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

III appello di Statistica per Ingegneria Fisica
11 Settembre 2020

Problema 1. Ogni giorno sul Monte Piovoso si abbatte un numero di temporali distribuito con densità poissoniana di parametro λ. Si sa inoltre che i temporali in giorni diversi sono tutti indipendenti fra loro. Infine, sul monte c'è una piccola stazione meteorologica che ogni domenica pubblica su Internet quanti temporali si sono verificati *in totale* nell'ultima settimana ($= 7$ giorni).
Siano X_1, X_2 e X_3 i valori che la stazione pubblicherà nelle prossime tre settimane.

(a) Calcolate in funzione di λ il valore atteso e la varianza di una qualsiasi delle variabili aleatorie X_i. Qual è la densità di X_i?

Aldo, Bruno e Carlo organizzano spesso gite in montagna insieme. Per la loro prossima escursione, i tre amici hanno in programma di trascorrere un giorno intero ($= 24$ ore) sul Monte Piovoso. Prima di partire, però, vorrebbero stimare la probabilità che in quel giorno si verifichi qualche temporale.
Anzitutto, i tre amici decidono di stimare il parametro λ, proponendo ognuno uno stimatore diverso:

$$T_{\text{Aldo}} = \frac{X_1 + X_2 + X_3}{3}\,, \qquad T_{\text{Bruno}} = \frac{X_1 + X_2 + X_3}{21}\,, \qquad T_{\text{Carlo}} = \frac{X_1 + X_2 - X_3}{7}\,.$$

(b) Calcolate la distorsione di ciascuno dei tre stimatori proposti per il parametro λ.

(c) Calcolate in funzione di λ l'errore quadratico medio dei tre stimatori.

(d) Chi ha proposto lo stimatore migliore? E chi invece quello peggiore?

(e) Calcolate in funzione di λ la probabilità p che si verifichi qualche temporale nel giorno della gita. Costruite inoltre uno stimatore per il parametro p utilizzando solo le variabili X_1, X_2 e X_3.

(f) Per lo stimatore di p costruito nel punto precedente, calcolate in modo approssimato la distorsione e l'errore quadratico medio.

Dopo che sono trascorse le tre settimane, gli amici vanno a leggere su Internet i valori di X_1, X_2 e X_3 pubblicati dalla stazione meterologica. Questi sono i dati trovati:

$$x_1 = 14\,, \qquad x_2 = 5\,, \qquad x_3 = 9\,.$$

(g) Fornite una stima per ciascuno dei parametri λ e p, e una stima dei rispettivi errori quadratici medi.

Risultati.

(a) Se $Z_{i,j}$ è il numero di temporali che si verificano nel j-esimo giorno dell'i-esima settimana, allora sappiamo dal testo che $Z_{i,j} \sim \mathcal{P}(\lambda)$ e che le v.a. $Z_{1,1}, \dots, Z_{3,7}$ sono tutte indipendenti tra loro. Poiché

$$X_i = \sum_{j=1}^{7} Z_{i,j}\,,$$

dalla proprietà di riproducibilità della densità poissoniana segue che

$$X_i \sim \mathcal{P}(7\lambda)\,.$$

In particolare,

$$\mathbb{E}[X_i] = \text{Var}(X_i) = 7\lambda\,.$$

(b) Calcoliamo i valori attesi

$$\mathbb{E}\left[T_{\text{Aldo}}\right] = \mathbb{E}\left[\frac{X_1+X_2+X_3}{3}\right] = \frac{1}{3}\left(\mathbb{E}\left[X_1\right]+\mathbb{E}\left[X_2\right]+\mathbb{E}\left[X_3\right]\right) = \frac{1}{3}\left(7\lambda+7\lambda+7\lambda\right) = 7\lambda\,,$$
$$\mathbb{E}\left[T_{\text{Bruno}}\right] = \mathbb{E}\left[\frac{X_1+X_2+X_3}{21}\right] = \frac{1}{21}\left(\mathbb{E}\left[X_1\right]+\mathbb{E}\left[X_2\right]+\mathbb{E}\left[X_3\right]\right) = \frac{1}{21}\left(7\lambda+7\lambda+7\lambda\right) = \lambda\,,$$
$$\mathbb{E}\left[T_{\text{Carlo}}\right] = \mathbb{E}\left[\frac{X_1+X_2-X_3}{7}\right] = \frac{1}{7}\left(\mathbb{E}\left[X_1\right]+\mathbb{E}\left[X_2\right]-\mathbb{E}\left[X_3\right]\right) = \frac{1}{7}\left(7\lambda+7\lambda-7\lambda\right) = \lambda\,.$$

