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

II appello di Statistica per Ingegneria Fisica
27 Luglio 2020

Problema 1. Quando Robin tira con l'arco, la minima distanza X (in centimetri) fra la traiettoria della sua freccia e il centro del bersaglio è una variabile aleatoria assolutamente continua con la seguente funzione di ripartizione:

$$F_X(t) = \begin{cases} \alpha & \text{se } t \le 0\,, \\ \dfrac{\beta t}{2\,t + 120} & \text{se } t > 0\,. \end{cases}$$

In questa espressione, α e β sono due parametri reali fissati.

(a) Determinate per quali valori dei parametri α e β la funzione F_X è effettivamente la funzione di ripartizione di una variabile aleatoria assolutamente continua. Giustificate opportunamente la vostra risposta.

D'ora in poi, usate i valori di α e di β trovati al punto precedente.
Il bersaglio è un disco col raggio di 60 cm. Robin realizza 5 punti quando lo colpisce a una distanza dal centro minore di 20 cm, mentre realizza 1 punto soltanto per distanze comprese tra 20 cm e 60 cm.

(b) Calcolate la probabilità che in un tiro Robin realizzi 5 punti e la probabilità che invece manchi il bersaglio.
(Se non ci riuscite, continuate assumendo che entrambe le probabilità siano pari a $1/3$*.)*

(c) Sia Y il punteggio ottenuto da Robin in un tiro. Determinate la funzione di massa di probabilità della variabile aleatoria Y, specificando l'insieme su cui è definita tale funzione.

(d) Calcolate il valore atteso e la varianza di Y.

Il Principe Giovanni ha indetto un torneo di tiro con l'arco, e il premio in palio è un bacio da Lady Marian. Vince chi totalizza più punti facendo 40 tiri complessivi. Robin è l'ultimo a cimentarsi, dopo che lo Sceriffo di Nottingham ha realizzato ben 72 punti e al momento è in testa su tutti.

(e) Qual è la probabilità, eventualmente approssimata, che Robin vinca il torneo?

Risultati.

(a) Per essere la funzione di ripartizione di una variabile aleatoria assolutamente continua, F_X deve innanzitutto soddisfare le due condizioni

$$\lim_{t\to-\infty} F_X(t) = 0\,, \qquad\qquad \lim_{t\to+\infty} F_X(t) = 1\,.$$

Per la funzione data, i due limiti in questione sono rispettivamente

$$\lim_{t\to-\infty} \alpha = \alpha\,, \qquad\qquad \lim_{t\to+\infty} \frac{\beta t}{2\,t + 120} = \frac{\beta}{2}\,.$$

Dunque deve essere $\alpha = 0$ e $\beta = 2$. Per tali valori di α e di β, abbiamo

$$F_X(t) = \begin{cases} 0 & \text{se } t \le 0\,, \\ \dfrac{t}{t + 60} & \text{se } t > 0\,, \end{cases}$$

che è una funzione nondecrescente e derivabile ovunque tranne in 0. Dunque, si tratta della funzione di ripartizione di una variabile aleatoria assolutamente continua.

(b) Con la funzione di ripartizione trovata in precedenza, le due probabilità richieste sono

$$\mathbb{P}(X \le 20) = F_X(20) = \frac{20}{20+60} = \frac{1}{4},$$
$$\mathbb{P}(X > 60) = 1 - F_X(60) = 1 - \frac{60}{60+60} = \frac{1}{2}.$$

(c) La funzione di massa di probabilità di Y è definita sull'insieme dei valori possibili per Y, e cioè su $S = \{0, 1, 5\}$. Indicandola con p_Y, essa è data da

$$p_Y(0) = \mathbb{P}(X > 60) = \frac{1}{2},$$
$$p_Y(5) = \mathbb{P}(X \le 20) = \frac{1}{4},$$
$$p_Y(1) = 1 - p_Y(0) - p_Y(5) = \frac{1}{4}.$$

(Coi valori nel suggerimento: $p_Y(0) = p_Y(1) = p_Y(5) = 1/3$.)

