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

IV APPELLO DI STATISTICA PER INGEGNERIA FISICA
16 Gennaio 2021

Problema 1. Orin Scrivello è un dentista di fama internazionale specializzato nella cura dei denti canini. A ogni paziente che visita, Orin Scrivello esegue l'otturazione di ciascun dente canino cariato. Sia X la variabile aleatoria che rappresenta il numero di denti canini cariati di un paziente. X ha la seguente funzione di massa di probabilità:

$$p_X(x) := \begin{cases} \frac{1}{2}a^2 & \text{se } x = 0 \\ \frac{1}{4} & \text{se } x = 1 \\ \frac{1}{2}a & \text{se } x = 2 \\ \frac{1}{2}a & \text{se } x = 3 \\ \frac{1}{2}a^2 & \text{se } x = 4 \end{cases}$$

dove $a \in \mathbb{R}$ è un parametro.

(a) Per quali valori del parametro a la funzione p_X è una funzione di massa di probabilità?

D'ora in poi, fissate a in modo che p_X sia una funzione di massa di probabilità.
[Se non siete riusciti a risolvere il punto precedente, supponete che la probabilità di avere 2 *denti canini cariati sia nulla e che le altre possibilità (*0, 1, 3 *o* 4 *denti canini cariati) siano equiprobabili.]*

(b) Calcolate $\mathbb{E}[X]$ e $\mathrm{Var}[X]$.

(c) Calcolate la probabilità che un paziente abbia almeno un dente canino cariato.

Orin Scrivello è molto veloce e per ogni otturazione impiega esattamente 5 minuti. Sia Y la variabile aleatoria che rappresenta il tempo che Orin Scrivello impiega a curare un paziente.

(d) Calcolate $\mathbb{E}[Y]$ e $\mathrm{Var}[Y]$.

Audrey, la segretaria di Orin Scrivello, ha fissato 42 appuntamenti per domani. Orin Scrivello è un appassionato di botanica e vuole finire di lavorare il prima possibile per tornare a casa a curare la sua rara pianta carnivora. Decide quindi che domani inizierà a lavorare alle ore 9:00 e non farà nessuna pausa finché non avrà curato tutti i pazienti a cui Audrey ha fissato un appuntamento.

(e) Calcolate la probabilità (eventualmente approssimata) che Orin Scrivello riesca a curare tutti i pazienti entro le ore 15:00.

Risultati.

(a) Affinché p_X sia una funzione di massa di probabilità si devono imporre le due condizioni:

$$\sum_{x=0}^{4} p_X(x) = 1 \qquad \text{e} \qquad p_X(x) \geq 0 \qquad \text{per} \qquad x \in \{0,1,2,3,4\}\,.$$

Imponendo la prima condizione si ottiene l'equazione di secondo grado

$$a^2 + a - \frac{3}{4} = 0$$

le cui due soluzioni sono $a_1 = -3/2$ e $a_2 = 1/2$. Imponendo la seconda condizione si ottiene $a \geq 0$, che porta a escludere la soluzione a_1. Si conclude che la funzione p_X è una funzione di massa di probabilità per $a = 1/2$.

(b) Col valore di a trovato al punto (a), abbiamo

$$p_X(0) = \frac{1}{8} \qquad p_X(1) = \frac{1}{4} \qquad p_X(2) = \frac{1}{4} \qquad p_X(3) = \frac{1}{4} \qquad p_X(4) = \frac{1}{8}$$

e dunque

$$\begin{aligned}
\mathbb{E}\left[X\right] &= \sum_{x=0}^{4} x\, p_X(x) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{4} + 4 \cdot \frac{1}{8} = 2\,,\\
\mathbb{E}\left[X^2\right] &= \sum_{x=0}^{4} x^2\, p_X(x) = 0^2 \cdot \frac{1}{8} + 1^2 \cdot \frac{1}{4} + 2^2 \cdot \frac{1}{4} + 3^2 \cdot \frac{1}{4} + 4^2 \cdot \frac{1}{8} = \frac{11}{2}\,,\\
\mathrm{Var}\left[X\right] &= \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2 = \frac{11}{2} - 2^2 = \frac{3}{2}\,.
\end{aligned}$$

