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

IV appello di Statistica per Ingegneria Fisica
27 Gennaio 2020



*Cognome, Nome e Numero di matricola:*

**Problema 1.** La Longobarda è una piccola e mediocre squadra di provincia che da tempo immemore staziona in Serie Z. Dalle statistiche degli anni passati, si sa che la Longobarda perde mediamente 2.40 volte ogni 6 partite disputate e che i risultati di partite diverse sono tutti indipendenti tra loro.

(a) Qual è la probabilità che la Longobarda perda almeno 5 delle prime 6 partite del prossimo campionato?

È noto inoltre che il punteggio che la Longobarda riesce a guadagnarsi in una partita qualsiasi è una variabile aleatoria $X$, con media pari a 0.80 punti e varianza di 0.76 punti$^2$. Ora, in base alle regole della Serie Z, il punteggio di una parita è (come per la Serie A):

- $X = 0$ in caso di sconfitta;
- $X = 1$ in caso di pareggio;
- $X = 3$ in caso di vittoria.

(b) Determinare la densità di probabilità della variabile aleatoria discreta $X =$ *punti guadagnati dalla Longobarda in una partita qualsiasi*, specificando l'insieme su cui tale densità è definita.

(c) Qual è la probabilità che la Longobarda vinca una partitia qualsiasi?

(d) Per essere promossa in Serie Y, la Longobarda deve riuscire a totalizzare almeno 39 punti nelle prossime 42 partite di campionato. Qual è la probabilità, eventualmente approssimata, che questo accada?

Deluso per quanto è esigua la probabilità di promozione trovata al punto precedente, il presidente della Longobarda ha finalmente licenziato il vecchio allenatore, e al suo posto ha inaggiato Oronzo Canà. Costui infatti gli ha promesso che, grazie al rivoluzionario modulo 5-5-5, il punteggio medio per partita non sarà più i miseri 0.80 punti di prima, ma raggiungerà un valore $\mu$ sufficiente a portare al 50% la probabilità di totalizzare almeno 39 punti in 42 partite. E tutto questo mantenendo $\mathrm{Var}(X) = 0.76$ inalterata!

(e) Quanto deve valere $\mu$ per rispettare la promessa di Canà?

**Risultati.**

(a) Introduciamo la v.a.

$$Y_6 = \text{numero di sconfitte in 6 partite} \sim B(6, p)\,,$$

dove $p$ è la probabilità di perdere una partita qualsiasi. Sappiamo dal testo che

$$\mathbb{E}[Y_6] = 2.40 \quad \Rightarrow \quad 6p = 2.40 \quad \Rightarrow \quad p = 0.40\,.$$

Dunque la probabilità cercata è

$$\begin{aligned}\mathbb{P}(Y_6 \geq 5) &= \sum_{k=5}^{6} p_{Y_6}(k) = \sum_{k=5}^{6} \binom{6}{k} p^k (1-p)^{6-k} = 6 \cdot 0.40^5 (1-0.40) + 1 \cdot 0.40^6 (1-0.40)^0 \\ &= 4.096\%\,.\end{aligned}$$

(b) Poiché $X$ può prendere solo i valori $S = \{0, 1, 3\}$, la densità discreta di $X$ è definita su tale insieme di valori: $p_X : S \to [0, 1]$. Sappiamo inoltre che deve essere:

$$
\begin{aligned}
& p_X(k) \geq 0 \quad \forall k \in S \\
& \sum_{k \in S} p_X(k) = 1 && \Longrightarrow \quad p_X(0) + p_X(1) + p_X(3) = 1 \\
& 0.80 = \mathbb{E}[X] = \sum_{k \in S} k\, p_X(k) && \Longrightarrow \quad 0 \cdot p_X(0) + 1 \cdot p_X(1) + 3 \cdot p_X(3) = 0.80 \\
& 0.76 = \mathrm{Var}(X) = \sum_{k \in S} k^2\, p_X(k) - \mathbb{E}[X]^2 && \Longrightarrow \quad 0^2 \cdot p_X(0) + 1^2 \cdot p_X(1) + 3^2 \cdot p_X(3) - 0.80^2 = 0.76 \\
& \mathbb{P}(X = 0) = 0.40 && \Longrightarrow \quad p_X(0) = 0.40\,.
\end{aligned}
$$

