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

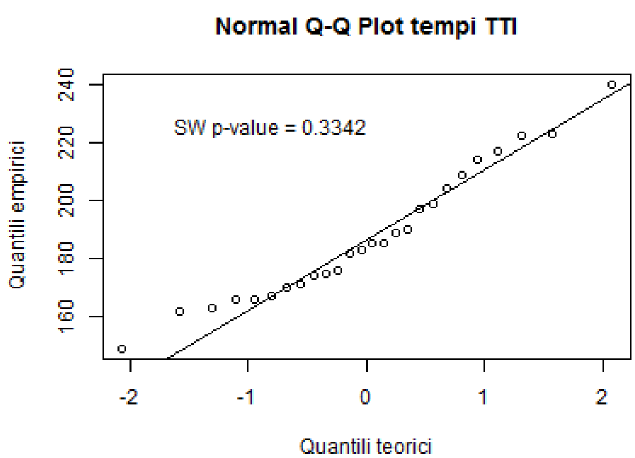
III appello di Statistica per Ingegneria Fisica
10 Settembre 2018

Cognome, Nome e Numero di matricola:

Problema 1. Un'azienda di abbigliamento per motociclisti sta collaudando un nuovo prototipo di giacca con airbag di protezione incorporato. Una delle caratteristiche essenziali di tale airbag è il Tempo Totale di Intervento (TTI), cioè l'intervallo tra l'istante in cui i sensori nella giacca rilevano l'incidente e quello in cui l'airbag completa il suo gonfiaggio.
Di seguito, sono riportati i TTI (in ms) ottenuti in 26 diversi crash test effettuati dall'azienda:

149	162	163	166	166	167	170	171	174	175	176	182	183
185	185	189	190	197	199	204	209	214	217	222	223	240 .

Per comodità, i dati sono già ordinati. Sotto si riportano anche il normal Q-Q plot col risultato del test di Shapiro-Wilk, e a fianco i valori di media e varianza campionarie ottenuti dai dati:



$$\bar{x} = 187.6154\,\mathrm{ms}$$
$$s^2 = 517.3662\,\mathrm{ms}^2$$

(a) Dopo aver suddiviso i dati in opportune classi, si costruisca la tabella di distribuzione delle frequenze assolute, relative e delle densità.

(b) Si rappresenti la distribuzione dei dati tramite istogramma.

(c) Si determinino la mediana e il primo e il terzo quartile dei dati.

(d) La distribuzione dei dati presenta una coda a destra o a sinistra?

(e) Si può assumere che la variabile aleatoria $X =$ *TTI rilevato in un crash test* abbia densità gaussiana?

Per essere omologato secondo la normativa europea prEN 1621.3/2010, un airbag da moto deve avere un TTI medio minore di 200 ms. Infatti, tempi medi maggiori di 200 ms rischiano di non essere sufficienti a proteggere il motociclista dai traumi di un incidente.

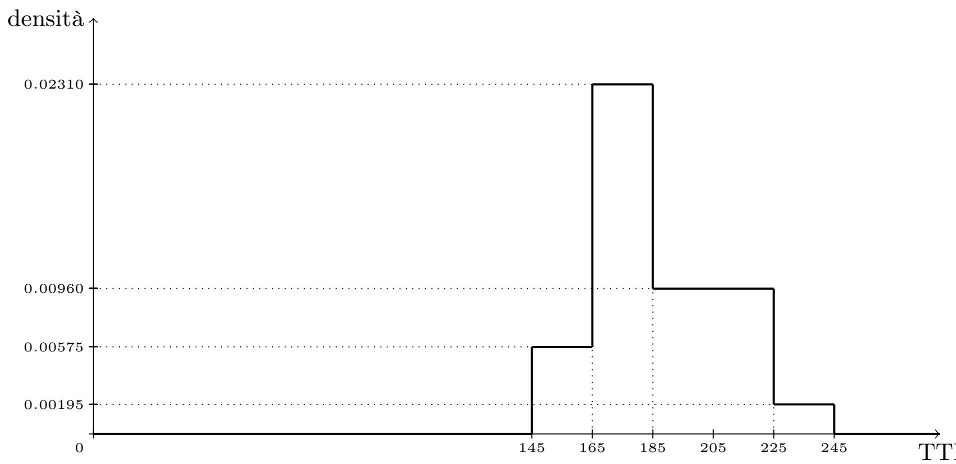
(f) Con un opportuno test d'ipotesi al livello di significatività del 2.5%, stabilite se l'airbag collaudato dall'azienda rispetta la normativa prEN 1621.3/2010. Nel test, l'errore più grave dev'essere (ovviamente!) concludere che l'airbag rispetta la normativa, quando in realtà ha un TTI medio maggiore o uguale a 200 ms.

