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

V APPELLO DI STATISTICA PER INGEGNERIA FISICA
15 Febbraio 2021

Problema 1. Philip J. Fry lavora come fattorino per la Pizzeria Panucci. La distanza X (in Km) che percorre in bicicletta fra una consegna e l'altra si può modellizzare con una variabile aleatoria assolutamente continua avente densità

$$f(x) = \begin{cases} \frac{2}{3}(2-x) & \text{se } 0 < x < a \\ 0 & \text{altrimenti} \end{cases}$$

dove $a > 0$ è un parametro reale fissato. Inoltre, le distanze percorse per effettuare consegne diverse sono indipendenti fra loro.

(a) Determinate tutti i valori del parametro a per cui la funzione f è effettivamente la densità di probabilità di una variabile aleatoria assolutamente continua. Per tali valori, tracciate un grafico qualitativo di f.

D'ora in poi, fissate a in modo che f sia la densità di probabilità di una variabile aleatoria assolutamente continua.

$$\left(\textit{Se non siete riusciti a risolvere il punto } \text{(a)}, \textit{ usate } \; f(x) = \begin{cases} \frac{4}{9}(3-2x) & \textit{se } 0 < x < \frac{3}{2} \\ 0 & \textit{altrimenti}. \end{cases} \right)$$

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

(c) Questa sera Fry deve consegnare 45 pizze. Calcolate la probabilità che per farlo debba pedalare in tutto per più di 22 Km.

(d) Fry ha appena consegnato la 44-esima pizza, e ora gli resta solo l'ultima, quella ordinata da un certo I.C. Wiener. Calcolate la probabilità che per consegnarla Fry debba pedalare ancora per più di 1 Km.

Risultati.

(a) Per essere la densità di probabilità di una v.a. assolutamente continua, la funzione f deve essere:

- *normalizzata*:

$$1 = \int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = \int_0^a \frac{2}{3}\,(2-x)\,\mathrm{d}x = \frac{2}{3}\left[2x - \frac{x^2}{2}\right]_{x=0}^{x=a} = \frac{1}{3}\,(4a - a^2)$$
$$\Rightarrow \quad a^2 - 4a + 3 = 0 \quad \Rightarrow \quad a \in \{1, 3\}\,;$$

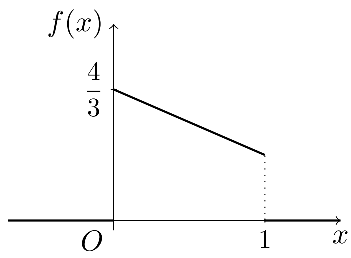
- *positiva*:

$$f(x) \geq 0 \quad \forall x \in (-\infty, +\infty) \quad \Rightarrow \quad \frac{2}{3}\,(2-x) \geq 0 \quad \forall x \in (0, a) \quad \Rightarrow \quad a \leq 2\,.$$

Combinando le due condizioni, troviamo che l'unico valore possibile è $a = 1$. Per tale valore,

$$f(x) = \begin{cases} \frac{2}{3}\,(2-x) & \text{se } 0 < x < 1 \\ 0 & \text{altrimenti} \end{cases}$$

e il grafico di f è



(b) Per la densità trovata al punto precedente, abbiamo

$$\mathbb{E}\,(X) = \int_{-\infty}^{+\infty} x f(x)\,\mathrm{d}x = \frac{2}{3}\int_0^1 (2x - x^2)\,\mathrm{d}x = \frac{2}{3}\left[x^2 - \frac{x^3}{3}\right]_{x=0}^{x=1} = \frac{4}{9}$$
$$\mathbb{E}\left(X^2\right) = \int_{-\infty}^{+\infty} x^2 f(x)\,\mathrm{d}x = \frac{2}{3}\int_0^1 (2x^2 - x^3)\,\mathrm{d}x = \frac{2}{3}\left[\frac{2x^3}{3} - \frac{x^4}{4}\right]_{x=0}^{x=1} = \frac{5}{18}$$
$$\mathrm{Var}\,(X) = \mathbb{E}\left(X^2\right) - \mathbb{E}\,(X)^2 = \frac{5}{18} - \left(\frac{4}{9}\right)^2 = \frac{13}{162}\,.$$

