Politecnico di Milano - Scuola di Ingegneria Industriale e dell'Informazione

III Appello di Statistica per Ingegneria Energetica
12 settembre 2016

Cognome, Nome e Numero di matricola:

Problema 1. Nel comune di Lambrate è stato aperto un corso di spagnolo con docente madre lingua, un costo contenuto ed un numero massimo di iscritti di 35 persone. La tabella seguente riporta le età degli iscritti, divisi per maschi e femmine.

F	24	24	24	25	21	22	23	24	25	25	25	26	26	27	28	28	28	29
M	22	23	24	24	24	26	28	28	30	31	31	39	18	19	20	21	21	

(a) Rappresentare graficamente la distribuzione del sesso degli iscritti.

(b) Rappresentare con due istogrammi separati la distribuzione dell'età degli iscritti per i maschi e per le femmine. Si suddivida in cinque classi equispaziate l'intervallo $[15, 40]$ per i maschi, l'intervallo $[20, 30]$ per le femmine.

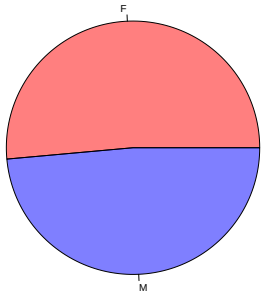
(c) Il numero di classi introdotte vi pare adeguato per i dati raccolti?

(d) Confrontare qualitativamente la distribuzione delle età nei due gruppi; in particolare, ipotizzare e motivare la relazione tra media e mediana.

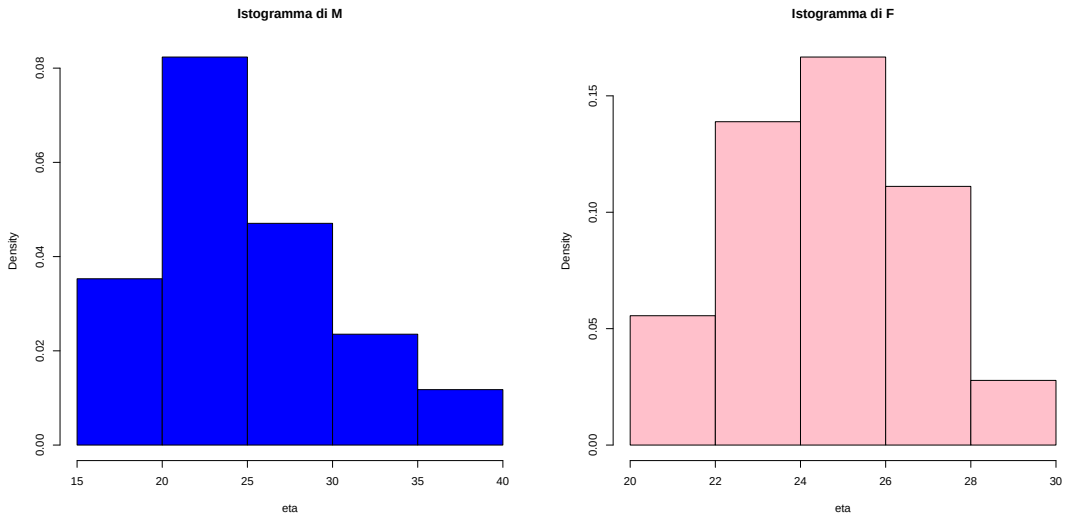
(e) Calcolare media, deviazione standard della popolazione e quartili dei due insiemi di dati.

Risultati.

(a) Si può costruire un semplice grafico a torta (oppure un grafico a barre):



(b) Istogrammi:



(c) I due gruppi sono formati da $n = 17, 18$ osservazioni. In entrambi i casi quindi le regole empiriche portano ad un numero di classi $k \simeq 5$: $\sqrt{n} = 4.1, 4.3$ mentre $1 + \log_2 18 = 5.1, 5.2$. Inoltre gli istogrammi ottenuti non presentano caratteristiche che inducano a variare il numero di classi.

(d) Esaminando gli istogrammi si nota una evidente asimmetria nella distribuzione dei dati maschili, con la classe modale spostata verso il margine sinistro del grafico e una coda superiore pesante (asimmetria a destra). In tale situazione la media sarà verosimilmente maggiore della mediana. La distribuzione dei dati nella categoria femminile, invece, è piuttosto simmetrica, pertanto non ci aspettiamo scostamenti particolari tra media e mediana.

(e) Per la popolazione maschile, si ha: media pari a 25.24 anni, dev. standard della popolazione pari a 5.25 anni e primo, secondo e terzo quartile pari, rispettivamente, a 21, 24 e 28 anni. Per la popolazione femminile, si ottengono nell'ordine i valori: 25.22, 2.12, 24, 25 e 27 anni. In particolare, si trova conferma della relazione tra medie e mediane ipotizzate al punto precedente.