Risultati.

(a) Il numero di classi si stabilisce in base a una delle due formule n. classi $= \sqrt{\text{n. dati}} = \sqrt{26} = 5.099$ oppure n. classi $= 1 + \log_2(\text{n. dati}) = 1 + \log_2 26 = 1 + \frac{\log_{10} 26}{\log_{10} 2} = 5.700$. Possiamo quindi scegliere n. classi $= 5$. Suddividendo l'intervallo $[145, 245]$ in 5 intervalli equispaziati, troviamo la tabella

classe	freq. assoluta	freq. relativa	densità
$(145, 165]$	3	0.115	0.00575
$(165, 185]$	12	0.462	0.02310
$(185, 205]$	5	0.192	0.00960
$(205, 225]$	5	0.192	0.00960
$(225, 245]$	1	0.039	0.00195

(b) In ascissa vanno i TTI, in ordinata le densità delle rispettive classi:



(c) Usando la formula

$$q_p = \begin{cases} x_{(\lfloor np \rfloor + 1)} & \text{se } np \notin \mathbb{N} \\ \frac{1}{2}\left(x_{(np)} + x_{(np+1)}\right) & \text{se } np \in \mathbb{N} \end{cases}$$

abbiamo

$$\begin{aligned} \text{1o quartile} = Q_1 = q_{0.25} &= x_{(\lfloor 6.5 \rfloor + 1)} = x_{(7)} = 170 \\ \text{mediana} = m = q_{0.50} &= \frac{1}{2}\left(x_{(13)} + x_{(13+1)}\right) = \frac{1}{2}\left(x_{(13)} + x_{(14)}\right) = \frac{1}{2}(183 + 185) = 184 \\ \text{3o quartile} = Q_3 = q_{0.75} &= x_{(\lfloor 19.5 \rfloor + 1)} = x_{(20)} = 204\,. \end{aligned}$$

Dal momento che $\bar{x} = 187.6154 > 184 = m$, la distribuzione dei dati presenta una coda a destra. Ciò è visibile anche dall'istogramma.

(d) I punti del normal Q-Q plot sono abbastanza allineati lungo la `qqline`, e il p-value del test di SW è relativamente alto $\Rightarrow$ non si può rigettare l'ipotesi nulla che la variabile aleatoria X abbia densità gaussiana.

(e) Abbiamo un campione aleatorio $X_1, \ldots, X_{26}$, in cui per il punto precedente possiamo assumere $X_i \sim N(\mu, \sigma^2)$, con μ e σ^2 entrambe incognite. Sappiamo che l'airbag rispetta la normativa europea se $\mu < 200\,\text{ms}$. Se vogliamo che l'errore più grave sia

concludere che l'airbag rispetta la normativa (cioè $\mu < 200\,\text{ms}$),
quando in realtà ha un TTI medio maggiore o uguale a $200\,\text{ms}$ (cioè $\mu \geq 200\,\text{ms}$)

questo deve diventare l'errore di I specie del nostro test, cioè accettare H_1, quando in realtà H_0 è vera.

Confrontando le due frasi, troviamo

$$H_0 : \mu \geq 200\,\mathrm{ms} =: \mu_0 \qquad \text{vs.} \qquad H_1 : \mu < \mu_0\,.$$

Decidiamo dunque tra queste due ipotesi statistiche con un T-test per una popolazione normale a varianza incognita. Al livello α, la regola è

$$\text{“ rifiuto } H_0 \text{ se } T_0 := \frac{\bar{X} - \mu_0}{S}\sqrt{n} < -t_{1-\alpha}(n-1)\text{ ”}\,.$$

Con i nostri dati,

$$t_0 = \frac{187.6154 - 200}{\sqrt{517.3662}}\sqrt{26} = -2.7763\,,$$

che al livello $\alpha = 2.5\%$ va confrontata con

$$-t_{1-\alpha}(n-1) = -t_{0.975}(25) = -2.0595\,.$$

Dal momento che $t_0 = -2.7763 < -2.0595 = -t_{0.975}(25)$, rifiutiamo H_0 e concludiamo che al 2.5% di significatività l'airbag rispetta la normativa europea.