Se invece usiamo la densità data nel suggerimento, abbiamo

$$p_X(0) = \frac{1}{4} \qquad p_X(1) = \frac{1}{4} \qquad p_X(2) = 0 \qquad p_X(3) = \frac{1}{4} \qquad p_X(4) = \frac{1}{4}$$

e con calcoli simili ai precedenti si trova

$$\mathbb{E}\left[X\right] = 2\,, \qquad \mathbb{E}\left[X^2\right] = \frac{13}{2}\,, \qquad \mathrm{Var}\left[X\right] = \frac{5}{2}\,.$$

(c) $\mathbb{P}\left(X \geq 1\right) = 1 - \mathbb{P}\left(X = 0\right) = 1 - p_X(0) = 1 - 1/8 = 7/8$ (con la p_X del suggerimento, invece, $\mathbb{P}\left(X \geq 1\right) = 3/4$).

(d)

$$\begin{aligned}
\mathbb{E}\left[Y\right] &= \mathbb{E}\left[5X\right] = 5\mathbb{E}\left[X\right] = 5 \cdot 2 = 10.0\,,\\
\mathrm{Var}\left[Y\right] &= \mathrm{Var}\left[5X\right] = 5^2\mathrm{Var}\left[X\right] = 25 \cdot \frac{3}{2} = \frac{75}{2} = 37.5\,.
\end{aligned}$$

(Con la p_X del suggerimento: $\mathbb{E}\left[Y\right] = 10.0$, $\mathrm{Var}\left[Y\right] = 125/2 = 62.5$.)

(e) Il tempo impiegato da Orin Scrivello per curare 42 pazienti è la variabile aleatoria

$$S_{42} = \sum_{i=1}^{42} Y_i \,,$$

dove Y_i è il tempo per curare l'i-esimo paziente, e le variabili $Y_1, \ldots, Y_{42}$ sono i.i.d. con la stessa media e la stessa varianza del punto precedente. Dalle 9:00 alle 15:00 trascorrono in tutto 360 minuti, dunque la probabilità richiesta è

$$\mathbb{P}\left(S_{42} \leq 360\right) \,.$$

Poiché il numero di pazienti è alto ($42 > 30$), si può usare il TCL e approssimare la distribuzione di S_{42} nel modo seguente:

$$S_{42} \approx N\left(42 \cdot 10,\ 42 \cdot 37.5\right) = N\left(420,\ 1575\right) \,.$$

Quindi

$$\begin{aligned}\mathbb{P}\left(S_{42} \leq 360\right) &\simeq \mathbb{P}\left(Z \leq \frac{360 - 420}{\sqrt{1575}}\right) = \mathbb{P}\left(Z \leq -1.512\right) = \Phi(-1.512) = 1 - \Phi(1.512) = 1 - 0.93448 \\ &= 6.552\% \,,\end{aligned}$$

dove $Z = (S_{42} - \mathbb{E}\left[S_{42}\right]) / \sqrt{\mathrm{Var}\left[S_{42}\right]} \approx N(0, 1)$. Con la correzione di continuità, tenendo conto che $S_{42}/5$ è ancora una v.a. discreta,

$$\begin{aligned}\mathbb{P}\left(S_{42} \leq 360\right) = \mathbb{P}\left(S_{42} \leq 362.5\right) &\simeq \mathbb{P}\left(Z \leq \frac{362.5 - 420}{\sqrt{1575}}\right) = \Phi(-1.4489) = 1 - 0.92647 = 0.07353 \\ &= 7.353\% \,.\end{aligned}$$

Se invece avessimo usato la p_X del suggerimento:

$$\mathbb{P}\left(S_{42} \leq 360\right) = \begin{cases} \Phi(-1.171) \simeq 1 - 0.87900 = 12.100\% & \text{senza la correzione}\,, \\ \Phi(-1.122) \simeq 1 - 0.88877 = 11.123\% & \text{con la correzione}\,. \end{cases}$$

Problema 2. La ACME è una grande multinazionale del tabacco, produttrice delle famose sigarette Zampiron. Per poter commercializzare la nuova versione Zampiron Extra-Strong, la ACME deve dimostrare alle autorità sanitarie che il contenuto medio di nicotina delle nuove sigarette è inferiore al limite massimo fissato per legge a 1 mg/sigaretta. Viene dunque analizzato in laboratorio un campione di 100 sigarette. Le concentrazioni di nicotina rilevate (in mg/sigaretta) sono poi raccolte nel vettore `ExtraStrong` ed elaborate con R come riportato nella pagina seguente (vedi Figura 1).

