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

III appello di Statistica per Ingegneria Fisica
13 Settembre 2019

Cognome, Nome e Numero di matricola:

Problema 1. Aldo ha da sempre due grandi passioni: la statistica e la cioccolata. Quando di recente ha scoperto la nuova marca di cioccolatini ACME, naturalmente non si è lasciato scappare l'occasione di combinare le due cose: ha comprato ben 60 scatole dei nuovi cioccolatini, le ha aperte tutte, e, a mano a mano che le apriva, ha contato e registrato il numero di cioccolatini fondenti (i suoi preferiti) trovati in ogni scatola. La tabella seguente riassume il dataset che Aldo si è costruito in questo modo:

Numero di cioccolatini fondenti trovati	Numero di scatole che li contenevano
0 o 1	0
2	4
3	12
4	20
5	11
6	7
7	5
8	0
9	1
10 o più	0

(a) Completate la tabella di distribuzione di frequenza con la Frequenza Relativa e la Frequenza Relativa Cumulata.

(b) Determinate la mediana, il primo e il terzo quartile e l'IQR dei dati di Aldo.

(c) Rappresentate la distribuzione dei dati con un boxplot.

(d) Determinate la media campionaria e la varianza campionaria dei dati.

(e) Sulle confezioni dei cioccolatini c'è scritto: "*La ACME garantisce che in ogni scatola ci sono in media almeno* 5 *cioccolatini fondenti*". Quest'affermazione è compatibile coi dati raccolti, o al contrario Aldo ha elementi sufficienti per poterla contestare? Impostate un opportuno test, calcolatene il p-value e fornite una risposta.

Aldo vuole inoltre stimare la probabilità che in una scatola a caso ci siano almeno 3 cioccolatini fondenti.

(e) Fornite una stima puntuale per la probabilità cercata da Aldo.

(f) Fornite un intervallo di confidenza bilatero a livello di confidenza del 90% per la stessa probabilità.

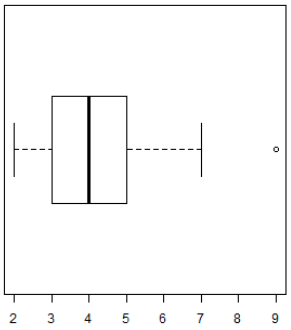
Risultati.

(a) La tabella completa è la seguente:

Classi	FA	FR (in %)	FC (in %)
0 o 1	0	0	0
2	4	6.67	6.67
3	12	20.00	26.67
4	20	33.33	60.00
5	11	18.33	78.33
6	7	11.67	90.00
7	5	8.33	98.33
8	0	0.00	98.33
9	1	1.67	100.00
10 o più	0	0	100.00

(b) Osserviamo dalla tabella delle frequenze cumulate che il 60% dei dati è ≤ 4 (FC(4) = 60%), mentre il 26.67% è ≤ 3 (FC(3) = 26.67%). Di conseguenza, $3 <$ Mediana ≤ 4, da cui Mediana = 4 perché i dati possono prendere solo valori interi. In modo simile si trova Q1 = 3, Q3 = 5, da cui IQR = Q3 − Q1 = 2.

(c) Il boxplot è riportato di seguito. C'è un solo outlier, in corrispondenza dell'unico dato pari a 9, in quanto $9 > \text{Q3} + 1.5 \cdot \text{IQR} = 8$.



(d) Si ha

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{k} k\,\text{FA}(k) = \frac{1}{60}\left(0\cdot 0 + 1\cdot 0 + 2\cdot 4 + 3\cdot 12 + \ldots + 9\cdot 1 + 10\cdot 0\right) \\
&= 4.4167\,, \\
s^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \frac{1}{n-1}\left(\sum_{k} k^2\text{FA}(k) - n\bar{x}^2\right) \\
&= \frac{1}{60-1}\left(0^2\cdot 0 + 1^2\cdot 0 + 2^2\cdot 4 + 3^2\cdot 12 + \ldots + 9^2\cdot 1 + 10^2\cdot 0 - 60\cdot 4.4167^2\right) \\
&= 2.1455\,.
\end{aligned}$$

