | var. | |
| 0.097 | dev.st | |
| 0.3545 | | |

*Errore sul volume come deviazione standard della distribuzione*

*Errore sulla densità usando la propagazione degli errori su V e M*

Proviamo a capire perché!
La formula utilizzata per la propagazione degli errori statistici:

$$\sigma_f^2 = \sum_i \left(\frac{\partial f}{\partial p_i}\right)^2 \sigma_{p_i}^2$$

è valida nell'ipotesi che le misure siano indipendenti. Se questo non è vero bisogna tener conto della correlazione tra le variabili e la formula corretta è:

$$\sigma_f^2 = \sum_{i,j} \frac{\partial f}{\partial p_i}\frac{\partial f}{\partial p_j} Cov_{ij}$$

dove $Cov_{ij}$ sono i coefficienti di covarianza, i termini diagonali $Cov_{ii} = \frac{1}{N}\sum_k (p_i - \bar{p})^2 = \sigma_{p_i}^2$

rappresentano la varianza della variabile i mentre i termini misti $Cov_{ij} = \frac{1}{N}\sum_k (p_i - \overline{p_i})(p_j - \overline{p_j})$ sono:

e si possono calcolare usando la funzione COVARIANZA tra due matrici di dati.

Quando le misure di D, H e W sono dominate da errori di lettura l'indipendenza dei dati di altezza, diametro e peso è plausibile. In questo caso la situazione è diversa: abbiamo visto che la dispersione dei risultati per D e H, molto maggiore dell'errore di lettura, suggerisce che i cilindretti siano effettivamente diversi l'uno dall'altro. Ora il peso e il volume dei cilindretti non sono variabili indipendenti, al contrario il peso di un cilindretto dipende dal suo volume, quindi peso e dimensioni non sono variabili indipendenti, anzi, esiste una legge fisica che lega il volume al peso:

$$W = \rho V$$

Il coefficiente di correlazione di Pearson r quantifica il grado di correlazione tra due grandezze X ($x_1, x_2, \ldots x_n$) e Y ($y_1, y_2, \ldots y_n$), indica quale frazione della variabilità osservata su Y dipenda dalla variabilità di X.

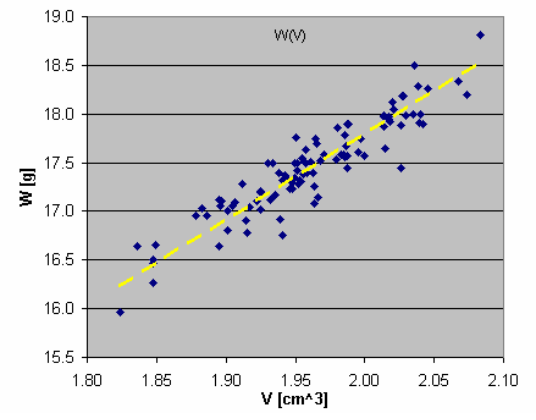
$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$$

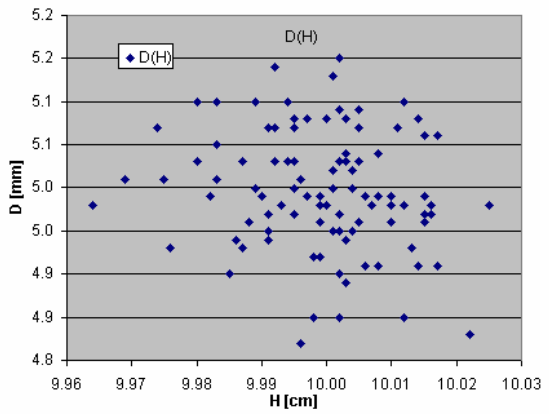
$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})$$

| | |
|---|---|
| correlazione M-V | 0.92 |
| correlazione D-H | -0.15 |

Si può usare la funzione CORRELAZIONE di Excel per valutare il grado di correlazione tra due insiemi di dati. La correlazione tra i valori di Volume (V) e Peso (W) è grande mentre tra Diametro (D) e Altezza (H) è piccola.

Si riporti su grafici i valori osservati di W in funzione di V e i valori di H in funzione di D




I grafici permettono di evidenziare il significato della "correlazione". I pesi osservati mostrano un andamento crescente in funzione dei pesi osservati. (correlazione elevata). Al contrario è difficile vedere una relazione tra i valori W e D (correlazione bassa).

Calcoliamo quindi l'errore sulla densità tenendo conto della correlazione tra le misure.
Nel nostro caso:

$$\sigma_\rho^2 = \left(\frac{1}{V}\sigma_M^2\right)^2 + \left(\frac{M}{V^2}\sigma_V^2\right)^2 + - \frac{1}{V}\cdot\frac{M}{V^2}Cov_{MV}$$

Usando la formula corretta per il calcolo della varianza, che tenga conto della correlazione tra le variabili (covarianza), i risultati sono del tutto consistenti: