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

II Appello di Statistica per Ingegneria Energetica
7 luglio 2015

Cognome, Nome e Numero di matricola:

Problema 1. Il Dottor H. vuole conoscere l'angolo θ compreso fra l'ipotenusa e il cateto orizzontale di un fregio sull'isola di Rodi. Ha a disposizione uno strumento per la misura delle lunghezze, in cm, affetto da errore (additivo) casuale $\epsilon \sim N(0\,,\,0.01)$. Ordina quindi che vengano effettuate 13 misure indipendenti $X_1, \ldots, X_{13}$ della lunghezza x del cateto orizzontale e 15 misure indipendenti $Y_1, \ldots, Y_{15}$ della lunghezza y del cateto verticale. Chiaramente il Dottor H. sa che

$$\theta = \arctan \frac{y}{x}.$$

(a) Esprimere il risultato X_k della k-esima misura in funzione di x e del corrispondente errore ϵ_k. Determinare quindi la distribuzione di X_k.

(b) Calcolare la probabilità di ottenere $|X_1 - x| > 0.1$ cm.

(c) Introdurre stimatori opportuni $\widehat{X}$ e $\widehat{Y}$ di x e y sulla base delle misure ordinate. Specificarne inoltre: distribuzione, distorsione, errore quadratico medio.

(d) Calcolare la probabilità di ottenere $|\widehat{X} - x| > 0.1$ cm.

(e) Proporre uno stimatore $\widehat{\Theta}$ per θ sulla base delle misure ordinate.

(f) Determinare, in funzione di x e y, eventualmente in modo approssimato: media, varianza, distorsione ed errore quadratico medio di $\widehat{\Theta}$.

Eseguite le misure, si trova

$$\overline{x}_{13} = 115.00 \text{ cm}, \qquad s_x^2 = 0.0103 \text{ cm}^2, \qquad \overline{y}_{15} = 186.64 \text{ cm}, \qquad s_y^2 = 0.00989 \text{ cm}^2.$$

(g) Fornire una stima puntuale di θ e del corrispondente errore quadratico medio.

Risultati.

(a) $X_k = x + \epsilon_k \sim N(x\,,\,0.01)$ per ogni k.

(b) $\mathbb{P}(|X_1 - x| > 0.1) = 2\left(1 - \Phi\left(\frac{0.1}{0.1}\right)\right) = 0.3173.$

(c)

$$\widehat{X} = \overline{X}_{13} \sim N\left(x\,,\,\frac{0.01}{13}\right), \qquad \text{bias}(\widehat{X}) = 0, \qquad \text{MSE}(\widehat{X}) = \frac{0.01}{13},$$

$$\widehat{Y} = \overline{Y}_{15} \sim N\left(y\,,\,\frac{0.01}{15}\right), \qquad \text{bias}(\widehat{Y}) = 0, \qquad \text{MSE}(\widehat{Y}) = \frac{0.01}{15}.$$

(b) $\mathbb{P}(|\widehat{X} - x| > 0.1) = 2\left(1 - \Phi\left(\frac{0.1}{0.1/\sqrt{13}}\right)\right) = 0.0003115.$

(e) $\widehat{\Theta} = \arctan\dfrac{\overline{Y}_{15}}{\overline{X}_{13}}.$

(f) Applicando il metodo delta, ovvero la formula di propagazione degli errori:

$$\mathbb{E}[\widehat{\Theta}] \simeq \theta, \qquad \text{Var}(\widehat{\Theta}) \simeq 0.01\left(\frac{1}{13}\frac{x^2}{(x^2+y^2)^2} + \frac{1}{15}\frac{y^2}{(x^2+y^2)^2}\right),$$

$$\text{bias}(\widehat{\Theta}) \simeq 0, \qquad \text{MSE}(\widehat{\Theta}) \simeq 0.01\left(\frac{1}{13}\frac{x^2}{(x^2+y^2)^2} + \frac{1}{15}\frac{y^2}{(x^2+y^2)^2}\right).$$

(g) $\widehat{\theta} = \arctan\dfrac{\overline{y}_{15}}{\overline{x}_{13}} = 1.0185795331 = 58.36°,$

$$\text{MSE}(\widehat{\Theta}) \simeq 0.01\left(\frac{1}{13}\frac{115^2}{(115^2+186.64^2)^2} + \frac{1}{15}\frac{186.64^2}{(115^2+186.64^2)^2}\right) = 1,4 \cdot 10^{-8}.$$

Problema 2. Supponendo di non conoscere alcuna cifra decimale di π, il matematico Mario Lazzarini vuole inferire statisticamente sul suo valore lanciando n aghi di lunghezza ℓ su un pavimento a strisce distanti $6\ell/5$, cosicché la probabilità che un singolo ago intersechi una qualche linea del pavimento valga

$$p = \frac{5}{3\pi}.$$

Sia X_k la variabile che indica se il k-esimo ago interseca una linea del pavimento.

(a) Come sono distribuite le variabili X_k?

In particolare il Lazzarini vuole confutare statisticamente la tesi, formulata da un progetto di legge del Parlamento dell'Indiana, secondo cui

$$\pi = 3.2.$$

(b) Introdurre un opportuno test statistico utile allo scopo del Lazzarini, specificando: ipotesi nulla, ipotesi alternativa, regione critica di livello α, condizioni di applicabilità.

(c) Calcolare la potenza del test introdotto nel caso di $n = 3{\cdot}408$ aghi, di un livello $\alpha = 0.01$, e del reale valore di $\pi = 3.141592920\ldots$

Il Lazzarini lancia quindi i suoi $3{\cdot}408$ aghi, contando $1{\cdot}808$ intersezioni.

(d) Stimare puntualmente p e π.

(e) Calcolare il p-value dei dati per il test introdotto.

(f) Trarre le debite conclusioni al livello scelto $\alpha = 0.01$.

(g) Se la conclusione tratta fosse errata, l'errore commesso sarebbe del primo o del secondo tipo?

Soluzione.

(a) $X_k \sim B(p)$ per ogni $k = 1, \ldots, n$.

(b) Posto $\pi_0 = 3.2$ e sua corrispondente probabilità $p_0 = \frac{5}{3 \cdot 3.2} = 0.520833$,

$$H_0 : p = p_0 \qquad \text{vs} \qquad H_1 : p \neq p_0$$

$$R_\alpha : |\overline{x}_n - p_0| > \sqrt{\frac{p_0(1-p_0)}{n}}\, z_{\alpha/2}$$

purché $\quad n \cdot p_0 > 5, \quad n\,(1-p_0) > 5.$

(c) Sapendo che $\pi = 3.141592920\ldots$, allora $p = 0.530516$ e la potenza richiesta vale

$$\begin{aligned}
&\mathbb{P}_p\left(\left|\overline{X}_n - p_0\right| > \sqrt{\frac{p_0(1-p_0)}{n}}\, z_{\alpha/2}\right) \\
&\qquad = \mathbb{P}_p\left(\overline{X}_n < p_0 - \sqrt{\frac{p_0(1-p_0)}{n}}\, z_{\alpha/2}\right) + \mathbb{P}_p\left(\overline{X}_n > p_0 + \sqrt{\frac{p_0(1-p_0)}{n}}\, z_{\alpha/2}\right) \\
&\qquad\quad = \Phi\left(\frac{p_0 - \sqrt{\frac{p_0(1-p_0)}{n}}\, z_{\alpha/2} - p}{\sqrt{\frac{p(1-p)}{n}}}\right) + 1 - \Phi\left(\frac{p_0 + \sqrt{\frac{p_0(1-p_0)}{n}}\, z_{\alpha/2} - p}{\sqrt{\frac{p(1-p)}{n}}}\right) \\
&\qquad\qquad = 0.0742.
\end{aligned}$$

(d) $\widehat{p} = \dfrac{1{\cdot}808}{3{\cdot}408} = 0.530516 \implies \widehat{\pi} = 3.141592920.$

(e)

$$z_{\alpha/2} = \frac{|\overline{x}_n - p_0|}{\sqrt{\frac{p_0(1-p_0)}{n}}} = 1.13$$