Problema 2. La quantità di vino delle bottiglie Château Maison ha distribuzione normale con deviazione standard di 1 ml. Tuttavia Omèr Cavalieri, un noto sommelier italo-francese, ha l'impressione che l'azienda vinicola Château Maison riempia le proprie bottiglie con una quantità di vino mediamente inferiore al contenuto dichiarato sull'etichetta, ovvero 750 ml. Per confermare il suo sospetto, decide di effettuare un'analisi statistica su un campione casuale di bottiglie.

(a) Impostare un opportuno test statistico per aiutare il sig. Cavalieri a validare la sua affermazione. In particolare introdurre il parametro di interesse e specificare ipotesi nulla, ipotesi alternativa e regione critica di livello α del test.

(b) Quante bottiglie di vino sarebbero necessarie al sig. Cavalieri per avere un test di significatività 5% con potenza almeno dell'80%, nel caso in cui la media fosse 749.5 ml?

Il sig. Cavalieri riesce però a controllare il contenuto di solo 10 bottiglie; di seguito i dati raccolti.

750.59 748.20 750.07 748.96 750.39 747.41 748.59 750.40 748.91 747.62

(c) Cosa può concludere il sig. Cavalieri con una significatività del 5%?

(d) Calcolare il p-value dei dati raccolti per il test del punto (a).

(e) Costruire un intervallo di confidenza al 95% per il contenuto medio delle bottiglie di vino Château Maison.

Risultati.

(a) Bisogna effettuare un test unilatero per la media μ di una popolazione normale con varianza σ_0^2 nota. Nel nostro caso

$$\mu = \text{contenuto medio delle bottiglie di vino Château Maison}$$

$$\sigma_0^2 = 1$$

$$\mu_0 = \text{valore di confronto} = 750$$

$$H_0 : \mu \geq 750 \qquad H_1 : \mu < 750 \qquad RC_\alpha : \frac{\bar{x} - 750}{1/\sqrt{n}} < -z_\alpha$$

dove n è la numerosità del campione casuale raccolto.

(b) Servono almeno 25 bottiglie: detta $\mu_1 = 749.5$ la media vera, si ha $Z = \dfrac{\bar{X} - \mu_1}{1/\sqrt{n}} \sim N(0,1)$, per cui

$$0.8 \leq P\Big(Z < -z_{0.05} + (\mu_0 - \mu_1)\sqrt{n}\Big) = P\Big(Z < -1.645 + 0.5\sqrt{n}\Big)$$

e quindi $-1.645 + 0.5\sqrt{n} \geq z_{0.2} \approx 0.85$, ovvero $\sqrt{n} \geq 4.99$;

(c) La media campionaria è $\bar{x} = 749.114$ con $n = 10$.
A livello 5% si rifiuta H_0, dato che $\frac{749.114-750}{1/\sqrt{10}} \approx -2.8 < -1.645 \approx -z_{0.05}$. I dati raccolti permettono di concludere ad un livello di significatività del 5% che il contenuto medio delle bottiglie di vino Château Maison è inferiore a 750 ml (conclusione forte).

(d) p-value= $P(Z < -2.8) \approx 0.0025$.

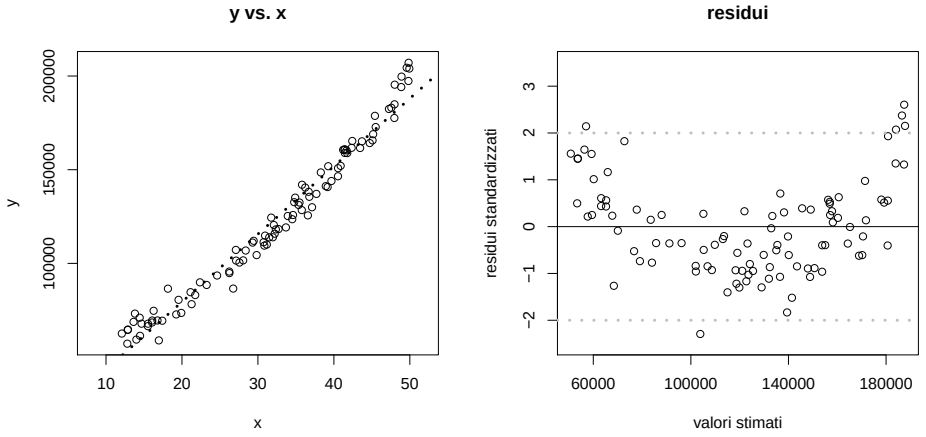
(e) IC= $(-\infty, 749.114 + \frac{1}{\sqrt{10}} z_{0.05}) \approx (-\infty, 749.63)$.