(d) Abbiamo

$$\mathbb{E}[Y] = \sum_{k\in\{0,1,5\}} k\, p_Y(k) = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} + 5 \cdot \frac{1}{4} = \frac{3}{2},$$
$$\mathbb{E}\left[Y^2\right] = \sum_{k\in\{0,1,5\}} k^2\, p_Y(k) = 0^2 \cdot \frac{1}{2} + 1^2 \cdot \frac{1}{4} + 5^2 \cdot \frac{1}{4} = \frac{13}{2},$$
$$\mathrm{Var}(Y) = \mathbb{E}\left[Y^2\right] - \mathbb{E}[Y]^2 = \frac{13}{2} - \left(\frac{3}{2}\right)^2 = \frac{17}{4}.$$

(Coi valori nel suggerimento: $\mathbb{E}[Y] = 2$, $\mathrm{Var}(Y) = 14/3$.)

(e) Sia $S_{40} = Y_1 + Y_2 + \ldots + Y_{40}$ il punteggio che otterrà Robin al termine dei suoi 40 tiri, in cui Y_i è il punteggio ottenuto nell'i-esimo tiro. La probabilità richiesta è

$$\mathbb{P}(S_{40} > 72) = \mathbb{P}(S_{40} \ge 72.5) = \mathbb{P}\left(\underbrace{\frac{S_{40} - n\,\mathbb{E}[Y]}{\sqrt{n\,\mathrm{Var}(Y)}}}_{\approx N(0,1)} \ge \frac{72.5 - 40 \cdot \frac{3}{2}}{\sqrt{40 \cdot \frac{17}{4}}} \right)$$
$$\simeq 1 - \Phi\left(\frac{72.5 - 40 \cdot \frac{3}{2}}{\sqrt{40 \cdot \frac{17}{4}}} \right) = 1 - \Phi(0.9587) = 1 - 0.83147 = 16.853\%.$$

Senza la correzione di continuità, invece,

$$\mathbb{P}(S_{40} > 72) \simeq 1 - \Phi\left(\frac{72 - 40 \cdot \frac{3}{2}}{\sqrt{40 \cdot \frac{17}{4}}} \right) = 1 - \Phi(0.9204) = 1 - 0.82121 = 17.879\%.$$

(Coi valori nel suggerimento: $\mathbb{P}(S_{40} > 72) \simeq 0.70884$ e $\mathbb{P}(S_{40} > 72) \simeq (0.71904 + 0.72240)/2 = 0.72072$, rispettivamente.)

Problema 2. Gemma ama Gianni. Secondo lei però Gianni non la ama abbastanza, perché quando chattano insieme al telefono i messaggi di lui sono sempre troppo corti.
Tina – che ha molta più esperienza di Gemma in queste cose – le ha rivelato che, quando un ragazzo chatta al telefono, la lunghezza di un suo messaggio qualunque è una variabile aleatoria X avente *deviazione standard nota* e pari a 23.5 caratteri. Inoltre, messaggi diversi dello stesso ragazzo hanno lunghezze tutte indipendenti fra loro e identicamente distribuite. È risaputo, però, che solo quando un ragazzo è veramente innamorato i suoi messaggi sono lunghi in media più di 120 caratteri.
Grazie ai consigli di Tina, Gemma ha dunque deciso che sposerà Gianni solo se i suoi prossimi 80 messaggi conterranno *in tutto* almeno 9700 caratteri, e che in caso contrario lo lascerà. E ovviamente, per Gemma l'errore molto più grave è sposare un ragazzo che non la ama veramente.

(a) Qual è la probabilità *massima* che, in base al test da lei impostato, Gemma sposi Gianni senza che egli la ami veramente?

(b) Se in realtà Gianni scrivesse a Gemma messaggi mediamente lunghi 126 caratteri, quale probabilità avrebbe di essere (ingiustamente!) lasciato da lei alla fine del test?

Ora il test si è concluso, e, analizzando gli 80 messaggi ricevuti, Gemma ha contato in tutto 9772 caratteri. Di conseguenza, Gemma sposerà Gianni.

(c) Siete d'accordo col modo in cui Gemma ha impostato il test, arrivando a questa conclusione? Perché?

(d) Impostate voi un test al livello di significatività α in cui, coerentemente con quanto detto sopra, l'errore più grave consista nel ritenere che Gianni ama Gemma quando in realtà ciò non è vero.

(e) In base ai dati di Gemma, qual è il p-value del vostro test? È compatibile col fatto che Gianni la ami veramente?