$$1 - \frac{\alpha}{2} = \Phi(1.13) = 0.8708 \implies \alpha = 0.2578$$

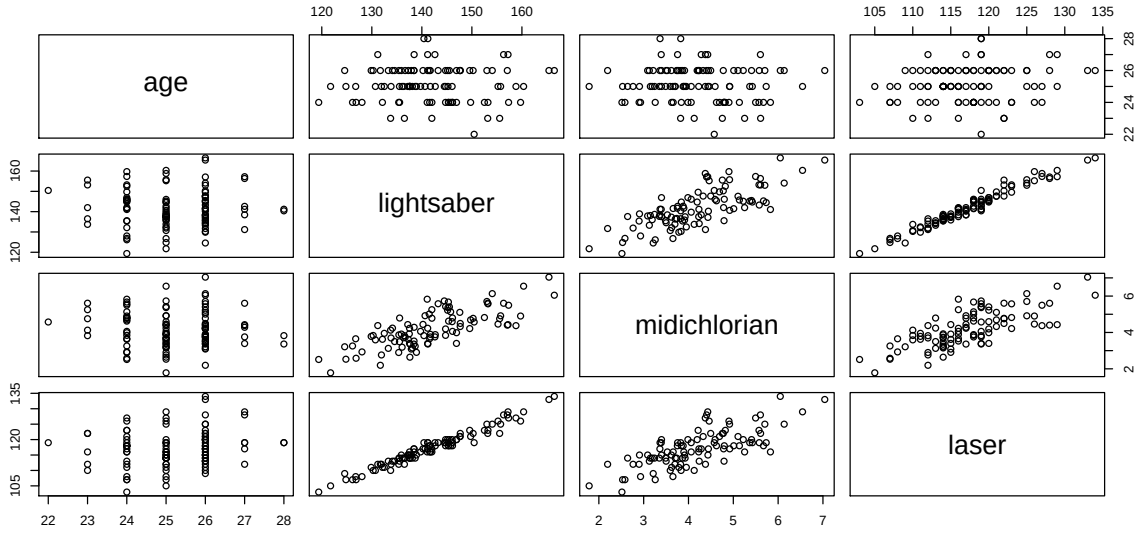
(f) I dati raccolti non consentono di rifiutare all'1% l'ipotesi nulla: potrebbe essere $\pi = 3.2$.

(g) Se la conclusione tratta fosse errata, l'errore commesso sarebbe del secondo tipo.

Problema 3. L'Ordine degli Jedi è da sempre attento all'addestramento dei sui allievi Padawan. Per questo motivo negli anni ha sviluppato tecniche di allenamento via via maggiormente sfidanti. Una delle più impegnative, consiste nel deviare il maggior numero di colpi di laser possibile in un minuto con il solo ausilio della spada laser per valutare la preparazione dell'allievo. Ogni singolo Padawan esegue questo esercizio in un'apposita arena. Il Maestro incaricato dell'esercizio raccoglie i seguenti dati per 100 allievi: età del Padawan (variabile `age`), lunghezza della lama della spada laser usata (variabile `lightsaber`), concentrazione di midichlorian nel sangue dell'allievo al momento dell'esame (variabile `midichlorian`) e in più registra il numero di colpi di laser deviati dal giovane Jedi sotto esame (variabile `laser`). I valori medi delle prime tre variabili risultano:

$$\overline{\texttt{midichlorian}} = 4.19, \qquad \overline{\texttt{age}} = 25.16, \qquad \overline{\texttt{lightsaber}} = 141.75.$$

La relazione tra tutti i dati invece è riportata nel seguente grafico. Si consideri il seguente modello



empirico gaussiano di regressione lineare semplice:

$$\texttt{laser} = \beta_0 + \beta_1 \cdot \texttt{midichlorian} + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2).$$

Si riportano di seguito il summary del modello, lo scatterplot dei dati con la retta di regressione, il grafico dei residui standardizzati e il loro normal probability plot.

(a) Si discuta la bontà del modello introdotto ponendo attenzione alla percentuale di variabilità di `laser` spiegata dal modello con la variabilità di `midichlorian`, alla bontà dell'ipotesi gaussiana, alla significatività globale del modello e alla significatività di `midichlorian`.

(b) Prevedere con un intervallo al 95% il numero di colpi deviati sapendo che sta per entrare nell'arena un Padawan con le seguenti caratteristiche: `midichlorian` $= 6.5$, `age` $= 23$, `lightsaber` $= 156$.

Al fine di creare esercizi "su misura" per gli allievi, alcuni Maestri Jedi, noti per le loro abilità statistiche, stanno elaborando un nuovo modello empirico gaussiano di regressione lineare multipla:

$$\texttt{laser} = \beta_0 + \beta_1 \cdot \texttt{midichlorian} + \beta_2 \cdot \texttt{age} + \beta_3 \cdot \texttt{lightsarber} + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2).$$

Si riportano di seguito il summary del modello e i grafici dei residui standardizzati e il loro normal probability plot.

(c) Discutere la significatività del predittore `midichlorian`. È cambiata? Perché?