Problema 3. Una particella carica accelerata radialmente produce una particolare radiazione magnetica che viene chiamata radiazione di sincrotrone. Questo è un fenomeno molto diffuso in numerose stelle che si trovano alla fine del loro ciclo vitale grazie alla presenza di elettroni che vengono diffusi nelle direzioni polari dell'oggetto. Sono state raccolte le velocità medie degli elettroni espresse in km/s (Y) emessi da 100 stelle di varia massa espressa in masse solari $M_\odot$ (X) ($1 = 1\ M_\odot$). L'intensità della radiazione è tanto più forte quanto più la velocità degli elettroni è prossima a quella della luce ($3 \cdot 10^8$ km/s). Di seguito si riportano media e deviazione standard campionarie di X e Y:

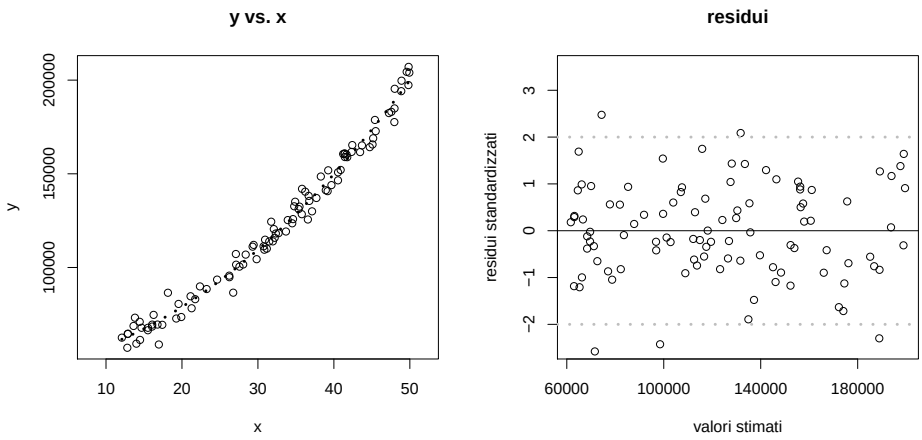
$$\bar{x}_{100} = 32.0142\ M_\odot \qquad s_{x_{100}} = 11.2284\ M_\odot$$
$$\bar{y}_{100} = 62081\ \text{km/s} \qquad s_{y_{100}} = 35875\ \text{km/s}$$

Un gruppo di astrofisici sta studiando un ammasso stellare in cui questo fenomeno è molto diffuso. Costruiscono ben 3 diversi modelli di regressione lineare di cui vengono riportati di seguito i p-value dei test di Shapiro-Wilk condotti sui relativi residui, alcuni scatterplot, i valori delle stime e degli indici principali.

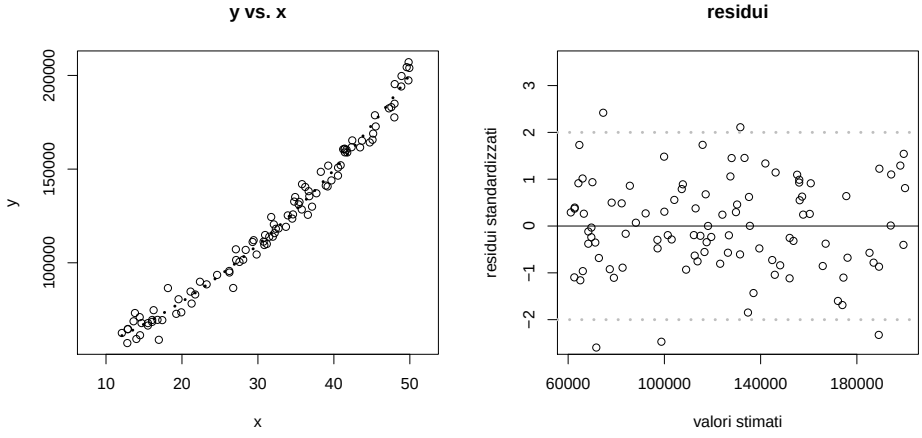
$$sw_1 = 0.0156 \qquad sw_2 = 0.8790 \qquad sw_3 = 0.7465$$




```
Residuals:
   Min     1Q Median     3Q    Max
-17297  -6099  -1091   3903  19645

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7058.95    2289.01   3.084  0.00265 **
x            3619.30      67.51  53.614  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7542 on 98 degrees of freedom
Multiple R-squared:  0.967,     Adjusted R-squared:  0.9667
F-statistic:  2874 on 1 and 98 DF,  p-value: < 2.2e-16
```




```
Residuals:
     Min       1Q   Median       3Q      Max
-12672.0  -3250.3   -314.3   3538.2  12167.5

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 47080.374   3767.906  12.495   <2e-16 ***
x             616.805    263.284   2.343   0.0212 *
I(x2)          48.795      4.219  11.567   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4915 on 97 degrees of freedom
Multiple R-squared:  0.9861,    Adjusted R-squared:  0.9859
F-statistic:  3451 on 2 and 97 DF,  p-value: < 2.2e-16
```




```
Residuals:
     Min       1Q   Median       3Q      Max
-12806.0  -3346.1   -369.5   3481.4  11942.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.191e+04  1.166e+04   3.594 0.000516 ***
x           1.219e+03  1.310e+03   0.930 0.354630
I(x2)       2.802e+01  4.448e+01   0.630 0.530184
I(x3)       2.199e-01  4.687e-01   0.469 0.640086
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4935 on 96 degrees of freedom
Multiple R-squared:  0.9862,    Adjusted R-squared:  0.9857
F-statistic:  2283 on 3 and 96 DF,  p-value: < 2.2e-16
```