(a) Impostate un test al livello di significatività α per decidere se c'è evidenza dai dati che le Zampiron Extra-Strong rispettino i limiti di legge. Scrivete le ipotesi nulla e alternativa e la regola di rifiuto del test al livello α. Ovviamente, l'errore più grave consiste nel ritenere che le sigarette sono conformi alla legge quando in realtà esse la violano.

(b) Nel test precedente, avete dovuto fare delle ipotesi sulla densità del campione? Se sì, quali e perché?

(c) Determinate il p-value del test del punto (a). In base al valore trovato, qual è la vostra conclusione? Si tratta di una conclusione debole o forte?

La ACME vende da tempo anche una versione Light delle Zampiron. Si sono fatte le stesse analisi pure su questa versione, raccogliendo i risultati nel vettore `Light`. I nuovi dati elaborati con R si trovano nella Figura 2 della pagina seguente.

(d) Con un test al livello del 5%, stabilite se c'è evidenza dai dati che le Zampiron Light abbiano un contenuto medio di nicotina effettivamente minore delle Extra-Strong.

(e) Nel test precedente, avete dovuto fare delle ipotesi sulla densità del campione? Se sì, quali e perché?

Non potete usare R per lo svolgimento di questi due esercizi. Potete usare le tavole delle distribuzioni (`Tavole distribuzioni.pdf`), il formulario (`Formulario.pdf`) e la calcolatrice. Nella form, dovete caricare solo la scansione del manoscritto con la vostra soluzione.

```
> ExtraStrong
  [1] 0.26 0.24 1.39 1.31 0.30 0.25 0.27 0.24 0.24 0.74
 [11] 0.36 0.54 0.26 1.03 0.95 0.32 0.32 0.27 0.24 0.39
 [21] 0.53 0.27 0.32 1.12 0.97 0.31 0.75 0.24 0.27 0.24
 [31] 0.70 1.36 0.88 0.89 0.37 0.26 0.42 0.41 0.36 0.29
 [41] 1.13 0.28 0.58 0.45 0.24 0.42 0.24 1.43 0.52 1.40
 [51] 0.24 1.33 0.49 1.50 0.98 1.15 1.16 1.11 0.80 0.35
 [61] 1.20 1.12 1.28 1.67 1.72 1.27 1.68 1.25 1.71 1.40
 [71] 1.56 1.72 1.47 1.29 1.34 1.43 1.68 1.03 1.67 1.44
 [81] 1.74 1.53 1.58 1.67 1.72 1.59 1.62 1.41 1.40 1.62
 [91] 1.47 1.60 1.43 1.52 1.07 1.22 0.74 1.61 1.58 0.63
> length(ExtraStrong)
[1] 100
> mean(ExtraStrong)
[1] 0.9436
> sd(ExtraStrong)
[1] 0.5410045
> shapiro.test(ExtraStrong)

        Shapiro-Wilk normality test

data:  ExtraStrong
W = 0.87676, p-value = 1.327e-07

> qqnorm(ExtraStrong)
> qqline(ExtraStrong)
```