Ne ricaviamo le distorsioni

$$\text{bias}(T_k;\lambda) = \mathbb{E}\left[T_k\right]-\lambda = \begin{cases} 7\lambda-\lambda = 6\lambda & \text{per } k = \text{Aldo} \\ \lambda-\lambda = 0 & \text{per } k = \text{Bruno} \\ \lambda-\lambda = 0 & \text{per } k = \text{Carlo} \end{cases}.$$

(c) Usiamo la formula

$$\text{mse}(T_k;\lambda) = \text{Var}\left(T_k\right) + \text{bias}(T_k;\lambda)^2$$

e calcoliamo dunque le varianze

$$\begin{aligned}
\text{Var}\left(T_{\text{Aldo}}\right) &= \text{Var}\left(\frac{X_1+X_2+X_3}{3}\right) = \left(\frac{1}{3}\right)^2\left(\text{Var}\left(X_1\right)+\text{Var}\left(X_2\right)+\text{Var}\left(X_3\right)\right) \\
&= \frac{1}{3^2}\left(7\lambda+7\lambda+7\lambda\right) = \frac{7}{3}\lambda\,, \\
\text{Var}\left(T_{\text{Bruno}}\right) &= \text{Var}\left(\frac{X_1+X_2+X_3}{21}\right) = \left(\frac{1}{21}\right)^2\left(\text{Var}\left(X_1\right)+\text{Var}\left(X_2\right)+\text{Var}\left(X_3\right)\right) \\
&= \frac{1}{21^2}\left(7\lambda+7\lambda+7\lambda\right) = \frac{1}{21}\lambda\,, \\
\text{Var}\left(T_{\text{Carlo}}\right) &= \text{Var}\left(\frac{X_1+X_2-X_3}{7}\right) = \left(\frac{1}{7}\right)^2\left(\text{Var}\left(X_1\right)+\text{Var}\left(X_2\right)+(-1)^2\text{Var}\left(X_3\right)\right) \\
&= \frac{1}{7^2}\left(7\lambda+7\lambda+7\lambda\right) = \frac{3}{7}\lambda\,.
\end{aligned}$$

Si ottiene

$$\text{mse}(T_k;\lambda) = \begin{cases} (7/3)\,\lambda + 36\,\lambda^2 & \text{per } k = \text{Aldo} \\ (1/21)\,\lambda & \text{per } k = \text{Bruno} \\ (3/7)\,\lambda & \text{per } k = \text{Carlo} \end{cases}.$$

(d) T_{Bruno} e T_{Carlo} sono stimatori non distorti per il parametro λ, mentre T_{Aldo} è distorto. Inoltre,

$$\text{mse}(T_{\text{Bruno}};\lambda) < \text{mse}(T_{\text{Carlo}};\lambda) < \text{mse}(T_{\text{Aldo}};\lambda)\,.$$

Dunque, lo stimatore migliore è quello proposto da Bruno, che è non distorto e ha l'mse minore. Il peggiore, invece, è quello di Aldo, che è distorto e ha l'mse più grande.

(e) Se Z è il numero di temporali che si verificheranno nel giorno della gita, sappiamo che $Z \sim \mathcal{P}(\lambda)$, e

$$p = \mathbb{P}\left(Z>0\right) = 1-\mathbb{P}\left(Z=0\right) = 1-p_Z(0) = 1-\mathrm{e}^{-\lambda}\,.$$

Essendo T_{Bruno} uno stimatore non distorto di λ, per il metodo delta uno stimatore approssimativamente non distorto di p è dato da

$$\hat{P} = 1-\mathrm{e}^{-T_{\text{Bruno}}}\,.$$

(f) Abbiamo già visto al punto precedente che $\hat{P}$ è uno stimatore approssimativamente non distorto, in quanto per il metodo delta

$$\mathbb{E}\left[\hat{P}\right] = \mathbb{E}\left[1-\mathrm{e}^{-T_{\text{Bruno}}}\right] \overset{\delta}{\simeq} 1-\mathrm{e}^{-\mathbb{E}[T_{\text{Bruno}}]} = 1-\mathrm{e}^{-\lambda} = p \qquad \Rightarrow \qquad \text{bias}(\hat{P};p) \simeq 0\,.$$

