

Capitolo 6

La verifica di ipotesi e gli intervalli di confidenza

6.1. La verifica di ipotesi

Sulla base del campione osservato si è interessati a stabilire se il vero valore del parametro appartiene ad un certo sottoinsieme dello spazio parametrico Θ , ovvero l'insieme di tutti i valori plausibili per il parametro θ . Dato un modello statistico, se gli insiemi Θ_0 e Θ_1 costituiscono una partizione di Θ , la verifica di ipotesi consiste in un procedimento decisionale di scelta fra l'ipotesi di base $H_0 : \theta \in \Theta_0$ e l'ipotesi alternativa $H_1 : \theta \in \Theta_1$. L'insieme delle ipotesi ammissibili e la sua partizione in H_0 e H_1 è detto sistema di ipotesi.

• **Esempio 6.1.1.** Nella semplice situazione in cui si ha un campione casuale da una variabile casuale X , il tipico modello classico assume che $X \sim N(\mu, \sigma^2)$ e l'usuale sistema di ipotesi consiste nel verificare $H_0 : \mu = \mu_0$ contro $H_1 : \mu \neq \mu_0$. Dal momento che lo spazio parametrico in questo caso è dato da

$$\Theta = \{(\mu, \sigma^2) : \mu \in]-\infty, \infty[, \sigma^2 \in]0, \infty[\},$$

risulta $\Theta_0 = \{(\mu, \sigma^2) : \mu = \mu_0, \sigma^2 \in]0, \infty[\}$ e $\Theta_1 = \{(\mu, \sigma^2) : \mu \neq \mu_0, \sigma^2 \in]0, \infty[\}$. Al contrario, in un approccio “distribution-free”, si assume solo che X sia una variabile casuale continua con funzione di ripartizione F e mediana λ , mentre il sistema di ipotesi risulta $H_0 : \lambda = \lambda_0$ contro $H_1 : \lambda \neq \lambda_0$. In questo caso, lo spazio “parametrico” è dato da

$$\Theta = \{(\lambda, F) : \lambda \in]-\infty, \infty[, F \in \mathcal{R}_c \},$$

dove \mathcal{R}_c è lo spazio delle funzioni di ripartizione continue. Si ha $\Theta_0 = \{(\lambda, F) : \lambda = \lambda_0, F \in \mathcal{R}_c \}$, mentre $\Theta_1 = \{(\lambda, F) : \lambda \neq \lambda_0, F \in \mathcal{R}_c \}$. \square

Lo strumento statistico che sulla base del campione consente di concludere in favore dell'una o dell'altra ipotesi è il test statistico. Scelta una opportuna statistica T definita sull'insieme \mathcal{T} , si dice test basato su T la funzione

$$D : \mathcal{T} \rightarrow \{H_0, H_1\},$$

mentre T è detta statistica test. Il test basato su T è una regola decisionale che suddivide \mathcal{T} negli insiemi complementari \mathcal{T}_0 e \mathcal{T}_1 , in modo tale che si accetta H_0 se la realizzazione t di T è tale che $t \in \mathcal{T}_0$, mentre si accetta H_1 se $t \in \mathcal{T}_1$. L'insieme \mathcal{T}_0 è detto regione di accettazione, mentre l'insieme \mathcal{T}_1 è detto regione critica del test basato su T . Un test è detto “distribution-free” se è basato su una statistica test “distribution-free”.

• **Esempio 6.1.2.** In un approccio classico, si consideri un campione casuale X_1, \dots, X_n da $X \sim N(\mu, 1)$ e il sistema di ipotesi $H_0 : \mu = \mu_0$ contro $H_1 : \mu \neq \mu_0$, dove μ_0 è una quantità nota. Se si suppone che la statistica test sia la media campionaria delle osservazioni \bar{X} , risulta $\mathcal{T} = \mathbb{R}$. Una possibile scelta per \mathcal{T}_0 potrebbe essere data da

$$\mathcal{T}_0 = \{\bar{x} : |\bar{x} - \mu_0| < a\},$$

con a costante. Di conseguenza, la regione critica del test basato su \bar{X} risulta

$$\mathcal{T}_1 = \{\bar{x} : |\bar{x} - \mu_0| \geq a\}.$$

Questa scelta di \mathcal{T}_1 appare logica, in quanto più la realizzazione della media campionaria differisce dal valore ipotizzato μ_0 per la media, più si è propensi ad accettare l'ipotesi alternativa. Rimane aperto il problema della scelta della costante a . Alternativamente, in un approccio “distribution-free”, si consideri un campione casuale da una variabile casuale continua X con mediana pari a λ . Si vuole verificare il sistema di ipotesi $H_0 : \lambda = \lambda_0$ contro $H_1 : \lambda \neq \lambda_0$. Il test basato sulla statistica

$$B = \sum_{i=1}^n \mathbf{1}_{]0, \infty[}(X_i - \lambda_0),$$

è “distribution-free” essendo una trasformata di variabili casuali segno (vedi Esempio 3.3.1). Si ha $\mathcal{T} = \{b : b = 0, 1, \dots, n\}$, mentre è ragionevole assumere $\mathcal{T}_0 = \{b : b = a + 1, a + 2, \dots, n - a\}$, dove a è un intero tale che $a < n/2$. In effetti, sotto ipotesi di base si attende che circa la metà delle osservazioni campionarie siano maggiori della mediana, ovvero una realizzazione b di B prossima a $n/2$. Di nuovo, si deve decidere il valore della costante a . \square

Uno strumento per misurare la capacità discriminativa del test basato su una statistica T è la funzione potenza. La funzione potenza del test basato su T è data da

$$P_T(\theta) = P(T \in \mathcal{T}_1),$$

dove la probabilità è indotta dalla distribuzione specificata dal modello quando il valore del parametro è pari a θ . Per ogni $\theta \in \Theta_0$ la funzione potenza $P_T(\theta)$ fornisce la probabilità di respingere H_0 quando questa è vera, ovvero la probabilità di commettere il cosiddetto errore di I specie. Analogamente, per ogni $\theta \in \Theta_1$ la quantità $(1 - P_T(\theta))$ fornisce la probabilità di accettare H_0 quando è vera H_1 , ovvero la probabilità di commettere il cosiddetto errore di II specie. Per ogni $\theta \in \Theta_1$, la funzione potenza $P_T(\theta)$ fornisce la probabilità di accettare H_1 quando questa è vera. Si dice infine che il test basato su T è al livello di significatività α se

$$\sup_{\theta \in \Theta_0} P_T(\theta) = \alpha.$$

Il livello di significatività α rappresenta dunque la massima probabilità di commettere un errore di I specie.

• **Esempio 6.1.3.** Dato un campione casuale da $X \sim N(\mu, 1)$, si consideri il sistema di ipotesi $H_0 : \mu \leq 0$ contro $H_1 : \mu > 0$. Se si suppone che la statistica test sia \bar{X} , dato che $\mathcal{T} = \mathbb{R}$ si può scegliere

$$\mathcal{T}_0 =] - \infty, z_{1-\alpha}/\sqrt{n}[$$

e

$$\mathcal{T}_1 = [z_{1-\alpha}/\sqrt{n}, \infty[.$$

Questa selezione di \mathcal{T}_1 appare logica, in quanto più la realizzazione della media campionaria è elevata, più si è propensi ad accettare l'ipotesi alternativa. Risulta $\bar{X} \sim N(\mu, 1/n)$ per ogni $\mu \in \mathbb{R}$ e la funzione potenza è data da

$$P_{\bar{X}}(\mu) = \Phi(\sqrt{n}\mu - z_{1-\alpha}).$$

Dal momento che $P_{\bar{X}}(\mu)$ è crescente e che

$$\sup_{\mu \leq 0} P_{\bar{X}}(\mu) = \alpha,$$

il test basato su \bar{X} è al livello di significatività α . Alternativamente, in un approccio “distribution-free”, si consideri un campione casuale da una variabile casuale continua X con mediana pari a λ e funzione di ripartizione $F(x) = G(x - \lambda)$ con $G \in \mathcal{R}_0$. Si consideri inoltre il sistema di ipotesi $H_0 : \lambda \leq 0$ contro $H_1 : \lambda > 0$. Se si suppone che la statistica test sia data da $B = \sum_{i=1}^n \mathbf{1}_{]0, \infty[}(X_i)$, dal momento che $\mathcal{T} = \{b : b = 0, 1, \dots, n\}$ si può scegliere

$$\mathcal{T}_0 = \{b : b = 0, 1, \dots, b_{n, 1-\alpha}\}$$

e

$$\mathcal{T}_1 = \{b : b = b_{n, 1-\alpha} + 1, \dots, n\},$$

dove $b_{n, \alpha}$ rappresenta il quantile di ordine α della distribuzione Binomiale $Bi(n, 1/2)$. Di nuovo, la scelta di \mathcal{T}_1 è evidente, in quanto più il numero di osservazioni maggiori di zero è elevato, più si è propensi ad accettare l'ipotesi alternativa. Se è vera l'ipotesi alternativa, risulta

$$P(X > 0) = 1 - G(-\lambda) = G(\lambda),$$

per cui B ha distribuzione Binomiale $Bi(n, G(\lambda))$. La funzione potenza è dunque data da

$$P_B(\lambda) = \sum_{b=b_{n, 1-\alpha}+1}^n \binom{n}{b} G(\lambda)^b (1 - G(\lambda))^{n-b}.$$

Essendo $G(0) = 1/2$, allora risulta $P_B(0) = \alpha$ per ogni $G \in \mathcal{R}_0$. Si può dimostrare che $P_B(\lambda)$ è crescente e si deve concludere quindi che il test è al livello di significatività α . Si osservi che il test basato su B ha livello di significatività α per ogni modello statistico con campionamento casuale da una variabile casuale continua, mentre il test basato su \bar{X} non mantiene in generale questo livello se viene a mancare l'assunzione di Normalità. \square

Dal momento che non si può rendere contemporaneamente pari a zero gli errori di I e II specie, si deve stabilire un insieme di proprietà desiderabili per un test. Una prima proprietà opportuna per un test è quella della correttezza. Un test basato su T al livello di significatività α con funzione potenza $P_T(\theta)$ è detto corretto al livello di significatività α se

$$P_T(\theta) \geq \alpha, \forall \theta \in \Theta_1.$$

La proprietà della correttezza permette di controllare l'errore di I specie e al tempo stesso assicura che la probabilità di accettare H_1 quando è vera risulta maggiore dell'errore di I specie.

Una seconda proprietà riguarda il comportamento per grandi campioni del test. Il test al livello di significatività α è detto coerente se

$$\lim_n P_T(\theta) = 1, \forall \theta \in \Theta_1.$$

La proprietà della coerenza assicura che la probabilità di commettere un errore di II specie tende a zero quando si dispone di grandi campioni.

• **Esempio 6.1.4.** Dato un campione casuale da $X \sim N(\mu, 1)$, si consideri il sistema di ipotesi $H_0 : \mu \leq 0$ contro $H_1 : \mu > 0$. Il test basato su \bar{X} è corretto in quanto si ha $P_{\bar{X}}(\mu) \leq \alpha$ per ogni $\mu \leq 0$, mentre $P_{\bar{X}}(\mu) > \alpha$ per ogni $\mu > 0$. Dal momento che la successione di funzioni $(\Phi(\sqrt{n}\mu - z_{1-\alpha}))_{n \geq 1}$ converge uniformemente ad una funzione costante pari ad 1 per $\mu > 0$, allora si ha

$$\lim_n P_{\bar{X}}(\mu) = 1, \forall \mu > 0,$$

ovvero il test è coerente. Alternativamente, in un approccio “distribution-free”, sotto le assunzioni dell'Esempio 6.1.3, si consideri inoltre il sistema di ipotesi $H_0 : \lambda \leq 0$ contro $H_1 : \lambda > 0$. Il test basato su B è corretto in quanto $P_B(\lambda) \leq \alpha$ per ogni $\lambda \leq 0$, mentre $P_B(\lambda) > \alpha$ per ogni $\lambda > 0$. Sulla base del Teorema Fondamentale del Limite, si può anche dimostrare che il test è anche coerente. Queste proprietà sono valide per ogni modello statistico con campionamento casuale da una variabile casuale continua. \square

Anche se si desidera che $P_T(\theta)$ sia più elevata possibile quando $\theta \in \Theta_1$ e più piccola possibile quando $\theta \in \Theta_0$, questi requisiti sono conflittuali tra loro. Un possibile modo di procedere è quello di fissare il livello di significatività α e scegliere quel test che ha più alta potenza per ogni $\theta \in \Theta_1$. Un test basato su T al livello di significatività α con funzione potenza $P_T(\theta)$ è detto uniformemente più potente al livello di significatività α se

$$P_T(\theta) \geq P_U(\theta), \forall \theta \in \Theta_1,$$

per ogni altro test basato su una qualsiasi statistica U al livello di significatività α . I test uniformemente più potenti esistono in alcuni casi solamente quando si considera un approccio classico. Al contrario, quando si considera un test “distribution-free” in generale non è possibile determinare il test uniformemente più potente.

Se X_1, \dots, X_n è un campione casuale da una variabile casuale X con funzione di ripartizione F , si consideri il sistema di ipotesi $H_0 : \theta = \theta_0$ contro $H_1 : \theta \neq \theta_0$, dove θ è un parametro reale. Se si vuole analizzare le prestazioni di due statistiche test T e U per questo sistema di ipotesi, si può considerare la cosiddetta efficienza relativa. Siano $N_T(\alpha, \beta, \theta, F)$ e $N_U(\alpha, \beta, \theta, F)$ le numerosità campionarie necessarie ai due test per raggiungere la potenza β al livello di significatività α quando si considera l'alternativa $\theta \neq \theta_0$. L'efficienza relativa è data dalla quantità

$$e_{T,U}(\alpha, \beta, \theta, F) = \frac{N_U(\alpha, \beta, \theta, F)}{N_T(\alpha, \beta, \theta, F)}.$$

Se si ha $e_{T,U}(\alpha, \beta, \theta, F) > 1$, allora il test basato su T è più efficiente di quello basato su U , mentre se $e_{T,U}(\alpha, \beta, \theta, F) < 1$ è vero l'opposto. L'efficienza relativa è una misura locale dell'efficienza, nel senso che dipende dalle quantità α, β, θ e F . Al fine di eliminare la dipendenza da α, β, θ e quindi di ottenere una misura più globale dell'efficienza, si adotta la cosiddetta efficienza asintotica relativa $ARE_{T,U}$, che è data dal limite del rapporto delle numerosità campionarie necessarie ad ottenere la medesima potenza dei test T e U per una successione di alternative che converge a θ_0 ad un livello di significatività costante. In effetti, si ha $ARE_{T,U} = ARE_{T,U}(F)$, ovvero l'efficienza asintotica relativa dipende comunque dalla struttura funzionale della funzione di ripartizione F . Se $ARE_{T,U} > 1$ il test basato su T è preferibile a quello basato su U , mentre se $ARE_{T,U} < 1$ è vero l'opposto.

• **Esempio 6.1.5.** Dato un campione casuale da una variabile casuale continua X con funzione di ripartizione F , si consideri il sistema di ipotesi $H_0 : \lambda = \lambda_0$ contro $H_1 : \lambda \neq \lambda_0$, dove λ rappresenta la mediana. La Tavola 6.1.1 riporta $ARE_{B,T}$ per alcune distribuzioni quando si confronta il classico test T di Student e il test B considerato nell'Esempio 6.1.2. Anche se il test basato su B dimostra scarse prestazioni per una distribuzione a code leggere come l'Uniforme, il test diventa infinitamente efficiente per una distribuzione a code pesanti come la Cauchy. Si può verificare inoltre che le prestazioni del test dei segni risultano generalmente superiori (talvolta in modo molto marcato) a quelle del test T di Student quando si considerano distribuzioni asimmetriche. In generale, si può dimostrare che $ARE_{B,T} \geq 1/3$. Dunque, il test T di Student può avere un comportamento pessimo quando viene a mancare l'ipotesi di Normalità. \square

Tavola 6.1.1.

distribuzione	$ARE_{B,T}$
$U(\lambda, \lambda + \delta)$	$1/3 \simeq 0.3333$
$N(\mu, \sigma^2)$	$2/\pi \simeq 0.6366$
$C(\lambda, \delta)$	∞

Un modo per costruire test ottimali quando si considera un approccio classico è mediante il cosiddetto rapporto delle verosimiglianze. Il test del rapporto delle verosimiglianze è basato sulla statistica test R , la cui realizzazione è data da

$$r = \frac{\max_{\theta \in \Theta_0} L(\theta)}{\max_{\theta \in \Theta} L(\theta)} .$$

Questa statistica test ha un'interpretazione intuitiva, nel senso che se si sta confrontando la "plausibilità" di un valore θ rispetto ad un altro sulla base di un campione, siamo portati a scegliere quel valore che fornisce la verosimiglianza più alta. Se non esiste un valore θ che fornisce una verosimiglianza sensibilmente più alta in Θ rispetto alla verosimiglianza massima in Θ_0 , siamo propensi ad accettare H_0 . Ovviamente, per una realizzazione r di R si ha $0 \leq r \leq 1$. Se la realizzazione r è prossima ad 1 si è più propensi ad accettare H_0 , mentre se la realizzazione r è prossima a 0 si è più propensi ad accettare H_1 . Di conseguenza, si può scegliere come regione critica al livello di significatività α l'insieme $\mathcal{T}_1 = \{r : r \leq r_\alpha\}$, dove r_α è il quantile di ordine α della distribuzione di R .