Problema 2. Un indice molto importante per classificare la qualità di un olio d'oliva è il suo livello di acidità, valutato come il contenuto percentuale di acido oleico: tanto minore è tale contenuto, tanto maggiore è la qualità dell'olio.

Per monitorare il livello di acidità dell'olio prodotto, una piccola cooperativa di olivicoltori fa analizzare ogni anno un certo numero di bottiglie della nuova spremitura. Dall'esperienza passata è noto che l'acidità che si misura in ciascuna bottiglia (espressa in ‰ sul contenuto d'olio dell'intera bottiglia) è una variabile aleatoria con densità gaussiana $N(\mu, \sigma^2)$. La deviazione standard σ è dovuta esclusivamente al limite di precisione delle misure in laboratorio: essa *è nota a priori* e pari a $\sigma = 1.5$ ‰, indipendendemente dall'anno del raccolto. La media μ, invece, è la vera acidità incognita dell'olio, ed è una caratteristica che può variare di anno in anno per via delle diverse condizioni meteorologiche.

Nel 2017, analizzando un totale di 9 bottiglie, la media campionaria delle acidità misurate era stata pari a $\bar{x} = 7.48$ ‰. Nel 2018, invece, misurando 12 bottiglie del nuovo raccolto, si è trovata la media campionaria $\bar{y} = 6.91$ ‰.

(a) Dopo il caldo anomalo di quest'estate, la cooperativa spera che l'acidità dell'olio del 2018 sia significativamente diminuita rispetto al 2017. Svolgendo un opportuno test al livello di significatività del 5%, stabilite se i dati dimostrano che sia effettivamente così.

(b) Calcolate il p-value del test del punto (a), e traetene una conclusione.

(c) In effetti, nel 2018 l'acidità incognita è realmente diminuita di 1.00 ‰ rispetto al 2017. Con quale probabilità il test al 5% del punto (a) farebbe concludere erroneamente che invece l'acidità è rimasta la stessa?

(d) Fornite un intervallo di confidenza bilatero al livello del 95% per l'acidità dell'olio prodotto nel 2018.

(e) Se la cooperativa volesse dimezzare l'ampiezza dell'intervallo precedente, quante ulteriori bottiglie dovrebbe far analizzare?

Risultati.

(a) Abbiamo due campioni aleatori indipendenti $X_1, \ldots, X_9$ e $Y_1, \ldots, Y_{12}$, in cui $X_i \sim N(\mu_X, 1.5^2)$ e $Y_j \sim N(\mu_Y, 1.5^2)$, e μ_X, μ_Y sono le acidità effettive dell'olio del 2017 e del 2018, rispettivamente. Dobbiamo fare un test per le ipotesi

$$H_0 : \mu_X = \mu_Y \qquad \text{vs.} \qquad H_1 : \mu_X > \mu_Y$$

(vogliamo vedere se μ_Y è significativamente minore di μ_X $\Rightarrow$ mettiamo $\mu_Y < \mu_X$ nell'ipotesi alternativa). Possiamo fare uno Z-test per la differenza delle medie di due campioni gaussiani indipendenti. La statistica test da usare è

$$Z_0 = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \,,$$

dove $\bar{X}$ e $\bar{Y}$ sono le rispettive medie campionarie. Con le ipotesi H_0 e H_1 precedenti, la regola di rifiuto al livello α è

$$\text{“ rifiuto } H_0 \text{ se } Z_0 > z_{1-\alpha} \text{”} \,.$$

Coi dati a disposizione,

$$z_0 = \frac{7.48 - 6.91}{\sqrt{\frac{1.5^2}{9} + \frac{1.5^2}{12}}} = 0.862 \,,$$

mentre, al livello $\alpha = 5\%$,

$$z_{1-\alpha} = z_{0.95} = 1.645 \,.$$

Dal momento che $z_0 = 0.862 \not> 1.645$, non possiamo rifiutare H_0.