(e) Sia μ il valore atteso di cioccolatini in una scatola qualunque. Poiché l'ipotesi di default è che la ACME dichiari il vero, scegliamo $\mu \geq 5$ come ipotesi nulla H_0. Aldo potrà contestare quanto

affermato dalla ACME solo se rifiuterà H_0 in un test che abbia $\mu < 5$ come ipotesi alternativa, cioè solo se otterrà forte evidenza dal suo test che $\mu < 5$. Riassumendo, le ipotesi del test devono essere

$$H_0 : \mu \geq 5 \qquad \text{contro} \qquad H_1 : \mu < 5\,.$$

Poiché abbiamo a disposizione un campione non gaussiano (i dati sono discreti, e in ogni caso non ne conosciamo la densità), ma comunque numeroso, possiamo fare un T-test per un campione numeroso a varianza incognita. Al livello di significatività α, abbiamo dunque la regola

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

Coi dati a disposizione, troviamo

$$t_0 = \frac{4.4167 - 5}{\sqrt{2.1455}}\sqrt{60} = -3.085\,.$$

Per calcolare il p-value del test, imponiamo l'uguaglianza nella regola di rifiuto ($\circ$):

$$\begin{aligned} t_0 \equiv -z_{1-\alpha} \quad &\Leftrightarrow \quad -3.085 = -z_{1-\alpha} \quad \Leftrightarrow \quad \Phi(3.085) = \Phi(z_{1-\alpha}) = 1 - \alpha \\ &\Leftrightarrow \quad \alpha = 1 - \Phi(3.085) = 1 - \frac{0.99896 + 0.99900}{2} = 0.00102 = 0.102\%\,. \end{aligned}$$

Cioè, p-value $= 0.102\%$. Con un valore così piccolo, rifiutiamo H_0 a ogni livello di significatività ragionevole. Possiamo dunque trarre la conclusione *forte* che la ACME dichiara il falso.

(f) Una stima puntuale per la probabilità p che in una scatola ci siano almeno 3 cioccolatini fondenti è data dalla frequenza empirica

$$\hat{p} = \frac{\#\,\text{scatole con almeno 3 cioccolatini}}{\#\,\text{scatole totali}} = 1 - \text{FC}(2) = 1 - 0.0667 = 0.9333 = 93.33\%\,.$$

(g) Un intervallo di confidenza bilatero a livello $\gamma = 90\%$ per la probabilità p del punto precedente è

$$\begin{aligned} p \in \left(\hat{p} \pm z_{\frac{1+\gamma}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) &= \left(0.9333 \pm z_{0.95}\sqrt{\frac{0.9333\,(1-0.9333)}{60}}\right) \\ = (0.9333 \pm 1.645 \cdot 0.0322) &= (0.8803\,,\, 0.9863) = (88.03\%\,,\, 98.63\%)\,. \end{aligned}$$

Problema 2. Aureliano è lo studente fuori sede che ormai da molti anni viene in università col treno da Macondo, il piccolo villaggio in cui vive sperduto nelle foreste pluviali a nord di Milano. In tutti questi anni, il servizio dei treni sulla linea Macondo-Milano è molto migliorato: ci sono due treni regionali, uno alle 7:00 (= 7.00) e un altro alle 7:30 (= 7.50), che sono entrambi sempre puntualissimi, mentre dopo le 7:30 circolano esclusivamente treni intercity, ugualmente puntuali ma più costosi dei due regionali.

Ovviamente, Aureliano preferisce risparmiare, e di conseguenza ogni mattina cerca di prendere uno dei due regionali. Tuttavia, non sempre ci riesce, perché l'ora X in cui egli arriva in stazione in un giorno qualsiasi non è una quantità deterministica, ma si può invece solo modellizzare con una variabile aleatoria *di densità gaussiana*. Le ore d'arrivo in stazione in giorni diversi, inoltre, si possono assumere tutte indipendenti tra loro.

Aureliano ha notato che riesce a prendere il regionale delle 7:00 mediamente una mattina su due, mentre ha una probabilità del 15% di arrivare dopo che sono partiti entrambi i regionali.

(a) Determinate media e varianza della variabile aleatoria X.