Facendo gli stessi calcoli per la densità data nel suggerimento, troviamo invece

$$\mathbb{E}\,(X) = \frac{1}{2}\,, \qquad \mathrm{Var}\,(X) = \frac{1}{8}\,.$$

(c) Sia $S_{45} = X_1 + X_2 + \ldots + X_{45}$ la distanza percorsa da Fry dopo 45 consegne, in cui X_i è la distanza corrispondente all'i-esima consegna. Allora le v.a. $X_1, X_2, \ldots, X_{45}$ sono i.i.d. e hanno tutte la stessa densità di X. Poiché $n = 45 > 30$, si può applicare il TLC alla loro somma S_{45}, ottenendo

$$\mathbb{P}\left(S_{45} > 22\right) = \mathbb{P}\Bigg(\underbrace{\frac{S_{45} - n\mathbb{E}\left(X_i\right)}{\sqrt{n\mathrm{Var}\left(X_i\right)}}}_{\approx N(0,1)} > \frac{22 - 45 \cdot (4/9)}{\sqrt{45 \cdot (13/162)}}\Bigg) \simeq 1 - \Phi\left(\frac{22 - 45 \cdot (4/9)}{\sqrt{45 \cdot (13/162)}}\right)$$
$$= 1 - \Phi(1.052470) = \begin{cases} 1 - 0.85314 = 14.686\% & \text{con le tavole} \\ 0.1462921 = 14.62921\% & \text{col comando } \texttt{1-pnorm} \text{ di R}\,. \end{cases}$$

Dal momento che la v.a. S_{45} è assolutamente continua (in quanto somma di v.a. continue), per calcolare la probabilità precedente *non* si deve fare nessuna correzione di continuità.
Gli stessi calcoli per la densità del suggerimento danno invece

$$\mathbb{P}\left(S_{45} > 22\right) \simeq 1 - \Phi\left(\frac{22 - 45 \cdot (1/2)}{\sqrt{45 \cdot (1/8)}}\right) = 1 - \Phi(-0.210819) = \Phi(0.210819)$$
$$= \begin{cases} 0.58317 = 58.317\% & \text{con le tavole} \\ 0.5834856 = 58.34856\% & \text{con R}\,. \end{cases}$$

(d) Per la densità corretta la probabilità richiesta è

$$\mathbb{P}\left(X_{45} > 1\right) = \int_1^{+\infty} f(x)\,\mathrm{d}x = \int_1^{+\infty} 0\,\mathrm{d}x = 0\,.$$

Per la densità del suggerimento invece

$$\mathbb{P}\left(X_{45} > 1\right) = \int_1^{+\infty} f(x)\,\mathrm{d}x = \int_1^{\frac{3}{2}} \frac{4}{9}\,(3 - 2x)\,\mathrm{d}x = \frac{4}{9}\left[3x - x^2\right]_{x=1}^{x=\frac{3}{2}} = \frac{1}{9}\,.$$

Problema 2. L'Associazione dei Librai ha commissionato un sondaggio per analizzare le abitudini di lettura nei giovani adolescenti. Ha partecipato al sondaggio un campione di 150 ragazzi. A ognuno di loro è stato chiesto il numero di libri letti nell'ultimo anno, ottenendo le risposte raggruppate nella tabella che segue:

libri letti	numero di risposte
0	42
1	51
2	15
3	21
4	12
5	6
6	0
7	1
8	2
9 o più	0

(a) Determinate la mediana, il primo e il terzo quartile e l'IQR dei dati precedenti. Rappresentate poi la distribuzione dei dati con un boxplot.

(b) Fornite una stima puntuale del valore atteso e della varianza della variabile aleatoria

$$X = \text{numero di libri letti da un giovane nell'ultimo anno}\,.$$

Fino a dieci anni fa, i giovani leggevano mediamente 1.95 libri a testa in un anno. Secondo i Librai, il sondaggio dimostra con forte evidenza che questo valore atteso è calato nell'ultimo anno.

(c) Impostate un opportuno test d'ipotesi al livello di significatività α per stabilire se i dati concordano con quanto sostenuto dall'Associazione dei Librai. Scrivete la statistica test e la regola di rifiuto, indicando in particolare se è necessario fare ipotesi sulla densità del campione.

(d) Calcolate il p-value del test precedente e traetene una conclusione. Si può essere d'accordo coi Librai nel sostenere che il numero medio di libri letti da un giovane è calato nell'ultimo anno?

(e) Fornite un limite superiore al livello di confidenza del 99% per la probabilità che un giovane non abbia letto nessun libro nell'ultimo anno.

Risultati.

(a) Per trovare i quantili richiesti, bisogna innanzitutto completare la tabella delle frequenze aggiungendo le frequenze relative e cumulate:

libri letti	FA	FR	FC
0	42	$42/150 = 0.28$	0.28
1	51	$51/150 = 0.34$	$0.34 + 0.28 = 0.62$
2	15	$15/150 = 0.1$	$0.62 + 0.1 = 0.72$
3	21	0.14	0.86
4	12	0.08	0.94
5	6	0.04	0.98
6	0	0	0.98
7	1	0.007	0.987
8	2	0.013	1
9 o più	0	0	1

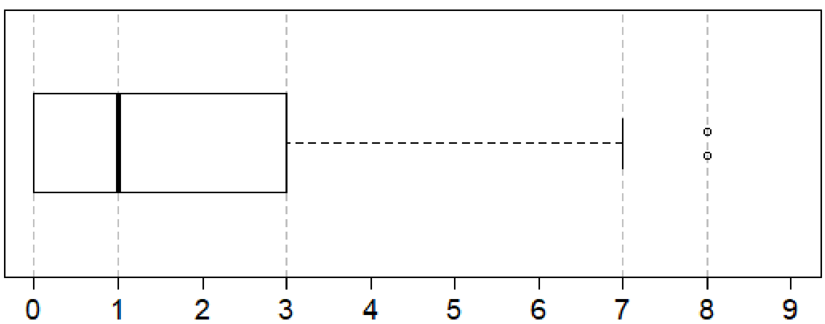
Dalla tabella vediamo che

$$\begin{array}{lcll} \mathrm{FC}(0) = 0.28 & & & \Rightarrow \quad \mathrm{Q1} = q_{0.25} = 0 \\ \mathrm{FC}(0) = 0.28 & \text{e} & \mathrm{FC}(1) = 0.62 & \Rightarrow \quad \mathrm{Q2} = m = q_{0.5} = 1 \\ \mathrm{FC}(2) = 0.72 & \text{e} & \mathrm{FC}(3) = 0.86 & \Rightarrow \quad \mathrm{Q3} = q_{0.75} = 3\,. \end{array}$$

Di conseguenza

$$\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1} = 3 \quad \Rightarrow \quad \mathrm{Q1} - 1.5 \cdot \mathrm{IQR} = -4.5\,, \quad \mathrm{Q3} + 1.5 \cdot \mathrm{IQR} = 7.5$$

e pertanto non ci sono outlier a sinistra, mentre i due dati nella classe $\{8\}$ sono outlier destri. Il boxplot è il seguente:



(b) Uno stimatore puntuale del valore atteso $\mu = \mathbb{E}(X)$ è la media campionaria, mentre uno stimatore della varianza $\sigma^2 = \mathrm{Var}(X)$ è la varianza campionaria. Le corrispondenti stime ricavate dai dati sono

$$\begin{aligned}
\bar{x} &= \sum_{k=0}^{8} k\,\mathrm{FR}(k) = 0 \cdot 0.28 + 1 \cdot 0.34 + \ldots + 8 \cdot 0.013 = 1.63333 \\
s^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \frac{1}{n-1}\left(\sum_{k=0}^{8} k^2\,\mathrm{FA}(k) - n\bar{x}^2\right) \\
&= \frac{1}{150-1}\left(0^2 \cdot 42 + 1^2 \cdot 51 + \ldots + 8^2 \cdot 2 - 150 \cdot 1.63333^2\right) = 2.81096\,.
\end{aligned}$$

(c) Se i librai vogliono sostenere senza ombra di dubbio che il valore atteso dei libri letti nell'ultimo anno è calato rispetto a dieci anni fa, devono mettere questa affermazione nell'ipotesi alternativa di un test:

$$H_0 : \mu = 1.95 =: \mu_0 \qquad \text{vs.} \qquad H_1 : \mu < \mu_0\,.$$

Abbiamo molti dati a disposizione ($n = 150 \gg 30$), dunque possiamo fare uno Z-test per un campione numeroso a varianza incognita *senza dover fare alcuna ipotesi sulla densità del campione.* La regola del test al livello α è