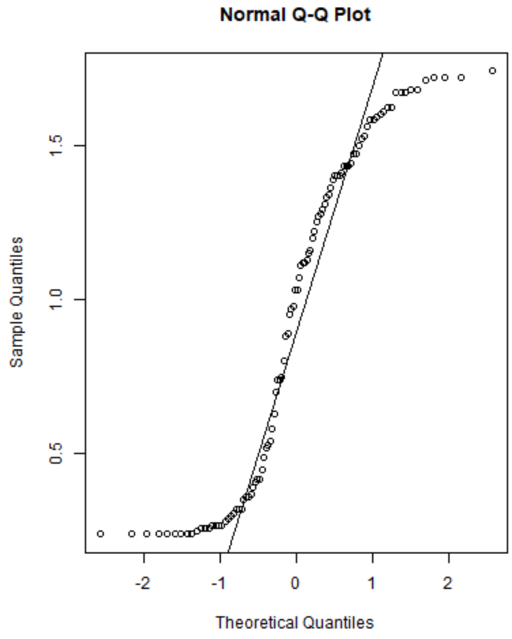


Figura 1: Console di R e relativo output per le sigarette ExtraStrong

---

```
> Light
 [1] 0.95 0.38 0.65 0.26 1.20 1.19 0.85 0.37 0.87 0.84
[11] 0.25 0.61 1.22 0.20 0.86 0.75 0.77 0.46 1.52 0.21
[21] 0.65 0.96 0.33 0.67 0.68 1.33 0.91 0.43 0.74 0.70
[31] 0.76 0.64 0.83 1.15 0.80 1.20 1.03 0.71 0.50 0.66
[41] 0.40 0.76 0.84 0.96 0.61 0.55 1.16 0.27 0.11 0.89
[51] 1.46 0.99 0.52 1.23 0.28 0.41 1.08 1.00 0.91 0.89
> length(Light)
[1] 60
> mean(Light)
[1] 0.7568333
> sd(Light)
[1] 0.3312661
> shapiro.test(Light)

        Shapiro-Wilk normality test

data:  Light
W = 0.9842, p-value = 0.6283

> qqnorm(Light)
> qqline(Light)
```

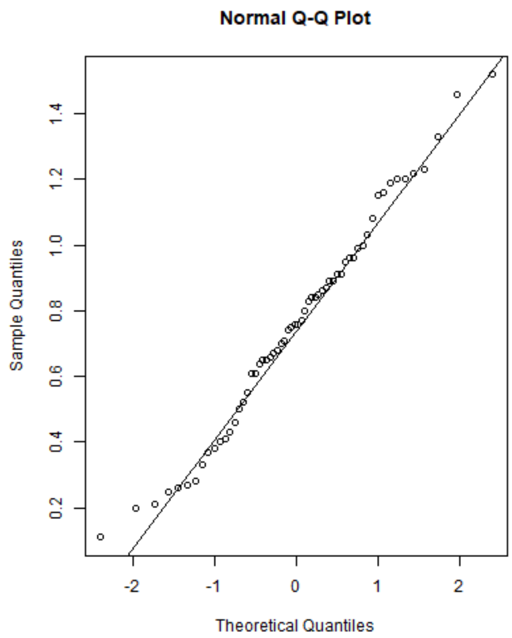


Figura 2: Console di R e relativo output per le sigarette Light

Risultati.

Il vettore `ExtraStrong` contiene una realizzazione da un campione aleatorio $X_1, \dots, X_n$ con $n = 100$. La densità delle X_i è incognita e sicuramente non gaussiana, come si vede dalla Figura 1 e dal p-value del test di Shapiro-Wilk. Indichiamo con $\mu_X = \mathbb{E}[X_i]$ il valore atteso della popolazione.

(a) Si tratta di impostare un test di livello α per le ipotesi statistiche

$$H_0 : \mu_X \geq 1 := \mu_0 \qquad \text{vs.} \qquad H_1 : \mu_X < \mu_0 \,.$$

Si noti che con questa scelta, l'errore più grave (cioè quello di primo tipo) consiste nel ritenere le sigarette conformi alla legge (equivalentemente: accettare H_1, rifiutando H_0 di conseguenza) quando in realtà esse la violano (equivalentemente: quando in realtà H_0 è vera). Siccome abbiamo già osservato che le X_i non sono gaussiane, necessariamente dovremo utilizzare un test asintotico; in questo caso possiamo farlo, perché $n = 100$ è un valore sufficientemente grande. Sul formulario, troviamo che la regola di un test al livello approssimativamente pari ad α è dunque

$$\text{“rifiuto } H_0 \text{ se e solo se } Z_0 := \frac{\overline{X} - \mu_0}{S_X}\sqrt{n} < -z_{1-\alpha}\text{”} \,.$$

(b) No, nessuna ipotesi.

(c) Con i dati a disposizione, risulta

$$z_0 = \frac{\overline{x} - \mu_0}{s_X}\sqrt{n} = \frac{0.9436 - 1}{0.5410045}\sqrt{100} = -1.042505 \,.$$

Il p-value si calcola dunque risolvendo in α l'equazione

$$z_0 \equiv -z_{1-\alpha} \quad \Leftrightarrow \quad 1.042505 = z_{1-\alpha} \quad \Leftrightarrow \quad \Phi(1.042505) = \Phi(z_{1-\alpha}) = 1 - \alpha$$
$$\Leftrightarrow \quad \alpha = 1 - \Phi(1.042505) = 1 - 0.85083 = 0.14917 \,.$$

Un p-value $\simeq 15\%$ è piuttosto alto, maggiore degli usuali livelli di significatività dell'1%, 5% e 10%. In altre parole, non si può rifiutare H_0 ai livelli di significatività dell'1%, del 5% e del 10%, e non c'è dunque nessuna evidenza contro l'ipotesi H_0 che le sigarette Zampiron Extra-Strong violino il limite di legge. Poiché accettiamo H_0, si tratta di una conclusione debole.

Ora invece abbiamo a disposizione un secondo campione aleatorio $Y_1, \dots, Y_m$ relativo al contenuto di nicotina delle Zampiron Light, indipendente da quello considerato in precedenza. Si noti che la numerosità di quest'ultimo campione è $m =$ `length(Light)` $= 60$. Dalla Figura 2 (sia dal normal Q-Q plot, sia dal p-value del test di Shapio-Wilk) si vede che $Y_i \sim N(\mu_Y, \sigma_Y^2)$ con varianza σ_Y^2 incognita.

(d) Si tratta di impostare un test di livello $\alpha = 5\%$ per la differenza dei valori attesi $\mu_X - \mu_Y$:

$$H_0 : \mu_X \leq \mu_Y \quad (\Leftrightarrow \mu_X - \mu_Y \leq 0) \qquad \text{vs.} \qquad H_1 : \mu_X > \mu_Y \quad (\Leftrightarrow \mu_X - \mu_Y > 0) \,.$$

Si noti che in H_1 abbiamo messo l'affermazione per cui cerchiamo evidenza dai dati.

Siccome il primo campione $X_1, \dots, X_n$ non è gaussiano, ma $n = 100$ e $m = 60$ (entrambi maggiori di 50) sono sufficientemente grandi, applicheremo un test asintotico sulla differenza delle medie di due campioni indipendenti. La regola di un tale test al livello approssimativamente pari ad α è

$$\text{“rifiuto } H_0 \text{ se e solo se } Z_0 := \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} > z_{1-\alpha}\text{”}$$

e con $\alpha = 5\%$ si ha

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

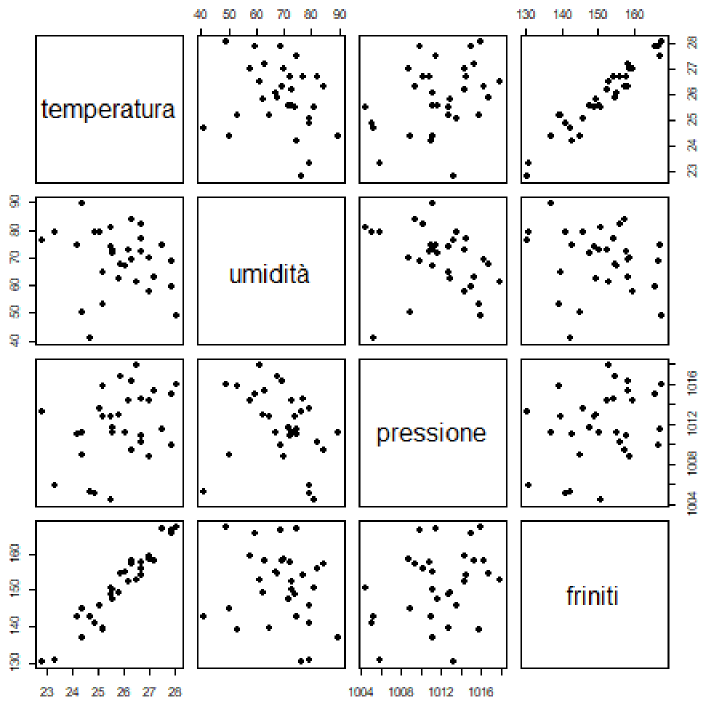
Con i dati a disposizione risulta

$$z_0 = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}} = \frac{0.9436 - 0.7568333}{\sqrt{\frac{0.5410045^2}{100} + \frac{0.3312661^2}{60}}} = 2.708240\,.$$

Poiché $z_0 = 2.708240 > 1.645 = z_{1-\alpha}$, dobbiamo rifiutare H_0 a livello di significatività del 5% e accettare H_1.