(b) Calcolate le probabilità:

  (i) che Aureliano riesca a prendere un regionale;

  (ii) che in un giorno qualsiasi Aureliano perda il regionale delle 7:00, ma riesca comunque a prendere quello delle 7:30.

(c) Calcolate la probabilità che lunedì prossimo Aureliano riesca a prendere un regionale, ma non ci riesca invece martedì.

(d) Calcolate la probabilità che in 5 giorni Aureliano riesca a prendere un regionale per almeno 4 mattine.

(e) Calcolate la probabilità (eventualmente approssimata) che in 50 giorni Aureliano riesca a prendere un regionale per almeno 40 mattine.

Risultati.

(a) Sappiamo che $X \sim N(\mu, \sigma^2)$ e che

$$\mathbb{P}(X \le 7) = 0.50\,, \qquad \mathbb{P}(X > 7.5) = 0.15\,.$$

Ne vogliamo ricavare i parametri μ e σ^2. Abbiamo

$$0.50 = \mathbb{P}(X \le 7) = \mathbb{P}\Big(\underbrace{\frac{X-\mu}{\sigma}}_{\sim N(0,1)} \le \frac{7-\mu}{\sigma}\Big) = \Phi\left(\frac{7-\mu}{\sigma}\right)$$

$$\Leftrightarrow \quad \frac{7-\mu}{\sigma} = z_{0.50} = 0$$

$$0.15 = \mathbb{P}(X > 7.5) = \mathbb{P}\Big(\underbrace{\frac{X-\mu}{\sigma}}_{\sim N(0,1)} > \frac{7.5-\mu}{\sigma}\Big) = 1 - \Phi\left(\frac{7.5-\mu}{\sigma}\right)$$

$$\Leftrightarrow \quad \Phi\left(\frac{7-\mu}{\sigma}\right) = 0.85 \quad \Leftrightarrow \quad \frac{7.5-\mu}{\sigma} = z_{0.85} = 1.04\,,$$

e quindi dobbiamo risolvere il sistema lineare

$$\begin{cases} \dfrac{7-\mu}{\sigma} = 0 \\ \dfrac{7.5-\mu}{\sigma} = 1.04 \end{cases} \quad \Leftrightarrow \quad \begin{cases} \mu = 7 \\ \sigma = 0.4808 \end{cases} \quad \Leftrightarrow \quad \sigma^2 = 0.2311 \quad .$$

(b) Abbiamo

  (i) $\mathbb{P}(X \le 7.50) = 1 - \mathbb{P}(X > 7.50) = 1 - 0.15 = 0.85 = 85\%$;

(ii) $\mathbb{P}(7 < X \leq 7.50) = \mathbb{P}(X \leq 7.50) - \mathbb{P}(X \leq 7) = 0.85 - 0.50 = 0.35 = 35\%$.

(c) Sia X_1 l'ora in cui Aureliano arriverà in stazione lunedì e X_2 l'ora in cui arriverà martedì. Allora dobbiamo calcolare la probabilità

$$\mathbb{P}(X_1 \leq 7.50\,,\, X_2 > 7.50) \underset{\substack{\text{indipendenza} \\ \text{di } X_1 \text{ e } X_2}}{=} \mathbb{P}(X_1 \leq 7.50)\,\mathbb{P}(X_2 > 7.50) = 0.85 \cdot 0.15 = 0.1275 = 12.75\%\,.$$

(d) Indicando con S_5 la variabile aleatoria che conta il numero di volte in cui Aureliano riesce a prendere un regionale in 5 giorni, abbiamo $S_5 \sim B(5\,,\,0.85)$, e quindi la probabilità richiesta è

$$\begin{aligned}\mathbb{P}(S_5 \geq 4) &= \sum_{k=4}^{5} p_{S_5}(k) = \sum_{k=4}^{5} \binom{5}{k} 0.85^k (1-0.85)^{5-k} = 5 \cdot 0.85^4 (1-0.85) + 1 \cdot 0.85^5 \cdot 1 \\ &= 0.83521 = 83.521\%\,.\end{aligned}$$

(e) In modo simile all'esercizio precedente, indichiamo ora con S_{50} il numero di volte in cui Aureliano riesce a prendere un regionale in 50 giorni. Ora abbiamo

$$S_{50} \sim B(50\,,\,0.85) \approx N(50 \cdot 0.85\,,\,50 \cdot 0.85 \cdot (1-0.85)) = N(42.5\,,\,6.375)\,,$$