• **Esempio 6.1.6.** Dato un campione casuale dalla variabile casuale $X \sim N(\mu, 1)$, si consideri il sistema di ipotesi $H_0 : \mu = \mu_0$ contro $H_1 : \mu \neq \mu_0$. La determinazione campionaria del rapporto delle verosimiglianze è data da

$$r = \frac{e^{-\frac{1}{2}n(s_x^2 + (\bar{x} - \mu_0)^2)}}{e^{-\frac{1}{2}ns_x^2}} = e^{-\frac{1}{2}n(\bar{x} - \mu_0)^2} .$$

Dal momento che r è una funzione biunivoca di $(\bar{x} - \mu_0)^2$, la regione critica indotta da R è la stessa di quella indotta da $(\bar{X} - \mu_0)^2$ e quindi i relativi test sono equivalenti. Se è vera H_0 si ha $\sqrt{n}(\bar{X} - \mu_0) \sim N(0, 1)$ e dunque $n(\bar{X} - \mu_0)^2 \sim \chi_1^2$. Tenendo presente che r è una funzione monotona decrescente di $(\bar{x} - \mu_0)^2$ e dunque si respinge H_0 per realizzazioni elevate di $(\bar{X} - \mu_0)^2$, allora la regione critica del test basato sul rapporto delle verosimiglianze è data da

$$\mathcal{T}_1 = \{\bar{x} : n(\bar{x} - \mu_0)^2 \geq \chi_{1,1-\alpha}^2\} .$$

Essendo $z_{1-\alpha/2} = \sqrt{\chi_{1,1-\alpha}^2}$ per la relazione fra le distribuzioni $N(0, 1)$ e χ_1^2 , la regione critica può essere anche espressa in modo equivalente come

$$\mathcal{T}_1 = \{\bar{x} : \sqrt{n} |\bar{x} - \mu_0| \geq z_{1-\alpha/2}\} . \quad \square$$

• **Esempio 6.1.7.** Si consideri un campionamento casuale da $X \sim N(\mu, \nu)$ e si supponga di voler verificare il sistema di ipotesi $H_0 : \mu = \mu_0$ contro $H_1 : \mu \neq \mu_0$. In questo caso si ha

$$\Theta = \{(\mu, \nu) : (\mu, \nu) \in \mathbb{R} \times \mathbb{R}^+\}$$

e

$$\Theta_0 = \{(\mu, \nu) : \mu = \mu_0, \nu \in \mathbb{R}^+\} .$$

Tenendo presente l'Esempio 4.3.1, si ottiene

$$\max_{(\mu, v) \in \Theta} l(\mu, v) = l(\bar{x}, s_x^2).$$

Inoltre, dalla disuguaglianza dell'Esempio 4.3.1 con $d = s_x^2 + (\bar{x} - \mu_0)^2$, si ha

$$\begin{aligned} l(\mu_0, v) &= \log c - \frac{n}{2} \log v - \frac{n}{2v} (s_x^2 + (\bar{x} - \mu_0)^2) \leq \log c - \frac{n}{2} \log(s_x^2 + (\bar{x} - \mu_0)^2) - \frac{n}{2} \\ &= l(\mu_0, s_x^2 + (\bar{x} - \mu_0)^2) = \max_{(\mu, v) \in \Theta_0} l(\mu, v). \end{aligned}$$

La determinazione campionaria di R è dunque data da

$$r = \frac{e^{-\frac{n}{2} \log(s_x^2 + (\bar{x} - \mu_0)^2)}}{e^{-\log s_x^2}} = \left(1 + \frac{n(\bar{x} - \mu_0)^2}{(n-1)s_{c,x}^2}\right)^{-\frac{1}{2}n} = \left(1 + \frac{t^2}{n-1}\right)^{-\frac{1}{2}n}$$

dove

$$t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s_{c,x}}.$$

Se è vera H_0 , t è la determinazione campionaria di una statistica T distribuita come una t di Student con $(n-1)$ gradi di libertà, dal momento che

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S_{c,x}} = \frac{\sqrt{n}(\bar{X} - \mu_0)/\sqrt{v}}{\sqrt{((n-1)S_{c,x}^2/v)/(n-1)}}$$

e che inoltre le statistiche $\sqrt{n}(\bar{X} - \mu_0)/\sqrt{v} \sim N(0, 1)$ e $(n-1)S_{c,x}^2/v \sim \chi_{n-1}^2$ sono indipendenti. Inoltre, r è una funzione monotona decrescente di $|t|$ e quindi i test costruiti su R e $|T|$ sono equivalenti. Tenendo presente la relazione fra le statistiche T e $|T|$ e dal momento che rifiutare H_0 per piccole determinazioni di R è equivalente a rifiutare H_0 per determinazioni elevate di $|T|$, la regione critica del test risulta

$$\mathcal{T}_1 = \{t : t \leq -t_{n-1, 1-\alpha/2}, t \geq t_{n-1, 1-\alpha/2}\}. \quad \square$$

Per quanto riguarda la distribuzione per grandi campioni del rapporto delle verosimiglianze, se k è il numero di parametri nel modello e q è il numero di parametri da stimare sotto ipotesi di base, assumendo alcune ipotesi di regolarità, la statistica test $-2 \ln R$ converge in distribuzione ad una variabile casuale con distribuzione χ_{k-q}^2 per $n \rightarrow \infty$. La regione critica per grandi campioni del test basato sul rapporto delle verosimiglianze è dunque data da

$$\mathcal{T}_1 = \{r : -2 \ln r \geq \chi_{k-q, 1-\alpha}^2\}.$$

• **Esempio 6.1.8.** Dato un campione casuale dalla variabile casuale $X \sim E(0, \sigma)$, si consideri il sistema di ipotesi $H_0 : \sigma = \sigma_0$ contro $H_1 : \sigma \neq \sigma_0$. La determinazione campionaria del rapporto delle verosimiglianze è data da

$$r = \frac{e^{-n \log \sigma_0 - n\bar{x}/\sigma_0}}{e^{-n \log \bar{x} - n}} = e^{n \log(\bar{x}/\sigma_0) - n\bar{x}/\sigma_0 + n},$$

ovvero r è una funzione non biunivoca di \bar{x} . Dunque, risulta laborioso determinare la distribuzione per campioni finiti del rapporto delle verosimiglianze. Tuttavia, se è vera H_0 risulta

$$-2 \log R = -2n \log \left(\frac{\bar{X}}{\sigma_0} \right) + \frac{2n\bar{X}}{\sigma_0} - 2n,$$

che converge in distribuzione ad una variabile casuale con distribuzione χ_1^2 per $n \rightarrow \infty$. Di conseguenza, la regione critica per grandi campioni del test basato sul rapporto delle verosimiglianze è data da

$$\mathcal{T}_1 = \left\{ \bar{x} : -2n \log \left(\frac{\bar{x}}{\sigma_0} \right) + \frac{2n\bar{x}}{\sigma_0} - 2n \geq \chi_{1,1-\alpha}^2 \right\}. \quad \square$$

• **Esempio 6.1.9.** Dato un campione casuale dalla variabile casuale $X \sim Bi(1, \pi)$, si consideri il sistema di ipotesi $H_0 : \pi = \pi_0$ contro $H_1 : \pi \neq \pi_0$. La determinazione campionaria del rapporto delle verosimiglianze è data da

$$r = \frac{e^{n\bar{x} \log \pi_0 + (n-n\bar{x}) \log(1-\pi_0)}}{e^{n\bar{x} \log \bar{x} + (n-n\bar{x}) \log(1-\bar{x})}} = e^{n\bar{x} \log(\pi_0/\bar{x}) + (n-n\bar{x}) \log((1-\pi_0)/(1-\bar{x}))},$$

ovvero r è anche in questo caso una funzione non biunivoca di \bar{x} e quindi non è semplice ottenere la relativa distribuzione per campioni finiti. Tuttavia, se è vera H_0 risulta

$$-2 \log R = -2n\bar{X}_n \log \left(\frac{\pi_0}{\bar{X}_n} \right) - 2(n - n\bar{X}_n) \log \left(\frac{1 - \pi_0}{1 - \bar{X}_n} \right),$$

che converge in distribuzione ad una variabile casuale con distribuzione χ_1^2 per $n \rightarrow \infty$. Di conseguenza, la regione critica per grandi campioni del test basato sul rapporto delle verosimiglianze è data da

$$\mathcal{T}_1 = \left\{ \bar{x} : -2n\bar{x} \log \left(\frac{\pi_0}{\bar{x}} \right) - 2(n - n\bar{x}) \log \left(\frac{1 - \pi_0}{1 - \bar{x}} \right) \geq \chi_{1,1-\alpha}^2 \right\}. \quad \square$$

Nel presente approccio, l'ipotesi di base e l'ipotesi alternativa vengono trattate in modo non simmetrico. In effetti, usualmente H_0 costituisce una affermazione privilegiata e si preferisce controllare il livello di significatività del test (ovvero l'errore di I specie) che comporta l'erroneo rifiuto di questa ipotesi privilegiata. Anche se per sviluppare la teoria è necessario fissare il livello di significatività α , quando si lavora operativamente non esiste nessuna regola per stabilirne la scelta. Questa considerazione porta al concetto di livello di significatività osservato o valore-P.

Se la regione critica del test basato su T è data da $\mathcal{T}_1 = \{t : t \geq c\}$, per un determinato valore campionario t si dice significatività osservata la quantità

$$\alpha_{oss} = \sup_{\theta \in \Theta_0} P(T \geq t),$$

mentre, se la regione critica è data da $\mathcal{T}_1 = \{t : t \leq c\}$, allora si dice significatività osservata la quantità

$$\alpha_{oss} = \sup_{\theta \in \Theta_0} P(T \leq t).$$

Quando invece la statistica test T ha una distribuzione simmetrica, se la regione critica del test basato su T è data da $\mathcal{T}_1 = \{t : t \leq c_1, t \geq c_2\}$, si dice significatività osservata la quantità

$$\alpha_{oss} = 2 \min \left\{ \sup_{\theta \in \Theta_0} P(T \leq t), \sup_{\theta \in \Theta_0} P(T \geq t) \right\}.$$

La significatività osservata rappresenta la probabilità di ottenere, quando H_0 è vera, un valore campionario t di T estremo (nella appropriata direzione) almeno quanto quello osservato. Dunque, la significatività osservata fornisce una misura su quanto l'ipotesi di base risulta compatibile con i dati campionari. Una significatività osservata bassa porta a ritenere poco compatibile con i dati campionari l'ipotesi di base, mentre con una significatività osservata elevata è vera l'affermazione contraria. In una verifica di ipotesi si può semplicemente riportare la significatività osservata, oppure si può arrivare ad una decisione sull'accettazione di H_0 fissando un livello di significatività α . Se il livello di significatività osservato è minore o uguale ad α , allora si respinge H_0 , altrimenti si accetta H_0 . Il livello di significatività osservato diventa in questo caso il più elevato livello di significatività per cui si accetta H_0 . In questo caso il livello di significatività osservato diventa non solo uno strumento per la decisione nella verifica di ipotesi, ma anche una misura quantitativa di questa decisione.

6.2. Gli intervalli di confidenza

Piuttosto che selezionare sulla base del campione un unico valore come nella stima per punti, può essere utile dal punto di vista operativo ottenere un insieme di valori plausibili del parametro. Considerato per semplicità il caso in cui si ha un singolo campione casuale X_1, \dots, X_n , sia $Q = Q(X_1, \dots, X_n; \theta)$ una quantità pivotale, ovvero una trasformata che dipende dal parametro ma con una distribuzione non dipende dal parametro stesso. Se c_1 e c_2 sono due valori tali che

$$P(c_1 < Q(X_1, \dots, X_n; \theta) < c_2) = 1 - \alpha,$$

con $0 < \alpha < 1$, e se $L = L(X_1, \dots, X_n)$ e $U = U(X_1, \dots, X_n)$ sono statistiche tali che per ogni θ

$$\{(x_1, \dots, x_n) : c_1 < Q(x_1, \dots, x_n; \theta) < c_2\} \Leftrightarrow \{(x_1, \dots, x_n) : L(x_1, \dots, x_n) < \theta < U(x_1, \dots, x_n)\},$$

allora l'intervallo casuale (L, U) è detto intervallo di confidenza di θ al livello di confidenza $(1 - \alpha)$. Se la quantità pivotale Q è "distribution-free", ovvero se la sua distribuzione rimane invariata per un modello "distribution-free", allora il relativo intervallo di confidenza è detto "distributon-free".

La nozione di intervallo di confidenza deve essere adoperata con cautela, ovvero non si deve affermare che il vero valore del parametro è contenuto in un intervallo con probabilità pari a $(1 - \alpha)$. In termini rigorosi di probabilità, una volta che l'intervallo di confidenza è stato determinato sul campione, questo contiene il vero valore con probabilità 0 o 1. Si può affermare invece che l'intervallo di confidenza è la determinazione di un procedimento casuale che seleziona intervalli in modo tale che la probabilità di ottenere un intervallo contenente il vero valore del parametro è pari a $(1 - \alpha)$.

• **Esempio 6.2.1.** Se si considera un campione casuale da $X \sim N(\mu, 1)$, una possibile quantità pivotale è data da

$$Q(X_1, \dots, X_n; \mu) = \sqrt{n}(\bar{X} - \mu).$$

Questa variabile casuale è in effetti una quantità pivotale, dal momento che la sua distribuzione non dipende da μ essendo $\sqrt{n}(\bar{X} - \mu) \sim N(0, 1)$. Scelto un livello di confidenza $(1 - \alpha)$, risulta

$$P(-z_{1-\alpha/2} < \sqrt{n}(\bar{X} - \mu) < z_{1-\alpha/2}) = 1 - \alpha.$$

L'insieme

$$\{(x_1, \dots, x_n) : -z_{1-\alpha/2} < \sqrt{n}(\bar{x} - \mu) < z_{1-\alpha/2}\}$$

è equivalente all'insieme

$$\{(x_1, \dots, x_n) : \bar{x} - z_{1-\alpha/2}/\sqrt{n} < \mu < \bar{x} + z_{1-\alpha/2}/\sqrt{n}\}$$

e si deve dunque concludere che

$$(\bar{X} - z_{1-\alpha/2}/\sqrt{n}, \bar{X} + z_{1-\alpha/2}/\sqrt{n})$$

è un intervallo di confidenza per μ al livello di confidenza $(1 - \alpha)$. \square

• **Esempio 6.2.2.** Se si consideri un campione casuale da una variabile casuale continua X con mediana λ , una quantità pivotale è data da

$$Q(X_1, \dots, X_n; \lambda) = B(X_1 - \lambda, \dots, X_n - \lambda) = \sum_{i=1}^n \mathbf{1}_{]0, \infty[}(X_i - \lambda).$$

La quantità pivotale $B(X_1 - \lambda, \dots, X_n - \lambda)$ è in effetti tale ed è “distribution-free”, dal momento che la sua distribuzione non dipende da λ e $B(X_1 - \lambda, \dots, X_n - \lambda) \sim Bi(n, 1/2)$. Dunque, scelto un livello di confidenza $(1 - \alpha)$, si ha

$$P(b_{n,\alpha/2} < B(X_1 - \lambda, \dots, X_n - \lambda) < n - b_{n,\alpha/2}) = 1 - \alpha.$$

Si osservi che la quantità pivotale $B(X_1 - \lambda, \dots, X_n - \lambda)$ rappresenta il numero di osservazioni campionarie maggiori di λ . Se $(X_{(1)}, \dots, X_{(n)})$ è la statistica ordinata, l'insieme

$$\{(x_1, \dots, x_n) : B(x_1 - \lambda, \dots, x_n - \lambda) < n - b_{n,\alpha/2}\}$$

è equivalente all'insieme

$$\{(x_1, \dots, x_n) : x_{(b_{n,\alpha/2}+1)} < \lambda\}.$$

Analogamente, l'insieme

$$\{(x_1, \dots, x_n) : b_{n,\alpha/2} < B(x_1 - \lambda, \dots, x_n - \lambda)\}$$

è equivalente all'insieme

$$\{(x_1, \dots, x_n) : \lambda < x_{(n-b_{n,\alpha/2})}\}.$$

Dunque, si ha $L = X_{(b_{n,\alpha/2}+1)}$ e $U = X_{(n-b_{n,\alpha/2})}$, ovvero $(X_{(b_{n,\alpha/2}+1)}, X_{(n-b_{n,\alpha/2})})$ è un intervallo di confidenza per λ “distribution-free” al livello di confidenza $(1 - \alpha)$. \square

Esiste una stretta connessione tra il problema della stima per intervalli e quello della verifica di ipotesi. Questa è anche la ragione per cui la stima per intervalli viene analizzata successivamente alla teoria relativa alla verifica di ipotesi. Più esattamente, se si considera il sistema di ipotesi $H_0 : \theta = \theta_0$ contro $H_1 : \theta \neq \theta_0$, esiste una equivalenza tra la regione critica della statistica test e l'intervallo di confidenza di θ . Si consideri un test corretto al livello di significatività α . Se C_0 è la regione di accettazione del test nello spazio campionario, ovvero l'insieme delle realizzazioni campionarie (x_1, \dots, x_n) che portano all'accettazione dell'ipotesi di base, è immediato verificare che C_0 dipende dal valore prefissato θ_0 di θ . Dunque risulta $C_0 = C_0(\theta)$. Inversamente, per un data realizzazione campionaria (x_1, \dots, x_n) nello spazio campionario deve esistere un insieme $E(x_1, \dots, x_n)$ di valori di θ per i quali si accetta l'ipotesi di base. I due insiemi sono equivalenti, ovvero

$$\{(x_1, \dots, x_n) : (x_1, \dots, x_n) \in C_0(\theta)\} \Leftrightarrow \{(x_1, \dots, x_n) : \theta \in E(x_1, \dots, x_n)\}.$$

Dal momento che

$$\{(x_1, \dots, x_n) : (x_1, \dots, x_n) \in C_0(\theta)\} = \{(x_1, \dots, x_n) : \mathbf{1}_{C_0(\theta)}(x_1, \dots, x_n) = 1\}$$

e

$$P((X_1, \dots, X_n) \in C_0(\theta)) = P(\mathbf{1}_{C_0(\theta)}(X_1, \dots, X_n) = 1) = 1 - \alpha,$$

se come quantità pivotale si considera la variabile casuale $\mathbf{1}_{C_0(\theta)}(X_1, \dots, X_n)$, allora $E(x_1, \dots, x_n)$ è un insieme di confidenza al livello di confidenza $(1 - \alpha)$. Quindi, l'insieme di tutti i valori θ per cui si accetta l'ipotesi di base costituisce un insieme di confidenza per θ . Nel caso che $E(x_1, \dots, x_n)$ sia un intervallo, con la precedente procedura si ottiene ovviamente un intervallo di confidenza e questa situazione è frequente con i modelli usualmente adoperati in pratica.

Partendo dunque da un test opportuno è possibile costruire un intervallo di confidenza al livello di confidenza scelto. Inoltre, più il test prescelto ha funzione potenza elevata più l'intervallo di confidenza risultante è desiderabile, dal momento che *a priori* dal campionamento la probabilità che contenga un qualunque valore θ diverso da θ_0 è bassa. È dunque buona regola costruire intervalli di confidenza a partire da un test che gode di buone proprietà. Queste considerazioni consentono di costruire un intervallo di confidenza per un dato parametro partendo da un opportuno sistema di ipotesi. Nel caso di un modello classico, l'intervallo di confidenza viene usualmente costruito a partire dal test del rapporto delle verosimiglianze. L'intervallo di confidenza risulta “distribution-free” se è costruito a partire da un test “distribution-free”. Esiste anche una connessione fra test e stima per punti. In effetti, la stima per punti risulta usualmente il valore centrale dell'intervallo di confidenza.

• **Esempio 6.2.3.** Dato un campione casuale da $X \sim U(0, \delta)$, si vuole costruire un intervallo di confidenza per δ . Se si considera il sistema di ipotesi $H_0 : \delta = \delta_0$ contro $H_1 : \delta \neq \delta_0$, tenendo presente l'Esempio 4.3.2, la determinazione campionaria del rapporto delle verosimiglianze è data da

$$r = \frac{\delta_0^{-n} \mathbf{1}_{(0,1)}(x_{(n)}/\delta_0)}{x_{(n)}^{-n}} = \left(\frac{x_{(n)}}{\delta_0} \right)^n \mathbf{1}_{(0,1)} \left(\frac{x_{(n)}}{\delta_0} \right),$$

Si noti che r è funzione monotona crescente di $x_{(n)}/\delta_0$ e quindi il test costruito sulla statistica $X_{(n)}/\delta_0$ è equivalente a quello del rapporto delle verosimiglianze. Quando H_0 è vera, la statistica $X_{(n)}/\delta_0$ ha distribuzione $Be(0, 1; n, 1)$ (vedi Esempio 4.3.2). Dal momento che il quantile di ordine α di una distribuzione $Be(0, 1; n, 1)$ è dato da $\alpha^{1/n}$, la regione di accettazione per il test del rapporto delle verosimiglianze risulta

$$\mathcal{T}_0 = \{(x_1, \dots, x_n) : \alpha^{1/n} < x_{(n)}/\delta < 1\},$$

che è equivalente all'insieme

$$\{(x_1, \dots, x_n) : x_{(n)} < \delta < x_{(n)}\alpha^{-1/n}\}.$$

Di conseguenza, $(X_{(n)}, X_{(n)}\alpha^{-1/n})$ è un intervallo di confidenza per δ al livello di confidenza $(1 - \alpha)$.

6.3. Riferimenti bibliografici

- Boos, D.D. e Stefanski, L.A. (2013) *Essential Statistical Inference*, Springer, New York.
- Chan, J.C.C. e Kroese, D.P. (2025) *Statistical Modeling and Computation*, seconda edizione, Springer, New York.
- Cox, D.R. e Hinkley, D.V. (1974) *Theoretical Statistics*, Chapman and Hall, London.
- Fay, P.M. e Brittain, E.H. (2022) *Statistical Hypothesis Testing in Context*, Cambridge University Press, Cambridge.
- Hettmansperger, T.P. e McKean, J.W. (2011) *Robust Nonparametric Statistical Methods*, seconda edizione, Chapman & Hall/CRC Press, Boca Raton.
- Hollander, M., Wolfe, D.A. e Chicken, E. (2014) *Nonparametric Statistical Methods*, terza edizione, Wiley, New York.

-
- Lauritzen, S. (2023) *Fundamentals of Mathematical Statistics*, Chapman & Hall/CRC Press, Boca Raton.
- Lehmann, E.L. (1999) *Elements of Large Sample Theory*, Springer, New York.
- Lehmann, E.L. e Romano, J.P. (2022) *Testing Statistical Hypothesis*, quarta edizione, Springer, Switzerland.
- Meeker, W.Q., Hahn, G.J. e Escobar, L.A. (2017) *Statistical Intervals*, seconda edizione, Wiley, New York.
- Rao, C.R. (1973) *Linear Statistical Inference and its Applications*, seconda edizione, Wiley, New York.
- Shao, J. (2003) *Mathematical Statistics*, seconda edizione, Springer, New York.
- Wilks, S.S. (1962) *Mathematical Statistics*, Wiley, New York.

Pagina intenzionalmente vuota

Capitolo 7

L'inferenza con una variabile

7.1. L'inferenza per un parametro di tendenza centrale

Il modello statistico più semplice assume un campionamento da una singola variabile, in modo tale che l'obiettivo è quello di produrre inferenza per un parametro di tendenza centrale. In un tipico approccio classico, si consideri inizialmente un campione casuale X_1, \dots, X_n da una variabile casuale $X \sim N(\mu, \sigma^2)$. Gli stimatori di massima verosimiglianza di μ e σ^2 risultano $\hat{\mu} = \bar{X}$ e $\hat{\sigma}^2 = S_x^2$. Se si considera il sistema di ipotesi $H_0 : \mu = \mu_0$ contro $H_1 : \mu \neq \mu_0$ (o una ipotesi direzionale), il test del rapporto delle verosimiglianze fornisce il test di t di Student basato sulla statistica test

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S_{c,x}},$$

che sotto ipotesi di base è distribuita come $T \sim t_{n-1}$. Nel caso dell'ipotesi alternativa bilaterale $H_1 : \mu \neq \mu_0$, si rifiuta H_0 per realizzazioni basse o elevate di T . Nel caso dell'ipotesi alternativa direzionale $H_1 : \mu > \mu_0$ ($H_1 : \mu < \mu_0$), si rifiuta H_0 per realizzazioni elevate (basse) di T . Infine, l'intervallo di confidenza per μ al livello di confidenza $(1 - \alpha)$ basato sul test T risulta

$$(\bar{X} - t_{n-1, 1-\alpha/2} S_{c,x} / \sqrt{n}, \bar{X} + t_{n-1, 1-\alpha/2} S_{c,x} / \sqrt{n}).$$

Il test basato su T è “distribution-free” per grandi campioni dal momento che T converge in distribuzione a una variabile casuale con distribuzione $N(0, 1)$ per $n \rightarrow \infty$, se $\sigma^2 < \infty$. In questo caso, l'intervallo di confidenza per grandi campioni per μ al livello di confidenza $(1 - \alpha)$ risulta

$$(\bar{X} - z_{1-\alpha/2} S_{c,x} / \sqrt{n}, \bar{X} + z_{1-\alpha/2} S_{c,x} / \sqrt{n}).$$

Il test t di Student è quindi “robusto” rispetto all'assunzione di Normalità.

• **Esempio 7.1.1.** Si considera di nuovo i dati relativi alle sfere di acciaio dell'Esempio 4.3.4. Dal momento che l'azienda che produce le sfere desidera produrre sfere con un diametro standard di un micron, il sistema di ipotesi risulta $H_0 : \mu = 1$ contro $H_1 : \mu \neq 1$. Il comando `t.test` fornisce l'implementazione del test t di Student e il relativo intervallo di confidenza:

```
> t.test(Diameter, alternative = "two.sided", mu = 1)
```

```
One Sample t-test
```

```
data: Diameter
t = 2.1178, df = 9, p-value = 0.06327
alternative hypothesis: true mean is not equal to 1
95 percent confidence interval:
 0.9867741 1.4012259
sample estimates:
mean of x
 1.194
```

La significatività osservata porta dunque ad ipotizzare che la produzione sia leggermente fuori controllo. \square

In un approccio “distribution-free”, si consideri un campione casuale da una variabile casuale continua X con funzione di ripartizione $F(x - \lambda)$, dove λ rappresenta la mediana. Si osservi che in questo caso l'inferenza è relativa alla mediana, invece che alla media, per non restringere eccessivamente il modello statistico. In effetti, esistono ampie classi di variabili casuali continue per cui la media non è finita o non è definita.

Se si considera il sistema di ipotesi $H_0 : \lambda = \lambda_0$ contro $H_1 : \lambda \neq \lambda_0$ (o una ipotesi direzionale), il test dei segni è basato sulla statistica test

$$B = \sum_{i=1}^n \mathbf{1}_{]0, \infty[}(X_i - \lambda_0),$$

che sotto ipotesi di base è tale che $B \sim Bi(n, 1/2)$. Nel caso dell'ipotesi alternativa bilaterale $H_1 : \lambda \neq \lambda_0$, si rifiuta H_0 per realizzazioni basse o elevate di B . Nel caso dell'ipotesi alternativa direzionale $H_1 : \lambda > \lambda_0$ ($H_1 : \lambda < \lambda_0$), si rifiuta H_0 per realizzazioni elevate (basse) di B . Lo stimatore di λ basato sul test B è la mediana campionaria $\tilde{X}_{0.5}$. Inoltre, se $b_{n,\alpha}$ rappresenta il quantile di ordine α di B , l'intervallo di confidenza per λ al livello di confidenza $(1 - \alpha)$ basato sul test B risulta

$$(X_{(b_{n,\alpha/2}+1)}, X_{(n-b_{n,\alpha/2})}).$$

• **Esempio 7.1.2.** Si considerano di nuovo i dati relativi alle sfere di acciaio dell'Esempio 4.3.4 e il sistema di ipotesi $H_0 : \lambda = 1$ contro $H_1 : \lambda \neq 1$. Anche se non esiste un comando specifico per il test dei segni, si può equivalentemente adottare il comando `binom.test` per verifiche di ipotesi sulle proporzioni:

```
> binom.test(length(Diameter[Diameter > 1]), length(Diameter),
+           p = 1/2, alternative = "two.sided")
```

```
Exact binomial test
```

```
data: length(Diameter[Diameter > 1]) and length(Diameter)
number of successes = 8, number of trials = 10, p-value = 0.1094
alternative hypothesis: true probability of success is not equal to
0.5
95 percent confidence interval:
 0.4439045 0.9747893
sample estimates:
probability of success
                0.8
```

Inoltre, la stima per punti e gli estremi dell'intervallo di confidenza per λ possono essere ottenuti mediante i seguenti comandi:

```
> median(Diameter)
[1] 1.185
> sd <- sort(Diameter)
> sd[qbinom(0.025, length(Diameter), 1 / 2)]
[1] 0.88
> sd[qbinom(0.975, length(Diameter), 1 / 2)]
[1] 1.42
```

Si osservi che in questo caso la significatività osservata e l'ampiezza dell'intervallo di confidenza sono maggiori di quelli ottenuti nell'Esempio 7.1.1. Questo fenomeno deriva dal fatto che le assunzioni fatte per il test dei segni sono molto più blande rispetto a quelle fatte per il test t di Student e quindi vi è una maggiore incertezza nel processo inferenziale. \square

Di nuovo con un approccio “distribution-free”, si consideri un campione casuale da una variabile casuale continua X , simmetrica rispetto alla mediana λ e con funzione di ripartizione $F(x - \lambda)$. La variabile casuale X è simmetrica rispetto alla mediana se la distribuzione di $(X - \lambda)$ coincide con quella di $(\lambda - X)$. Quindi, rispetto alle assunzioni fatte per il test dei segni, si assume questa ulteriore ipotesi.

Se si considera il sistema di ipotesi $H_0 : \lambda = \lambda_0$ contro $H_1 : \lambda \neq \lambda_0$ (o una ipotesi direzionale), il test di Wilcoxon è basato sulla statistica test

$$W^+ = \sum_{i=1}^n \mathbf{1}_{]0, \infty[}(X_i - \lambda_0) R_i^+,$$

dove R_1^+, \dots, R_n^+ rappresentano i ranghi assegnati alle trasformate $|X_1 - \lambda_0|, \dots, |X_n - \lambda_0|$. La distribuzione di W^+ sotto ipotesi di base può essere tabulata anche se non può essere espressa in forma chiusa. Nel caso dell'ipotesi alternativa bilaterale $H_1 : \lambda \neq \lambda_0$, si rifiuta H_0 per realizzazioni basse o elevate di W^+ . Nel caso dell'ipotesi alternativa direzionale $H_1 : \lambda > \lambda_0$ ($H_1 : \lambda < \lambda_0$), si rifiuta H_0 per realizzazioni elevate (basse) di W^+ .

Se $k = n(n+1)/2$, siano W_1, \dots, W_k le medie di Walsh, ovvero tutte le possibili k semisomme distinte delle n osservazioni. La stima di λ basata sul test W^+ è la mediana delle medie di Walsh (la cosiddetta pseudomediana). Se $w_{n,\alpha}^+$ rappresenta il quantile di ordine α di W^+ e se $W_{(1)}, \dots, W_{(k)}$ è la statistica ordinata relativa alle medie di Walsh, allora l'intervallo di confidenza per λ al livello di confidenza $(1 - \alpha)$ basato sul test di Wilcoxon risulta

$$(W_{(w_{n,\alpha/2}^++1)}, W_{(k-w_{n,\alpha/2}^+)}) .$$

• **Esempio 7.1.3.** Si considerano di nuovo i dati relativi alle sfere di acciaio dell'Esempio 4.3.4 e il sistema di ipotesi $H_0 : \lambda = 1$ contro $H_1 : \lambda \neq 1$. Il comando `wilcox.test` fornisce l'implementazione del test di Wilcoxon:

```
> wilcox.test(Diameter, alternative = "two.sided",
+   mu = 1, conf.int = TRUE)

Wilcoxon signed rank test

data: Diameter
V = 46, p-value = 0.06445
alternative hypothesis: true location is not equal to 1
95 percent confidence interval:
 0.985 1.405
sample estimates:
(pseudo)median
 1.19
```

In questo caso, la significatività osservata e l'ampiezza dell'intervallo di confidenza sono quasi identici a quelli ottenuti nell'Esempio 7.1.1. Dunque, l'ipotesi di simmetria aumenta l'efficienza della procedura “distribution-free” in modo sostanziale. \square

Con un approccio “distribution-free”, si consideri un campione scambiabile X_1, \dots, X_n da una variabile casuale continua e simmetrica X con funzione di ripartizione $F(x - \lambda)$, dove λ rappresenta di nuovo la mediana. Un campione è detto scambiabile se la distribuzione di X_1, \dots, X_n coincide con quella di X_{i_1}, \dots, X_{i_n} per qualsiasi permutazione i_1, \dots, i_n nell'insieme \mathcal{S}_n delle permutazioni dei primi n interi. Dunque, un campione scambiabile non è generalmente un campione casuale, in quanto le osservazioni campionarie possono essere dipendenti.

Se si considera il sistema di ipotesi $H_0 : \lambda = \lambda_0$ contro $H_1 : \lambda \neq \lambda_0$ (o una ipotesi direzionale), condizionatamente alla realizzazione x_1, \dots, x_n del campione, sotto ipotesi di base i valori $-|x_i - \lambda_0|$ e $+|x_i - \lambda_0|$ sono ugualmente probabili. Il test di permutazione è basato sulle 2^n (ugualmente probabili) permutazioni dei segni delle osservazioni trasformate $|x_1 - \lambda_0|, \dots, |x_n - \lambda_0|$. Se S_1, \dots, S_n rappresenta un vettore scambiabile di variabili casuali di Bernoulli ognuna con parametro $1/2$ e supporto $\{-1, 1\}$, il test di permutazione dei segni è basato sulla statistica test

$$T = \sum_{i=1}^n |x_i - \lambda_0| S_i.$$

La distribuzione della statistica test sotto ipotesi di base può essere calcolata, anche se non è possibile esprimerla in forma chiusa. Tuttavia, per ogni differente realizzazione campionaria, si deve effettuare un nuovo calcolo, che comunque è proibitivo anche per numerosità campionarie modeste. Quindi, la distribuzione della statistica test è usualmente approssimata mediante il metodo Monte Carlo. Nel caso dell'ipotesi alternativa bilaterale $H_1 : \lambda \neq \lambda_0$, si rifiuta H_0 per realizzazioni basse o elevate di T . Nel caso dell'ipotesi alternativa direzionale $H_1 : \lambda > \lambda_0$ ($H_1 : \lambda < \lambda_0$), si rifiuta H_0 per realizzazioni elevate (basse) di T .

Si deve osservare che l'approccio basato sui test di permutazione è radicale, nel senso che l'inferenza è completamente condizionata al campione osservato. Questo paradigma può essere evidentemente in contrasto con un approccio come quello Bayesiano.

