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

I APPELLO DI STATISTICA PER INGEGNERIA FISICA
30 Giugno 2020

Problema 1. Anche alla Scuola di Magia e Stregoneria di Hogwarts quest'anno gli esami finali si svolgeranno in modalità a distanza. Durante le interrogazioni, gli studenti comunicheranno col loro professore attraverso la Metropolvere, cioè la rete che collega i camini di tutte le abitazioni della comunità magica. Per connettersi, gli studenti getteranno una manciata di Polvere Volante nel focolare di casa, ci infileranno la testa dentro e... in un attimo, la Metropolvere li metterà in contatto coll'insegnante.

Lo svantaggio di questo sistema è che ogni camino perde la connessione dalla Metropolvere dopo un tempo aleatorio. Di tale tempo, è noto soltanto che **ha densità esponenziale** e che **il suo valore atteso è pari a un'ora esatta.** Si sa anche che i tempi di connessioni diverse sono **tutti indipendenti fra loro.**

(a) Domani Harry Potter sarà interrogato a distanza dal prof. Filius Vitious. Per ogni $t > 0$, calcolate la probabilità che, trascorse t ore dall'inizio dell'interrogazione,

   (i) la connessione di Harry funzioni ancora;

   (ii) sia la connessione di Harry sia quella del prof. Vitious funzionino ancora **entrambe**.

(b) Sia T il tempo trascorso dall'inizio dell'interrogazione al momento in cui si interrompe il collegamento tra Harry e il professore perché cade la connessione di almeno uno dei due. Determinate la funzione di ripartizione F_T della variabile aleatoria T.

$$\left(\textit{Se non ci riuscite, d'ora in poi usate } \; F_T(t) = \begin{cases} 0 & \textit{se } t < 0\,, \\ 2\,t & \textit{se } 0 \leq t < 1/2\,, \\ 1 & \textit{se } t \geq 1/2\,. \end{cases} \right)$$

(c) Ricavate la densità di T e determinatene il valore atteso.

(d) Il prof. Vitious vuole ridurre al 30% la probabilità che il collegamento con Harry si interrompa mentre lo sta ancora interrogando. Se è così, quale dev'essere la durata massima t_0 dell'interrogazione?

(e) Domani il prof. Vitious esaminerà in tutto 50 studenti, e ha già deciso che l'interrogazione di ognuno durerà esattamente t_0 ore (dove t_0 è fissato al punto precedente). Qual è la probabilità approssimata che con almeno 20 di loro il collegamento si interrompa prima della fine dell'interrogazione?

Risultati.

(a) (i) Sia T_1 l'istante in cui Harry perde la connessione. Sappiamo che $T_1 \sim \mathcal{E}(1/1) = \mathcal{E}(1)$, e quindi la probabilità richiesta è

$$\mathbb{P}(T_1 > t) = \int_t^{+\infty} f_{T_1}(z)\mathrm{d}z = \int_t^{+\infty} \mathrm{e}^{-z}\mathrm{d}z = \mathrm{e}^{-t}\,.$$

(ii) Sia T_2 l'istante in cui il prof. Vitious perde la connessione. Sappiamo che anche $T_2 \sim \mathcal{E}(1/1) = \mathcal{E}(1)$ e che T_1 e T_2 sono indipendenti. Quindi, la probabilità richiesta è

$$\mathbb{P}(\text{“}\,T_1 > t\text{”} \wedge \text{“}\,T_2 > t\text{”}) = \mathbb{P}(T_1 \geq t) \cdot \mathbb{P}(T_2 \geq t) = \mathrm{e}^{-t} \cdot \mathrm{e}^{-t} = \mathrm{e}^{-2t}\,.$$

Alternativamente, sia X_t il numero di connessioni ancora attive al tempo t tra le due di Harry e del prof. Vitious. Allora, $X_t \sim B(2, q)$, dove $q = \mathrm{e}^{-t}$ è la probabilità trovata al punto (i). La probabilità richiesta si può dunque calcolare anche come

$$\mathbb{P}(X_t = 2) = p_{X_t}(2) = \binom{2}{2} q^2(1-q)^{2-2} = q^2 = \mathrm{e}^{-2t}\,.$$