Avendo ben 4 equazioni per sole 3 incognite, possiamo mettere a sistema solo le 3 equazioni 'più facili', e cioè

$$
\begin{cases} p_X(0) + p_X(1) + p_X(3) = 1 \\ 0 \cdot p_X(0) + 1 \cdot p_X(1) + 3 \cdot p_X(3) = 0.80 \\ p_X(0) = 0.40 \end{cases} \quad \Longrightarrow \quad \begin{cases} p_X(0) = 0.40 \\ p_X(1) = 0.50 \\ p_X(3) = 0.10\,. \end{cases}
$$

(c) La probabilità di vincere una partita qualsiasi è

$$
\mathbb{P}(X = 3) = p_X(3) = 0.10\,.
$$

(d) Sia ora $S_{42}$ la v.a.

$$
S_{42} = \text{punti totalizzati in 42 partite} = X_1 + X_2 + \ldots + X_{42}\,,
$$

dove $X_i$ è il risultato dell'$i$-esima partita. Allora, $X_1, \ldots, X_{42}$ sono i.i.d., con $\mathbb{E}[X_i] = 0.80$ e $\mathrm{Var}(X_i) = 0.76$. Perciò,

$$
\begin{aligned}
\mathbb{P}(S_{42} \geq 39) \underset{\substack{\text{correzione} \\ \text{di continuità}}}{=} \mathbb{P}(S_{42} \geq 38.5) &= \mathbb{P}\left( \frac{S_{42} - 42\mathbb{E}[X_i]}{\sqrt{42\mathrm{Var}(S_{42})}} \geq \frac{38.5 - 42 \cdot 0.80}{\sqrt{42 \cdot 0.76}} \right) = \\
&\underset{\text{TLC}}{\simeq} 1 - \Phi\left( \frac{38.5 - 42 \cdot 0.80}{\sqrt{42 \cdot 0.76}} \right) = 1 - \Phi(0.867) \simeq 1 - 0.80785 \\
&= 19.215\%
\end{aligned}
$$

(senza correzione di continuità $\mathbb{P}(S_{42} \geq 39) = 1 - \Phi(0.956) \simeq 1 - (0.82894 + 0.83147)/2 = 16.980\%$).

(e) Ora vogliamo scegliere $\mu = \mathbb{E}[X_i]$ in modo tale che

$$
\begin{aligned}
0.50 \equiv \mathbb{P}(S_{42} \geq 39) = \mathbb{P}(S_{42} \geq 38.5) &= \mathbb{P}\left( \frac{S_{42} - 42\mathbb{E}[X_i]}{\sqrt{42\mathrm{Var}(S_{42})}} \geq \frac{38.5 - 42\mu}{\sqrt{42 \cdot 0.76}} \right) \simeq 1 - \Phi\left( \frac{38.5 - 42\mu}{\sqrt{42 \cdot 0.76}} \right) \\
\Rightarrow \quad \Phi\left( \frac{38.5 - 42\mu}{\sqrt{42 \cdot 0.76}} \right) = 0.50 \quad &\Rightarrow \quad \frac{38.5 - 42\mu}{\sqrt{42 \cdot 0.76}} = z_{0.50} = 0 \\
\Rightarrow \quad \mu = 0.9167 &
\end{aligned}
$$

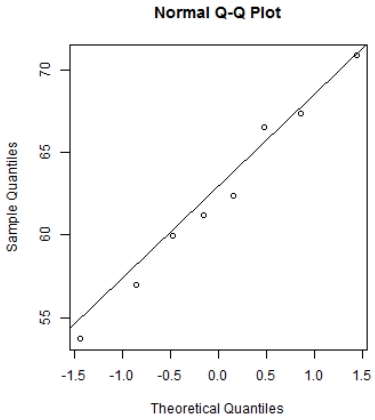
(senza correzione di continuità $\mu = 0.9286$).

**Problema 2.** Il Signor Gio. Batta è un uomo estremamente parsimonioso. Così parsimonioso che, da quando si è trasferito a Milano dall'amata Genova, egli ha rinunciato al suo piatto preferito – le trenette al pesto col basilico e i pinoli – perché secondo lui i pinoli venduti a Milano sono troppo più costosi di quelli che comperava a Genova. A sostegno della propria convinzione, il Signor Gio. Batta si basa su una sua personale indagine dei prezzi dei pinoli rilevati in 8 fruttivendoli di Milano e in 15 diversi *bisagnini* di Genova, che abbiamo riportato qui di seguito (tutti i valori sono espressi in €/Kg):

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Milano | 56.99 | 70.85 | 66.50 | 61.19 | 59.93 | 53.72 | 67.33 | 62.38 |
| Genova | 48.30 | 48.69 | 56.50 | 42.99 | 40.00 | 44.15 | 28.19 | 46.61 |
| | 48.82 | 37.36 | 49.41 | 30.25 | 43.15 | 41.29 | 53.79 | |

Abbiamo inoltre elaborato questi dati con R, ottenendo l'output seguente:

```
> Mi <- c( 56.99, 70.85, 66.50, 61.19, 59.93, 53.72, 67.33, 62.38 )
> Ge <- c( 48.30, 48.69, 56.50, 42.99, 40.00, 44.15, 28.19, 46.61,
+ 48.82, 37.36, 49.41, 30.25, 43.15, 41.29, 53.79 )
> mean(Mi); mean(Ge)
[1] 62.36125
[1] 43.96667
> sd(Mi); sd(Ge)
[1] 5.66545
[1] 7.838064
> shapiro.test(Mi); qqnorm(Mi); qqline(Mi)

        Shapiro-Wilk normality test

data:  Mi
W = 0.98188, p-value = 0.9716

> shapiro.test(Ge); qqnorm(Ge); qqline(Ge)

        Shapiro-Wilk normality test

data:  Ge
W = 0.95414, p-value = 0.5918

> var.test(Mi, Ge, ratio = 1, alternative = "two.sided")

        F test to compare two variances

data:  Mi and Ge
F = 0.52246, num df = 7, denom df = 14, p-value = 0.3931
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1545764 2.4012654
sample estimates:
ratio of variances
         0.5224578
```



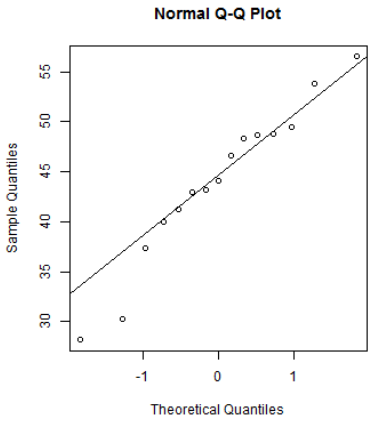




Figura 1: Console di R coi comandi utilizzati e a fianco l'output grafico corrispondente

(a) Secondo il Signor Gio. Batta, la sua indagine dimostra in modo evidente e inconfutabile che i pinoli venduti a Milano sono mediamente più cari di oltre 10 €/Kg rispetto a quelli comperati a Genova. Impostate un opportuno test al livello di significatività $\alpha$ per stabilire se dai dati c'è evidenza che il Signor Gio. Batta abbia ragione.

(b) Quali condizioni sono richieste ai due campioni per poter effettuare il test precedente? Tali condizioni sono tutte soddisfatte dai dati del Signor Gio. Batta? Giustificate la risposta.

(c) Determinate un intervallo in cui cade il $p$-value del test del punto (a) e traetene una conclusione.

(d) Calcolate un intervallo di confidenza bilatero al livello del 95% per il prezzo atteso (in €/Kg) dei pinoli comperati in un qualsiasi fruttivendolo di Milano.

**Risultati.**

(a) Siano $\mu_{\text{Mi}}$ e $\mu_{\text{Ge}}$ i prezzi attesi dei pinoli comperati rispettivamente a Milano e a Genova. Il Signor Gio. Batta afferma che secondo lui $\mu_{\text{Mi}} - \mu_{\text{Ge}} > 10$€/Kg. Vogliamo stabilire se c'è evidenza dai dati a favore di questa affermazione $\Rightarrow$ la mettiamo nell'ipotesi alternativa di un test sulla differenza delle medie:

$$H_0 : \mu_{\text{Mi}} - \mu_{\text{Ge}} \leq 10 =: \delta_0 \qquad \text{vs.} \qquad H_1 : \mu_{\text{Mi}} - \mu_{\text{Ge}} > \delta_0 \,.$$

I due campioni $X_1, \ldots, X_8$ e $Y_1, \ldots, Y_{15}$ hanno entrambi varianza incognita e non sono numerosi. L'unico test per le ipotesi statistiche precedenti che conosciamo è dunque il test per la differenza delle medie di due campioni gaussiani indipendenti e a varianze incognite ma uguali. La sua regola di rifiuto è

$$\text{“ rifiuto } H_0 \text{ se } T_0 := \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{S_p^2 \left(\frac{1}{m} + \frac{1}{n}\right)}} > t_{1-\alpha}(m+n-2)\text{”}\,, \qquad (**)$$

dove $S_p^2$ è la varianza pooled

$$S_p^2 := \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$$

e $\bar{X}$, $\bar{Y}$ e $S_X^2$, $S_Y^2$ sono rispettivamente le medie e le varianze campionarie dei due campioni, mentre $m$ e $n$ sono le loro numerosità.

(b) Oltre all'(ovvia) indipendenza, le condizioni richieste ai due campioni sono la gaussianità e l'uguaglianza delle varianze. La gaussianità è soddisfatta da entrambi i campioni, in quanto

- i $p$-value dei test di Shapiro-Wilk sono elevati ($p$-value$_{\text{Mi}} = 97.16\%$ e $p$-value$_{\text{Ge}} = 59.18\%$, entrambi ben al di sopra delle usuali soglie al 5% - 10%)
- nei normal Q-Q plot i punti sono ben allineati lungo la Q-Q line.

Anche la condizione di uguaglianza delle varianze non può essere rifiutata, in quanto l'$F$-test bilatero visibile nell'output di R ha restituito un $p$-value del 39.31%, troppo elevato per poter rifiutare l'ipotesi nulla $H_0 : \sigma_{\text{Mi}}^2 = \sigma_{\text{Ge}}^2$.

(c) Per calcolare il $p$-value del test del punto (a), calcoliamo la realizzazione della statistica test $T_0$ sui dati raccolti dal Signor Gio. Batta e poi imponiamo l'uguaglianza nella regola di rifiuto $(**)$:

$$s_p^2 = \frac{(8-1)\cdot 5.66545^2 + (15-1)\cdot 7.838064^2}{8+15-2} = 51.65594$$
$$t_0 = \frac{62.36125 - 43.96667 - 10}{\sqrt{51.65594 \cdot \left(\frac{1}{8} + \frac{1}{15}\right)}} = 2.66788\,,$$

dove abbiamo usato i valori $\bar{x} = 62.36125$, $\bar{y} = 43.96667$ e $s_X = 5.66545$, $s_Y = 7.838064$ leggibili sull'output di R. Il $p$-value è dunque il valore di $\alpha$ che soddisfa l'equazione

$$t_0 \equiv t_{1-\alpha}(m+n-2) \quad \Leftrightarrow \quad 2.66788 \equiv t_{1-\alpha}(21)\,.$$

Dalle tavole vediamo che

$$t_{0.99}(21) = 2.5176 < 2.66788 < 2.8314 = t_{0.995}(21) \quad \Rightarrow \quad 0.99 < 1-\alpha < 0.995 \quad \Rightarrow \quad 0.005 < \alpha < 0.01\,.$$