(f) Quanto siete confidenti che Gianni scriva a Gemma messaggi mediamente più lunghi di 117 caratteri?

Risultati.

(a) Sia X_i la lunghezza (in caratteri) dell'i-esimo messaggio di Gianni, con $i = 1, \dots, 80$. Da quanto dice Tina, sappiamo che $X_1, \dots, X_{80}$ è un campione aleatorio con $\sigma := \sqrt{\mathrm{Var}(X_i)} = 23.5$. Inoltre, Gianni è veramente innamorato solo se $\mu := \mathbb{E}[X_i] > 120$.
Gemma ha deciso che sposerà Gianni solo se il numero totale di caratteri negli 80 messaggi sarà pari ad almeno 9700, cioè solo se $N_{80} := X_1 + \ldots + X_{80} \geq 9700$. Per il TLC, ciò accadrà con probabilità

$$\begin{aligned}\mathbb{P}_\mu(N_{80} \geq 9700) &= \mathbb{P}_\mu\left(\underbrace{\frac{N_{80} - \mathbb{E}[N_{80}]}{\sqrt{\mathrm{Var}(N_{80})}}}_{\approx N(0,1)} \geq \frac{9700 - n\mu}{\sigma\sqrt{n}}\right) \simeq 1 - \Phi\left(\frac{9700 - n\mu}{\sigma\sqrt{n}}\right) \\ &= 1 - \Phi\left(\frac{9700 - 80\mu}{23.5\sqrt{80}}\right)\end{aligned}$$

in quanto $\mathbb{E}[N_{80}] = n\mu$ e $\mathrm{Var}(N_{80}) = n\sigma^2$. L'ultima funzione è crescente in μ. Pertanto, se Gianni non è veramente innamorato di Gianna, la probabilità precedente è massima quando $\mu = 120$, cioè

$$\begin{aligned}\max_{\mu \leq 120} \mathbb{P}_\mu(N_{80} \geq 9700) &\simeq 1 - \Phi\left(\frac{9700 - 80 \cdot 120}{23.5\sqrt{80}}\right) = 1 - \Phi(0.4758) = 1 - \frac{0.68082 + 0.68439}{2} \\ &= 31.7395\%\,.\end{aligned}$$

(b) Se i messaggi mandati da Gianni sono mediamente lunghi $\mu = 126$ caratteri, allora la probabilità che Gemma lo lasci (ingiustamente!) come conseguenza del test è pari a

$$\begin{aligned}\mathbb{P}_{\mu=126}(N_{80} < 9700) &= 1 - \mathbb{P}_{\mu=126}(N_{80} \geq 9700) = 1 - \left[1 - \Phi\left(\frac{9700 - 80 \cdot 126}{23.5\sqrt{80}}\right)\right] \\ &= \Phi(-1.8079) = 1 - \Phi(1.8079) = 1 - 0.96485 \\ &= 3.515\%\,.\end{aligned}$$

(c) Non possiamo essere d'accordo col modo in cui Gemma ha impostato il test, perché abbiamo visto nel punto (a) che la probabilità dell'errore per lei più grave (sposare Gianni quando in realtà lui non la ama) è troppo elevata ($31.7395\% \gg 5\%$).

(d) L'errore più grave dev'essere l'errore di I specie del nostro test, cioè rifiutare H_0 quando in realtà H_0 è vera. Se scegliamo

$$H_0 : \mu \leq 120 \qquad \text{vs.} \qquad H_1 : \mu > 120$$

allora tale errore è proprio ritenere che Gianni ama Gemma (= rifiutare H_0) quando in realtà egli non è innamorato (= H_0 è vera).
Possiamo dunque decidere tra H_0 e H_1 impostando uno Z-test per un campione numeroso a varianza nota:

$$\text{“ rifiuto } H_0 \text{ se } Z_0 := \frac{\bar{X}_{80} - 120}{\sigma}\sqrt{n} > z_{1-\alpha} \text{ ”} .$$

(e) Per determinare il p-value, calcoliamo la realizzazione della statistica test coi dati di Gemma:

$$z_0 = \frac{\bar{x}_{80} - 120}{\sigma}\sqrt{n} = \frac{\frac{9772}{80} - 120}{23.5}\sqrt{80} = 0.8183$$

(dove abbiamo usato il fatto che $\bar{x}_{80} = n_{80}/80$) e imponiamo l'equazione

$$\begin{aligned} z_0 \equiv z_{1-p\text{-value}} \quad &\Leftrightarrow \quad \Phi(z_0) = 1 - p\text{-value} \\ &\Leftrightarrow \quad p\text{-value} = 1 - \Phi(0.8183) = 1 - 0.79389 = 20.611\% . \end{aligned}$$

Con un p-value così elevato, NON possiamo concludere che Gianni sia veramente innamorato.