(b) Per l'istante T in cui si interrompe il collegamento tra Harry e il professore, abbiamo l'uguaglianza di eventi

$$\text{“} T > t \text{”} = \text{“} T_1 > t \text{”} \wedge \text{“} T_2 > t \text{”}$$

e dunque per $t > 0$ la funzione di ripartizione è

$$F_T(t) = \mathbb{P}(T \le t) = 1 - \mathbb{P}(T > t) = \mathbb{P}(\text{“} T_1 > t \text{”} \wedge \text{“} T_2 > t \text{”}) = 1 - \mathrm{e}^{-2t}.$$

Alternativamente, si può usare l'uguaglianza di eventi

$$\text{“} T \le t \text{”} = \text{“} X_t \le 1 \text{”},$$

dove X_t è il numero di connessioni ancora attive al tempo t introdotto al punto precedente. Di conseguenza,

$$F_T(t) = \mathbb{P}(T \le t) = \mathbb{P}(X_t \le 1) = 1 - \mathbb{P}(X_t = 2) = 1 - q^2 = 1 - \mathrm{e}^{-2t},$$

che è lo stesso risultato. Ovviamente, per $t \le 0$ abbiamo

$$F_T(t) = \mathbb{P}(T \le t) = \mathbb{P}(\emptyset) = 0.$$

(c) Per ricavare la densità di T, deriviamo la sua funzione di ripartizione:

$$f_T(t) = \frac{\mathrm{d}}{\mathrm{d}t} F_T(t) = \begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}(1 - \mathrm{e}^{-2t}) = 2\mathrm{e}^{-2t} & \text{se } t > 0 \\ \frac{\mathrm{d}}{\mathrm{d}t} 0 = 0 & \text{se } t \le 0 \end{cases}$$

Nella funzione $f_T(t) = 2\mathrm{e}^{-2t}\mathbb{1}_{(0,+\infty)}(t)$, riconosciamo la densità esponenziale di parametro 2, cioè $T \sim \mathcal{E}(2)$. Dal formulario, troviamo che il suo valore atteso è $\mathbb{E}[T] = 1/2$.
Se invece avessimo usato la funzione di ripartizione data nel suggerimento, avremmo ottenuto

$$f_T(t) = \frac{\mathrm{d}}{\mathrm{d}t} F_T(t) = \begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} 0 = 0 & \text{se } t < 0, \\ \frac{\mathrm{d}}{\mathrm{d}t} 2\,t = 2 & \text{se } 0 \le t < 1/2, \\ \frac{\mathrm{d}}{\mathrm{d}t} 1 = 0 & \text{se } t \ge 1/2, \end{cases}$$

che è la densità uniforme continua $\mathcal{U}([0,\, 1/2])$. In questo caso, il valore atteso di T è

$$\mathbb{E}[T] = \frac{1}{2}\left(0 + \frac{1}{2}\right) = \frac{1}{4}.$$

(d) Il prof. Vitious vuole trovare t_0 tale che $\mathbb{P}(T \le t_0) = 30\%$. Per farlo, deve risolvere l'equazione

$$F_T(t_0) = 0.3 \quad \Rightarrow \quad \begin{cases} 1 - \mathrm{e}^{-2t} = 0.3 \quad \Rightarrow \quad t = -\frac{1}{2}\ln(1 - 0.3) = 0.1783 \quad \text{con la } F_T \text{ corretta,} \\ 2\,t = 0.3 \quad \Rightarrow \quad t = \frac{0.3}{2} = 0.15 \quad \text{con la } F_T \text{ del suggerimento.} \end{cases}$$