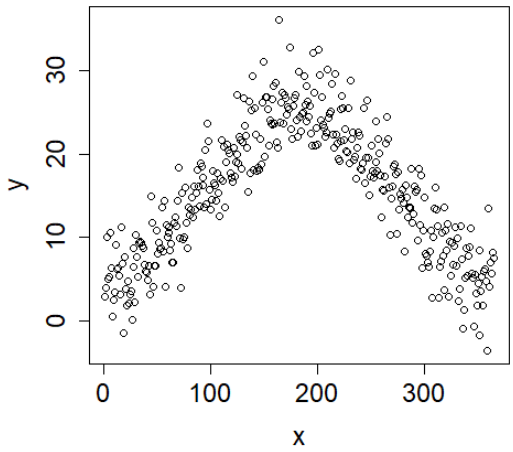
Il $p$-value del test è dunque compreso tra lo 0.5% e l'1%. Con un $p$-value così piccolo, non si può accettare $H_0$ e c'è invece evidenza a favore di $H_1$. Ne concludiamo che il Signor Gio. Batta ha ragione (conclusione forte).

(d) Dobbiamo calcolare un intervallo di confidenza per una popolazione gaussiana a varianza incognita. Al livello di confidenza $\gamma = 95\%$, tale intervallo è dato da

$$\mu_{\text{Mi}} \in \left(\bar{x} \pm t_{\frac{1+\gamma}{2}}(m-1)\frac{s_X}{\sqrt{m}}\right) = \left(62.36125 \pm 2.3646 \cdot \frac{5.66545}{\sqrt{8}}\right) = (57.6249\,,\, 67.0976)$$

dove $t_{\frac{1+\gamma}{2}}(m-1) = t_{0.975}(7) = 2.3646$.

**Problema 3.** Si vuole studiare l'andamento delle temperature medie giornaliere registrate dalla stazione meteorologica di Milano Lambrate nell'anno 2019. Sia $Y$ la variabile aleatoria che rappresenta la temperatura media giornaliera e $x$ la variabile che rappresenta il giorno. I giorni dell'anno 2019 sono stati numerati da 1 a 365 ($x_i$ con $i = 1, ..., 365$) e, per ogni giorno, è disponibile il valore medio giornaliero di temperatura in gradi centigradi ($y_i$ con $i = 1, ..., 365$). I dati raccolti sono rappresentati nel diagramma di dispersione qui sotto.




Sono state eseguite tre diverse regressioni lineari semplici con risposta $Y$:

- la prima con regressore $x$ (Modello 1);
- la seconda con regressore una trasformazione cosinusoidale di $x$ avente pulsazione $\frac{2\pi}{365} \simeq 0.01721$ (Modello 2);
- la terza con regressore una trasformazione quadratica di $x$ centrata in $\frac{365}{2} = 182.5$ (Modello 3).

Gli output dei tre diversi modelli sono riportati nella Figura 2 della pagina seguente. Nella successiva Figura 3 sono riportati anche i grafici dei residui per i tre modelli.

(a) Scrivere la relazione ipotizzata fra $Y$ e $x$ per ciascuno dei tre modelli.

(b) Indicare, per ciascun modello, se sono soddisfatte le ipotesi di gaussianità e di omoschedasticità dei residui. Giustificare la risposta.

(c) Quale dei tre modelli è preferibile? Giustificare la risposta.

(d) Qual è la percentuale di variabilità spiegata dal modello scelto al punto precedente?

(e) Il modello scelto è globalmente significativo?

(f) Scrivere l'equazione stimata per il modello scelto al punto (c).

(g) Utilizzare il modello scelto per fornire una previsione puntuale della temperatura media giornaliera del giorno 1/1/2020.