(e) No, neanche qui nessuna ipotesi.

Problema 3. Il professor Dolbear vuole stabilire come una determinata specie di grilli (*Oecanthus fultoni*) frinisce in relazione alle condizioni meteorologiche. Allo scopo ha raccolto i valori di `temperatura` (in $°C$), `umidità` (in %) e `pressione` (in hPa) di 31 sere d'agosto, e il numero medio di `friniti` al minuto rilevati in ciascuna di queste sere. Gli scatterplot tra le variabili considerate sono mostrati nella figura sottostante. L'area di lavoro allegata contiene il *data frame* `dati` con le misure ottenute. *(È un file `.RData`. Potete caricarlo selezionando `File` → `Carica area di lavoro...` dal menù di R.)*



Per descrivere la relazione tra le variabili precedenti, Dolbear propone il modello di regressione lineare di cui trovate l'output e la diagnostica dei residui nella pagina seguente.

(a) Scrivete la relazione tra le variabili `temperatura`, `umidità`, `pressione` e `friniti` ipotizzata dal modello di Dolbear.

(b) Il modello di Dolbear spiega bene la variabilità della risposta? Perché?

(c) Il modello rispetta le ipotesi alla base del modello lineare gaussiano? Perché?

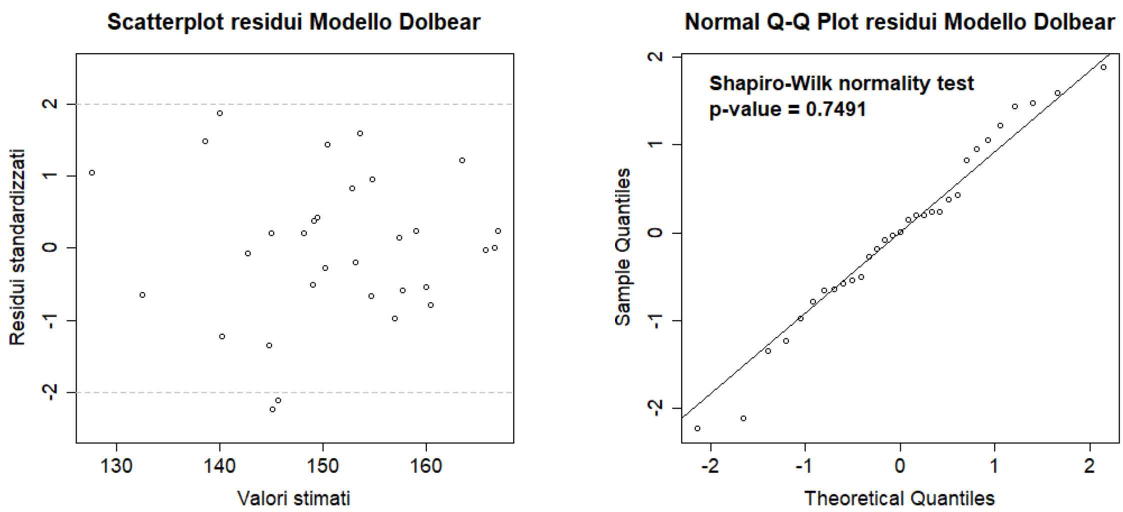
(d) Nel modello i regressori sono tutti significativi? Perché?

(e) In base ai dati a disposizione, proroponete voi un nuovo modello di regressione lineare gaussiano che abbia in risposta la variabile `friniti` e che sia migliore di quello di Dolbear. In particolare, spiegate in che modo siete arrivati a proporre il vostro modello.

(f) Sottoponete il nuovo modello alle stesse verifiche che avete fatto per quello di Dolbear nei punti (b), (c) e (d). In cosa è superiore il vostro modello?