(e) Sia Y il numero di studenti con cui si interrompe il collegamento prima della fine dell'interrogazione. Allora $Y \sim B(50, 0.3)$, dove 0.3 è la probabilità del punto precedente. Vogliamo calcolare

$$\mathbb{P}(Y \ge 20) = \mathbb{P}(Y \ge 19.5) = \mathbb{P}\left(\underbrace{\frac{Y - np}{\sqrt{np(1-p)}}}_{\approx N(0,1)} \ge \frac{19.5 - 50 \cdot 0.3}{\sqrt{50 \cdot 0.3 \cdot (1 - 0.3)}}\right)$$

$$\simeq 1 - \Phi\left(\frac{19.5 - 50 \cdot 0.3}{\sqrt{50 \cdot 0.3 \cdot (1 - 0.3)}}\right) = 1 - \Phi(1.3887) = 1 - 0.91774 = 8.226\%$$

(senza correzione di continuità: $\mathbb{P}(Y \ge 20) = 1 - \Phi(1.5430) = 1 - 0.93822 = 6.178\%$), dove abbiamo potuto usare l'approssimazione gaussiana della binomiale perché $n > 20$, $np = 15 > 5$ e $n(1-p) = 35 > 5$.

Problema 2. Carlo è un appassionato corridore. Di recente, però, ha dovuto sospendere gli allenamenti per circa due mesi, e di conseguenza teme che adesso le sue prestazioni siano sensibilmente peggiorate.

D'altra parte, Carlo è bravo in Statistica almeno quanto lo era nella corsa, e per questo tiene un *database* aggiornato di tutti gli allenamenti. Per lui, un allenamento si svolge sempre correndo 30 giri di pista consecutivi e senza interruzioni. Carlo ha una grandissima resistenza, dunque le durate dei vari giri si possono supporre indipendenti e identicamente distribuite fra loro. Terminati i 30 giri, Carlo registra nel suo *database* il tempo complessivamente impiegato per correrli tutti, ottenendo così un dato solo per ogni allenamento.

Ecco i tempi registrati in questo modo negli ultimi 5 allenamenti fatti prima della sosta forzata, con sotto quelli ottenuti invece nei primi 4 allenamenti subito dopo la sosta (in minuti):

	Tempo di un allenamento (minuti complessivi)				
Prima della sosta	63.5	58.0	65.7	61.9	67.1
Dopo la sosta	68.9	67.1	72.0	70.3	

(a) Potete fare delle ipotesi sulla densità da cui sono stati estratti i dati precedenti? Se sì, quali ipotesi e perché?

(b) Impostate un test per stabilire dai dati se la varianza dei tempi di allenamento sia significativamente cambiata dopo la sosta. Qual è il risultato del test al livello di significatività del 20%?

(c) Determinate un limite inferiore per il p-value del test precedente, e traetene una conclusione.

(d) Impostate un altro test, questa volta sul valore atteso, per stabilire se il timore di Carlo è fondato, cioè se vi è evidenza dai dati che le sue prestazioni siano significativamente peggiorate dopo la sosta.

(e) Calcolate un intervallo in cui cade il p-value del test precedente, e traetene una conclusione.

(f) Carlo vuol stimare il tempo medio μ che impiega per completare un allenamento adesso che la sosta è finita. Coi dati a disposizione, aiutatelo a fornire una stima puntuale $\hat{\mu}$ per il parametro μ. Per tale stima, quanto siete confidenti che $|\hat{\mu} - \mu| < 1.704\,\mathrm{min}$?

Risultati.

(a) È ragionevole supporre che entrambi i campioni siano stati estratti da una densità gaussiana. Infatti, il tempo T di un allenamento è la somma $T = T_1 + T_2 + \ldots + T_{30}$, in cui T_i è il tempo impiegato per percorrere l'i-esimo giro. Poiché le variabili aleatorie $T_1, \ldots, T_{30}$ sono i.i.d. e $n = 30$ è grande, la densità di T è approssimativamente gaussiana per il TLC.

(b) Indichiamo con $X_1, \ldots, X_m$ il campione dei tempi registrati negli ultimi $m = 5$ allenamenti prima della sosta e con $Y_1, \ldots, Y_n$ quello dei primi $n = 4$ allenamenti dopo la sosta. Sappiamo dal punto precedente che $X_i \sim N(\mu_X, \sigma_X^2)$ e $Y_j \sim N(\mu_Y, \sigma_Y^2)$. Ora vogiliamo decidere tra le ipotesi statistiche

$$H_0 : \sigma_X^2 = \sigma_Y^2 \qquad \text{contro} \qquad H_1 : \sigma_X^2 \neq \sigma_Y^2 \,.$$

Un test di livello α per tali ipotesi è dato dalla regola

$$\text{“ rifiuto } H_0 \text{ se trovo } F_0 := \frac{S_X^2}{S_Y^2} < f_{\frac{\alpha}{2}}(m-1, n-1) \text{ oppure } F_0 > f_{1-\frac{\alpha}{2}}(m-1, n-1)\text{”}.$$

Coi dati a disposizione, troviamo le realizzazioni

$$\bar{x} = \frac{63.5 + \ldots + 67.1}{5} = 63.24 \qquad s_X^2 = \frac{(63.5 - 63.24)^2 + \ldots + (67.1 - 63.24)^2}{5-1} = 12.548$$

$$\bar{y} = \frac{68.9 + \ldots + 70.3}{4} = 69.575 \qquad s_Y^2 = \frac{(68.9 - 69.575)^2 + \ldots + (70.3 - 69.575)^2}{4-1} = 4.329 \,.$$

e dunque

$$f_0 = \frac{s_X^2}{s_Y^2} = \frac{12.548}{4.329} = 2.903\,.$$

Al livello $\alpha = 20\%$, dobbiamo confrontare quest'ultimo valore coi quantili

$$\begin{aligned} f_{\frac{\alpha}{2}}(m-1, n-1) &= f_{0.10}(4,3) = \frac{1}{f_{0.90}(3,4)} = \frac{1}{4.191} = 0.239\,, \\ f_{1-\frac{\alpha}{2}}(m-1, n-1) &= f_{0.90}(4,3) = 5.343\,. \end{aligned}$$

Poiché

$$0.239 < f_0 = 2.903 < 5.343\,,$$

non possiamo rifiutare H_0.

(c) Nel punto precedente, abbiamo accettato H_0 al livello $\alpha = 20\%$. Ne segue che p-value $> 20\%$. Con un p-value così elevato, non c'è nessuna evidenza a favore di H_1 (conclusione debole).

(d) Vogliamo verificare se vi è evidenza dai dati che le prestazioni di Carlo siano significativamente peggiorate dopo la sosta, cioè che $\mu_X < \mu_Y \Rightarrow$ mettiamo quest'ultima affermazione nell'ipotesi alternativa del test:

$$H_0 : \mu_X = \mu_Y \qquad \text{contro} \qquad H_1 : \mu_X < \mu_Y\,.$$

Per le conclusioni che abbiamo tratto nei punti (a) e (b), possiamo fare un test per la differenza dei valori attesi di due campioni gaussiani a varianze incognite ma uguali. Un tale test al livello α ha la seguente regola:

$$\text{“ rifiuto } H_0 \text{ se trovo } T_0 := \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}} < -t_{1-\alpha}(m+n-2)\text{”}$$

in cui

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

(e) Per calcolare il p-value dei dati trovati, calcoliamo la realizzazione delle statistica test del punto precedente:

$$s_p^2 = \frac{(5-1)\cdot 12.548 + (4-1)\cdot 4.329}{5+4-2} = 9.037\,, \qquad t_0 = \frac{63.24 - 69.575}{\sqrt{9.037\cdot\left(\frac{1}{5}+\frac{1}{4}\right)}} = -3.141$$

e uguagliamola al quantile nella regione di rifiuto:

$$t_0 \equiv -t_{1-\alpha}(m+n-2) \quad \Leftrightarrow \quad -3.141 = -t_{1-\alpha}(7) \quad \Leftrightarrow \quad 3.141 = t_{1-\alpha}(7)\,.$$

Sulle tavole troviamo che

$$t_{0.99}(7) = 2.9979 < 3.141 < 3.4995 = t_{0.995}(7)$$

e quindi

$$0.99 < 1-\alpha < 0.995 \quad \Leftrightarrow \quad 0.5\% < \alpha < 1\%\,.$$

In altre parole, il p-value è compreso fra lo 0.5% e l'1%. Con un p-value così piccolo, possiamo tranquillamente rifiutare H_0 a tutti i livelli di significatività sensati (conclusione forte).