*Suggerimento: Nel secondo output in Figura 2, la funzione* `cos` *di R calcola il coseno del suo argomento. Per esempio,*

```
> z <- c( 0, 3.1416/2, 3.1416, 2*3.1416 )
> cos(z)
[1]  1.000000e+00 -3.673205e-06 -1.000000e+00  1.000000e+00
```

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-18.8881  -6.5908   0.3585   6.3506  20.7599

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.524e+01  8.365e-01  18.215   <2e-16 ***
x           3.883e-04  3.961e-03   0.098    0.922
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.974 on 363 degrees of freedom
Multiple R-squared:  2.647e-05, Adjusted R-squared:  -0.002728
F-statistic: 0.00961 on 1 and 363 DF,  p-value: 0.922
```

---

```
Call:
lm(formula = y ~ I(cos(0.01721 * x)))

Residuals:
    Min      1Q  Median      3Q     Max
-8.6148 -2.1584 -0.2069  2.0981 11.0143

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         15.3046     0.1711   89.45   <2e-16 ***
I(cos(0.01721 * x)) -10.2588     0.2420  -42.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.269 on 363 degrees of freedom
Multiple R-squared:  0.8319,    Adjusted R-squared:  0.8315
F-statistic:  1797 on 1 and 363 DF,  p-value: < 2.2e-16
```

---

```
Call:
lm(formula = y ~ I((x - 182.5)^2))

Residuals:
     Min       1Q   Median       3Q      Max
-10.5220  -2.6005  -0.1473   2.5919  13.1795