(f) Un $IC_\mu(\gamma)$ unilatero sinistro per un campione numeroso a varianza nota è

$$\left(\bar{x}_n - z_\gamma \frac{\sigma}{\sqrt{n}}\, , +\infty\right) .$$

In altre parole,

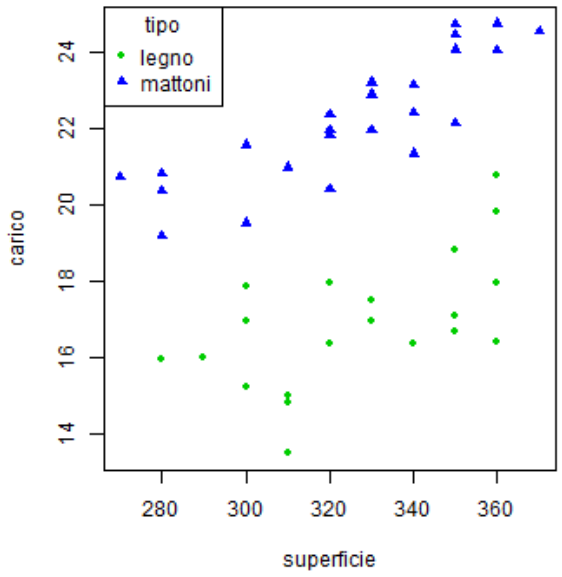
$$\mu > \bar{x}_n - z_\gamma \frac{\sigma}{\sqrt{n}} \qquad \text{al livello di confidenza } \gamma .$$

Imponiamo

$$\begin{aligned} \bar{x}_n - z_\gamma \frac{\sigma}{\sqrt{n}} \equiv 117 \quad &\Leftrightarrow \quad z_\gamma = \frac{\bar{x}_n - 117}{\sigma}\sqrt{n} = \frac{\frac{9772}{80} - 117}{23.5}\sqrt{80} = 1.9601 \\ &\Leftrightarrow \quad \gamma = \Phi(1.9601) = 0.97500 . \end{aligned}$$

Pertanto, coi dati di Gianna siamo confidenti al 97.5% che $\mu > 117$.

Problema 3. L'Ingegner Cane sta studiando l'efficienza energetica di due diversi tipi di case di montagna: quelle costruite in legno e quelle in mattoni. Per il suo studio, ha fatto misurare il carico termico di 45 di tali edifici (variabile `carico`, in W/m^2). Di ogni edificio, egli conosce inoltre la misura della superficie muraria (variabile `superficie`, in m^2) e il tipo di materiale con cui è stato costruito (variabile categorica `tipo`, coi due livelli `legno` e `mattoni`). I tre vettori coi dati dell'Ingener Cane sono stati raggruppati nel *data frame* `df` e salvati nell'area di lavoro di R che trovate allegata. *(È un file `.RData`. Potete caricarlo selezionando `File` → `Carica area di lavoro...` dal menù di R.)* Gli stessi dati sono anche rappresentati nella figura seguente:



Per descrivere la relazione fra `carico`, `superficie` e `tipo`, l'Ingegner Cane propone il modello di regressione lineare di cui trovate l'output nella pagina seguente. Sotto sono riportati anche i grafici di diagnostica dei residui standardizzati.

(a) Per il *modello proposto dall'Ingegner Cane*, scrivete la relazione ipotizzata tra l'efficienza energetica e i regressori utilizzati. Questo modello si adatta bene ai dati raccolti? Perché?

(b) Scrivete un *nuovo modello* di regressione lineare gaussiano con risposta il `carico` termico e regressori la `superficie` muraria, il `tipo` di edificio e l'interazione tra `superficie` e `tipo`.