(f) Il tempo medio μ di cui si richiede una stima non è altro che il parametro μ_Y. Un suo stimatore non distorto e consistente in media quadratica è la media campionaria $\bar{Y}$. La stima corrispondente è la realizzazione $\bar{y} = 69.575$.
Un intervallo di confidenza di livello γ per il parametro μ_Y è

$$\begin{aligned} \mu_Y \in &\left(\bar{y} - t_{\frac{1+\gamma}{2}}(n-1)\,\frac{s_Y}{\sqrt{n}}\,,\, \bar{y} + t_{\frac{1+\gamma}{2}}(n-1)\,\frac{s_Y}{\sqrt{n}}\right) \\ \Leftrightarrow &\quad |\bar{y} - \mu_Y| \leq t_{\frac{1+\gamma}{2}}(n-1)\,\frac{s_Y}{\sqrt{n}}\,. \end{aligned}$$

Se imponiamo l'uguaglianza

$$t_{\frac{1+\gamma}{2}}(n-1)\,\frac{s_Y}{\sqrt{n}} \equiv 1.704\,,$$

allora siamo confidenti al livello γ che $|\bar{y}-\mu_Y| \leq 1.704$. Ne ricaviamo

$$\begin{aligned} t_{\frac{1+\gamma}{2}}(n-1) = 1.704 \cdot \frac{\sqrt{n}}{s_Y} \quad &\Leftrightarrow \quad t_{\frac{1+\gamma}{2}}(3) = 1.704 \cdot \sqrt{\frac{4}{4.329}} = 1.6379 \simeq 1.6377 = t_{0.9}(3) \\ &\Leftrightarrow \quad \frac{1+\gamma}{2} \simeq 0.9 \quad \Leftrightarrow \quad \gamma = 0.8\,. \end{aligned}$$

In altre parole, coi dati trovati siamo confidenti all'80% che $|\bar{y}-\mu_Y| \leq 1.704$.

Problema 3. Il signor Hamlin è il titolare di *Safarà*, una catena di negozi di bric-à-brac sparsi nel centro di Londra. Per valutare l'apertura di nuove sedi, Hamlin sta cercando un legame tra l'area espositiva (variabile `area`, in metri quadrati) e l'incasso settimanale (variabile `incasso`, in migliaia di sterline) dei suoi punti-vendita. A questo scopo, ha fatto raccogliere i dati della settimana scorsa in tutti i negozi della catena. I due vettori coi valori trovati sono 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)

(a) Disegnate lo scatterplot dei dati di Hamlin, considerando `incasso` come variabile di output.

Dopo aver visto lo scatterplot che avete disegnato, Hamlin ipotizza il seguente legame tra le variabili `area` e `incasso`:

$$\texttt{incasso} = \beta_0 + \beta_1\,\texttt{area} + \beta_2\,\texttt{area}^2 + E \quad \text{con} \quad E \sim N(0, \sigma^2)\,,$$

in cui ovviamente gli incassi di negozi diversi sono tutti indipendenti tra loro.

(b) Il modello di Hamlin spiega bene la variabilità dei dati? Perché?

(c) I dati rispettano le ipotesi gaussiane nel modello di Hamlin? Perché?

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

(e) Proponete voi un nuovo modello, e sottoponetelo alle stesse verifiche che avete fatto per quello di Hamlin nei punti (b), (c) e (d). È migliore il vostro modello o quello di Hamlin?

(f) Fornite una previsione puntuale per l'incasso settimanale di un negozio di $95\,\mathrm{m}^2$.