• **Esempio 7.1.4.** Si considerano di nuovo i dati relativi alle sfere di acciaio dell'Esempio 4.3.4. Richiamando la libreria `exactRankTests`, il comando `wilcox.test` fornisce l'implementazione del test di permutazione dei segni. Le osservazioni sono state moltiplicate per cento al fine di ottenere valori interi per ottenere una elaborazione più rapida. Di conseguenza, anche il sistema di ipotesi risulta $H_0 : \lambda = 100$ contro $H_1 : \lambda \neq 100$. Il test di permutazione viene implementato mediante i seguenti comandi:

```
> library(exactRankTests)
> perm.test(round(100 * Diameter), paired = FALSE,
+          alternative = "two.sided", mu = 100)

1-sample Permutation Test

data: round(100 * Diameter)
T = 237, p-value = 0.07031
alternative hypothesis: true mu is not equal to 100
```

La significatività osservata risulta essere quasi identica a quella ottenuta nell'Esempio 7.1.1. Tuttavia, in questo approccio non è stata fatta neanche l'assunzione di indipendenza delle osservazioni rispetto al test classico. \square

Con un approccio “distribution-free”, si supponga che x_1, \dots, x_n sia la determinazione di un campione casuale da una variabile casuale X con media μ . Di nuovo in un approccio condizionato alla realizzazione del campione x_1, \dots, x_n , la cosiddetta distribuzione “bootstrap” della media campionaria può essere ottenuta considerando tutti i campioni con ripetizione di ordine n dai valori x_1, \dots, x_n . Se B_1, \dots, B_n rappresenta un vettore di variabili casuali con distribuzione Binomiale (ognuna con

parametri n e $1/n$) e tali che $\sum_{i=1}^n B_i = n$, allora la media campionaria bootstrap è data dalla statistica test

$$\bar{X}_{\text{BOOT}} = \frac{1}{n} \sum_{i=1}^n x_i B_i .$$

Se si considera il sistema di ipotesi $H_0 : \mu = \mu_0$ contro $H_1 : \mu \neq \mu_0$ (o una ipotesi direzionale), il test bootstrap e il relativo intervallo di confidenza bootstrap della media possono essere basati sulla statistica \bar{X}_{BOOT} . La distribuzione bootstrap della statistica può essere calcolata, anche se ovviamente non è possibile esprimerla in forma chiusa. Anche in questo caso, la distribuzione bootstrap della statistica è usualmente approssimata mediante il metodo Monte Carlo. Nel caso dell'ipotesi alternativa bilaterale $H_1 : \mu \neq \mu_0$, si rifiuta H_0 per realizzazioni basse o elevate della statistica test. Nel caso dell'ipotesi alternativa direzionale $H_1 : \mu > \mu_0$ ($H_1 : \mu < \mu_0$), si rifiuta H_0 per realizzazioni elevate (basse) della statistica test.

• **Esempio 7.1.5.** Si considerano di nuovo i dati relativi alle sfere di acciaio dell'Esempio 4.3.4. La distribuzione bootstrap mediante il metodo Monte Carlo, e il relativo istogramma, della media campionaria si possono ottenere mediante i seguenti comandi:

```
> Boot.mean <- numeric(10000)
> for (i in 1:10000) Boot.mean[i] <- mean(sample(Diameter,
+   replace = T))
> hist(Boot.mean, xlab = "Bootstrap sample mean",
+   ylab = "Density", main = "Histogram")
```

Il precedente comando fornisce il grafico della Figura 7.1.1.

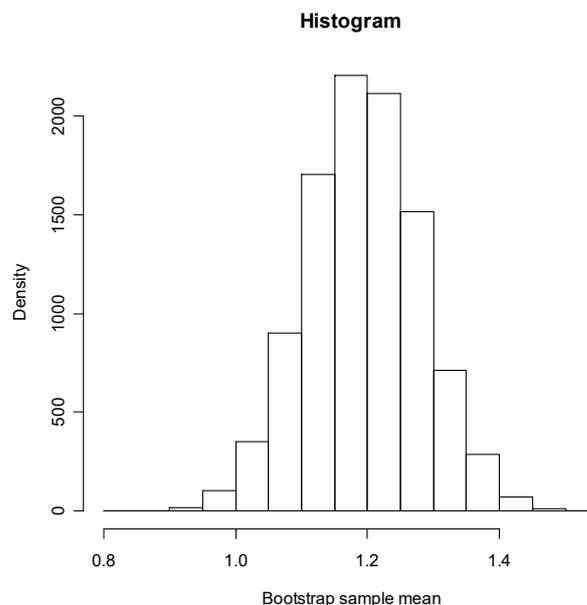


Figura 7.1.1.

Inoltre, la significatività osservata del test bootstrap può essere ottenuta mediante il seguente comando

```
> 2 * length(Boot.mean[Boot.mean < 1]) / 10000
[1] 0.0218
```

Richiamando la libreria `exactRankTests`, il comando `boot` fornisce l'implementazione della stima bootstrap della distorsione e della varianza della media campionaria:

```
> library(boot)
> m <- function(x, w) sum(x$Diameter * w)
> boot(d, m, R = 9999, stype = "w")
```

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:

```
boot(data = d, statistic = m, R = 9999, stype = "w")
```

Bootstrap Statistics :

```
      original      bias      std. error
t1*      1.194 0.0006592659  0.08684963
```

Inoltre, il comando `boot.ci` fornisce l'implementazione dell'intervallo di confidenza bootstrap per la media con quattro metodi differenti:

```
> boot.ci(boot(d, m, R = 9999, stype = "w"), conf = 0.95,
+         type = c("norm", "basic", "perc", "bca"))
```

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS

Based on 9999 bootstrap replicates

CALL :

```
boot.ci(boot.out = boot(d, m, R = 9999, stype = "w"), conf = 0.95,
        type = c("norm", "basic", "perc", "bca"))
```

Intervals :

```
Level      Normal              Basic
95%      ( 1.025,  1.363 )      ( 1.028,  1.363 )
```

```
Level      Percentile          BCa
95%      ( 1.025,  1.360 )      ( 1.021,  1.356 )
```

Calculations and Intervals on Original Scale

Dunque, la significatività osservata e l'ampiezza dell'intervallo di confidenza con l'approccio bootstrap sono minori di quelli ottenuti nell'Esempio 7.1.1 con un test classico. \square

Si considerino le osservazioni relative a n soggetti su cui è stata osservata una certa variabile prima e dopo un trattamento, ovvero si hanno le osservazioni x_{11}, \dots, x_{n1} prima del trattamento e le osservazioni x_{12}, \dots, x_{n2} dopo il trattamento. L'obiettivo è quello di valutare l'efficacia del trattamento e i dati di questo tipo sono detti appaiati. Al fine di analizzare queste osservazioni si costruiscono le differenze delle osservazioni d_1, \dots, d_n , dove $d_i = x_{i2} - x_{i1}$. Supponendo che queste differenze siano realizzazioni di un campione casuale proveniente da una variabile casuale da $D \sim N(\mu, \sigma^2)$, la verifica dell'efficacia del trattamento si riduce a considerare il sistema di ipotesi sulla media $H_0 : \mu = 0$ contro $H_1 : \mu \neq 0$ (o una ipotesi direzionale). Alternativamente, assumendo D come una variabile casuale con funzione di ripartizione (non nota) $F(x - \lambda)$ dove λ rappresenta la mediana, si può considerare il sistema di ipotesi sulla mediana $H_0 : \lambda = 0$ contro $H_1 : \lambda \neq 0$ (o una ipotesi direzionale). In questo caso, è sufficiente applicare le procedure di verifica di ipotesi viste in precedenza.

• **Esempio 7.1.6.** Su 8 pazienti con anemia cronica grave è stato misurato l'indice di infarto (in ml/battito/m²) prima e dopo un trattamento medico (Fonte: Bhatia, M.L., Manchanda, S.C. e Roy, S.B., 1969, Coronary haemodynamic studies in chronic severe anaemia, *British Heart Journal* **31**, 365-374). I dati sono contenuti nel file `stroke.txt` e vengono letti e resi disponibili mediante i seguenti comandi:

```
> d <- read.table("c:\\Rwork\\examples\\stroke.txt", header = T)
> attach(d)
> Difference <- Post - Pre
```

La stima di nucleo della funzione di densità viene ottenuta mediante i seguenti comandi:

```
> library(sm)
> sm.density(Difference, hcv(Difference, hstart = 0.01, hend = 100),
+   yht = 0.02, xlim = c(-70, 30),
+   xlab = "Stroke index difference")
> title(main = "Kernel density estimation ('CV' h = 18.272)")
```

I precedenti comandi forniscono il grafico della Figura 7.1.2.

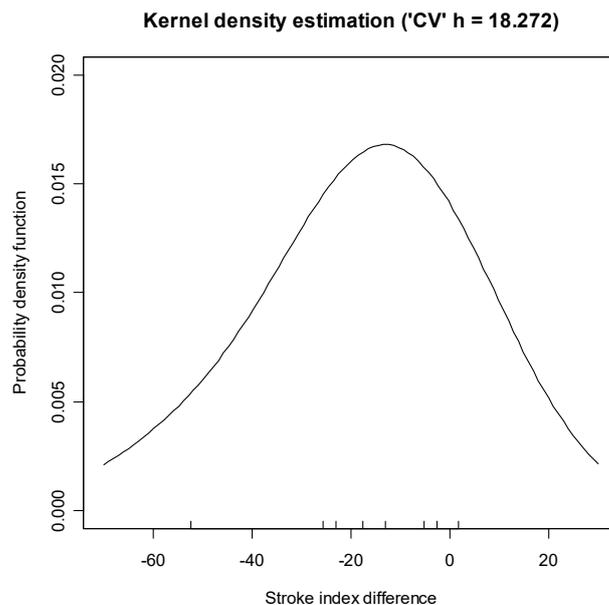


Figura 7.1.2.

Il diagramma quantile-quantile per verificare empiricamente l'ipotesi di normalità viene ottenuto mediante i seguenti comandi:

```
> qqnorm(Difference)
> qqline(Difference)
```

I precedenti comandi forniscono il grafico della Figura 7.1.3.

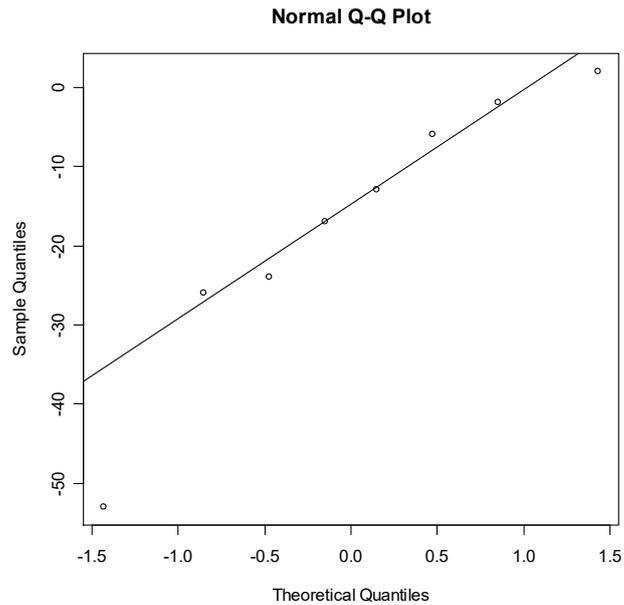


Figura 7.1.3.

Il sistema di ipotesi $H_0 : \mu = 0$ contro $H_1 : \mu < 0$ può essere verificato con il test t di Student mediante il seguente comando:

```
> t.test(Difference, alternative = "less", mu = 0)
```

One Sample t-test

```
data: Difference
t = -2.8055, df = 7, p-value = 0.01316
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
 -Inf -5.64162
sample estimates:
mean of x
 -17.375
```

Il sistema di ipotesi $H_0 : \lambda = 0$ contro $H_1 : \lambda < 0$ può essere verificato con il test di Wilcoxon mediante il seguente comando:

```
> wilcox.test(Difference, alternative = "less", mu = 0,
+             conf.int = F)
```

Wilcoxon signed rank test with continuity correction

```
data: Difference
V = 1.5, p-value = 0.01244
alternative hypothesis: true location is less than 0
```

Le significatività osservate ottenute con l'approccio classico e “distribution-free” sono dunque molto simili e portano a ritenere il trattamento efficace, in quanto l'indice di infarto tende a diminuire dopo il trattamento. \square

7.2. I test funzionali

Assumendo un modello statistico per un campionamento da una singola variabile, l'obiettivo è spesso quello di produrre inferenza per l'intera distribuzione. Questi metodi inferenziali sono anche detti test per la bontà d'adattamento. Si consideri inizialmente un campione casuale da una variabile casuale continua X con funzione di ripartizione F . Si desidera verificare il sistema di ipotesi funzionale $H_0 : F(x) = F_0(x)$ per ogni x , contro $H_1 : F(x) \neq F_0(x)$ per qualche x , dove F_0 è una funzione di ripartizione completamente specificata. Il test si basa sulla statistica di Kolmogorov data da

$$D = \sup_x |\hat{F}(x) - F_0(x)|.$$

Si può dimostrare che la distribuzione di D non dipende da F e quindi il test è “distribution-free”. Inoltre, la distribuzione della statistica test sotto ipotesi di base è nota e può essere tabulata. Se la funzione di ripartizione empirica si discosta molto da quella ipotizzata, allora si hanno valori elevati di D che conseguentemente portano a respingere l'ipotesi di base in favore dell'ipotesi alternativa. Per un calcolo pratico, la statistica D può essere convenientemente espressa come

$$D = \max_{1 \leq i \leq n} (\max(|i/n - F_0(X_{(i)})|, |(i-1)/n - F_0(X_{(i)})|)).$$

• **Esempio 7.2.1.** Sono stati determinati i carichi da applicare a un campione di fibre di poliestere al fine di provocarne il cedimento (Fonte: Quesenberry, C.P. e Hales, C., 1980, Concentration bands for uniformity plots, *Journal of Statistical Computation and Simulation* **11**, 41-53). Si sospetta che la distribuzione dei carichi segua una distribuzione Log-Normale. Le osservazioni originali sono state ricalcolate mediante una trasformazione che conduce all'uniformità, ovvero il nuovo campione deve provenire da una variabile casuale con distribuzione $U(0, 1)$ se l'ipotesi di log-normalità è vera. In questo caso si ha

$$F_0(x) = x\mathbf{1}_{]0,1[}(x) + \mathbf{1}_{[1,\infty[}(x).$$

I dati sono contenuti nel file `tensile.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\tensile.txt", header = T)
> attach(d)
```

Il grafico della funzione di ripartizione empirica e di quella teorica viene ottenuto mediante i seguenti comandi:

```
> plot(ecdf(Strength), do.points = F, verticals = T,
+      xlab = "Transformed tensile strength", ylab = "Probability",
+      main = "Distribution function")
> rug(Strength)
> plot(function(x) punif(x), -0.05, 1.05, lty = 3,
+      ylab = "Probability", add = T)
> legend(0.6, 0.3, c("Empirical", "Theoretical"), lty = c(1, 3))
```

I precedenti comandi forniscono il grafico della Figura 7.2.1.

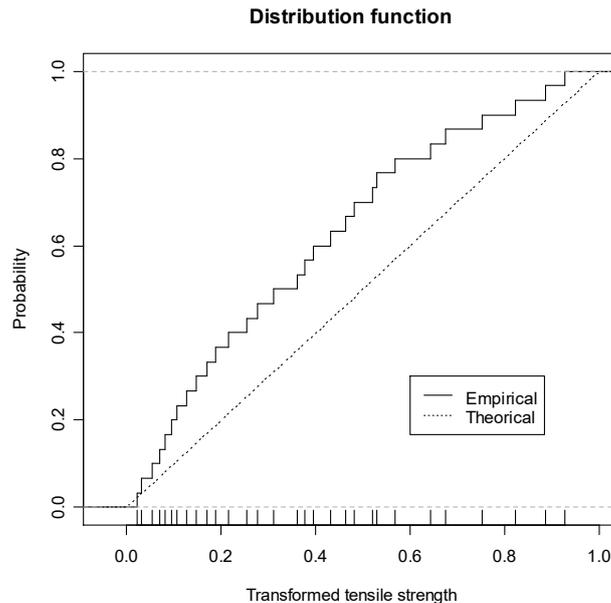


Figura 7.2.1.

Il test di Kolmogorov viene eseguito mediante il seguente comando:

```
> ks.test(Strength, "punif", 0, 1)

One-sample Kolmogorov-Smirnov test

data: Strength
D = 0.2377, p-value = 0.05644
alternative hypothesis: two-sided
```

La significatività osservata porta ad ipotizzare che l'uniformità della distribuzione considerata sia questionabile, anche se non è abbastanza bassa per permettere un rifiuto netto dell'ipotesi di base. \square

Se si considera un campionamento casuale da una variabile casuale discreta a supporto finito o da una variabile qualitativa X , allora le osservazioni campionarie possono essere espresse attraverso le frequenze osservate n_1, \dots, n_r delle realizzazioni distinte c_1, \dots, c_r della variabile in analisi. Se la funzione di probabilità di X è data da $p(c_j) = \pi_j$ per $j = 1, \dots, r$, le quantità $n\pi_1, \dots, n\pi_r$ sono dette frequenze attese. Si è interessati a verificare il sistema di ipotesi $H_0 : \pi_j = \pi_{0j}$ per ogni j , contro $H_1 : \pi_j \neq \pi_{0j}$ per qualche j . Dal momento che le probabilità π_j specificano completamente la distribuzione di X , la precedente ipotesi è a tutti gli effetti una ipotesi funzionale. Per verificare questo sistema di ipotesi si adotta la statistica test Chi-quadrato per la bontà d'adattamento data da

$$\chi^2 = \sum_{j=1}^r \frac{(n_j - n\pi_{0j})^2}{n\pi_{0j}}.$$

La distribuzione per grandi campioni di χ^2 non dipende dai valori π_{0j} e quindi il test è “distribution-free” per grandi campioni. In effetti, sotto ipotesi di base, per $n \rightarrow \infty$ la statistica test χ^2 converge in distribuzione ad una variabile casuale con distribuzione χ_{r-1}^2 . L'approssimazione è valida per campioni finiti se $n > 30$ e se tutte le frequenze attese sono maggiori di uno. Se le frequenze osservate si discostano molto dalle frequenze attese, si ottengono determinazioni elevate della statistica test che portano a respingere l'ipotesi di base.

• **Esempio 7.2.2.** È stata osservata la prima cifra dei numeri contenuti in un volume della rivista Reader's Digest scelto casualmente (Fonte: Benford, F., 1938, The law of anomalous numbers, *Proceedings of the American Philosophical Society* **78**, 551-572). Un modello teorico per questi dati è la cosiddetta distribuzione anomala con funzione di probabilità

$$p_0(x) = \log_{10} \left(1 + \frac{1}{x} \right) \mathbf{1}_{\{1, \dots, 9\}}(x).$$

I dati sono contenuti nel file `benford.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\benford.txt", header = T)
> attach(d)
```

Le probabilità teoriche vengono calcolate mediante i seguenti comandi:

```
> Theory.Probs <- numeric(9)
> for (i in 1:9) Theory.Probs[i] <- logb(1 + 1 / i, 10)
```

Il grafico della distribuzione di probabilità osservata e di quella teorica viene ottenuto mediante i seguenti comandi:

```
> h <- list(Digit = c(Digit, Digit), Type = c(rep("Theoretical", 9),
+ rep("Observed", 9)),
+ Probs = c(Theory.Probs, Counts / sum(Counts)))
> class(Table <- xtabs(Probs ~ ., h))
[1] "xtabs" "table"
> barplot(t(Table), beside = T, legend = colnames(Table),
+ xlab = "Digit", ylab = "Probability",
+ main = "Probability distribution")
```

I precedenti comandi forniscono il grafico della Figura 7.2.2.