(c) In termini di variabilità spegata, il nuovo modello è migliore di quello dell'Ingegner Cane? Perché?

(d) Il nuovo modello rispetta le ipotesi gaussiane? Perché?

(e) Il nuovo modello è globalmente significativo? Perché?

(f) Nel nuovo modello, tutti i regressori sono significativi? Perché?

(g) Proponete voi un *terzo modello* di regressione lineare, e confrontatelo coi due precedenti sottoponendolo alle stesse verifiche dei punti (c), (d), (e) e (f). Qual è il migliore fra tutti i tre modelli?

(h) C'è evidenza dai dati che, a parità di superficie muraria, il tipo di materiale influenzi il carico termico medio di un edificio? Perché?

(i) Fornite una previsione puntuale per il carico termico di un edificio costruito in mattoni la cui superficie muraria misuri $300\,m^2$.

```
Call:
lm(formula = df$carico ~ df$superficie)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5925 -2.4188  0.8894  2.8469  3.9550

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.14081    5.46681   1.123   0.2675
df$superficie  0.04181    0.01673   2.500   0.0163 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.959 on 43 degrees of freedom
Multiple R-squared:  0.1269,    Adjusted R-squared:  0.1066
F-statistic: 6.249 on 1 and 43 DF,  p-value: 0.01632
```

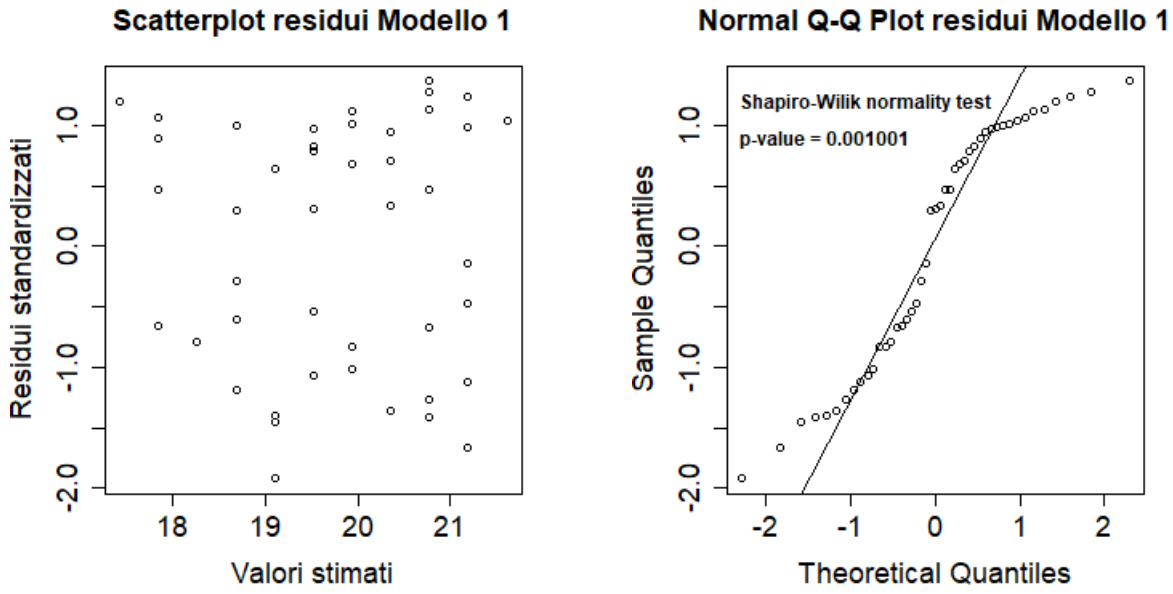


Figura 1: Summary e diagnostica dei residui standardizzati per il modello dell'Ingegner Cane

Risultati.

(a) La relazione ipotizzata dall'Ingegner Cane è

$$\textbf{Modello 1:} \qquad \texttt{carico} = \beta_0 + \beta_1\, \texttt{superficie} + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

Tale modello non si adatta affatto ai dati raccolti, in quanto nell'output di R vediamo che la variabilità spiegata è estremamente bassa ($r^2 = 12.69\% \ll 81\%$) e i residui, benché siano omoschedastici, non si possono considerare gaussiani, dal momento che nel normal Q-Q plot i quantili teorici ed empirici non seguono un andamento lineare e il p-value dello Shapiro-test è molto basso ($0.1001\% \ll 5\text{-}10\%$).