(b) Il p-value è il valore di α che risolve l'equazione

$$\begin{aligned} z_0 \equiv z_{1-\alpha} \quad &\Leftrightarrow \quad \Phi(z_0) \equiv \Phi(z_{1-\alpha}) = 1 - \alpha \\ &\Leftrightarrow \quad \alpha = 1 - \Phi(z_0) = 1 - \Phi(0.862) = 1 - 0.80511 = 0.19489 = 19.489\% \,. \end{aligned}$$

Con un p-value così elevato, a tutti i livelli di significatività sensati non si può rifiutare H_0. Non c'è pertanto alcuna evidenza che l'acidità nell'olio del 2018 sia minore di quella del 2017.

(c) La probabilità richiesta è la probabilità di errore di II specie

$$\begin{aligned}
\mathbb{P}_{\mu_Y=\mu_X-1}(\text{“accetto } H_0\text{”}) &= \mathbb{P}_{\mu_Y=\mu_X-1}\left(\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}} \leq z_{0.95}\right) \\
&= \mathbb{P}_{\mu_X-\mu_Y=1}\left(\underbrace{\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}}}_{\substack{\sim N(0,1)\text{ perché} \\ \bar{X}-\bar{Y}\sim N\left(\mu_X-\mu_Y\,,\,\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}\right)}} \leq z_{0.95} - \frac{\mu_X-\mu_Y}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}}\right) \\
&= \Phi\left(z_{0.95} - \frac{\mu_X-\mu_Y}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}}\right) \qquad \text{con } \mu_X-\mu_Y=1 \\
&= \Phi\left(1.645 - \frac{1}{\sqrt{\frac{1.5^2}{9}+\frac{1.5^2}{12}}}\right) = \Phi(0.133) = 0.55172 \\
&= 55.172\%\,.
\end{aligned}$$

(d) Un intervallo di confidenza bilatero al livello $\gamma = 95\%$ per la media μ_Y del campione gaussiano a varianza nota $Y_1, \ldots, Y_{12}$ è

$$\mu_Y \in \left(\bar{y} - z_{\frac{1+\gamma}{2}}\frac{\sigma}{\sqrt{n_Y}}\,,\,\bar{y} + z_{\frac{1+\gamma}{2}}\frac{\sigma}{\sqrt{n_Y}}\right),$$

dove

$$z_{\frac{1+\gamma}{2}} = z_{\frac{1+0.95}{2}} = z_{0.975} = 1.96\,.$$

Pertanto,

$$\mu_Y \in \left(6.91 - 1.96\,\frac{1.5}{\sqrt{12}}\,,\,6.91 + 1.96\,\frac{1.5}{\sqrt{12}}\right) = (6.0613\,,\,7.7587)\,.$$

(e) L'ampiezza dell'intervallo precedente è

$$2\,z_{\frac{1+\gamma}{2}}\frac{\sigma}{\sqrt{n_Y}}\,.$$

Dunque, se vogliamo dimezzarla, dobbiamo fare in tutto almeno n misure, con n tale che

$$\begin{aligned}
&2\,z_{\frac{1+\gamma}{2}}\frac{\sigma}{\sqrt{n}} = \frac{1}{2}\left(2\,z_{\frac{1+\gamma}{2}}\frac{\sigma}{\sqrt{n_Y}}\right) = z_{\frac{1+\gamma}{2}}\frac{\sigma}{\sqrt{n_Y}} \\
&\Rightarrow \quad n = 4n_Y\,.
\end{aligned}$$

Servono pertanto almeno altre

$$n - n_Y = 3n_Y = 3\cdot 12 = 36$$

misure.

Problema 3. Il famoso esploratore e statistico Jack Bettazzi, recatosi in Groenlandia per condurre la sua ricerca sulle specie locali di foche, dopo mesi di rilievi e osservazioni, riesce a misurare i seguenti caratteri di $n = 215$ esemplari: il numero di `anni` di vita dell'esemplare; il `peso` in Kg dell'esemplare; la `specie` dell'esemplare, dove la specie A è bianca e pelosa mentra la B è grigia e lucida. Decide di impostare un modello linare per spiegare la variabile `peso` tramite l'età e la specie. In Figura sono riportati gli output dei due modelli considerati.