Allo stesso modo, usiamo il metodo delta per calcolare l'MSE:

$$\mathrm{mse}(\hat{P};p) \;=\; \mathrm{Var}(\hat{P}) + \mathrm{bias}(\hat{P};p)^2 \;\simeq\; \mathrm{Var}(\hat{P}) \;\overset{\delta}{\simeq}\; \left[\left.\frac{\mathrm{d}(1-\mathrm{e}^{-t})}{\mathrm{d}t}\right|_{t=\mathbb{E}[T_{\mathrm{Bruno}}]}\right]^2 \mathrm{Var}\left(T_{\mathrm{Bruno}}\right) \;=\; (\mathrm{e}^{-t})^2 \frac{1}{21}\,\lambda$$
$$= \mathrm{e}^{-2\lambda}\,\frac{1}{21}\,\lambda\,.$$

(g) Stima puntuale di λ:

$$t_{\mathrm{Bruno}} = \frac{1}{21}(x_1+x_2+x_3) = \frac{1}{21}(14+5+9) = \frac{4}{3} = 1.33333\,.$$

Stima puntuale di p:

$$\hat{p} = 1-\mathrm{e}^{-t_{\mathrm{Bruno}}} = 1-\mathrm{e}^{-\frac{4}{3}} = 1-0.26360 = 0.7364 = 73.64\%\,.$$

Stimatori approssimativamente non distorti dei rispettivi mse:

$$\widehat{\mathrm{MSE}}(T_{\mathrm{Bruno}};\lambda) = \frac{1}{21}\,T_{\mathrm{Bruno}}\,,$$
$$\widehat{\mathrm{MSE}}(\hat{P};p) = \mathrm{e}^{-2T_{\mathrm{Bruno}}}\,\frac{1}{21}\,T_{\mathrm{Bruno}}\,.$$

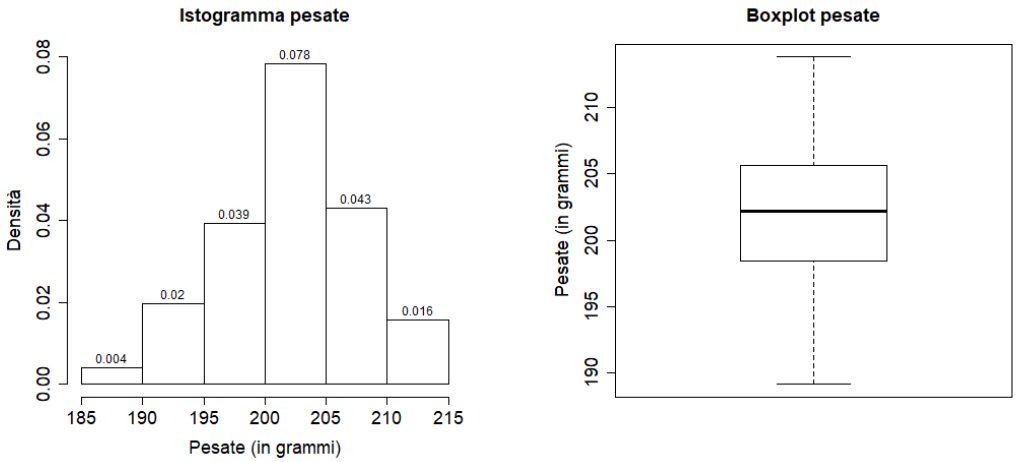
Stime corrispondenti:

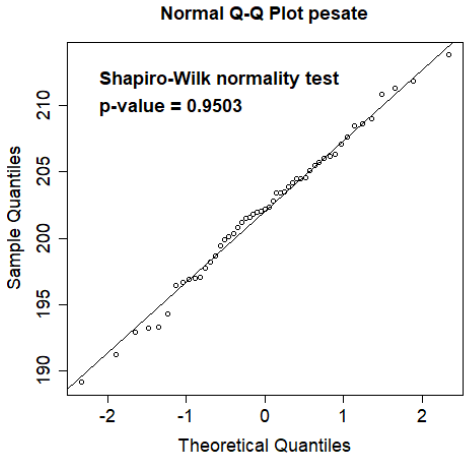
$$\widehat{\mathrm{mse}}(T_{\mathrm{Bruno}};\lambda) = \frac{1}{21}\,t_{\mathrm{Bruno}} = \frac{1}{21}\cdot\frac{4}{3} = \frac{4}{63} = 0.06349\,,$$
$$\widehat{\mathrm{mse}}(\hat{P};p) = \mathrm{e}^{-2t_{\mathrm{Bruno}}}\frac{1}{21}\,t_{\mathrm{Bruno}} = \mathrm{e}^{-2\cdot\frac{4}{3}}\cdot\frac{1}{21}\cdot\frac{4}{3} = 0.00441\,.$$

Problema 2. Nando è il macellaio sotto casa da cui vado sempre a fare la spesa. Negli ultimi tempi, però, mi è venuto il sospetto che, quando compro qualcosa da lui, egli mi pesi sistematicamente di più di quanto gli chiedo. D'altra parte, Nando, prima di fare il macellaio aveva studiato Statistica all'università. Perciò, quando mi sono permesso di esporgli i miei dubbi, lui offeso mi ha risposto che non mi devo basare sul risultato delle singole pesate, ma piuttosto su quello che è il loro valore atteso. Secondo Nando, infatti, ogni sua pesata è una variabile aleatoria centrata sul valore che gli chiedo io, con una varianza a lui nota e pari a $25\,\mathrm{g}^2$.
Non essendo convinto dalla spiegazione di Nando, ho deciso di metterlo alla prova, e così, nelle ultime 51 volte in cui ho fatto la spesa da lui, gli ho sempre chiesto di pesarmi 200 g precisi di mortadella. Queste sono la media e la varianza campionarie dei 51 dati così ottenuti:

$$\overline{x}_{51} = 202.0853\,\mathrm{g}\,, \qquad\qquad s^2_{51} = 29.7508\,\mathrm{g}^2\,.$$

Ecco anche il loro istogramma (sopra ogni barra è riportata la densità corrispondente), il boxplot e il normal Q-Q plot con indicato il p-value del test di Shapiro-Wilk:





(a) Con un opportuno test al livello di significatività del 10%, stabilite se i dati sono compatibili con la varianza dichiarata da Nando. C'è sufficiente evidenza per affermare che sia diversa?

(b) Impostate un altro test al livello di significatività α per stabilire se i dati confermano i miei sospetti. Ovviamente, accusare Nando quando in realtà dice il vero dev'essere l'errore più grave di questo test.

(c) Qual è il p-value del test del punto (b)? Quale conclusione ne traete?

(d) Quali condizioni vanno rispettate per poter svolgere i test dei punti (a) e (b)? I dati le soddisfano?

(e) Domani tornerò da Nando, e come al solito gli chiederò di pesarmi 200 g precisi di mortadella. Voglio stimare la probabilità che invece lui me ne pesi di più. Fornite una stima puntuale e un intervallo di confidenza bilatero al livello del 95% per tale probabilità.

Risultati.

(a) Sia σ^2 la varianza di una qualsiasi pesata di Nando. Si richiede di fare un test bilatero per le ipotesi

$$H_0 : \sigma^2 = \sigma_0^2 \qquad \text{contro} \qquad H_1 : \sigma^2 \neq \sigma_0^2$$

con $\sigma_0^2 := 25$ come dichiara Nando. Per poterlo fare, dobbiamo assumere che le pesate di Nando siano gaussiane, perché solo in questo caso possiamo usare la seguente regola al livello di significatività α:

$$\text{“rifiuto } H_0 \text{ se } X_0^2 < \chi^2_{\frac{\alpha}{2}}(n-1) \text{ oppure } X_0^2 > \chi^2_{1-\frac{\alpha}{2}}(n-1)\text{”}\,,$$

dove

$$X_0^2 := \frac{(n-1)S^2}{\sigma_0^2}\,.$$

Dai dati ricaviamo la realizzazione

$$x_0^2 = \frac{(51-1)\cdot 29.7508}{25} = 59.5016\,,$$

mentre al livello $\alpha = 10\%$

$$\chi^2_{\frac{\alpha}{2}}(n-1) = \chi^2_{0.05}(50) = 34.7642\,, \qquad \chi^2_{1-\frac{\alpha}{2}}(n-1) = \chi^2_{0.95}(50) = 67.5048\,.$$

Poiché

$$59.5016 \in (34.7642\,,\, 67.5048)\,,$$

la regola del test ci impone di non rifiutare H_0 al livello del 10%. Dal momento che il 10% è un livello di significatività alto, non abbiamo nessuna evidenza che $\sigma^2 \neq 25\,\mathrm{g}^2$ (conclusione debole).