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

(d) Per trovare il p-value, dobbiamo innanzitutto realizzare la statistica test sui dati del sondaggio:

$$t_0 = \frac{\overline{x} - \mu_0}{s}\sqrt{n} = \frac{1.63333 - 1.95}{\sqrt{2.81096}}\sqrt{150} = -2.31324$$

e poi imporre l'uguaglianza nella regola di rifiuto:

$$\begin{aligned}
t_0 \equiv -z_{1-\alpha} \quad &\Leftrightarrow \quad -2.31324 = -z_{1-\alpha} \quad \Leftrightarrow \quad 2.31324 = z_{1-\alpha} \quad \Leftrightarrow \quad \Phi(2.31324) = \Phi(z_{1-\alpha}) = 1 - \alpha \\
&\Leftrightarrow \quad \alpha = 1 - \Phi(2.31324) = 1 - 0.98956 = 1.044\%\,.
\end{aligned}$$

Un p-value $= 1.044\%$ è molto basso, al di sotto degli usuali livelli di significatività del 5% e del 2.5%. Possiamo dunque rifiutare H_0 e concludere che $\mu < 1.95$, come sostenuto dai librai (conclusione forte).

(e) Si tratta ora di costruire un $IC_p(99\%)$ unilatero destro per il parametro p di un campione bernoulliano numeroso $Y_1, \ldots, Y_n$, dove $n = 150$ e

$$Y_i = \begin{cases} 1 & \text{se l'}i\text{-esimo intervistato non ha letto nessun libro} \\ 0 & \text{altrimenti} \end{cases} \qquad \text{con} \qquad Y_i \sim B(1,p)\,.$$

Al livello di confidenza $\gamma = 0.99$, troviamo dal formulario

$$p \in \left(0\,,\, \overline{y} + z_\gamma\sqrt{\frac{\overline{y}(1-\overline{y})}{n}}\right) = \left(0\,,\, 0.28 + 2.33\sqrt{\frac{0.28(1-0.28)}{150}}\right) = (0\,,\, 0.36542)$$

dove $\overline{y} = \mathrm{FR}(0) = 0.28$ non è altro che la frequenza relativa della classe $\{0\}$. Dunque il limite superiore cercato è il 36.542%.

Problema 3. Il Barone di Münchhausen vuole verificare sperimentalmente la formula che lega la gittata d di un cannone – cioè la distanza orizzontale percorsa dal proiettile prima di toccare terra – all'angolo di alzo θ. È noto infatti dalla Fisica che tale formula è

$$d = \frac{v_0^2}{g} \sin(2\,\theta)$$

dove $g = 9.81\,\mathrm{m/s^2}$ è il valore noto dell'accelerazione di gravità e v_0 è la velocità iniziale del proiettile (la medesima velocità per ogni tiro). Il Barone effettua dunque 50 tiri di prova, ognuno con un angolo di alzo diverso, e ne misura le corrispondenti gittate. Ottiene così i due vettori `d` (in metri) e `theta` (in radianti), dei quali si riporta lo scatterplot in Figura 1 della pagina seguente. I due vettori sono stati poi 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.)*
Il Barone ipotizza il seguente modello lineare gaussiano per legare tra loro le variabili `d` e `theta`:

$$\mathtt{d} = \beta_0 + \beta_1 \sin(2\cdot\mathtt{theta}) + E \quad \text{con} \quad E \sim N(0,\sigma^2)\,.$$

(Suggerimento: la funzione 'seno' è il comando `sin` di R.)

(a) Il modello del Barone spiega bene la variabilità dei dati? Perché?

(b) I dati rispettano le ipotesi gaussiane nel modello del Barone? Perché?

(c) Durante una delle 50 misure si è alzato un forte vento che ha deviato la traiettoria del proiettile. Individuate nello scatterplot dei residui l'outlier estremo che corrisponde a questa misura, indicando qual è la sua posizione $i \in \{1, 2, \ldots, 50\}$ nel vettore dei dati.