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.312e+01  2.986e-01   77.42   <2e-16 ***
I((x - 182.5)^2) -7.038e-04  2.005e-05  -35.10   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.804 on 363 degrees of freedom
Multiple R-squared:  0.7725,    Adjusted R-squared:  0.7718
F-statistic:  1232 on 1 and 363 DF,  p-value: < 2.2e-16
```

Figura 2: Summary di R per i tre modelli studiati

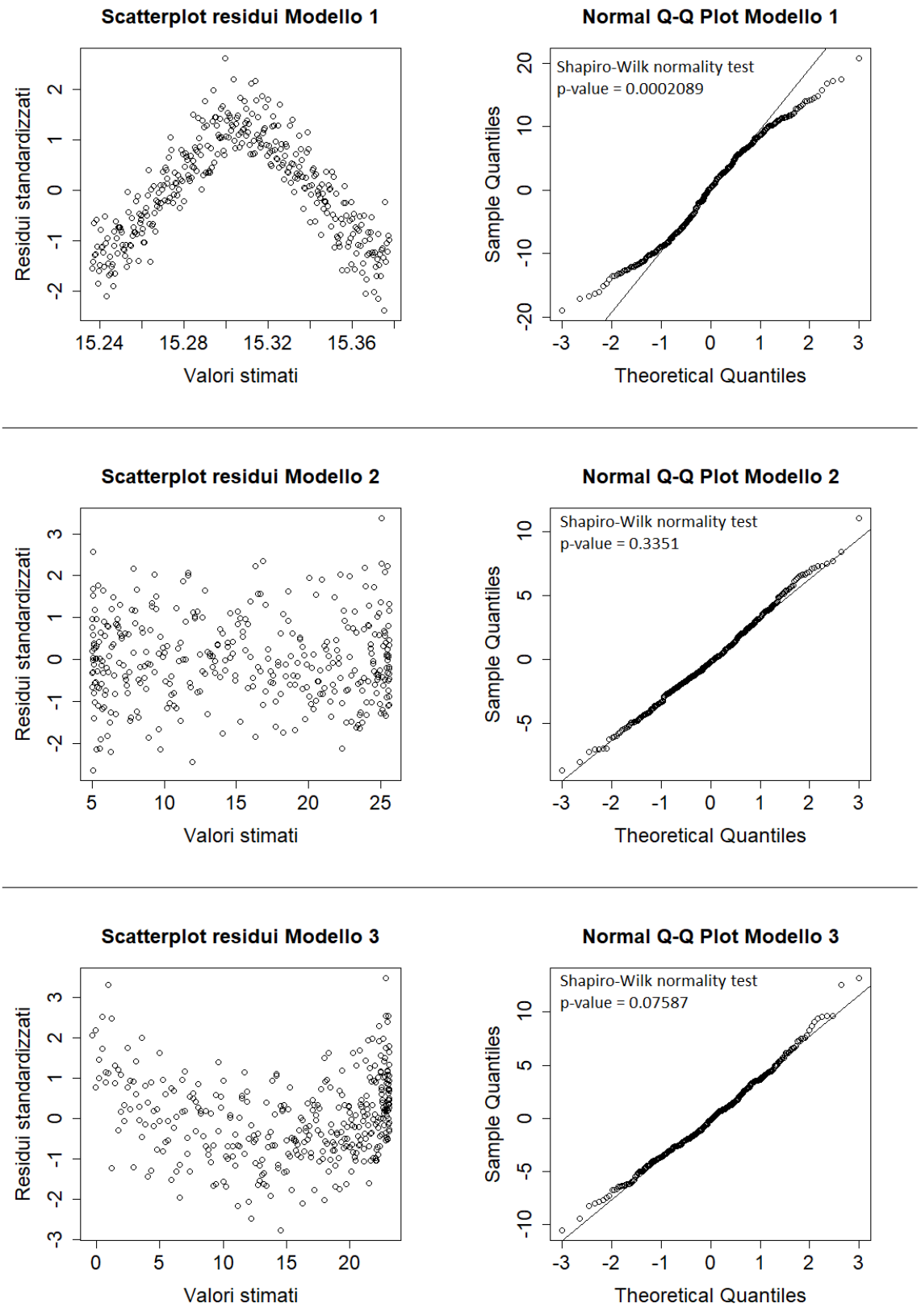



Figura 3: Scatterplot e normal Q-Q plot dei residui per i tre modelli studiati

**Risultati.**

(a)
- Modello 1: $Y = \beta_0 + \beta_1 x + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$
- Modello 2: $Y = \beta_0 + \beta_1 \cos(2\pi \frac{x}{365}) + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$
- Modello 3: $Y = \beta_0 + \beta_1 (x - \frac{365}{2})^2 + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$

(b) L'ipotesi di gaussianità dei residui è soddisfatta per i Modelli 2 e 3 (nei Q-Q plot dei quantili teorici contro quelli empirici l'andamento lineare è rispettato e i $p$-value degli Shapiro-test sono maggiori di 0.05), mentre non è rispettata per il Modello 1 (nel Q-Q plot dei quantili teorici contro quelli empirici l'andamento lineare non è rispettato e il $p$-value degli Shapiro-test è inferiore a 0.05). L'ipotesi di omoschedasticità dei residui è rispettata solo per il Modello 2 (scatterplot dei residui che non presenta particolari pattern). Infatti, nei Modelli 1 e 3 è presente un chiaro pattern negli scatterplot dei residui (a forma di U con concavità rispettivamente negativa e positiva per i due modelli).

(c) Come visto al punto precedente, il Modello 2 è l'unico che rispetta le ipotesi di gaussianità e omoschedasticità dei residui. Perciò, tale modello è preferibile agli altri due, che presentano residui eteroschedastici.

(d) La percentuale di variabilità spiegata dal Modello 2 è 83.19% ($R^2 = 0.8319$).

(e) Il Modello 2 è globalmente significativo: infatti, il $p$-value dell'$F$-test sulla significatività globale della regressione (che corrisponde in questo caso al test sul coefficiente $\beta_1$) è molto piccolo, minore di $2.2 \cdot 10^{-16}$ (`<2.2e-16`).

(f) $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cos(2\pi \frac{x}{365}) = 15.3046 - 10.2588 \cdot \cos(2\pi \frac{x}{365})$

(g) Il giorno 1/1/2020 corrisponde al valore del regressore $x_{new} = 366$. La previsione puntuale della temperatura media giornaliera del giorno 1/1/2020 fornita dal Modello 2 è quindi

$$\hat{y}_{\text{new}} = 15.3046 - 10.2588 \cdot \cos(2\pi \frac{x_{\text{new}}}{365}) = 15.3046 - 10.2588 \cdot \cos(2\pi \frac{366}{365}) = 5.1078$$