(a) Scrivere la relazione tra le variabili ipotizzata dai due modelli. *[Si associ alla specie A il valore 0 e alla specie B il valore 1.]*

(b) Quali sono le ipotesi sui residui alla base del modello? Sono rispettate dai due modelli?

(c) I singoli regressori nei due modelli sono tutti significativi?

(d) Quale dei due modelli spiega meglio la variabilità della risposta?

(e) Sulla base delle risposte precedenti, scegliere il modello migliore e scrivere il modello stimato per ciascuna delle due specie di foche.

(f) Stimare il peso di un nuovo esemplare di foca di 10 anni e 6 mesi appartenente alla specie A.

Risultati.

(a) Le relazioni ipotizzate dai due modelli sono:

$$\begin{array}{l}\texttt{peso}_i = \beta_0 + \beta_1\texttt{anni}_i + \beta_2\texttt{specie}_i + \beta_3\texttt{specie}_i \cdot \texttt{anni}_i + \varepsilon_i \\ \texttt{peso}_i = \beta_0 + \beta_1\texttt{anni}_i + \beta_2\texttt{specie}_i + \varepsilon_i\end{array} \qquad \varepsilon_1, \ldots, \varepsilon_{215} \ \text{i.i.d.} \ \sim N(0, \sigma^2)\,.$$

(b) I residui devono essere omoschedastici e normali. Dagli scatterplot dei residui di entrambi i modelli vediamo che i residui si dispongono a nuvoletta e non presentano pattern particolali, di conseguenza i residui possono essere considerati omoschedastici. I normal probability plot dei residui di entrambi i modelli ci suggeriscono di accettare la normalità dei residui, dato che i puntini si distribuiscono intorno alla retta. Inoltre il p-value alto dello Shapiro-Wilk test non ci consente di rifiutare l'ipotesi di normalità in entrambi i casi. Concludiamo che i residui possono essere supposti normali.

(c) I regressori del modello ridotto sono tutti significativi, mentre il termine con l'interazione nel modello completo non è singolarmente significativo, visto il p-value del t-test su β_3.

(d) I due modelli hanno lo stesso R^2, però per confrontare due modelli con un numero diverso di regressori è più opportuno confrontare i valori dell'R^2 aggiustato. L'R^2 aggiustato del modello ridotto è leggermente più alto di quello del modello completo, per cui è da preferire il modello ridotto (anche se molto debolmente).

(e) I due modelli sono equivalenti in base alle considerazioni del punto (b), il modello ridotto è leggermente migliore per il punto (d), ma è decisamente da preferirsi in base a quanto detto sulla significatività dei regressori nel punto (c). Dalle stime dei coefficienti otteniamo i seguenti modelli fittati:

Modello per le foche della specie A:	$\texttt{peso}_i = 182.0599 + 7.3668\,\texttt{anni}_i$
Modello per le foche della specie B:	$\texttt{peso}_i = 182.0599 + 98.4484 + 7.3668\,\texttt{anni}_i$

(f) La stima puntuale per il peso di un nuovo esemplare di foca di 10 anni e 6 mesi appartenente alla specie A è:

$$\widehat{\texttt{peso}} = 182.0599 + 7.3668 \cdot 10.5 = 259.4113\,.$$

```
Call:
lm(formula = peso ~ anni + specie + anni:specie)

Residuals:
    Min      1Q  Median      3Q     Max 
-37.071  -6.270   0.269   7.074  28.022 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  182.6787     2.0503  89.096   <2e-16 ***
anni           7.3078     0.1709  42.750   <2e-16 ***
specieB       97.1913     2.8834  33.708   <2e-16 ***
anni:specieB   0.1217     0.2455   0.496    0.621    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.04 on 211 degrees of freedom
Multiple R-squared:  0.9758,	Adjusted R-squared:  0.9755 
F-statistic:  2840 on 3 and 211 DF,  p-value: < 2.2e-16
Call:
lm(formula = peso ~ anni + specie)

Residuals:
    Min      1Q  Median      3Q     Max 
-37.042  -6.161   0.381   6.923  27.815 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 182.0599     1.6236  112.14   <2e-16 ***
anni          7.3668     0.1225   60.14   <2e-16 ***
specieB      98.4484     1.3693   71.89   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.02 on 212 degrees of freedom
Multiple R-squared:  0.9758,	Adjusted R-squared:  0.9756 
F-statistic:  4275 on 2 and 212 DF,  p-value: < 2.2e-16
```