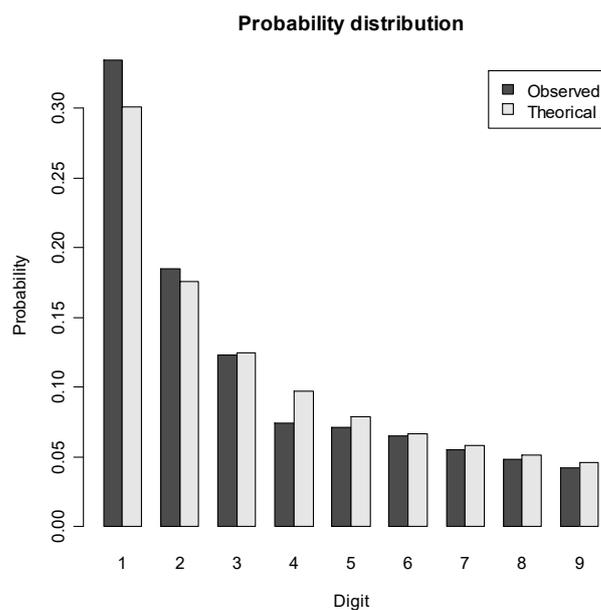


Figura 7.2.2.

Il test χ^2 viene ottenuto mediante il seguente comando:

```
> chisq.test(xtabs(Counts ~ ., d), p = Theory.Probs)
```

```
Chi-squared test for given probabilities
```

```
data: xtabs(Counts ~ ., d)
X-squared = 3.2735, df = 8, p-value = 0.916
```

Dal momento che la significatività osservata è molto elevata, si può accettare l'ipotesi di base. \square

• **Esempio 7.2.3.** In un esperimento di genetica sono stati considerati ibridi di pomodoro con un rapporto atteso di quattro fenotipi pari a 9 : 3 : 3 : 1 ottenendo le frequenze del numero di piante generate per ogni fenotipo (Fonte: McArthur, J.W., 1931, Linkage studies with the tomato III. Fifteen factors in six groups, *Transaction of the Royal Canadian Institute* **18**, 1-19. Si vuole verificare sperimentalmente i risultati della teoria genetica, ovvero l'ipotesi di base $H_0: \pi_1 = 9/16, \pi_2 = 3/16, \pi_3 = 3/16, \pi_4 = 1/16$. I dati sono contenuti nel file `tomato.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\tomato.txt", header = T)
> attach(d)
```

Le probabilità teoriche vengono calcolate mediante i seguenti comandi:

```
> Theory.Probs <- c(9 / 16, 3 / 16, 3 / 16, 1 / 16)
```

Il grafico delle distribuzioni di probabilità osservata e di quella teorica viene ottenuto mediante i seguenti comandi:

```
> h <- list(Phenotype = c(Phenotype, Phenotype),
+   Type = c(rep("Theoretical", 4), rep("Observed", 4)),
+   Probs = c(Theory.Probs, Counts / sum(Counts)))
> class(Table <- xtabs(Probs ~ ., h))
[1] "xtabs" "table"
> barplot(t(Table), beside = T, legend = colnames(Table),
+   names.arg = c("Tall cut-leaf", "Tall potato-leaf",
+   "Dwarf cut-leaf", "Dwarf potato-leaf"),
+   xlab = "Phenotype", ylab = "Probability",
+   main = "Probability distribution")
```

I precedenti comandi forniscono il grafico della Figura 7.2.3. Il test χ^2 viene implementato mediante il seguente comando:

```
> chisq.test(xtabs(Counts ~ ., d), p = Theory.Probs)
```

```
Chi-squared test for given probabilities
```

```
data: xtabs(Counts ~ ., d)
X-squared = 1.4687, df = 3, p-value = 0.6895
```

Dal momento che la significatività osservata è elevata, si può accettare l'ipotesi di base. \square

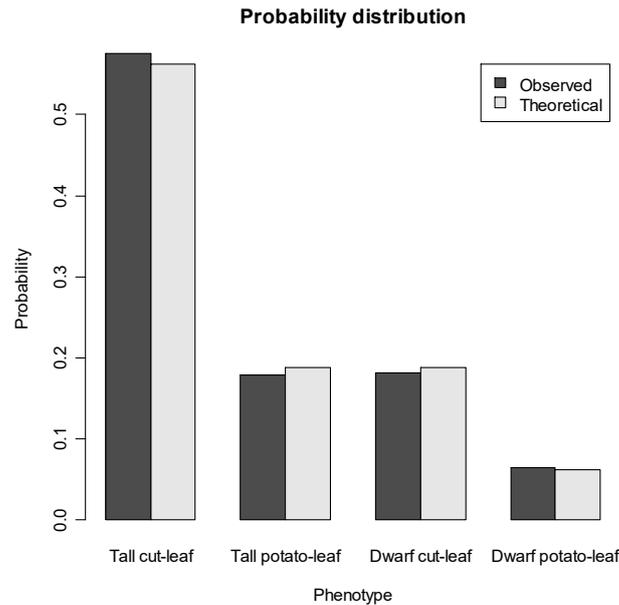


Figura 7.2.3.

Se la funzione di probabilità di X dipende da un insieme di k parametri (non noti) θ , ovvero se $p(c_j) = \pi_j(\theta)$, il sistema di ipotesi diventa $H_0 : \pi_j = \pi_{0j}(\theta)$ per ogni j , contro $H_1 : \pi_j \neq \pi_{0j}(\theta)$ per qualche j . Si assuma l'esistenza uno stimatore $\hat{\Theta}$ per θ che sia coerente, efficiente per grandi campioni e distribuito normalmente per grandi campioni. Lo stimatore $\hat{\Theta}$ viene usualmente ottenuto con il metodo della massima verosimiglianza e le quantità $n\pi_1(\hat{\Theta}), \dots, n\pi_r(\hat{\Theta})$ sono dette frequenze attese stimate. Per verificare questo sistema di ipotesi si adotta una opportuna modifica della statistica test Chi-quadrato data da

$$\chi^2 = \sum_{j=1}^r \frac{(n_j - n\pi_{0j}(\hat{\Theta}))^2}{n\pi_{0j}(\hat{\Theta})}.$$

La distribuzione per grandi campioni di χ^2 non dipende dai valori π_j e quindi il test è “distribution-free” per grandi campioni. Sotto ipotesi di base, per $n \rightarrow \infty$ la statistica test χ^2 converge in distribuzione ad una variabile casuale con distribuzione χ_{r-k-1}^2 . L'approssimazione è valida per campioni finiti se $n > 30$ e se tutte le frequenze attese stimate sono maggiori di uno. Se le frequenze osservate si discostano molto dalle frequenze attese stimate, si ottengono determinazioni elevate della statistica test che portano a respingere l'ipotesi di base.

• **Esempio 7.2.4.** È stato osservato il numero di taxi arrivati in ogni intervalli di un minuto alla stazione di Euston a Londra fra le 9.00 e le 10.00 di una mattina del 1950 (Fonte: Kendall, D.G., 1951, Some problems in the theory of queues, *Journal of the Royal Statistical Society* **B13**, 151-185). Se gli arrivi sono casuali, per la teoria dei processi stocastici, le osservazioni provengono da $X \sim Po(\mu)$. La distribuzione di Poisson è definita sui numeri naturali, e quindi si devono raggruppare le osservazioni maggiori di un predeterminato valore (in questo caso 5) in una unica classe. I dati sono contenuti nel file `taxi.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\taxi.txt", header = T)
> attach(d)
```

Le probabilità teoriche stimate vengono calcolate mediante i seguenti comandi:

```
> mu <- sum(Taxi * Counts) / sum(Counts)
> Theory.Probs <- numeric(6)
> for (i in 0:4) Theory.Probs[i + 1] <- dpois(i, mu)
> Theory.Probs[6] <- 1 - sum(Theory.Probs)
```

Il grafico delle distribuzioni di probabilità osservata e di quella teorica viene ottenuto mediante i seguenti comandi:

```
> h <- list(Taxi = c(Taxi, Taxi), Type = c(rep("Estimated", 6),
+   rep("Observed", 6)),
+   Probs = c(Theory.Probs, Counts / sum(Counts)))
> class(Table <- xtabs(Probs ~ ., h))
[1] "xtabs" "table"
> barplot(t(Table), beside = T, legend = colnames(Table),
+   names.arg = c("0", "1", "2", "3", "4", "+5"),
+   xlab = "Taxi", ylab = "Probability",
+   main = "Probability distribution")
```

I precedenti comandi forniscono il grafico della Figura 7.2.3.

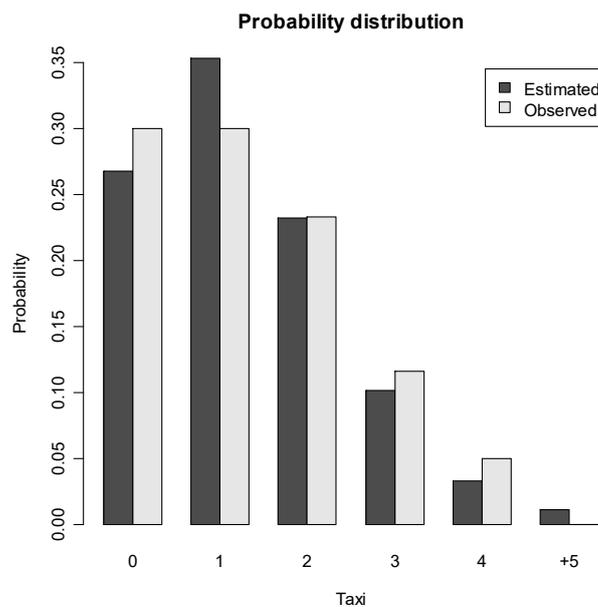


Figura 7.2.4.

Il test χ^2 viene ottenuto mediante il seguente comando:

```
> 1 - pchisq(sum((Counts - sum(Counts) * Theory.Probs)^2 /
+   (sum(Counts) * Theory.Probs)), 4)
[1] 0.7380024
```

Dal momento che la significatività osservata è elevata, si può accettare l'ipotesi di base. □

7.3. Riferimenti bibliografici

Agresti, A. (2019) *An Introduction to Categorical Data Analysis*, terza edizione, Wiley, New York.
 Chernick, M.R. (2008) *Bootstrap Methods*, seconda edizione, Wiley, New York.

- Davison, A.C. e Hinkley, D.V. (1997) *Bootstrap Methods and their Application*, Cambridge University Press, Cambridge.
- Efron, B. e Tibshirani, R. (1993) *An Introduction to the Bootstrap*, Chapman & Hall, London.
- Good (2005) *Permutation, Parametric and Bootstrap Test of Hypotheses*, terza edizione, Springer, New York.
- Hájek, J. (1969) *Nonparametric Statistics*, Holden Day, San Francisco.
- Hájek, J. e Šidák, Z. (1967) *Theory of Rank Tests*, Academic Press, New York.
- Hettmansperger, T.P. e McKean, J.W. (2011) *Robust Nonparametric Statistical Methods*, seconda edizione, Chapman & Hall/CRC Press, Boca Raton.
- Hollander, M., Wolfe, D.A. e Chicken, E. (2014) *Nonparametric Statistical Methods*, terza edizione, Wiley, New York.
- Kloke, J e McKean, J. (2024) *Nonparametric Statistical Methods Using R*, seconda edizione, CRC Press, Boca Raton.
- Kolassa, J.E. (2020) *An Introduction to Nonparametric Statistics*, CRC Press, Boca Raton.
- Kvam, P., Vidakovic, B. e Kim, S. (2023) *Nonparametric Statistics with Applications to Science and Engineering with R*, seconda edizione, Wiley, New York.
- Lehmann, E.L. e Romano, J.P. (2022) *Testing Statistical Hypothesis*, quarta edizione, Springer, Switzerland.
- Shao, J. e Tu, D. (1995) *The Jackknife and Bootstrap*, Springer, New York.

Pagina intenzionalmente vuota

Capitolo 8

L'inferenza con due variabili

8.1. L'inferenza per i parametri di tendenza centrale

Un modello statistico più elaborato di quello visto nel precedente Capitolo assume un campionamento da due variabili. Se la prima variabile è qualitativa e sotto controllo dello sperimentatore (ovvero è un fattore), mentre l'altra variabile è quantitativa, si ha la tipica situazione usualmente descritta come analisi con due campioni (se il fattore assume due determinazioni) o analisi della varianza (se il fattore assume più determinazioni). Ovviamente, queste strutture campionarie possono essere analizzate sia con un approccio classico che con un approccio “distribution-free”. L'analisi della dipendenza è finalizzata invece all'indagine dell'esistenza di un legame fra le due variabili, che possono essere entrambe quantitative o entrambe qualitative.

Risulta importante avere cautela per quanto riguarda l'uso della notazione nel caso in cui si analizzi una variabile qualitativa e una quantitativa. Ad esempio, quando si considera l'analisi con due campioni, viene assunta l'esistenza di due campioni casuali di numerosità n_1 e n_2 , tali che $n = n_1 + n_2$, provenienti da due differenti variabili casuali. Tuttavia, si dovrebbe affermare più correttamente che si dispone di n_1 osservazioni della variabile quantitativa Y al livello c_1 della variabile qualitativa X e di n_2 osservazioni della variabile quantitativa Y al livello c_2 della variabile qualitativa X . Dunque, si dovrebbe adoperare la notazione $(Y | X = c_1)$ e $(Y | X = c_2)$ per specificare le due variabili casuali da cui si sta campionando. Per parsimonia e con un lieve abuso di notazione, queste due variabili casuali vengono denotate con Y_{c_1} e Y_{c_2} nel seguito. Analogamente, se il fattore X assume r livelli c_1, \dots, c_r (ovvero quando si considera l'analisi della varianza) si adotta la notazione Y_{c_1}, \dots, Y_{c_r} . In questi casi, si ha comunque un'analisi statistica relativa a due variabili, ovvero una variabile qualitativa X sotto controllo e una variabile quantitativa Y su cui si desidera effettuare l'inferenza.

In un tipico approccio classico, si considerino inizialmente due campioni casuali indipendenti di numerosità n_1 e n_2 , con $n = n_1 + n_2$, da una variabile casuale Y a due livelli differenti c_1 e c_2 di un fattore, tali che $Y_{c_1} \sim N(\mu_1, \sigma^2)$ e $Y_{c_2} \sim N(\mu_2, \sigma^2)$. Il modello statistico è quindi indicizzato dai parametri μ_1, μ_2 e σ^2 . Quando si assume l'omogeneità delle varianze di Y_{c_1} e Y_{c_2} come in questo caso, si ha la cosiddetta ipotesi di omoschedasticità. Si indichi con \bar{Y}_{c_1} e $S_{c_1}^2$ la media campionaria e la varianza campionaria delle osservazioni provenienti da Y_{c_1} . Analogamente, si indichi con \bar{Y}_{c_2} e $S_{c_2}^2$ la media campionaria e la varianza campionaria delle osservazioni provenienti da Y_{c_2} . Si denoti inoltre con

$$S_w^2 = \frac{1}{n} (n_1 S_{c_1}^2 + n_2 S_{c_2}^2)$$

la media ponderata delle varianze campionarie. Si può facilmente dimostrare che \bar{Y}_{c_1} , \bar{Y}_{c_2} e S_w^2 sono gli stimatori di massima verosimiglianza di μ_1 , μ_2 e σ^2 . Se si considera il sistema di ipotesi $H_0 : \mu_1 = \mu_2$ contro $H_1 : \mu_1 \neq \mu_2$ (o una ipotesi direzionale), il test del rapporto delle verosimiglianze fornisce il test t di Student a due campioni basato sulla statistica test

$$T = \sqrt{\frac{n_1 n_2 (n-2)}{n^2}} \frac{\bar{Y}_{c_2} - \bar{Y}_{c_1}}{S_w},$$

che sotto ipotesi di base è distribuita come $T \sim t_{n-2}$. Nel caso dell'ipotesi alternativa bilaterale $H_1: \mu_1 \neq \mu_2$, si rifiuta H_0 per realizzazioni basse o elevate di T . Nel caso dell'ipotesi alternativa direzionale $H_1: \mu_1 > \mu_2$ ($H_1: \mu_1 < \mu_2$), si rifiuta H_0 per realizzazioni basse (elevate) di T . Il test basato su T è “distribution-free” per grandi campioni dal momento che T converge in distribuzione a una variabile casuale con distribuzione $N(0, 1)$ per $n \rightarrow \infty$. Il test t di Student a due campioni è quindi “robusto” rispetto all'assunzione di normalità.