(g) Scrivete la relazione tra le variabili `temperatura`, `umidità`, `pressione` e `friniti` ipotizzata dal vostro modello. Fornite inoltre una stima puntuale di *tutti* i parametri incogniti che intervengono in tale relazione.

```
Call:
lm(formula = friniti ~ temperatura + umidità + pressione, data = dati)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.0973 -1.6694  0.0306  1.7678  4.9123 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 106.87196  167.62862   0.638    0.529    
temperatura   7.58993    0.45443  16.702 9.29e-16 ***
umidità       0.01492    0.05106   0.292    0.772    
pressione    -0.15148    0.16840  -0.900    0.376    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.981 on 27 degrees of freedom
Multiple R-squared:  0.9207,	Adjusted R-squared:  0.9119 
F-statistic: 104.6 on 3 and 27 DF,  p-value: 5.615e-15
```



Oltre a quanto già usato nella prima prova, ora potete usare R e il file `Riepilogo+R.pdf`*. Nei tre slot della form dell'esame, dovete caricare:*

- *la scansione del manoscritto con la vostra soluzione;*
- *la copia della console di R contenente la parte svolta al computer, coi comandi che avete usato e la relativa risposta del programma (basta selezionare tutte le righe col mouse e copia-incollarle su un file di Word);*
- *i file grafici ottenuti con R (basta cliccare col tasto destro del mouse nella finestra in cui R visualizza ciascuna figura, selezionare* `Copia come metafile` *e incollare su un file di Word; usate lo stesso file di Word per tutte le figure, aggiungendo sotto ciascuna una breve didascalia).*

Risultati.

(a) Dalla seconda riga della *summary* di R vediamo che il modello ipotizzato è

$$\texttt{friniti}_i = \beta_0 + \beta_1\,\texttt{temperatura}_i + \beta_2\,\texttt{umidità}_i + \beta_3\,\texttt{pressione}_i + E_i$$

con $i = 1, \ldots, 31$, $E_1, \ldots, E_{31}$ i.i.d. e $E_i \sim N(0, \sigma^2)$ per ogni i.

(b) Il modello di Dolbear spiega molto bene la variabilità della risposta, perché la sua percentuale di variabilità spiegata (che nella regressione multipla è data dall'r^2-adjusted) è $r^2_{\text{adj}} = 0.9119 = 91.19\% > 80\%$.

(c) Il modello di Dolbear rispetta le ipotesi alla base del modello lineare gaussiano, in quanto i suoi residui standardizzati:

- sono omoschedastici (cioè disposti a nuvola e senza nessun pattern particolare nel relativo scatterplot),
- si allineano bene coi quantili teorici della normale standard nel normal Q-Q plot,
- forniscono un elevato p-value nel test di Shapiro-Wilk ($p\text{-value}_{\text{SW}} = 74.91\%$),
- a ulteriore conferma della gaussianità, presentano solo il $2/31 = 6.45\%$ di outlier fuori dall'intervallo $[-2, +2]$, che è una frequenza prossima alla probabilità teorica del 5% fornita dalla $N(0, 1)$.

(d) Nel modello di Dolbear non tutti i regressori sono significativi. Non sono infatti siglificativi i regressori `umidità` e `pressione`, per i quali il T-test sui relativi coefficienti fornisce $p\text{-value}_{\texttt{umidità}} = 77.2\%$ e $p\text{-value}_{\texttt{pressione}} = 37.6\%$, entrambi ben maggiori dell'usuale 5%. È invece molto significativo il regressore `temperatura`, con $p\text{-value}_{\texttt{temperatura}} = 9.29 \cdot 10^{-16} \ll 5\%$.

(e) Rimuoviamo innanzitutto il regressore meno significativo dal modello di Dolbear, cioè il regressore `umidità`, e rilanciamo su R il modello così ridotto:

```
> mod2 <- lm(friniti ~ temperatura + pressione, data = dati)
> summary(mod2)

Call:
lm(formula = friniti ~ temperatura + pressione, data = dati)

Residuals:
    Min      1Q  Median      3Q     Max
-6.2705 -1.5973  0.0405  1.8035  4.5812

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 117.4082   161.0061   0.729    0.472
temperatura   7.5713     0.4425  17.109 2.33e-16 ***
pressione    -0.1604     0.1629  -0.985    0.333
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.932 on 28 degrees of freedom
Multiple R-squared:  0.9205,    Adjusted R-squared:  0.9148
F-statistic: 162.1 on 2 and 28 DF,  p-value: 4.03e-16
```