(d) Alla luce di quanto osservato, i Maestri Jedi possono ritenersi soddisfatti? Suggerireste l'elaborazione di un ulteriore modello? Quale?

```
Call:
lm(formula = y.2 ~ midichlorian, data = dati.2)

Residuals:
   Min     1Q Median     3Q    Max
-8.283 -3.197 -0.136  2.753 10.793

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   98.9755     1.7631   56.14   <2e-16 ***
midichlorian   4.3421     0.4087   10.62   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.038 on 98 degrees of freedom
Multiple R-squared:  0.5353,    Adjusted R-squared:  0.5306
F-statistic: 112.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

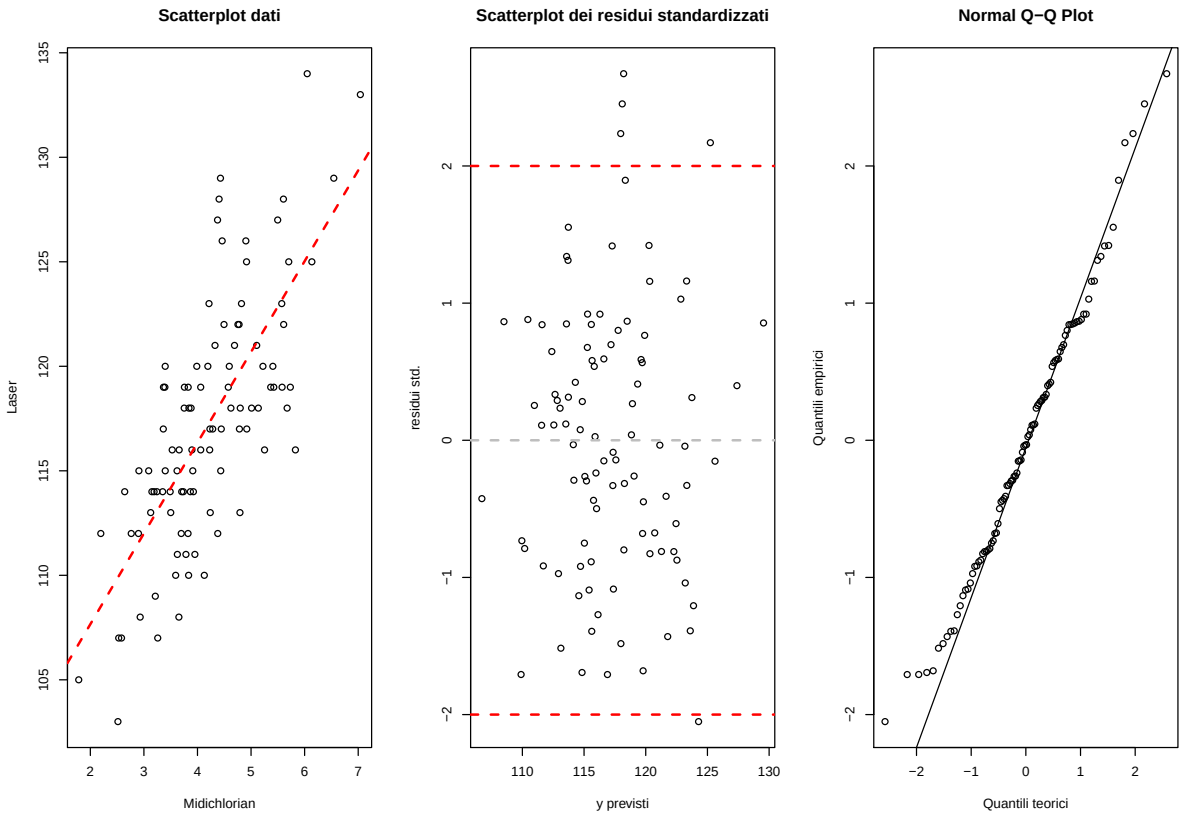


Figura 1: Output e grafici dei residui del modello semplice

```
Call:
lm(formula = y.2 ~ age + lightsaber + midichlorian, data = dati.2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.48673 -0.43364  0.03045  0.42104  1.43984