Si è evidenziato che il test t di Student a due campioni assume l'omoschedasticità. Al fine di verificare l'omogeneità delle varianze è opportuno assumere $Y_{c_1} \sim N(\mu_1, \sigma_1^2)$ e $Y_{c_2} \sim N(\mu_2, \sigma_2^2)$ e verificare il sistema di ipotesi $H_0: \sigma_1^2 = \sigma_2^2$ contro $H_1: \sigma_1^2 \neq \sigma_2^2$. In questo caso, il test del rapporto delle verosimiglianze fornisce il test di Fisher basato sulla statistica test

$$F = \frac{S_{c_1}^2}{S_{c_2}^2},$$

che sotto ipotesi di base si distribuisce come $F \sim F_{n_1-1, n_2-1}$. L'ipotesi di base viene rifiutata per realizzazioni elevate di F . Nel caso che l'ipotesi di omoschedasticità venga respinta, per verificare l'omogeneità delle medie si può comunque adoperare la statistica test

$$T = \frac{\bar{Y}_{c_2} - \bar{Y}_{c_1}}{\sqrt{S_{c_1}^2/n_1 + S_{c_2}^2/n_2}},$$

che converge in distribuzione a una variabile casuale con distribuzione $N(0, 1)$ per $n \rightarrow \infty$ ed è quindi “distribution-free” per grandi campioni.

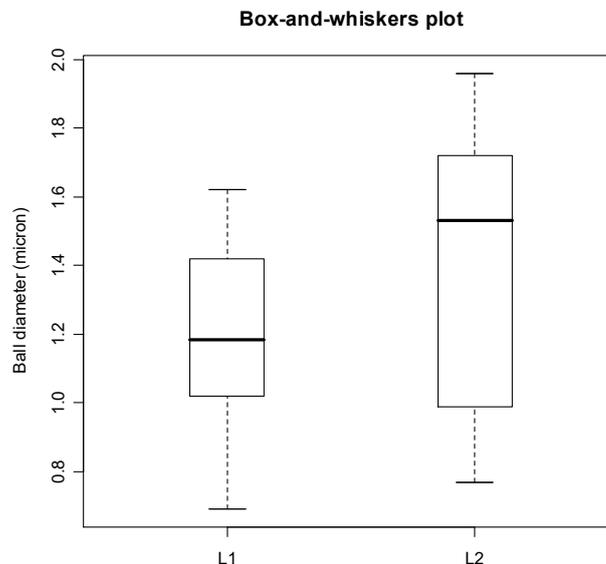


Figura 8.1.1.

• **Esempio 8.1.1.** Si dispone delle osservazioni di due campioni casuali di diametri di sfere misurate in micron provenienti da due differenti linee di produzione (Fonte: Romano, A., 1977, *Applied Statistics for Science and Industry*, Allyn and Bacon, Boston). I dati sono contenuti nel file `ball12.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\ball2.txt", header = T)
> attach(d)
```

I dati originali vengono ripartiti nei due campioni mediante i comandi:

```
> Diameter.1 <- split(Diameter, Line)[[1]]
> Diameter.2 <- split(Diameter, Line)[[2]]
```

Il diagramma a scatola e baffi condizionato viene ottenuto mediante il comando:

```
> boxplot(Diameter ~ Line, boxwex = 0.3,
+         ylab = "Ball diameter (micron)",
+         main = "Box-and-whiskers plot")
```

Il precedente comando fornisce il grafico della Figura 8.1.1. Le stime di nucleo delle funzioni di densità vengono ottenute mediante i comandi:

```
> library(sm)
> hcv(Diameter.1, hstart = 0.01, hend = 1)
> sm.density(Diameter.1, hcv(Diameter, hstart = 0.01, hend = 1),
+           yht = 0.92, xlim = c(-0.35, 2.65),
+           xlab = "Line 1 ball diameter (micron)")
> title(main = "Kernel density estimation ('CV' h = 0.32)")
> hcv(Diameter.2, hstart = 0.01, hend = 1)
> sm.density(Diameter.2, hcv(Diameter, hstart = 0.01, hend = 1),
+           yht = 0.92, xlim = c(-0.35, 2.65),
+           xlab = "Line 2 ball diameter (micron)")
> title(main = "Kernel density estimation ('CV' h = 0.37)")
```

I precedenti comandi forniscono il grafico della Figura 8.1.2.

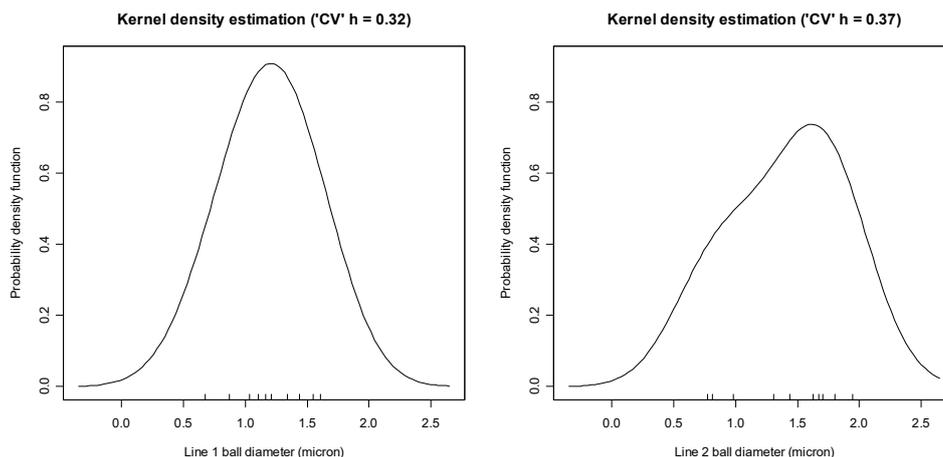


Figura 8.1.2.

Il sistema di ipotesi $H_0 : \sigma_1^2 = \sigma_2^2$ contro $H_1 : \sigma_1^2 \neq \sigma_2^2$ può essere verificato mediante il comando `var.test` che fornisce l'implementazione del test F di Fisher:

```
> var.test(Diameter.1, Diameter.2, paired = F,
+         alternative = "two.sided")
```

F test to compare two variances

```

data: Diameter.1 and Diameter.2
F = 0.4574, num df = 9, denom df = 9, p-value = 0.2595
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1136199 1.8416221
sample estimates:
ratio of variances
 0.4574329

```

Il sistema di ipotesi $H_0 : \mu_1 = \mu_2$ contro $H_1 : \mu_1 \neq \mu_2$ può essere verificato mediante il comando `t.test` che fornisce l'implementazione del test t di Student per due campioni:

```

> t.test(Diameter.1, Diameter.2, var.equal = T,
+        alternative = "two.sided")

```

Two Sample t-test

```

data: Diameter.1 and Diameter.2
t = -1.2965, df = 18, p-value = 0.2112
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5555277 0.1315277
sample estimates:
mean of x mean of y
 1.194      1.406

```

Il test per l'omogeneità delle medie senza assunzione di omoschedasticità viene ottenuto mediante i seguenti comandi:

```

> 2 * pnorm((mean(Diameter.1) - mean(Diameter.2)) /
+          (variance(Diameter.1) / length(Diameter.1) +
+          variance(Diameter.2) / length(Diameter.2))^(1 / 2))
[1] 0.1717296

```

Le significatività osservate portano dunque ad accettare l'omoschedasticità e successivamente l'omogeneità delle medie. Anche il test senza l'ipotesi di omoschedasticità permette di accettare l'ipotesi di omogeneità delle medie. \square

In un approccio “distribution-free”, si consideri di nuovo due campioni casuali indipendenti da una variabile casuale Y a due livelli differenti c_1 e c_2 di un fattore, tali che Y_{c_1} ha funzione di ripartizione $F(y - \lambda_1)$ e Y_{c_2} ha funzione di ripartizione $F(y - \lambda_2)$, mentre λ_1 e λ_2 rappresentano le rispettive mediane. Si indichi come campione misto l'insieme di tutte le n osservazioni senza considerare l'effetto del fattore. Inoltre, si assuma che R_1, \dots, R_{n_1} siano i ranghi assegnati alle osservazioni campionarie da Y_{c_1} nel campione misto, mentre siano R_{n_1+1}, \dots, R_n i ranghi assegnati alle osservazioni campionarie da Y_{c_2} nel campione misto. Se si considera il sistema di ipotesi $H_0 : \lambda_1 = \lambda_2$ contro $H_1 : \lambda_1 \neq \lambda_2$ (o una ipotesi direzionale), il test di Mann-Whitney è basato sulla statistica test

$$W = \sum_{i=1}^{n_1} R_i .$$

La distribuzione di W sotto ipotesi di base può essere tabulata, anche se non può essere espressa in forma chiusa. Nel caso dell'ipotesi alternativa bilaterale $H_1 : \lambda_1 \neq \lambda_2$, si rifiuta H_0 per realizzazioni

basse o elevate di W . Nel caso dell'ipotesi alternativa direzionale $H_1 : \lambda_1 > \lambda_2$ ($H_1 : \lambda_1 < \lambda_2$), si rifiuta H_0 per realizzazioni basse (elevate) di W .

• **Esempio 8.1.2.** Si considera di nuovo i dati relativi alle sfere di acciaio dell'Esempio 8.1.1. Il sistema di ipotesi $H_0 : \lambda_1 = \lambda_2$ contro $H_1 : \lambda_1 \neq \lambda_2$ può essere verificato mediante il comando `wilcox.test` che fornisce l'implementazione del test di Mann-Whitney:

```
> wilcox.test(Diameter ~ Line, alternative = "two.sided")

Wilcoxon rank sum test with continuity correction

data: Diameter by Line
W = 32.5, p-value = 0.1986
alternative hypothesis: true location shift is not equal to 0
```

Il test di Mann-Whitney produce una significatività simile a quella del test t di Student e si può dunque accettare l'ipotesi di omogeneità delle mediane. \square

In un approccio “distribution-free”, si consideri due campioni scambiabili da una variabile casuale Y a due livelli differenti c_1 e c_2 di un fattore, tali che Y_{c_1} ha funzione di ripartizione $F(y - \lambda_1)$ e Y_{c_2} ha funzione di ripartizione $F(y - \lambda_2)$, mentre λ_1 e λ_2 rappresentano le rispettive mediane. Se si considera il sistema di ipotesi $H_0 : \lambda_1 = \lambda_2$ contro $H_1 : \lambda_1 \neq \lambda_2$ (o una ipotesi direzionale), condizionatamente alla realizzazione del campione misto, sotto ipotesi di base ogni partizione del campione misto in due gruppi di numerosità n_1 e n_2 è ugualmente probabile. In questo caso, un test di permutazione è basato sulle $\binom{n_1+n_2}{n_1}$ (ugualmente probabili) permutazioni di livelli del fattore assegnati al campione misto. Se S_1, \dots, S_n rappresenta un vettore scambiabile di variabili casuali ognuna con supporto $\{-1, 1\}$, tali che $P(S_i = -1) = P(S_i = 1) = 1/2$ e $\sum_{i=1}^n S_i = n_1 - n_2$, il test di permutazione dei segni è basato sulla statistica test

$$T = \sum_{i=1}^n x_i S_i .$$

La distribuzione della statistica test T è usualmente approssimata mediante il metodo Monte Carlo. Nel caso dell'ipotesi alternativa bilaterale $H_1 : \lambda_1 \neq \lambda_2$, si rifiuta H_0 per realizzazioni basse o elevate di T . Nel caso dell'ipotesi alternativa direzionale $H_1 : \lambda_1 > \lambda_2$ ($H_1 : \lambda_1 < \lambda_2$), si rifiuta H_0 per realizzazioni basse (elevate) di T .

• **Esempio 8.1.3.** Si considera di nuovo i dati relativi alle sfere di acciaio dell'Esempio 8.1.1. Richiamando la libreria `exactRankTests`, il comando `wilcox.test` fornisce l'implementazione del test di permutazione (le osservazioni sono state moltiplicate per cento al fine di ottenere valori interi per ottenere una elaborazione più rapida):

```
> perm.test(round(100 * Diameter.1), round(100 * Diameter.2),
+   paired = F, alternative = "two.sided")

2-sample Permutation Test

data: round(100 * Diameter.1) and round(100 * Diameter.2)
T = 1194, p-value = 0.2105
alternative hypothesis: true mu is not equal to 0
```

Il test di permutazione produce una significatività simile a quella del test di Mann-Whitney e si può di nuovo accettare l'ipotesi di omogeneità delle mediane. \square

In un approccio “distribution-free”, si consideri due campioni casuali da una variabile casuale Y a due livelli differenti c_1 e c_2 di un fattore, tali che Y_{c_1} ha funzione di ripartizione $F(y - \mu_1)$ e Y_{c_2} ha funzione di ripartizione $F(y - \mu_2)$, dove μ_1 e μ_2 rappresentano le rispettive medie. Condizionatamente alla realizzazione del campione misto, la distribuzione bootstrap della differenza delle medie campionarie indicata con T_{BOOT} può essere ottenuta considerando tutti i campioni con ripetizione di ordine n dal campione misto che vengono successivamente ripartiti in due campioni di numerosità n_1 e n_2 . Se si considera il sistema di ipotesi $H_0 : \mu_1 = \mu_2$ contro $H_1 : \mu_1 \neq \mu_2$ (o una ipotesi direzionale), il test bootstrap e il relativo intervallo di confidenza bootstrap può essere basato sulla statistica test T_{BOOT} . Anche in questo caso, la distribuzione della statistica test T_{BOOT} è usualmente approssimata mediante il metodo Monte Carlo.

• **Esempio 8.1.4.** Si considera di nuovo i dati relativi alle sfere di acciaio. La significatività osservata del test bootstrap può essere ottenuta mediante i seguenti comandi:

```
> Boot.meandif <- numeric(10000)
> Boot.sample <- numeric(length(Diameter))
> for (i in 1:10000) {Boot.sample <- sample(Diameter, replace = T);
+   Boot.diameter1 <- Boot.sample[c(1:length(Diameter.1))];
+   Boot.diameter2 <-
+   Boot.sample[c((length(Diameter.1) + 1):length(Diameter))];
+   Boot.meandif[i] <- mean(Boot.diameter1) - mean(Boot.diameter2)}
> 2 * length(Boot.meandif[Boot.meandif <
+   mean(Diameter.1) - mean(Diameter.2)]) / 10000
[1] 0.191
```

Il test bootstrap produce a sua volta una significatività simile a quella del test di Mann-Whitney e si può accettare l'ipotesi di omogeneità delle medie. \square

8.2. I test funzionali

In un approccio “distribution-free”, si consideri due campioni casuali indipendenti, di numerosità n_1 e n_2 con $n = n_1 + n_2$, da una variabile casuale Y a due livelli differenti c_1 e c_2 di un fattore e tali che Y_{c_1} ha funzione di ripartizione F_{c_1} e Y_{c_2} ha funzione di ripartizione F_{c_2} . Si assuma che \hat{F}_{c_1} e \hat{F}_{c_2} siano rispettivamente le funzioni di ripartizione empiriche relative alle osservazioni provenienti da Y_{c_1} e Y_{c_2} . Se si considera il sistema di ipotesi $H_0 : F_{c_1}(y) = F_{c_2}(y)$ per ogni y , contro $H_1 : F_{c_1}(y) \neq F_{c_2}(y)$ per qualche y , il test di Kolmogorov-Smirnov è basato sulla statistica test

$$D = \sup_y |\hat{F}_{c_1}(y) - \hat{F}_{c_2}(y)|.$$

Si può dimostrare che la distribuzione di D non dipende da F_{c_1} e F_{c_2} sotto ipotesi di base e quindi il test è “distribution-free”. La distribuzione della statistica test sotto ipotesi di base è nota e può essere tabulata. Evidentemente, se le due funzioni di ripartizione empiriche si discostano molto fra loro, allora si hanno valori elevati di D che conseguentemente portano a respingere l'ipotesi di base in favore dell'ipotesi alternativa. Per un calcolo pratico, se $Y_{(1)}, \dots, Y_{(n)}$ è la statistica ordinata relativa al campione misto, la statistica D può essere espressa come

$$D = \max_{1 \leq i \leq n} |\hat{F}_{c_1}(Y_{(i)}) - \hat{F}_{c_2}(Y_{(i)})|.$$

• **Esempio 8.2.1.** Si considera di nuovo i dati relativi alle sfere di acciaio dell'Esempio 8.1.1. I grafici delle due funzioni di ripartizione empiriche sono ottenute mediante i seguenti comandi:

```
> plot(ecdf(Diameter.1), do.points = F, verticals = T,
+      xlim = c(0.5, 2.0), lty = 1,
+      xlab = "Ball diameter (micron)", ylab = "Probability",
+      main = "Empirical distribution function")
> plot(ecdf(Diameter.2), do.points = F, verticals = T,
+      lty = 3, add = T)
> legend(1.5, 0.3, c("Line 1", "Line 2"), lty = c(1, 3))
```

I precedenti comandi forniscono il grafico della Figura 8.2.1.

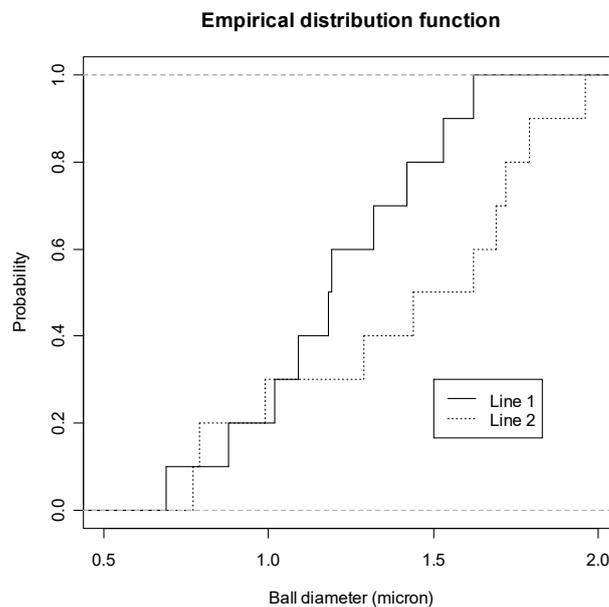


Figura 8.2.1.

Il sistema di ipotesi $H_0 : F_{c_1}(y) = F_{c_2}(y)$ per ogni y , contro $H_1 : F_{c_1}(y) \neq F_{c_2}(y)$ per qualche y , può essere verificato mediante il seguente comando:

```
> ks.test(Diameter.1, Diameter.2)
```

Two-sample Kolmogorov-Smirnov test

```
data: Diameter.1 and Diameter.2
D = 0.4, p-value = 0.4005
alternative hypothesis: two-sided
```

Il test di Kolmogorov-Smirnov produce porta dunque ad accettare l'omogeneità delle distribuzioni. Si osservi che la significatività del test è più alta rispetto a quella che è stata ottenuta nel caso dei test per l'omogeneità della tendenza centrale. Questo fatto è dovuto alla generalità del test di Kolmogorov-Smirnov rispetto alla specificità dei test visti nella precedente Sezione. \square

8.3. L'analisi della varianza

In un tipico approccio classico, si consideri r campioni casuali indipendenti, ciascuno di numerosità n_j e tali che $\sum_{j=1}^r n_j = n$, da una variabile casuale Y a r livelli differenti c_1, \dots, c_r di un fattore e tali che $Y_{c_j} \sim N(\mu_j, \sigma^2)$. Siano \bar{Y}_{c_j} e $S_{c_j}^2$ la media campionaria e la varianza campionaria delle osservazioni provenienti da \bar{Y}_{c_j} , mentre siano

$$S_b^2 = \frac{1}{n} \sum_{j=1}^r n_j (\bar{Y}_{c_j} - \bar{Y})^2$$

e

$$S_w^2 = \frac{1}{n} \sum_{j=1}^r n_j S_{c_j}^2$$

la cosiddetta varianza “between” (ovvero fra i gruppi) e la cosiddetta varianza “within” (ovvero all'interno dei gruppi). Si può verificare che gli stimatori di massima verosimiglianza di μ_1, \dots, μ_r e σ^2 sono dati rispettivamente da $\bar{Y}_{c_1}, \dots, \bar{Y}_{c_r}$ e S_w^2 . Se si considera il sistema di ipotesi $H_0 : \mu_1 = \dots = \mu_r$ contro $H_1 : \mu_j \neq \mu_l$ per qualche (j, l) , il test del rapporto delle verosimiglianze fornisce il test F di Fisher per l'analisi della varianza basato sulla statistica test

$$F = \frac{(n - m) S_b^2}{(m - 1) S_w^2},$$

che sotto ipotesi di base si distribuisce come $F \sim F_{r-1, n-r}$. L'ipotesi di base viene rifiutata per realizzazioni elevate di F .

Quando l'ipotesi di base viene rifiutata, si desidera conoscere da quale coppia di medie dipende il rifiuto. Non è opportuno effettuare singoli test per la verifica dell'omogeneità di coppie di medie, in quanto le statistiche test sono dipendenti e quindi la significatività globale non può essere calcolata a partire da quella dei singoli test. In questo caso, la procedura di Tukey genera un insieme di intervalli di confidenza simultanei per le $r(r-1)/2$ possibili differenze $(\mu_j - \mu_l)$. Questa procedura è “distribution-free” per grandi campioni. Data la dualità esistente fra intervallo di confidenza e test statistico, analizzando gli intervalli di confidenza simultanei che non contengono il valore zero, si può risalire alle coppie di medie che hanno causato il rifiuto dell'ipotesi di base nell'analisi della varianza.

• **Esempio 8.3.1.** In un famoso esperimento di Michelson e Morley sono stati fatti 5 esperimenti di 20 prove ognuno per determinare la velocità della luce (Fonte: Weekes, A.J., 1986, *A Genstat Primer*, Arnold, London). Queste misurazioni riportano solo le ultime 3 cifre (senza decimali) della velocità della luce (in km/sec). Si noti che la moderna misurazione risulta 299,792.458 (km/sec). I dati sono contenuti nel file `light.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\light.txt", header = T)
> attach(d)
```

I dati originali vengono ripartiti nei cinque campioni mediante i comandi:

```
> Speed.1 <- split(Speed, Trial)[[1]]
> Speed.2 <- split(Speed, Trial)[[2]]
> Speed.3 <- split(Speed, Trial)[[3]]
> Speed.4 <- split(Speed, Trial)[[4]]
> Speed.5 <- split(Speed, Trial)[[5]]
```

Il diagramma a scatola e baffi condizionato viene ottenuto mediante il comando:

```
> boxplot(Speed ~ Trial, main = "Box-and-whiskers plot",
+         xlab = "Experiment Number")
```

Il precedente comando fornisce il grafico della Figura 8.3.1.

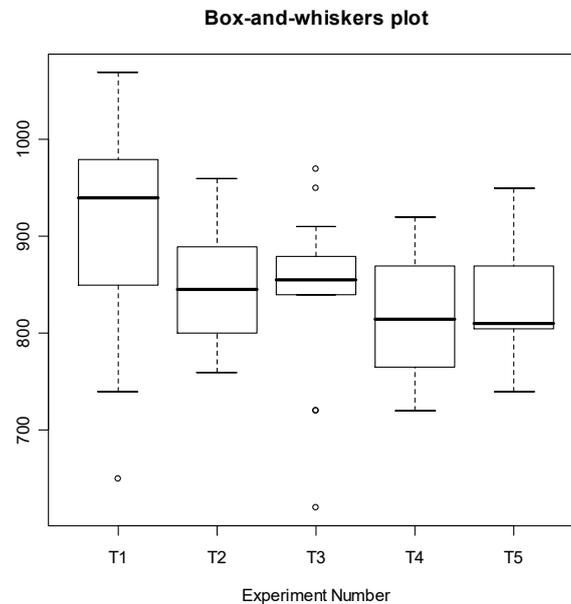


Figura 8.3.1.

Le stime di nucleo delle funzioni di densità vengono ottenute mediante i comandi:

```
> library(sm)
> par(mfrow = c(3, 2))
> sm.density(Speed.1, hnorm(Speed.1), yht = 0.008,
+           xlim = c(580, 1120), xlab = "Ligth speed (Trial 1)")
> title(main = "Kernel density estimation")
> sm.density(Speed.2, hnorm(Speed.2), yht = 0.008,
+           xlim = c(580, 1120), xlab = "Ligth speed (Trial 2)")
> title(main = "Kernel density estimation")
> sm.density(Speed.3, hnorm(Speed.3), yht = 0.008,
+           xlim = c(580, 1120), xlab = "Ligth speed (Trial 3)")
> title(main = "Kernel density estimation")
> sm.density(Speed.4, hnorm(Speed.4), yht = 0.008,
+           xlim = c(580, 1120), xlab = "Ligth speed (Trial 4)")
> title(main = "Kernel density estimation")
> sm.density(Speed.5, hnorm(Speed.5), yht = 0.008,
+           xlim = c(580, 1120), xlab = "Ligth speed (Trial 5)")
> title(main = "Kernel density estimation")
> par(mfrow = c(1, 1))
```

I precedenti comandi forniscono il grafico della Figura 8.3.2.

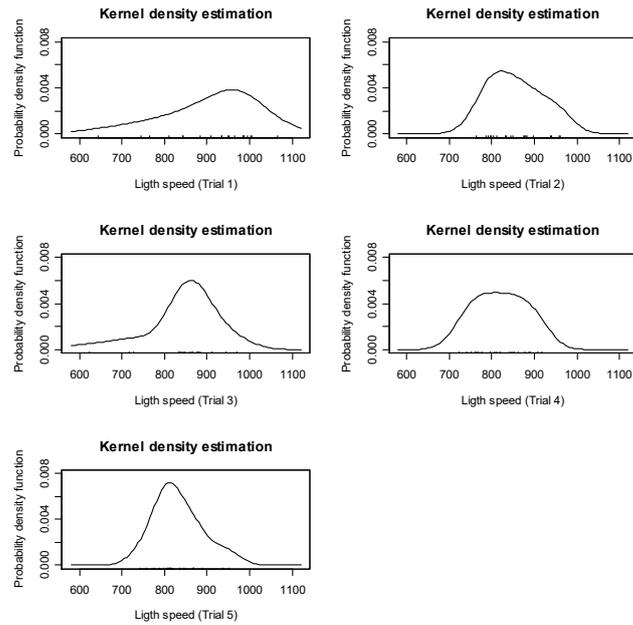


Figura 8.3.2.

L'analisi della varianza viene implementata mediante il comando:

```
> summary(aov(Speed ~ Trial))
              Df Sum Sq Mean Sq F value    Pr(>F)
Trial          4  94514   23629   4.2878 0.003114 **
Residuals     95 523510    5511
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La procedura di Tukey può essere implementata mediante il seguente comando:

```
> TukeyHSD(aov(Speed ~ Trial), "Trial")
  Tukey multiple comparisons of means
  95% family-wise confidence level

Fit: aov(formula = Speed ~ Trial)

$Trial
      diff       lwr       upr     p adj
T2-T1 -53.0 -118.28006  12.280058 0.1679880
T3-T1 -64.0 -129.28006   1.280058 0.0574625
T4-T1 -88.5 -153.78006 -23.219942 0.0025733
T5-T1 -77.5 -142.78006 -12.219942 0.0115793
T3-T2 -11.0  -76.28006  54.280058 0.9899661
T4-T2 -35.5 -100.78006  29.780058 0.5571665
T5-T2 -24.5  -89.78006  40.780058 0.8343360
T4-T3 -24.5  -89.78006  40.780058 0.8343360
T5-T3 -13.5  -78.78006  51.780058 0.9784065
T5-T4  11.0  -54.28006  76.280058 0.9899661
```

L'analisi della varianza porta a rifiutare l'ipotesi di omogeneità delle medie. Inoltre, la procedura di Tukey evidenzia che il rifiuto risulta principalmente da $\mu_1 > \mu_4$ e $\mu_1 > \mu_5$. \square

In un approccio “distribution-free”, si consideri r campioni casuali indipendenti, ciascuno di numerosità n_j e tali che $\sum_{j=1}^r n_j = n$, da una variabile casuale Y a r livelli differenti c_1, \dots, c_r di un fattore, tali che Y_{c_j} ha funzione di ripartizione $F(y - \lambda_j)$ e λ_j rappresenta la rispettiva mediana. Si assuma che R_{c_j} sia la somma dei ranghi assegnati alle osservazioni provenienti da Y_{c_j} nel campione misto. Se si considera il sistema di ipotesi $H_0 : \lambda_1 = \dots = \lambda_r$ contro $H_1 : \lambda_j \neq \lambda_l$ per qualche (j, l) , il test di Kruskal-Wallis è basato sulla statistica test

$$H = \frac{12}{n(n+1)} \sum_{j=1}^r n_j (R_{c_j}/n_j - (n+1)/2)^2.$$

La distribuzione di H sotto ipotesi di base può essere tabulata anche se non può essere espressa in forma chiusa. Inoltre, per grandi campioni H converge in distribuzione ad una variabile casuale con distribuzione χ_{r-1}^2 . Evidentemente, valori elevati della realizzazione di F portano al rifiuto dell'ipotesi di base.

• **Esempio 8.3.2.** Si considerano di nuovo i dati all'esperimento sulla velocità della luce dell'Esempio 8.3.1. Il test di Kruskal-Wallis viene implementato mediante il seguente comando:

```
> kruskal.test(Speed ~ Trial)

Kruskal-Wallis rank sum test

data: Speed by Trial
Kruskal-Wallis chi-squared = 15.0221, df = 4, p-value = 0.004656
```

Il test di Kruskal-Wallis produce a sua volta una significatività simile a quella ottenuta nell'analisi della varianza e si rifiuta quindi l'ipotesi di omogeneità delle medie. \square

8.4. L'inferenza per l'associazione

In un tipico approccio classico, si consideri un campione casuale $(X_1, Y_1), \dots, (X_n, Y_n)$ da una variabile casuale Normale bivariata (X, Y) . Per una proprietà di caratterizzazione della Normale bivariata, esclusivamente per questa distribuzione si ha che la nullità del coefficiente di correlazione, ovvero

$$\rho_{xy} = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}},$$

implica l'indipendenza delle componenti marginali X e Y . Dunque la verifica dell'indipendenza si riduce alla verifica dell'ipotesi $H_0 : \rho_{xy} = 0$ contro $H_1 : \rho_{xy} \neq 0$. Il test del rapporto delle verosimiglianze fornisce il test basato sul rapporto di correlazione campionario $R_{xy} = S_{xy}/(S_x S_y)$, ovvero

$$T = \sqrt{n-2} \frac{R_{xy}}{\sqrt{1-R_{xy}^2}},$$

che sotto ipotesi di base si distribuisce come $T \sim t_{n-2}$. Si noti inoltre che $F = T^2 \sim F_{1, n-2}$. Evidentemente, valori elevati della realizzazione di F portano al rifiuto dell'ipotesi di base.

• **Esempio 8.4.1.** Si sono considerate le misure (in metri) fatte nel lancio del peso e del giavellotto dalle 25 atlete partecipanti alla gara di eptathlon femminile alle Olimpiadi del 1988 (Fonte: Lunn, A.D.

e McNeil, D.R., 1991, *Computer-interactive Data Analysis*, Wiley, New York). I dati sono contenuti nel file `heptathlon.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\heptathlon.txt", header = T)
> attach(d)
```

Il diagramma di dispersione con la stima di nucleo della funzione di densità bivariata viene ottenuto mediante i comandi:

```
> library(sm)
> plot(Shot, Javelin, xlab = "Shot (meters)",
+      ylab = "Javelin (meters)")
> sm.density(d[, c(1, 2)], hcv(d[, c(1, 2)]), display = "slice",
+      props = c(75, 50, 25), add = T)
> title(main = "Kernel density estimation
+       ('CV' h1 = 1.11, h2 = 2.64)")
```

I precedenti comandi forniscono il grafico della Figura 8.4.1.

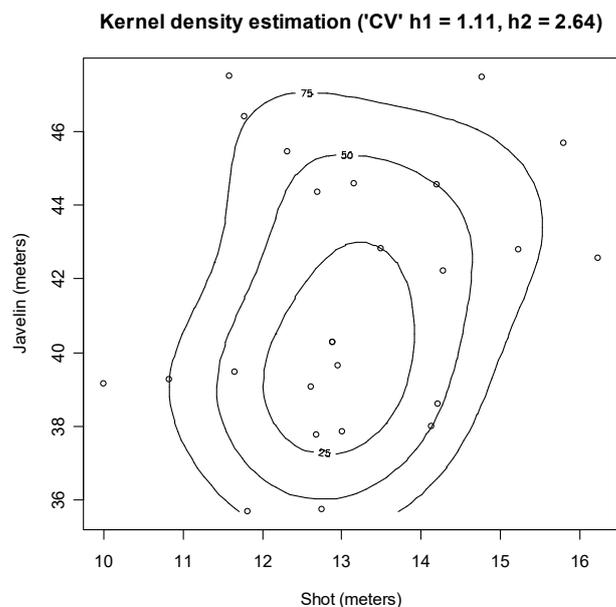


Figura 8.4.1.

Il sistema di ipotesi $H_0 : \rho_{xy} = 0$ contro $H_1 : \rho_{xy} \neq 0$ può essere verificato mediante il comando `cor.test` che fornisce l'implementazione del test per l'indipendenza:

```
> cor.test(Shot, Javelin, method = "pearson")
```

Pearson's product-moment correlation

```
data: Shot and Javelin
t = 1.3394, df = 23, p-value = 0.1935
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.1411436  0.6003148
sample estimates:
cor
0.2689888
```

Sulla base della significatività osservata si accetta dunque l'ipotesi di indipendenza fra i risultati nel lancio del peso e del giavellotto. \square

In un ambito generale non è possibile verificare l'ipotesi di indipendenza basandosi su ipotesi relative ad un singolo parametro che descrive la dipendenza fra le componenti della coppia di variabili casuali. In questo caso, ci si deve limitare a verificare la presenza o l'assenza di associazione, ovvero dell'esistenza di una relazione di dipendenza diretta o inversa tra le variabili. Si consideri un campione casuale $(X_1, Y_1), \dots, (X_n, Y_n)$ da una variabile casuale bivariata (X, Y) . Il coefficiente di correlazione campionario di Spearman è il coefficiente di correlazione campionario calcolato sui ranghi relativi a X_1, \dots, X_n e sui ranghi relativi a Y_1, \dots, Y_n . Il coefficiente di correlazione di Spearman può essere ottenuto semplicemente ordinando rispetto alle realizzazioni di Y_1, \dots, Y_n e successivamente assegnando i ranghi R_1, \dots, R_n alle realizzazioni di X_1, \dots, X_n . Il coefficiente di correlazione di Spearman risulta dunque

$$\rho_S = \frac{12}{n(n^2 - 1)} \sum_{i=1}^n iR_i - \frac{3(n+1)}{n-1}.$$

La statistica ρ_S è basata sui ranghi e quindi è “distribution-free”. La statistica ρ_S gode ovviamente di tutte le proprietà di un coefficiente di correlazione campionario. Dunque, valori di ρ_S intorno allo zero denotano mancanza di associazione, mentre valori vicino ad 1 o -1 denotano presenza di associazione monotona diretta e inversa. La statistica test ρ_S può essere dunque adottata per la verifica dell'ipotesi di associazione. In questo caso sotto ipotesi di base viene verificata la mancanza di associazione. La distribuzione di ρ_S sotto ipotesi di base può essere tabulata anche se non può essere espressa in forma chiusa.