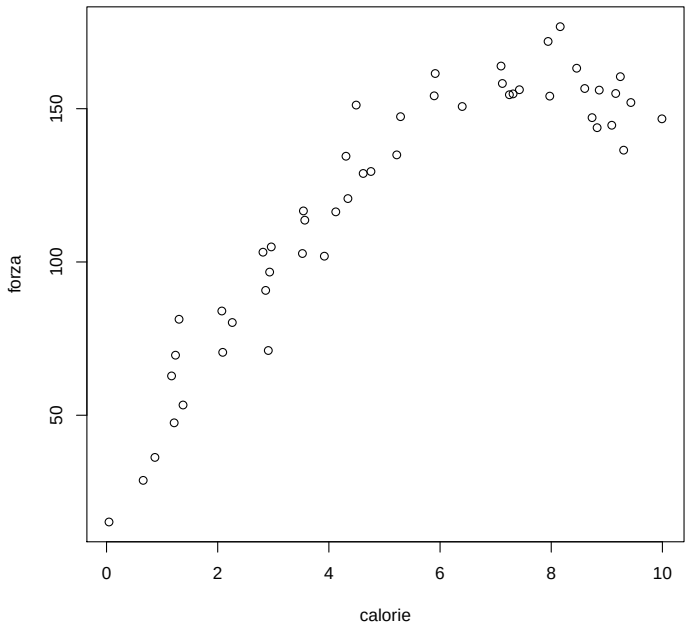
in cui abbiamo potuto usare l'approssimazione gaussiana della binomiale in quanto $n = 50 > 20$, $np = 50 \cdot 0.85 = 42.5 > 5$ e $n(1-p) = 50 \cdot (1-0.85) = 7.5 > 5$. La probabilità richiesta è quindi

$$\begin{aligned}\mathbb{P}(S_{50} \geq 40) &= \mathbb{P}\Bigg(\underbrace{\frac{S_{50} - \mathbb{E}[S_{50}]}{\sqrt{\mathrm{Var}(S_{50})}}}_{\approx N(0,1)} > \frac{40 - np}{\sqrt{np(1-p)}}\Bigg) = 1 - \Phi\left(\frac{40 - np}{\sqrt{np(1-p)}}\right) \\ &= 1 - \Phi\left(\frac{40 - 50 \cdot 0.85}{\sqrt{50 \cdot 0.85 \cdot (1-0.85)}}\right) = 1 - \Phi(-0.990) = \Phi(0.990) \\ &= 0.83891 = 83.891\%\,.\end{aligned}$$

Se avessimo voluto trovare un'approssimazione più precisa avremmo potuto usare la correzione di continuità:

$$\begin{aligned}\mathbb{P}(S_{50} \geq 40) = \mathbb{P}(S_{50} \geq 39.5) = \ldots &= 1 - \Phi\left(\frac{39.5 - 50 \cdot 0.85}{\sqrt{50 \cdot 0.85 \cdot (1-0.85)}}\right) = 1 - \Phi(-1.188) = \Phi(1.188) \\ &= 0.88298 = 88.298\%\,.\end{aligned}$$

Problema 3. Monkey D. Rufy. è il capitano della ciurma dei pirati di Cappello di Paglia e, come noto, ama abbuffarsi ai banchetti. Al fine di migliorare le capacità combattive del proprio capitano, il dottore della ciurma, TonyTony Chopper, decide di studiare la relazione tra la quantità di `calorie` (in migliaia di kcal) assunte da Rufy e la potenza del suo pugno nell'ora successiva al banchetto. Decide quindi, dopo ogni abbuffata, di misurare la `forza` (in migliaia di Newton) esercitata da un pugno del capitano sull'albero maestro della nave. In Figura 1 sono riportati gli output di tre diversi modelli lineari stimati da TonyTony Chopper. I dati raccolti sono i seguenti:



I p-value dello Shapiro-test sui residui dei tre modelli sono:

$$p\text{-value}_1 = 0.224 \qquad p\text{-value}_2 = 0.198 \qquad p\text{-value}_3 = 0.465.$$

(a) Fornire le equazioni per i tre modelli teorici ipotizzati da TonyTony Chopper.