(b) Sia μ il valore atteso di una qualsiasi pesata di Nando. Ora si richiede di fare un test per le ipotesi

$$H_0 : \mu = \mu_0 \qquad \text{contro} \qquad H_1 : \mu > \mu_0$$

dove $\mu_0 = 200$. Sempre assumendo che le pesate di Nando siano gaussiane, possiamo usare uno Z-test per la media di un campione gaussiano a varianza nota e pari a $\sigma^2 = 25\,\mathrm{g}^2$ (vedi conclusione del test precedente). La regola al livello α è dunque la seguente:

$$\text{“rifiuto } H_0 \text{ se } Z_0 > z_{1-\alpha}\text{”}\,, \qquad (*)$$

dove

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n}\,.$$

(c) Dai dati ricaviamo

$$z_0 = \frac{202.0853 - 200}{5}\sqrt{51} = 2.9784\,.$$

Dunque, imponendo l'uguaglianza nella regola $(*)$, troviamo

$$z_0 = z_{1-\alpha} \quad \Leftrightarrow \quad \Phi(z_0) = 1-\alpha \quad \Leftrightarrow \quad \alpha = 1 - \Phi(z_0) = 1 - \Phi(2.9784) = 1 - 0.99856 = 0.00144\,,$$

cioè p-value $= 0.144\%$. Con un p-value così piccolo, siamo costretti ad accettare l'ipotesi alternativa $\mu > 200$ a tutti i livelli di significatività sensati (conclusione forte).

(d) L'ipotesi di entrambi i test precedenti è che le 51 misure costituiscano un campione gaussiano. Solo il test sulla media sarebbe valido (ma in modo approssimato) anche se il campione non fosse gaussiano; per il test sulla varianza, invece, la condizione di normalità è indispensabile. Tale ipotesi, tuttavia, è soddisfatta in virtù dell'elevato valore del p-value del test di Shapiro-Wilk (addirittura maggiore del 90%) e della buona aderenza dei quantili empirici alla Q-Q line nel normal Q-Q plot.

(e) Si tratta di trovare una stima puntuale e un IC bilatero per il parametro p di un campione bernoulliano numeroso $Y_1, \dots, Y_{51}$, dove

$$Y_i = \begin{cases} 1 & \text{se nell'}i\text{-esimo giorno Nando pesa più di } 200\,\text{g}\,, \\ 0 & \text{altrimenti}\,. \end{cases}$$

Una stima puntuale è data dalla frequenza di successi empirica

$$\begin{aligned} \overline{y}_{51} = \mathrm{FR}((200, +\infty)) &= \sum_{\text{classi} > 200} \text{ampiezza classe} \cdot \text{densità classe} \\ &= 5 \cdot 0.078 + 5 \cdot 0.043 + 5 \cdot 0.016 = 0.685 = 68.5\%\,. \end{aligned}$$

Un IC bilatero al livello $\gamma = 95\%$ è invece dato da

$$p \in \left(\overline{y}_{51} \pm z_{\frac{1+\gamma}{2}} \sqrt{\frac{\overline{y}_{51}(1-\overline{y}_{51})}{n}} \right) = \left(0.685 \pm 1.96 \sqrt{\frac{0.685(1-0.685)}{51}} \right) = (55.751\%\,,\, 81.249\%)$$

dove

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

Problema 3. Benedetta sta imparando a preparare il pane fatto in casa. Per ottenere il risultato migliore, sta effettuando alcuni esperimenti per valutare l'effetto del tempo di attesa sulla lievitazione dell'impasto. Nelle diverse prove, ha determinato il rapporto tra il volume dell'impasto lievitato e quello iniziale (variabile `lievitazione`, numero puro), misurando questo rapporto al variare del tempo (variabile `tempo`, espressa in ore). I due vettori coi dati ottenuti sono stati raggruppati nel *data frame* `dati` 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.)*
Per descrivere la relazione fra il `tempo` (in input) e la `lievitazione` (in output), Benedetta è indecisa su quale scegliere tra i due seguenti modelli lineari gaussiani:

$$\texttt{lievitazione} = \beta_0 + \beta_1\,\texttt{tempo} + \epsilon \qquad \epsilon \sim N(0, \sigma^2)\,,$$
$$\texttt{lievitazione} = \beta_0 + \beta_1 \log(\texttt{tempo}) + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