(b) Il nuovo modello è

$$\textbf{Modello 2:} \qquad \texttt{carico} = \beta_0 + \beta_1\, \texttt{superficie} + \beta_2\, \texttt{tipo} + \beta_3\, \texttt{superficie}\cdot\texttt{tipo} + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)\,.$$

Il corrispondente output di R si ottiene con

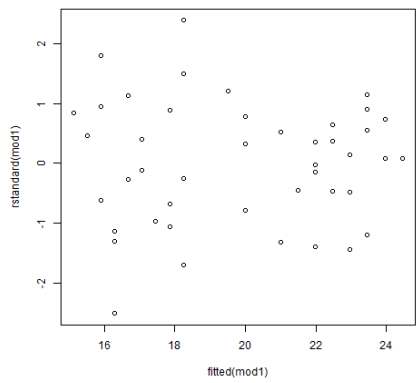
```
> mod2 <- lm( df$carico ~ df$superficie + df$tipo + df$superficie:df$tipo )
> summary(mod2)

Call:
lm(formula = df$carico ~ df$superficie + df$tipo + df$superficie:df$tipo)

Residuals:
     Min       1Q   Median       3Q      Max
-2.78696 -0.74940  0.09126  0.80468  2.56249

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    4.26559    3.32275   1.284 0.206437
df$superficie                  0.03881    0.01014   3.827 0.000436 ***
df$tipomattoni                 1.92647    4.32367   0.446 0.658257
df$superficie:df$tipomattoni   0.01053    0.01322   0.797 0.430306
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.15 on 41 degrees of freedom
Multiple R-squared:  0.8742,    Adjusted R-squared:  0.8649
F-statistic: 94.93 on 3 and 41 DF,  p-value: < 2.2e-16
```

(c) In termini di variabilità spegata, il nuovo modello è di gran lunga migliore di quello dell'Ingegner Cane. Infatti, il suo r^2_{adj} è pari all'86.49%, che si confronta col misero 10.66% dell'Ingegnere.

(d) Per decidere se il nuovo modello verifica le ipotesi gaussiane, dobbiamo vedere se i suoi residui standardizzati risultano omoschedastici e gaussiani. Per l'omoschedasticità, dobbiamo guardare lo scatterplot dei residui, che otteniamo col comando

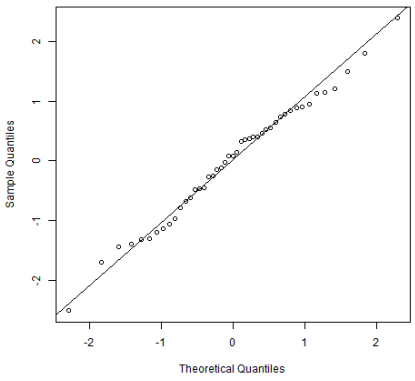
```
> plot(fitted(mod2), rstandard(mod2))
```

Ecco il risultato:



da cui si vede che i residui sono omoschedastici. Per quanto riguarda invece la gaussianità, dobbiamo anzitutto osservare il normal Q-Q plot dei residui standardizzati, ottenuto coi comandi

```
> qqnorm(rstandard(mod2))
> qqline(rstandard(mod2))
```



Vediamo che i punti si allineano bene sulla Q-Q line. Per confermare questo fatto, svolgiamo il test di Shapiro Wilks sui residui

```
> shapiro.test(rstandard(mod2))

        Shapiro-Wilk normality test

data:  rstandard(mod2)
W = 0.98977, p-value = 0.9578
```

Il p-value del 95.78% è molto elevato, dunque non abbiamo nessun motivo per rifiutare l'ipotesi nulla di gaussianità dei residui. Concludiamo che il nuovo modello soddisfa le ipotesi gaussiane.

(e) Il nuovo modello è globalmente significativo, in quanto vediamo dall'output di R che il p-value dell'F-test è estremamente basso (p-value $< 2.2 \cdot 10^{-16}$).