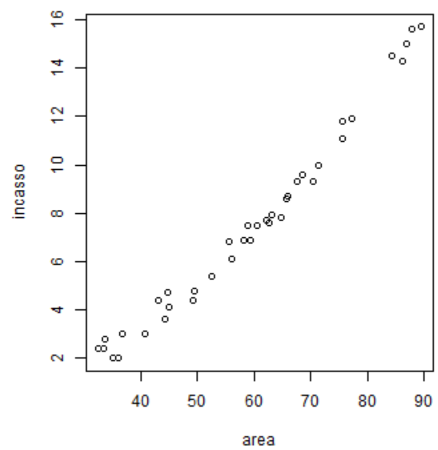
(g) Naturalmente, in un modello realistico l'incasso settimanale atteso per un negozio con area pari a $0\,\mathrm{m}^2$ dovrebbe essere del tutto nullo. Impostate un opportuno test per stabilire se questa condizione è rispettata dal modello che avete proposto. Determinate il p-value del test e traetene una conclusione.

Risultati.

(a) Col comando

```
> plot(incasso, area)
```

otteniamo il grafico



(b) Inserendo i comandi

```
> fit <- lm(incasso ~ area + I(area^2))
> summary(fit)
```

otteniamo l'output

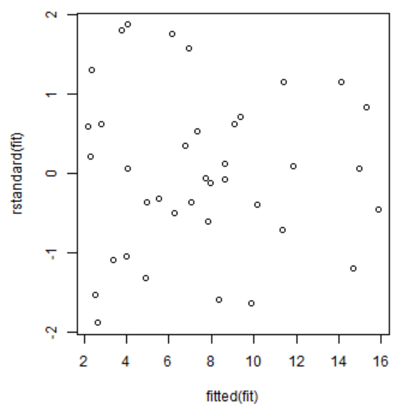
```
Call:
lm(formula = incasso ~ area + I(area^2))

Residuals:
     Min       1Q   Median       3Q      Max
-0.65555 -0.21143 -0.00216  0.22189  0.67153

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1631380  0.7378518   0.221    0.826
area        -0.0017146  0.0256998  -0.067    0.947
I(area^2)    0.0019816  0.0002123   9.333 4.99e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3689 on 35 degrees of freedom
Multiple R-squared:  0.992,     Adjusted R-squared:  0.9916
F-statistic:  2184 on 2 and 35 DF,  p-value: < 2.2e-16
```

La percentuale di variabilità spiegata dal modello di Hamlin è data dall'r^2_{adj}, che nell'output è pari al 99.16%. Dunque il modello di Hamlin spiega estremamente bene la variabilità dei dati.

(c) Le ipotesi gaussiane sono verificate se i residui standardizzati del modello di Hamlin risultano omoschedastici e gaussiani. Per l'omoschedasticità, dobbiamo guardare lo scatterplot dei residui, che otteniamo col comando

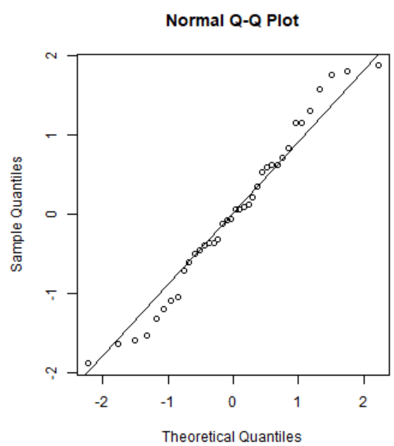
```
> plot(fitted(fit), rstandard(fit))
```

Ecco il risultato:



da cui si vede che i residui sono omoschedastici. Per quanto riguarda invece la gaussianità, dobbiamo anzitutto osservare il normal Q-Q plot dei residui standardizzati, ottenuto coi comandi

```
> qqnorm(rstandard(fit))
> qqline(rstandard(fit))
```



Vediamo che i punti si allineano bene sulla Q-Q line. Per confermare questo fatto, svolgiamo il test di Shapiro Wilks sui residui