(d) Eliminate dal modello l'outlier che avete trovato nel punto precedente. Come cambiano le risposte ai punti (a) e (b) per il modello senza outlier?

Se avete risolto i punti (c), (d), proseguite col modello senza outlier. Altrimenti, continuate con tutti i dati.

(e) In un modello realistico, un tiro verticale (cioè con $\theta = \pi/2$) dovrebbe avere gittata attesa pari a zero. Questa condizione è rispettata dai dati nel modello del Barone? Decidetelo impostando un test opportuno, trovandone il p-value e commentando il risultato.

(f) Fornite un intervallo di confidenza bilatero al livello del 95% per la velocità iniziale v_0 dei proiettili sparati dal Barone.

(g) Per il prossimo tiro, l'alzo del cannone è stato regolato a $\theta = \pi/4$. Con questa inclinazione, il Barone vuol sapere che distanza può raggiungere salendo a cavallo del proiettile. Aiutatelo fornendo una previsione puntuale per tale distanza. *(Trascurate l'effetto del Barone sul moto del proiettile).*

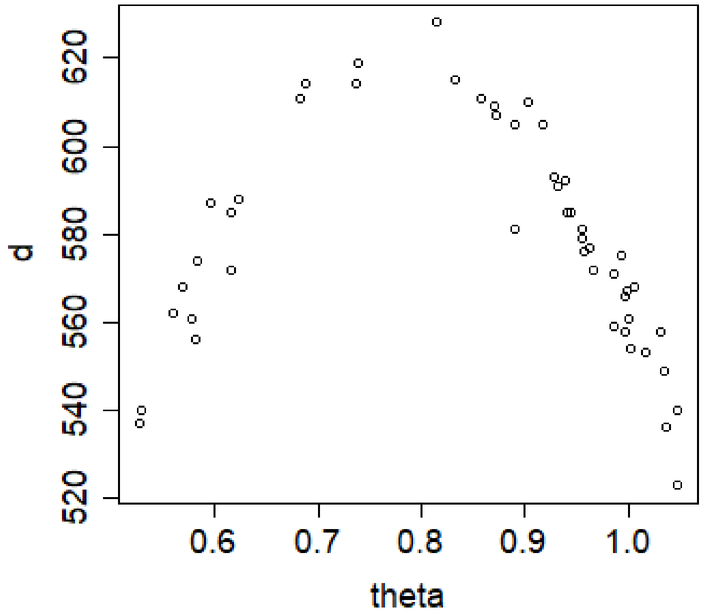


Figura 1: Scatterplot dei dati

Risultati.

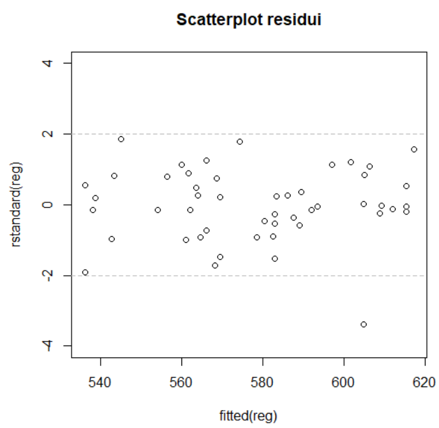
(a) La percentuale di variabilità spiegata dal modello di regressione semplice è il suo r^2. Per trovarlo, dobbiamo innanzitutto fare il fit dei dati con R e visualizzarne la summary:

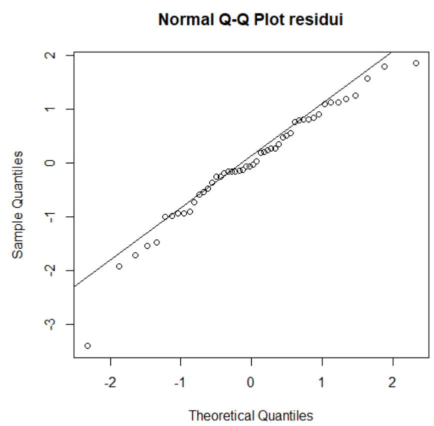
```
> reg <- lm( d ~ sin(2*theta), data = dati )
> summary(reg)

Call:
lm(formula = d ~ sin(2 * theta), data = dati)

Residuals:
     Min       1Q   Median       3Q      Max
-23.8028  -3.6994  -0.3012   5.5579  12.9027

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.027     24.398   0.206    0.838
sin(2 * theta)  613.144     26.107  23.485   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.175 on 48 degrees of freedom
Multiple R-squared:  0.9199,    Adjusted R-squared:  0.9183
F-statistic: 551.6 on 1 and 48 DF,  p-value: < 2.2e-16
```

Vediamo che la percentuale di variabilità spiegata è $r^2 = 0.9199 = 91.99\%$, che è un valore elevato. Dunque il modello spiega molto bene la variabilità dei dati.