(nel secondo modello, log è il logaritmo naturale in base $e = 2.718\ldots$).

(a) Disegnate lo scatterplot dei dati di Benedetta.

(b) Indicate la percentuale di variabilità spiegata da ciascuno dei due modelli proposti.

(c) Nei due modelli, i dati rispettano le ipotesi gaussiane? Perché?

(d) I due modelli sono globalmente significativi? Perché?

(e) Stabilite quale dei due modelli è preferibile, giustificando la risposta.

(f) Scrivete la relazione stimata fra `tempo` e `lievitazione` per il modello scelto al punto precedente.

(g) Secondo quanto c'è scritto sulla confezione del lievito di Benedetta, dopo 1 ora esatta il volume medio dell'impasto dovrebbe essere pari ad almeno 1.52 volte quello iniziale. In base ai suoi dati, Benedetta ha elementi sufficienti per contestare questa affermazione? Impostate un test opportuno, calcolate un intervallo in cui cade il suo p-value e traetene una conclusione.

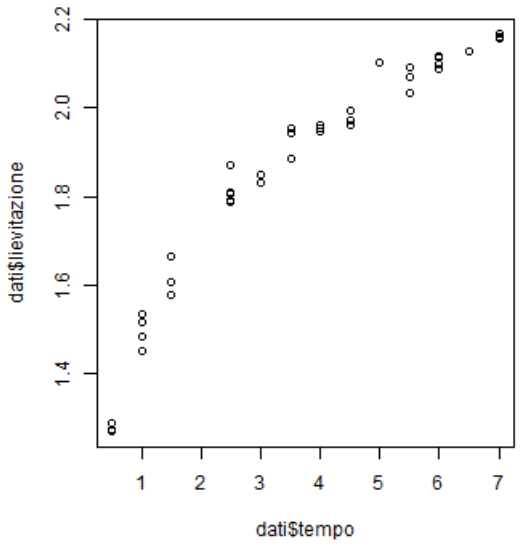
(h) Benedetta ha appena finito di preparare $10\,\text{cm}^3$ d'impasto, e ora lo lascerà lievitare per ben 8 ore. Fornite un intervallo di previsione bilatero al livello del 99% per il volume dell'impasto (in cm^3) quando sarà trascorso tale tempo.

Risultati.

(a) Col comando

```
> plot(dati\$tempo, dati\$lievitazione)
```

otteniamo il grafico



(b) Inserendo i comandi

```
> fit1 <- lm(lievitazione ~ tempo, data = dati)
> fit2 <- lm(lievitazione ~ log(tempo), data = dati)
> summary(fit1)
> summary(fit2)
```

otteniamo gli output dei due modelli:

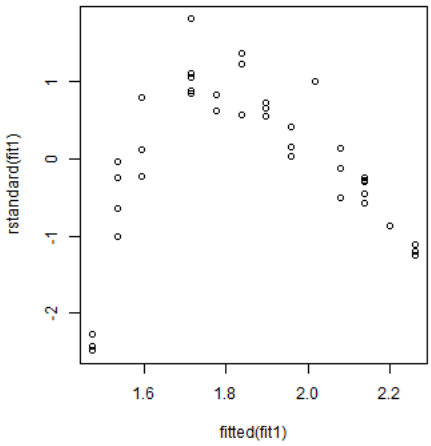
```
Call:
lm(formula = lievitazione ~ tempo)

Residuals:
      Min        1Q    Median        3Q       Max
-0.207028 -0.044265  0.006649  0.068470  0.155514

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.41466    0.02835   49.90   <2e-16 ***
tempo        0.12073    0.00679   17.78   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08711 on 38 degrees of freedom
Multiple R-squared:  0.8927,    Adjusted R-squared:  0.8899
F-statistic: 316.2 on 1 and 38 DF,  p-value: < 2.2e-16
```