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.770962   1.708513   5.134 1.48e-06 ***
age          0.934049   0.054466  17.149  < 2e-16 ***
lightsaber   0.595047   0.009614  61.896  < 2e-16 ***
midichlorian 0.141322   0.092294   1.531    0.129
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6023 on 96 degrees of freedom
Multiple R-squared:  0.9899,    Adjusted R-squared:  0.9896
F-statistic:  3128 on 3 and 96 DF,  p-value: < 2.2e-16
```

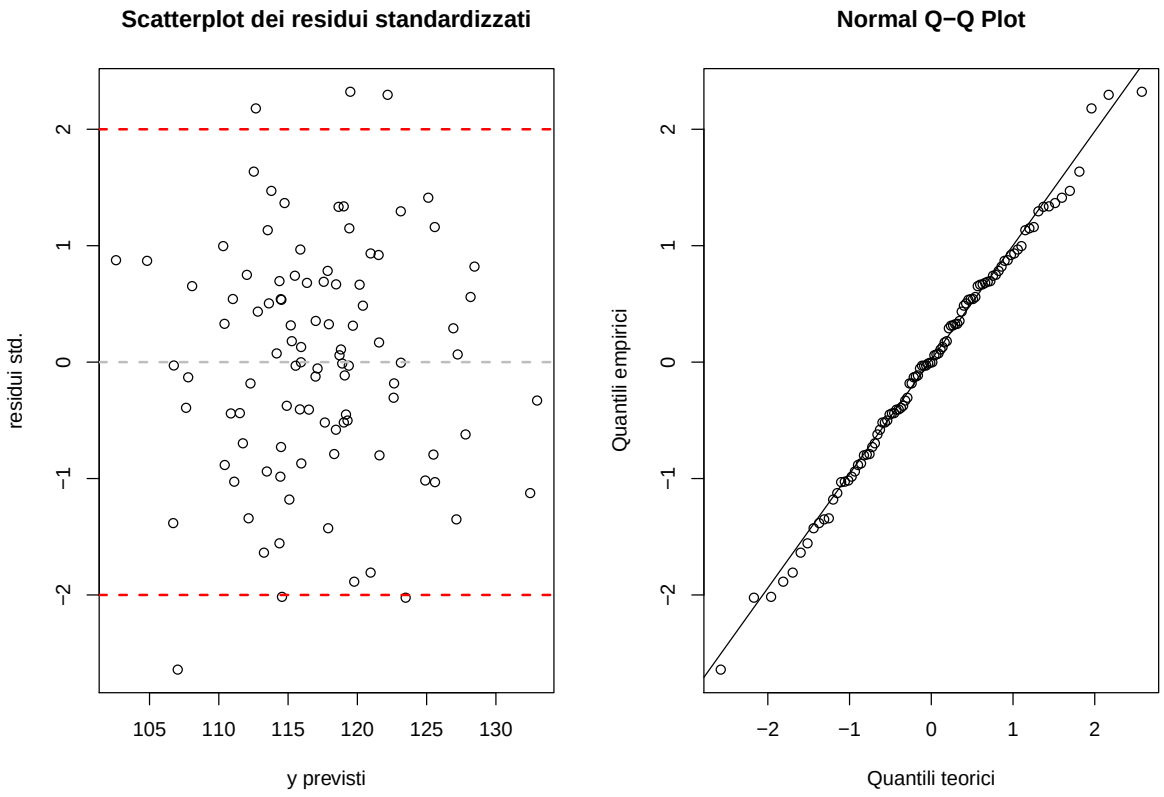


Figura 2: Output e grafici dei residui del modello completo

Risultati

(a) Il modello spiega circa il 53% della variabilità totale usando un solo predittore.

Lo scatterplot dei residui e il loro Normal Probability Plot confermano l'ipotesi di gaussiana: lo scatterplot dei residui non presenta pattern particolari (il che conferma come il modello sia omoschedastico) il Normal Probability Plot mostra valori ben allineati, anche se si osserva la presenza di una coda più pesanti ai valori più bassi. Anche il numero di outlier rispetta l'ipotesi di gaussianità.

La significatività globale del modello, essendo semplice, è identica a quella del predittore che risulta estremamente significativo con p-value $< 2 \cdot 10^{-16}$.

Giudichiamo il modello buono dal punto di vista della validità delle ipotesi gaussiane e medio dal punto di vista della capacità di intercettare il fenomeno osservato.

(b) L'intervallo di previsione è definito nel seguente modo:

$$\hat{y}_0 \pm t_{\alpha/2;n-2}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}\right]} = 127.1317 \pm 8.5409 = (118.5908\ ,\ 135.6726)$$

(c) Nel modello multiplo, la significatività del predittore `midichlorian` è molto diminuita, con un p-value salito oltre il 10%. Evidentemente la sua utilità diminuisce in presenza degli altri predittori. Se osserviamo il primo grafico, infatti, notiamo come i due predittori `lightsaber` e `midichlorian` siano fortemente correlati. Questo rende il predittore `midichlorian` non significativo se già si usa `lightsaber`.

(d) Con il modello completo la percentuale di variabilità di `laser` sale dal 53% al 99%, senza aver esagerato nel numero di predittori (R^2_{corretto} ha la stessa variazione), e senza aver perso l'ipotesi gaussiana (anzi, il normal probability plot è migliorato).

Questo rende il modello completo preferibile al modello semplice.

Tuttavia la scarsa significatività del predittore `midichlorian` indica l'opportunità di elaborare e analizzare un nuovo modello usando come predittori solo `age` e `lightsaber`.