Figura 1: Summary dei modelli di regressione

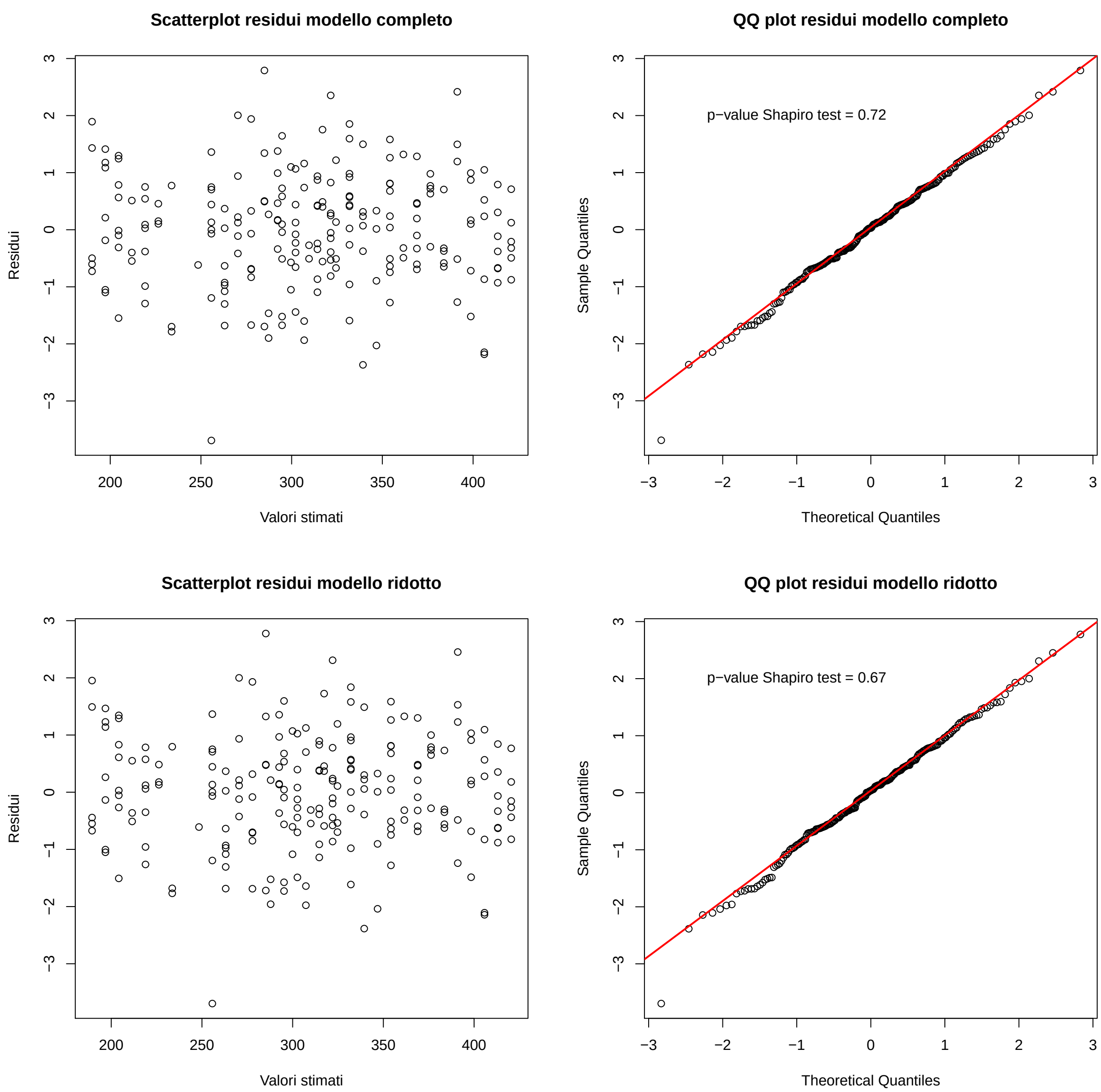


Figura 2: Grafici dei residui