(b) Quale dei tre modelli è migliore in termini di variabilità spiegata?

(c) Quali modelli presentano residui omoschedastici?

(d) Quali modelli presentano residui normali?

(e) Scegliere il modello migliore sulla base delle risposte precedenti e fornire l'equazione stimata per il modello scelto.

(f) Fornire una stima puntuale per la `forza` di un pugno di Monkey D. Rufy. dopo un pasto da 4000 kcal [`calorie` $= 4$].

```
Call:
lm(formula = forza ~ calorie)

Residuals:
    Min      1Q  Median      3Q     Max
-40.575 -14.332   2.979  13.951  39.502

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  55.2019     5.5806   9.892 3.61e-13 ***
calorie      12.5771     0.9397  13.384  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.37 on 48 degrees of freedom
Multiple R-squared:  0.7887,    Adjusted R-squared:  0.7843
F-statistic: 179.1 on 1 and 48 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = forza ~ I(calorie^2))

Residuals:
    Min      1Q  Median      3Q     Max
-68.655 -17.331   4.285  21.473  46.513

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   83.8415     5.6642  14.802  < 2e-16 ***
I(calorie^2)   1.0329     0.1205   8.575 3.02e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.49 on 48 degrees of freedom
Multiple R-squared:  0.605,     Adjusted R-squared:  0.5968
F-statistic: 73.53 on 1 and 48 DF,  p-value: 3.023e-11
```

```
Call:
lm(formula = forza ~ calorie + I(calorie^2))

Residuals:
     Min       1Q   Median       3Q      Max
-28.3081  -6.0334  -0.5375   5.9852  25.1987

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   12.0800     4.8351   2.498    0.016 *
calorie       36.7940     2.2124  16.631  < 2e-16 ***
I(calorie^2)  -2.3298     0.2075 -11.230 6.64e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.2 on 47 degrees of freedom
Multiple R-squared:  0.9426,    Adjusted R-squared:  0.9402
F-statistic: 386.1 on 2 and 47 DF,  p-value: < 2.2e-16
```

Figura 1: Summary dei tre modelli stimati.