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

        Shapiro-Wilk normality test

data:  rstandard(fit)
W = 0.97538, p-value = 0.5555
```

Il p-value del 55.55% è molto elevato, dunque non abbiamo nessun motivo per rifiutare l'ipotesi nulla di gaussianità dei residui. Concludiamo che il modello di Hamlin soddisfa le ipotesi gaussiane.

(d) Nel modello di Hamlin, il regressore `area` non è significativo, in quanto il p-value del T-test corrispondente è pari al 94.7% e dunque non possiamo rifiutare l'ipotesi nulla $H_0 : \beta_1 = 0$.

(e) Riproponiamo il modello di Hamlin, ma questa volta eliminandone il regressore `area` non significativo:

$$\texttt{incasso} = \beta_0 + \beta_1\, \texttt{area}^2 + E \quad \text{con} \quad E \sim N(0, \sigma^2)\,.$$

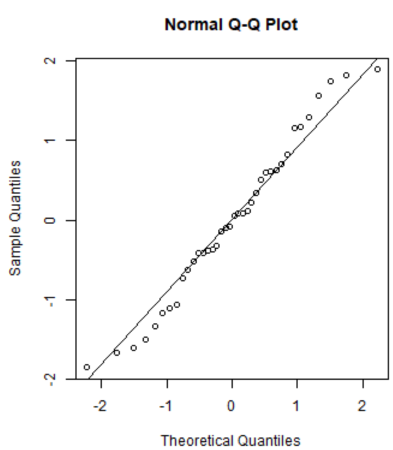
Anche per questo nuovo modello, visualizziamo il summary di R:

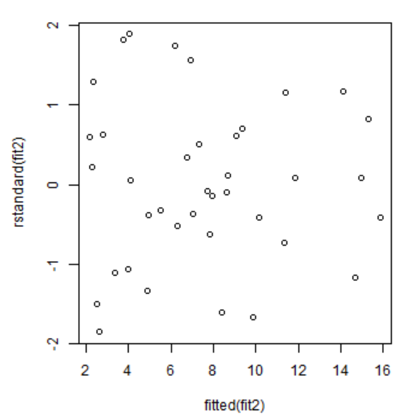
```
> fit2 <- lm(incasso ~ I(area^2))
> summary(fit2)
```

```
Call:
lm(formula = incasso ~ I(area^2))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.65054 -0.21523 -0.00422  0.22064  0.67144 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.147e-01  1.257e-01   0.912    0.368    
I(area^2)   1.968e-03  2.936e-05  67.018   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3637 on 36 degrees of freedom
Multiple R-squared:  0.992,	Adjusted R-squared:  0.9918 
F-statistic:  4491 on 1 and 36 DF,  p-value: < 2.2e-16
```

Si vede che il nuovo modello ha praticamente la stessa percentuale di variabilità spiegata del modello di Hamlin, con la differenza però che in questo caso l'unico regressore è significativo. Dobbiamo ancora verificare le ipotesi gaussiane per il nuovo modello:

```
> plot(fitted(fit2), rstandard(fit2))
> qqnorm(rstandard(fit2))
> qqline(rstandard(fit2))
> shapiro.test(rstandard(fit2))

        Shapiro-Wilk normality test

data:  rstandard(fit2)
W = 0.97541, p-value = 0.5565
```





Vediamo che gli scatterplot, il normal Q-Q plot e il p-value del test di Shapiro-Wilk sui residui sono praticamente gli stessi del modello di Hamlin. Ne concludiamo che, a parità di tutte le altre caratteristice, il nuovo modello è migliore di quello di Hamlin in quanto ora l'unico regressore è significativo.

(f) Usiamo il nostro modello per ottenere una previsione puntuale in corrispondenza del nuovo input $x^* = 95$:

$$y^* = \hat{\beta}_0 + \hat{\beta}_1 \, {x^*}^2 = 0.1147 + 0.001968 \cdot 95^2 = 17.8759\,.$$

(g) In base al nostro modello, l'incasso settimanale atteso per un negozio con area pari a zero è

$$\mathbb{E}[Y \mid x = 0] = \beta_0 + \beta_1 0^2 = \beta_0 .$$

Si chiede dunque di fare un test per le ipotesi

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

Il p-value di un T-test per tali ipotesi si trova subito nel summary di R: p-value $= 36.80\%$. Con un p-value così elevato, non possiamo rifiutare l'ipotesi nulla $\beta_0 = 0$ a nessun livello di significatività ragionevole.