(a) Scrivere la relazione fra Y e X ipotizzata da ciascuno dei tre modelli di regressione empirici gaussiani.

(b) Identificare quali modelli presentano residui omoschedastici. Perché?

(c) Indicare, fra i modelli selezionati al punto precedente, per quali è soddisfatta l'ipotesi gaussiana. Perché?

(d) Indicare, fra i modelli selezionati ai punti precedenti, quale modello spiega meglio la relazione tra la velocità media degli elettroni e la massa stellare. Perché?

(e) Prevedere puntualmente la velocità media degli elettroni emessi da una stella da 25 $M_{\odot}$.

Si vuole capire se l'intercetta del modello selezionato sia uguale o meno a 5000.

(f) Si imposti un test statistico specificando ipotesi nulla, alternativa, regione critica.

(g) Si calcoli il p-value dei dati e si tragga una conclusione. La conclusione è forte o debole?

Risultati.

(a) Indicando con $\epsilon \sim N(0, \sigma^2)$ l'errore gaussiano presente in ciascun modello, possiamo scrivere:

$$\begin{aligned} &\text{modello 1:} \quad Y = \beta_0 + \beta_1 X + \epsilon \\ &\text{modello 2:} \quad Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon \\ &\text{modello 3:} \quad Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon \end{aligned}$$

(b) Sia il modello 2 che 3 presentano residui omoschedastici. Non si evidenziano pattern particolari, al contrario del modello 1 dove si nota chiaramente un comportamento quadratico.

(c) Sia il modello 2 che 3 soddisfano l'ipotesi gaussiana sui residui come riportato dai testi di Shapiro-Wilk.

(d) Sicuramente il modello 2 è il migliore.

Come il modello 3 verifica l'ipotesi gaussiana ed è un modello globalmente significativo, ma il modello 2 presenta un R^2-adjusted leggermente superiore e soprattutto tutti i suoi predittori risultano significativi (mentre il modello 3, che pur si comporta bene dal punto di vista dei residui della significatività globale e di R^2-adjusted, non possiede alcun predittore significativo; anzi i p-value di significatività dei predittori indicano proprio l'eliminazione di X^3, ovvero il passaggio al modello 2).

(e) Utilizzando il modello 2:

$$\widehat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X + \widehat{\beta_2} X^2 = 47080.37428 + 616.80504 \cdot 25 + 48.79517 \cdot 25^2 = 92997.48$$

(f) Impostiamo un test bilatero come segue:

$$\begin{aligned} H_0 &: \beta_0 = 5000 \\ H_1 &: \beta_0 \neq 5000 \end{aligned}$$

con regione critica di livello α

$$RC : |t_0| > t_{\alpha/2}(97), \qquad \text{dove } t_0 = \frac{\widehat{\beta_0} - 5000}{se(\widehat{\beta_0})}.$$

(g) Per i dati raccolti risulta

$$t_0 = \frac{47080.37 - 5000}{3767.906} = 11.16811$$

per cui

$$|t_0| = t_{\alpha/2}(97) \implies t_{\alpha/2}(97) = 11.16811 > t_{0.0005}(97) \simeq t_{0.0005}(120) = 3.291 \implies \alpha < 0.001$$

quindi i dati consentono di rifiutare H_0 in favore di H_1 a tutti gli usuali livelli di significatività: $\beta_0 \neq 5000$ (conclusione forte).