(f) Nel nuovo modello, il regressore `tipo` e il termine di interazione tra `superficie` e `tipo` non sono significativi, in quanto i p-value dei rispettivi T-test sono molto elevati (nell'ordine, 65.8257% e 43.0306%). È invece significativo il regressore `superficie` (p-value = 0.0436%).

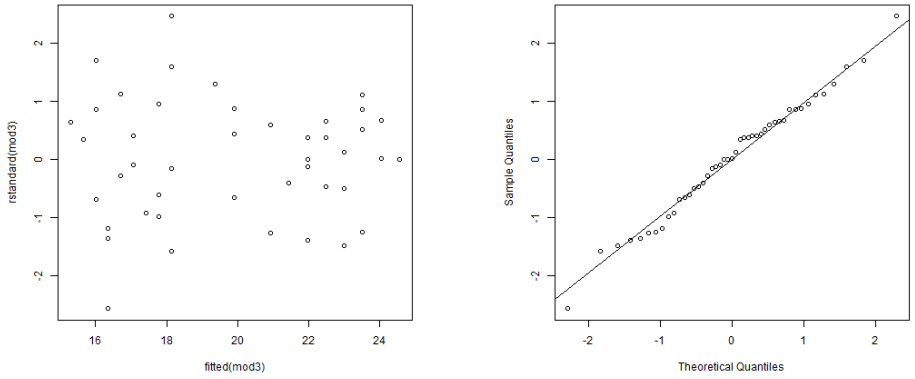
(g) Col metodo dell'eliminazione a ritroso, eliminiamo il regressore meno significativo dal modello precedente, cioè `tipo`. Otteniamo il terzo modello

Modello 3: $\texttt{carico} = \beta_0 + \beta_1\,\texttt{superficie} + \beta_2\,\texttt{superficie}\cdot\texttt{tipo} + \epsilon, \qquad \epsilon \sim N(0,\sigma^2)\,.$

Per quest'ultimo modello, ripetiamo tutte le analisi che abbiamo fatto col precedente, ricavandone il summary e la diagnostica dei residui seguenti:

```
Call:
lm(formula = df$carico ~ df$superficie + df$superficie:df$tipo)

Residuals:
     Min       1Q   Median       3Q      Max
-2.85117 -0.71382  0.00604  0.72604  2.67143

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   5.403355   2.105668   2.566   0.0139 *
df$superficie                 0.035348   0.006454   5.477 2.24e-06 ***
df$superficie:df$tipomattoni 0.016404   0.001042  15.747  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 1.139 on 42 degrees of freedom
Multiple R-squared:  0.8735,    Adjusted R-squared:  0.8675
F-statistic: 145.1 on 2 and 42 DF,  p-value: < 2.2e-16
```



Il test di Shapiro-Wilk sui residui standardizzati restituisce un p-value del 93.42%. Vediamo quindi che il Modello 3 è praticamente equivalente al Modello 2 in termini di variabilità spiegata ($r^2_{\text{adj}} = 86.75\%$), rispetta le ipotesi gaussiane (residui omoschedastici, ben allineati lungo la Q-Q line e con elevato p-value del test di Shapiro-Wilk) ed è globalmente significativo. In più, però, ora tutti i regressori sono significativi, con p-value dei rispettivi T-test entrambi molto bassi. Ne concludiamo che il Modello 3 è migliore del Modello 2 e dunque di tutti i tre modelli.

(h) Vediamo che nel modello migliore (Modello 3) il p-value del termine di interazione tra le variabili `superficie` e `tipo` è molto basso (p-value del T-test su $\beta_2 < 2 \cdot 10^{-16}$). Dunque c'è fortissima evidenza che, a parità di `superficie`, il `tipo` di materiale influenzi il `carico` termico medio di un edificio.

(i) Usando il Modello 3, la previsione puntuale per il carico termico di un edificio costruito in mattoni con misura di superficie muraria pari a $300\,\text{m}^2$ è

$$\begin{aligned}\widehat{\texttt{carico}} &= \hat\beta_0 + \hat\beta_1\,\texttt{superficie} + \hat\beta_2\,\texttt{superficie}\cdot\texttt{tipo}\\ &= 5.40335 + 0.035348\cdot 300 + 0.016404\cdot 300\cdot 1\\ &= 20.92895\end{aligned}$$

in quanto vediamo dall'output di R che il programma ha assegnato valore 1 al livello `mattoni`.