(b) Per vedere se i dati rispettano le ipotesi gaussiane nel modello del Barone, bisogna verificare l'omoschedasticità e la gaussianità dei residui. Disegniamo innanzitutto lo scatterplot e il normal Q-Q plot dei residui standardizzati:

```
> plot(fitted(reg), rstandard(reg), ylim = c(-4,4), main = "Scatterplot residui")
> abline(h = c(-2,2), col = "gray75", lty = 2)
> qqnorm(rstandard(reg), main = "Normal Q-Q Plot residui")
> qqline(rstandard(reg))
```





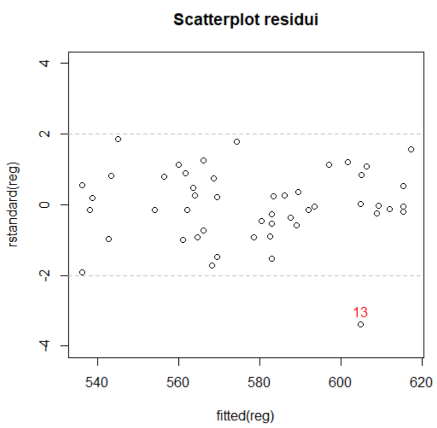
Dallo scatterplot vediamo che l'omoschedasticità è soddisfatta. Nel normal Q-Q plot i punti si allineano abbastanza bene sulla Q-Q line (a eccezione del punto in basso a sinistra, che verosimilmente corrisponde all'unico outlier nello scatterplot) e l'ipotesi di gaussianità sembra dunque soddisfatta. A conferma di questo, svolgiamo il test di Shapiro-Wilk sui residui:

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

        Shapiro-Wilk normality test

data:  rstandard(reg)
W = 0.9645, p-value = 0.1371
```

Benché il p-value del 13.71% non sia altissimo, è comunque maggiore delle usuali soglie del 5 - 10%. Dunque non possiamo rifiutare l'ipotesi nulla di gaussianità dei residui.

(c) Identifichiamo l'outlier estremo nello scatterplot dei residui utilizzando il comando `identify`:

```
> plot(fitted(reg), rstandard(reg), ylim = c(-4,4), main = "Scatterplot residui")
> abline(h = c(-2,2), col = "gray75", lty = 2)
> identify(fitted(reg), rstandard(reg), col = "red")
```



Quindi, l'outlier corrisponde alla misura di posto $i = 13$. Alternativamente, possiamo usare il comando `which` per trovare l'indice i del residuo r_i con $|r_i| > 2$:

```
> which(abs(rstandard(reg)) > 2)
13
13
```

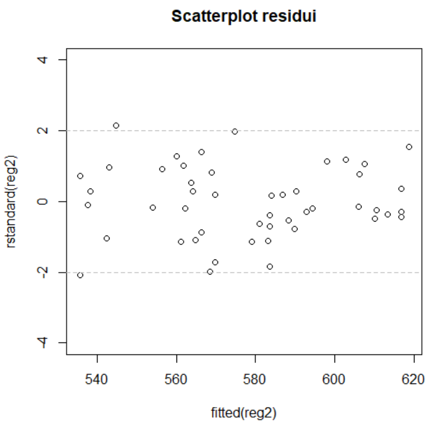
(13 è ripetuto due volte perché `rstandard(reg)` è un *named vector*).

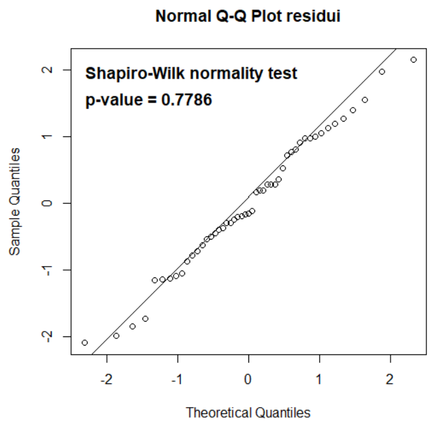
(d) Rifacciamo il fit dei dati eliminando il dato di posto 13:

```
> reg2 <- lm( d[-13] ~ sin(2*theta[-13]), data = dati )
> summary(reg2)
```

```
Call:
lm(formula = d[-13] ~ sin(2 * theta[-13]), data = dati)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.6541  -3.8874  -0.9551   5.0959  13.1809 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)           -8.191     21.761  -0.376    0.708    
sin(2 * theta[-13])  627.835     23.308  26.936   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.32 on 47 degrees of freedom
Multiple R-squared:  0.9392,    Adjusted R-squared:  0.9379 
F-statistic: 725.6 on 1 and 47 DF,  p-value: < 2.2e-16
```

Vediamo che la variabilità spiegata è leggermente aumentata rispetto al modello con l'outlier (93.92% contro 91.99%). Tuttavia, ora non c'è più alcun dubbio sulla gaussianità dei residui, come si vede dallo scatterplot senza outlier, dal normal Q-Q plot aderente alla Q-Q line e dal p-value nettamente più alto del test di Shapiro-Wilk:





(e) Quando $\theta = \pi/2$, il modello prevede

$$D = \beta_0 + \beta_1 \sin(2 \cdot \pi/2) + E = \beta_0 + E\,, \qquad E \sim N(0, \sigma^2)$$

e dunque $\mathbb{E}(D) = \beta_0 + \mathbb{E}(E) = \beta_0$. Si tratta dunque di fare un test per le ipotesi

$$H_0 : \beta_0 = 0 \qquad \text{vs.} \qquad H_1 : \beta_0 \neq 0\,.$$

Dall'output di R, leggiamo che il p-value del T-test che confronta queste due ipotesi è $0.708 = 70.8\%$ nel modello senza outlier (con l'outlier: 83.8%). Essendo un valore estremamente alto, non abbiamo nessun motivo per rifiutare H_0, e quindi il modello è realistico.

(f) Nella relazione fra D e θ, il coefficiente angolare è

$$\beta_1 = \frac{v_0^2}{g} \qquad \Rightarrow \qquad v_0 = \sqrt{g\beta_1}\,.$$

Poiché $g = 9.81$ è nota, si tratta di ricavare un intervallo di confidenza per β_1 e di trasformarlo poi in uno per v_0. Dal formulario sappiamo che un $IC_{\beta_1}(95\%)$ è

$$\beta_1 \in \left(\hat{\beta}_1 \pm t_{\frac{1+\gamma}{2}}(n-k-1)\,\mathrm{se}(\hat{\beta}_1)\right) = \left(\hat{\beta}_1 \pm t_{\frac{1+0.95}{2}}(50-1-1)\,\mathrm{se}(\hat{\beta}_1)\right) = (627.835 \pm 2.0106 \cdot 23.308)$$
$$= (580.972\,,\, 674.698)$$

dove $\gamma = 0.95$, $k = 1$ è il numero di predittori e $t_{\frac{1+0.95}{2}}(50-1-1) = t_{0.975}(48) = 2.0106$ (ottenuto col comando `qt(p=0.975, df=48)` di R) oppure $\simeq t_{0.975}(50) = 2.0086$ (con le tavole). L'intervallo per v_0 è dunque

$$v_0 \in \left(\sqrt{9.81 \cdot 580.972}\,,\, \sqrt{9.81 \cdot 674.698}\right) = (75.494\,,\, 81.356)$$

ovviamente in m/s (con l'outlier: $(74.162\,,\, 80.808)$).

(g) Usando il modello proposto, una previsione puntuale per la gittata D in corrispondenza dell'alzo $\theta = \pi/4$ è

$$\hat{d} = \hat{\beta}_0 + \hat{\beta}_1 \sin\left(2 \cdot \frac{\pi}{4}\right) = -8.191 + 627.835 \cdot 1 = 619.644$$

(con l'outlier: 618.171).