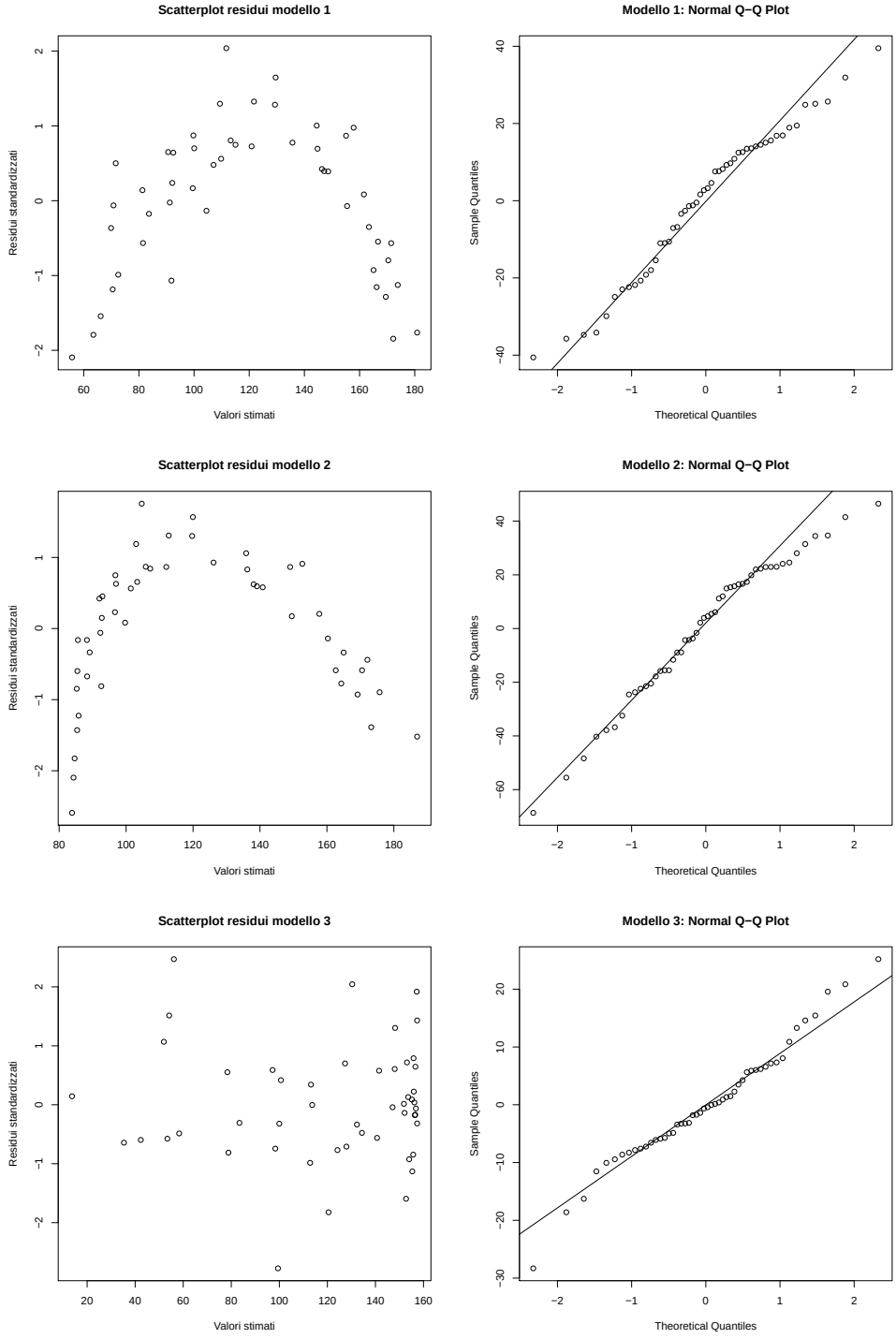


Figura 2: Scatterplot e qq-plot dei residui dei tre modelli stimati.

Risultati.

(a)
- modello 1: $\texttt{forza} = \beta_0 + \beta_1\,\texttt{calorie} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2),$
- modello 2: $\texttt{forza} = \beta_0 + \beta_1\,\texttt{calorie}^2 + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2),$
- modello 3: $\texttt{forza} = \beta_0 + \beta_1\,\texttt{calorie} + \beta_2\,\texttt{calorie}^2 + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2).$

(b) Per determinare il modello migliore in termini di variabilità spiegata dobbiamo confrontare l'R^2_{adj} dei tre modelli, in quanto due di essi coinvolgono più di un predittore. Abbiamo

$$R^2_{\text{adj }1} = 0.7843\,, \qquad R^2_{\text{adj }2} = 0.5968\,, \qquad R^2_{\text{adj }3} = 0.9402\,.$$

Quindi, il migliore in termini di variabilità spiegata risulta essere il modello 3. A seguire c'è il modello 1 e per ultimo il 2.

(c) L'ipotesi di omoschedasticità dei residui è rispettata solo per il terzo modello. Infatti, nei primi due è presente un chiaro pattern parabolico nello scatterplot dei residui.

(d) L'ipotesi di normalità dei residui è verificata per tutti e tre i modelli, dato che nel qq-plot dei quantili teorici contro quelli empirici l'andamento lineare è abbastanza rispettato e i p-value degli Shapiro-test sono tutti maggiori di 0.05 - 0.10.

(e) Il modello migliore è il modello 3, dato che è l'unico per cui è rispettata l'ipotesi di omoschedasticità, e inoltre ha un valore di R^2 molto alto. L'equazione stimata è:

$$\begin{aligned}\widehat{\texttt{forza}} &= \hat\beta_0 + \hat\beta_1\,\texttt{calorie} + \hat\beta_2\,\texttt{calorie}^2 \\ &= 12.08 + 36.794 \cdot \texttt{calorie} - 2.3298 \cdot \texttt{calorie}^2\end{aligned}$$

(f) Scegliendo il modello 3 la previsione è:

$$\widehat{\texttt{forza}} = 12.0800 + 36.7940 \cdot 4 - 2.3298 \cdot 4^2 = 121.9792.$$