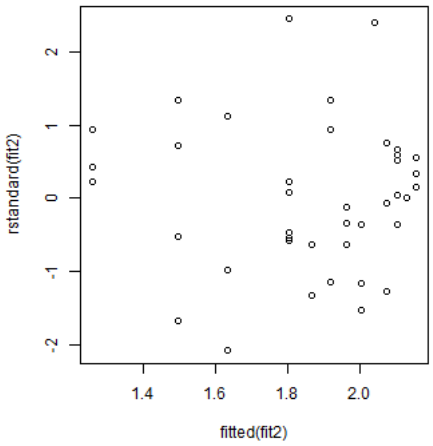
```
Call:
lm(formula = lievitazione ~ log(tempo))

Residuals:
      Min        1Q    Median        3Q       Max
-0.056540 -0.016105  0.000635  0.016627  0.067256

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.495853   0.007575   197.5   <2e-16 ***
log(tempo)  0.337110   0.005773    58.4   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02792 on 38 degrees of freedom
Multiple R-squared:  0.989,     Adjusted R-squared:  0.9887
F-statistic:  3410 on 1 and 38 DF,  p-value: < 2.2e-16
```

Trattandosi di due modelli di regressione semplice, la percentuale di variabilità spiegata è data direttamente dall'r^2, che negli output è pari all'89.27% per il primo modello e al 98.9% per il secondo.

(c) Le ipotesi gaussiane sono verificate se i residui standardizzati risultano omoschedastici e gaussiani. Dagli scatterplot dei residui, ottenuti tramite i comandi

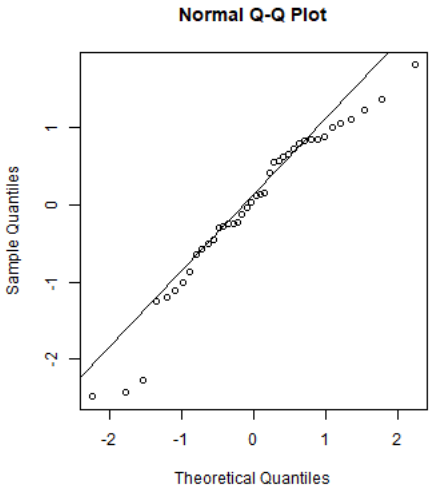
```
> plot(fitted(fit1), rstandard(fit1))
> plot(fitted(fit2), rstandard(fit2))
```

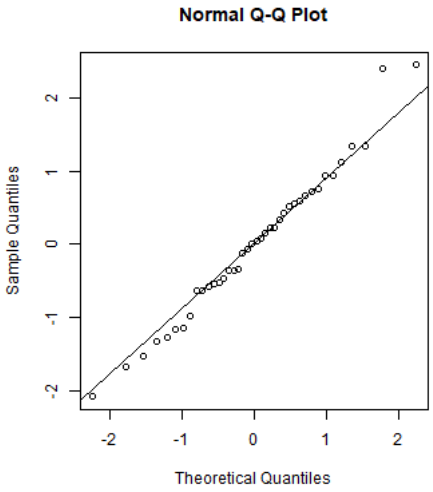




vediamo che solo nel secondo modello è rispettata l'omoschedasticità, mentre nel primo modello si riscontra un chiaro pattern parabolico. Per quanto riguarda invece la gaussianità, dobbiamo anzitutto osservare i normal Q-Q plot dei residui, ottenuti coi comandi

```
> qqnorm(rstandard(fit1)); qqline(rstandard(fit1))
> qqnorm(rstandard(fit2)); qqline(rstandard(fit2))
```





Per gli stessi residui, il test di Shapiro Wilks dà inoltre i seguenti p-value:

```
> shapiro.test(rstandard(fit1))

        Shapiro-Wilk normality test

data:  rstandard(fit1)
W = 0.94762, p-value = 0.0628

> shapiro.test(rstandard(fit2))

        Shapiro-Wilk normality test

data:  rstandard(fit2)
W = 0.98245, p-value = 0.7791
```

Vediamo che per il primo modello l'ipotesi di normalità è dubbia, dato che nel normal Q-Q plot i

quantili teorici e quelli empirici si discostano moderatamente dall'andamento lineare e il p-value dello Shapiro-test è vicino al 5%. Non c'è nessun dubbio, invece, sul fatto che l'ipotesi di normalità sia verificata dal secondo modello, dato che nel normal Q-Q plot i quantili teorici ed empirici seguono l'andamento lineare e il p-value dello Shapiro-test è molto alto (77.91%). In conclusione, le ipotesi gaussiane non sono verificate dal primo modello, mentre lo sono dal secondo.