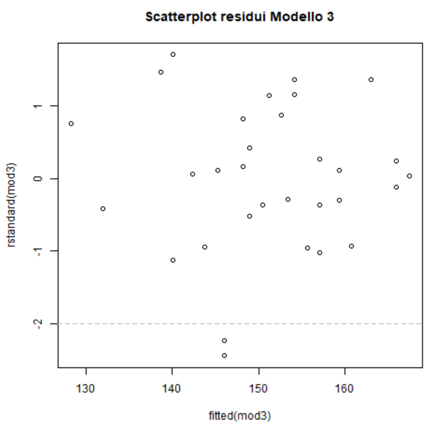
Il modello ottenuto spiega sempre molto bene la variabilità della risposta ($r^2_{\text{adj}} = 91.48\%$), tuttavia il suo regressore `pressione` è ancora non significativo (p-value$_{\texttt{pressione}} = 33.3\% > 5\%$). Rimuoviamo anche questo regressore riducendo ulteriormente il modello:

```
> mod3 <- lm(friniti ~ temperatura, data = dati)
> summary(mod3)

Call:
lm(formula = friniti ~ temperatura, data = dati)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0105 -1.3007  0.1961  2.1733  4.8200

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -40.7993    10.6776  -3.821 0.000649 ***
temperatura   7.4131     0.4121  17.988  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.93 on 29 degrees of freedom
Multiple R-squared:  0.9177,    Adjusted R-squared:  0.9149
F-statistic: 323.6 on 1 and 29 DF,  p-value: < 2.2e-16
```

Adesso il modello spiega ancora benissimo la variabilità della risposta ($r^2_{\text{adj}} = 91.49\%$, persino meglio del modello completo di Dolbear), e l'unico regressore `temperatura` è estremamente significativo (p-value$_{\texttt{temperatura}} < 2 \cdot 10^{-16}$). Perciò, possiamo proporre quest'ultimo come modello migliore. Il modo in cui siamo arrivati alla nostra conclusione è il metodo di eliminazione a ritroso dei predittori meno significativi, che è il metodo corretto per semplificare un modello con regressori ridondanti.

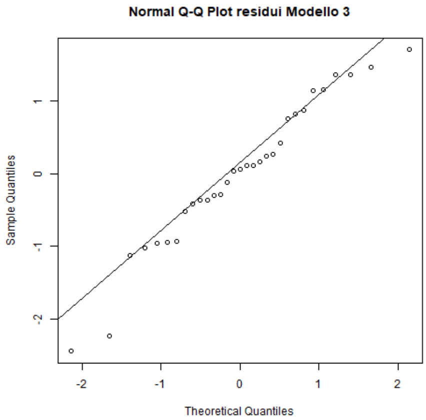
(f) Abbiamo già visto che il modello proposto (Modello 3) ha un'elevata percentuale di variabilità spiegata ($r^2_{\text{adj}} = 91.49\%$), come del resto il modello di Dolbear. Lo scatterplot dei suoi residui standardizzati è

```
> plot(fitted(mod3), rstandard(mod3), main="Scatterplot residui Modello 3")
> abline(h=c(-2,2), col="gray75", lty=2)
```



da cui si vede che i residui sono omoschedastici. Il test di Shapiro-Wilk e il normal Q-Q plot dei residui standardizzati danno invece

```
> shapiro.test(rstandard(mod3))

        Shapiro-Wilk normality test

data:  rstandard(mod3)
W = 0.96315, p-value = 0.3526

> qqnorm(rstandard(mod3), main="Normal Q-Q Plot residui Modello 3")
> qqline(rstandard(mod3))
```



Col p-value$_{\text{SW}}$ = 35.26% e i quantili così ben allineati lungo la Q-Q line (tranne forse i primi due), non c'è nessuna evidenza per rifiutare l'ipotesi nulla di gaussianità. Infine, per il modo stesso in cui è stato costruito, il Modello 3 ha l'unico regressore molto significativo, ed è in questo superiore al modello completo di Dolbear.

(g) Il Modello 3 è il seguente

$$\texttt{friniti}_i = \beta_0 + \beta_1 \texttt{temperatura}_i + E_i$$

con $i = 1, \dots, 31$, $E_1, \dots, E_{31}$ i.i.d. e $E_i \sim N(0, \sigma^2)$ per ogni i. I suoi parametri sono β_0, β_1 e σ. Le rispettive stime dai dati sono

$$\widehat{\beta}_0 = -40.7993\,, \qquad \widehat{\beta}_1 = 7.4131\,, \qquad \widehat{\sigma} = 2.93\,.$$