(d) Entrambi i modelli sono globalmente significativi. Infatti, gli F-test sulla significatività globale della regressione (che coincidono con i test sulla significatività dell'unico regressore, in quanto modelli di regressione lineare semplice) hanno p-value molto bassi, di valore inferiore a $2.2 \cdot 10^{-16}$ (`p-value: < 2.2e-16`).

(e) È preferibile il secondo modello, in quanto è il solo che soddisfa le ipotesi di normalità e di omoschedasticità dei residui. Inoltre, è globalmente significativo e spiega una percentuale di variabilità elevata.

(f) Relazione stimata:

$$\widehat{\texttt{lievitazione}} \;=\; \widehat{\beta}_0 + \widehat{\beta}_1 \log(\texttt{tempo}) \;=\; 1.495853 + 0.337110 \log(\texttt{tempo})\,.$$

(g) Il rapporto tra il volume medio dell'impasto dopo $x^* = 1$ ora e il volume iniziale è

$$\mathbb{E}\left[\texttt{lievitazione} \mid x = x^*\right] = \beta_0 + \beta_1 \log(x^*) = \beta_0 + \beta_1 \log(1) = \beta_0\,.$$

Vogliamo verificare se vi è evidenza dai dati che questo valore sia minore di 1.52, cioè $\beta_0 < 1.52 \;\Rightarrow$ mettiamo quest'ultima affermazione nell'ipotesi alternativa del test:

$$H_0 : \beta_0 \geq 1.52 \qquad \text{contro} \qquad H_1 : \beta_0 < 1.52\,.$$

Un test per tali ipotesi al livello α ha la seguente regola:

$$\text{“ rifiuto } H_0 \text{ se } t_0 := \frac{\hat{\beta}_0 - 1.52}{\mathrm{se}(\hat{\beta}_0)} < -t_{1-\alpha}(n-2) \text{”}\,.$$

Calcoliamo la realizzazione della statistica test: $t_0 := \frac{1.495853 - 1.52}{0.007575} = -3.1877$. Per calcolare il p-value, uguagliamo quest'ultimo valore al quantile nella regione di rifiuto:

$$-3.1877 = -t_{1-\alpha}(38) \iff 3.1877 = t_{1-\alpha}(38) \simeq t_{1-\alpha}(40)\,,$$

in cui abbiamo trovato il numero di dati $n = 40$ usando il comando

```
> length(dati$tempo)
```

Sulle tavole troviamo che

$$t_{0.995}(40) = 2.7045 < 3.1877 < 3.3069 = t_{0.999}(40) \quad \Leftrightarrow \quad 0.995 < 1 - \alpha < 0.999$$
$$\Leftrightarrow \quad 0.001 < \alpha < 0.005\,,$$

cioè $0.1\% < p\text{-value} < 0.5\%$. Con un p-value così piccolo, possiamo rifiutare H_0 a tutti i livelli di significatività sensati (conclusione forte).

(h) Usando il modello migliore, troviamo innanzitutto un intervallo di previsione al 99% per il rapporto di lievitazione dopo che sono trascorse $x^{**} = 8$ ore

```
> tempo_nuovo <- data.frame(tempo = 8)
> predict(fit2, newdata = tempo_nuovo, interval = "prediction", level = 0.99)
       fit      lwr      upr
1 2.196854 2.118592 2.275115
```

Vediamo che R restituisce l'intervallo $(2.118592\,,\, 2.275115)$. Se il volume iniziale è pari a $10\,\mathrm{cm}^3$, il volume finale sarà dunque compreso nell'intervallo

$$(2.118592\,,\, 2.275115) \cdot 10\,\mathrm{cm}^3 = \left(21.18592\,\mathrm{cm}^3\,,\, 22.75115\,\mathrm{cm}^3\right)$$