• **Esempio 8.4.2.** Si considera di nuovo i dati dell'epathlon dell'Esempio 8.4.1. Il comando `cor.test` fornisce l'implementazione del test di Spearman:

```
> cor.test(Shot, Javelin, method = "spearman")

Spearman's rank correlation rho

data: Shot and Javelin
S = 2062.793, p-value = 0.3217
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.2066179
```

Sulla base della significatività osservata si accetta dunque l'ipotesi di mancanza di associazione fra i risultati nel lancio del peso e del giavellotto. \square

Si consideri di nuovo un campione casuale $(X_1, Y_1), \dots, (X_n, Y_n)$ da una variabile casuale bivariata (X, Y) . Il cosiddetto coefficiente di correlazione di Kendall è dato dalla percentuale di coppie campionarie concordanti (ovvero coppie campionarie con lo stesso segno), ovvero

$$\tau = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{segn}(X_j - X_i) \text{segn}(Y_j - Y_i),$$

dove $\text{segn}(x) = 2\mathbf{1}_{]0, \infty[}(x) - 1$. La statistica τ gode delle proprietà di un indice di dipendenza. Evidentemente, valori intorno a 0 denotano mancanza di associazione, mentre valori vicino ad 1 o -1

denotano presenza di associazione monotona diretta e inversa. La statistica test τ può essere dunque adottata per la verifica dell'ipotesi di associazione. La distribuzione di τ sotto ipotesi di base può essere tabulata anche se non può essere espressa in forma chiusa.

• **Esempio 8.4.3.** Si considera di nuovo i dati dell'epathlon dell'Esempio 8.4.1. Il comando `cor.test` fornisce l'implementazione del test di Kendall:

```
> cor.test(Shot, Javelin, method = "kendall")

Kendall's rank correlation tau

data: Shot and Javelin
z = 1.0515, p-value = 0.293
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
0.1505017
```

Il test basato sul coefficiente di correlazione di Kendall fornisce una significatività simile a quella del test basato sul coefficiente di correlazione campionario di Spearman e si accetta dunque l'ipotesi di mancanza di associazione fra i risultati nel lancio del peso e del giavellotto. \square

Se si considera un campionamento casuale da una variabile qualitativa bivariata (X, Y) , allora le osservazioni campionarie forniscono le frequenze osservate congiunte n_{jl} delle realizzazioni distinte (c_j, d_l) della variabile. Si assuma inoltre che la funzione di probabilità congiunta di (X, Y) sia data da $p_{X,Y}(c_j, d_l) = \pi_{jl}$, dove $j = 1, \dots, r, l = 1, \dots, s$, mentre la distribuzione marginale di probabilità di X sia data da $p_X(c_j) = \pi_{j+}$, dove $j = 1, \dots, r$ con $\pi_{j+} = \sum_{l=1}^s \pi_{jl}$, e quella di Y sia data da $p_Y(d_l) = \pi_{+l}$, $l = 1, \dots, s$ con $\pi_{+l} = \sum_{j=1}^r \pi_{jl}$. Si è interessati a verificare l'indipendenza di X e Y , ovvero il sistema di ipotesi $H_0 : \pi_{jl} = \pi_{j+}\pi_{+l}$ per ogni (j, l) , contro $H_1 : \pi_{jl} \neq \pi_{j+}\pi_{+l}$ per qualche (j, l) . Per verificare questo sistema di ipotesi si adotta la statistica test Chi-quadrato per l'indipendenza data da

$$\chi^2 = \sum_{j=1}^r \sum_{l=1}^s \frac{(n_{jl} - n_{j+}n_{+l}/n)^2}{n_{j+}n_{+l}/n}.$$

Le quantità $n_{j+}n_{+l}/n$ sono le frequenze attese stimate sotto ipotesi d'indipendenza. La distribuzione per grandi campioni di χ^2 non dipende dai valori π_{jl} e quindi il test è “distribution-free” per grandi campioni. Sotto ipotesi di base, per $n \rightarrow \infty$ la statistica test χ^2 converge in distribuzione ad una variabile casuale con distribuzione $\chi^2_{(r-1)(s-1)}$. L'approssimazione è valida per campioni finiti se $n > 30$ e se tutte le frequenze attese stimate sono maggiori di uno. Se le frequenze osservate si discostano molto dalle frequenze attese stimate, si ottengono determinazioni elevate della statistica test che portano a respingere l'ipotesi di base.

• **Esempio 8.4.4.** Durante uno studio della malattia di Hodgkin sono stati considerati 538 malati, ognuno dei quali è stato classificato per tipologie istologiche (indicate con le sigle LP, NS, MC e LD) e per la risposta al trattamento dopo tre mesi di cura (Fonte: Dunsmore, I.R. e Daly, F., 1987, *M345 Statistical Methods, Unit 9: Categorical Data*, The Open University, Milton Keynes). I dati sono contenuti nel file `hodgkin.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\hodgkin.txt", header = T)
> attach(d)
```

La tabella a doppia entrata viene ottenuta mediante il comando:

```
> xtabs(Count ~ Type + Response)
      Response
Type None Part Pos
LD    44   10  18
LP    12   18  74
MC    58   54 154
NS    12   16  68
```

Il diagramma a nastri condizionato viene ottenuto mediante il comando:

```
> library(lattice)
> barchart(xtabs(Count ~ Type + Response),
+         ylab = "Hystological type",
+         auto.key = list(title = "Response", cex = 0.8))
```

I precedenti comandi forniscono il grafico della Figura 8.4.2.

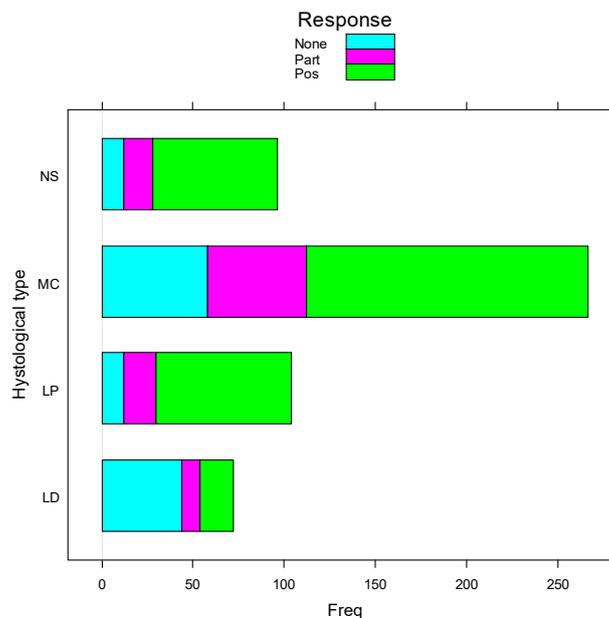


Figura 8.4.2.

Il comando `chisq.test` fornisce l'implementazione del test χ^2 per l'indipendenza:

```
> chisq.test(xtabs(Count ~ Type + Response))
```

```
      Pearson's Chi-squared test
```

```
data:  xtabs(Count ~ Type + Response)
X-squared = 75.8901, df = 6, p-value = 2.517e-14
```

Sulla base della significatività osservata si deve dunque rifiutare l'ipotesi di indipendenza delle tipologie istologiche dagli esiti del trattamento. \square

Considerando un test di permutazione, condizionatamente alla realizzazione del campione, sotto ipotesi di base (ovvero di indipendenza delle variabili) ogni partizione delle osservazioni relative alla seconda variabile in gruppi di numerosità n_{1+}, \dots, n_{r+} è ugualmente probabile. Dunque, si può

costruire un test di permutazione basato sulle $\binom{n}{n_{1+} \dots n_{r+}}$ (ugualmente probabili) permutazioni di livelli della prima variabile rispetto alla seconda variabile. Il test esatto di Fisher è basato sulla statistica test χ^2 di permutazione che si ottiene calcolando la statistica χ^2 sulle $\binom{n}{n_{1+} \dots n_{r+}}$ tabelle a doppia entrata, ognuna relativa ad una delle possibili permutazioni di gruppi. La distribuzione della statistica test sotto ipotesi di base può essere anche in questo caso approssimata mediante un metodo Monte Carlo. Determinazioni elevate della statistica test portano a respingere l'ipotesi di base.

• **Esempio 8.4.5.** In un famoso studio sono state considerate coppie di gemelli ognuna delle quali è stata classificata per tipologia (ovvero se la coppia è costituita da gemelli omozigoti o eterozigoti) e per propensione alla criminalità (ovvero se entrambi i gemelli sono stati detenuti in prigione) (Fonte: Fisher, R. A., 1970, *Statistical Methods for Research Workers*, quattordicesima edizione, Oliver & Boyd, London). I dati sono contenuti nel file `twins.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\twins.txt", header = T)
> attach(d)
```

La tabella a doppia entrata viene ottenuta mediante il comando:

```
> xtabs(Count ~ Type + Conviction)
      Conviction
Type          No Yes
Dizygotic    15  2
Monozygotic   3 10
```

Il diagramma a nastri condizionato viene ottenuto mediante i comandi:

```
> library(lattice)
> barchart(xtabs(Count ~ Type + Conviction), ylab = "Type",
+          auto.key = list(title = "Conviction", cex = 0.8))
```

I precedenti comandi forniscono il grafico della Figura 8.4.3.

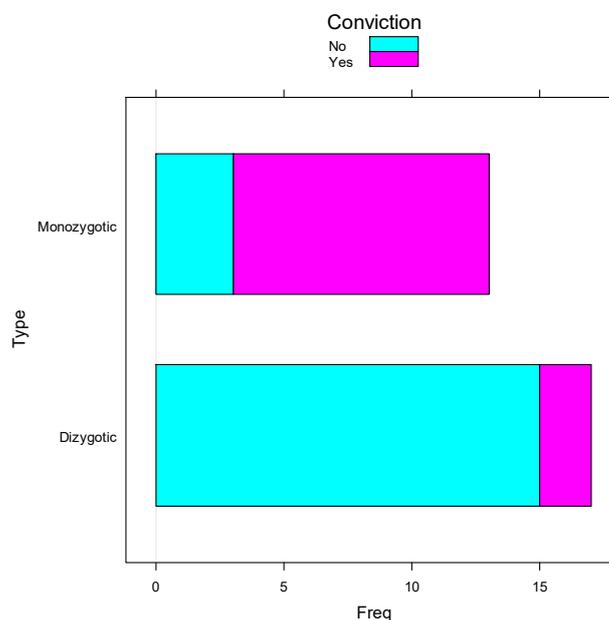


Figura 8.4.3.

Il comando `fisher.test` fornisce l'implementazione del test esatto di Fisher:

```
> fisher.test(xtabs(Count ~ Type + Conviction))

Fisher's Exact Test for Count Data

data:  xtabs(Count ~ Type + Conviction)
p-value = 0.0005367
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 2.753438 300.682787
sample estimates:
odds ratio
 21.30533
```

Sulla base della significatività osservata si deve dunque rifiutare l'ipotesi di indipendenza fra la tipologia dei gemelli e la propensione alla criminalità. \square

• **Esempio 8.4.6.** Per evidenziare la relazione fra il test χ^2 e il test esatto di Fisher, si assuma per semplicità che $r = s = 2$. A priori dal campionamento i dati possono essere equivalentemente organizzati nella tabella a doppia entrata di Tavola 8.4.1.

Tavola 8.4.1

	d_1	d_2	
c_1	N_{11}	N_{12}	N_{1+}
c_2	N_{21}	N_{22}	N_{2+}
	N_{+1}	N_{+2}	N

Se si assume che N_{11} , N_{12} , N_{21} e N_{22} sono variabili casuali indipendenti, rispettivamente con distribuzione di Poisson $Po(\lambda_{11})$, $Po(\lambda_{12})$, $Po(\lambda_{21})$ e $Po(\lambda_{22})$, la funzione di probabilità congiunta del vettore casuale $(N_{11}, N_{12}, N_{21}, N_{22})$ è data da

$$p_{N_{11}, N_{12}, N_{21}, N_{22}}(n_{11}, n_{12}, n_{21}, n_{22}) = \prod_{j,l=1}^2 e^{-\lambda_{jl}} \frac{\lambda_{jl}^{n_{jl}}}{n_{jl}!} \mathbf{1}_{\mathbb{N}}(n_{jl}).$$

Per le proprietà della distribuzione di Poisson, N ha distribuzione di Poisson $Po(\lambda)$ dove $\lambda = \sum_{j,l=1}^2 \lambda_{jl}$. Inoltre, si ha la seguente funzione di probabilità condizionata

$$p_{N_{11}, N_{12}, N_{21}, N_{22}}(n_{11}, n_{12}, n_{21}, n_{22} \mid N = n) = \binom{n}{n_{11} \ n_{12} \ n_{21} \ n_{22}} \prod_{j,l=1}^2 \pi_{jl}^{n_{jl}} \mathbf{1}_S(n_{11}, n_{12}, n_{21}, n_{22}),$$

dove si è posto $\pi_{jl} = \lambda_{jl}/\lambda$ e

$$S = \{(n_{11}, n_{12}, n_{21}, n_{22}) : n_{jl} \in \{0, 1, \dots, n\}, \sum_{j,l=1}^2 n_{jl} = n\}.$$

Dal momento che $\pi_{jl} \geq 0$ e $\sum_{j,l=1}^2 \pi_{jl} = 1$, è evidente che questa distribuzione condizionata è Multinomiale $M_4(n, (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}))$. Inoltre, la distribuzione di N_{1+} condizionata all'evento $\{N = n\}$ è Binomiale $Bi(n, \pi_{1+})$ con $\pi_{1+} = \pi_{11} + \pi_{12}$, mentre la distribuzione di N_{+1} condizionata all'evento $\{N = n\}$ è Binomiale $Bi(n, \pi_{+1})$ con $\pi_{+1} = \pi_{11} + \pi_{21}$. Quindi, si hanno le seguenti funzioni di probabilità condizionate

$$p_{N_{1+}|N=n}(n_{1+}) = \binom{n}{n_{1+}} \pi_{1+}^{n_{1+}} (1 - \pi_{1+})^{n-n_{1+}} \mathbf{1}_{\{0,1,\dots,n\}}(n_{1+}) = \binom{n}{n_{1+}} \pi_{1+}^{n_{1+}} \pi_{2+}^{n-n_{1+}} \mathbf{1}_{\{0,1,\dots,n\}}(n_{1+})$$

e

$$p_{N_{+1}|N=n}(n_{+1}) = \binom{n}{n_{+1}} \pi_{+1}^{n_{+1}} (1 - \pi_{+1})^{n-n_{+1}} \mathbf{1}_{\{0,1,\dots,n\}}(n_{+1}) = \binom{n}{n_{+1}} \pi_{+1}^{n_{+1}} \pi_{+2}^{n-n_{+1}} \mathbf{1}_{\{0,1,\dots,n\}}(n_{+1}).$$

Supponendo l'indipendenza delle distribuzioni marginali, ovvero assumendo che $\pi_{jl} = \pi_j \pi_{+l}$ per $j, l = 1, 2$, queste due distribuzioni condizionate sono indipendenti. In questo caso, si ha la funzione di probabilità condizionata

$$\begin{aligned} p_{N_{11}}(n_{11} | N = n, N_{1+} = n_{1+}, N_{+1} = n_{+1}) &= \frac{p_{N_{11}, N_{12}, N_{21}, N_{22}}(n_{11}, n_{12}, n_{21}, n_{22} | N = n)}{p_{N_{1+}|N=n}(n_{1+}) p_{N_{+1}|N=n}(n_{+1})} \\ &= \frac{\binom{n_{1+}}{n_{11}} \binom{n_{2+}}{n_{21}}}{\binom{n}{n_{+1}}} \mathbf{1}_A(x), \end{aligned}$$

dove $A = \{\max(0, n_{+1} - n + n_{1+}), \dots, \min(n_{+1}, n_{1+})\}$. La distribuzione di N_{11} condizionata all'evento

$$\{N = n, N_{1+} = n_{1+}, N_{+1} = n_{+1}\} = \{N_{1+} = n_{1+}, N_{2+} = n_{2+}, N_{+1} = n_{+1}, N_{+2} = n_{+2}\}$$

è dunque data dalla legge Ipergeometrica $I(n_{+1}, n_{1+}, n)$. Quindi, partendo dallo schema probabilistico di Poisson nella tabella a doppia entrata e condizionando sul totale $N = n$, si ottiene lo schema probabilistico Multinomiale. Il test χ^2 per l'indipendenza viene sviluppato nello schema probabilistico Multinomiale. Partendo invece dallo schema probabilistico Multinomiale nella tabella a doppia entrata e condizionando sui totali $N_{1+} = n_{1+}, N_{2+} = n_{2+}, N_{+1} = n_{+1}, N_{+2} = n_{+2}$, si ottiene lo schema probabilistico Ipergeometrico. Il test esatto di Fisher viene sviluppato nello schema probabilistico Ipergeometrico. Il test possiede questa denominazione in quanto può essere effettivamente basato sulla statistica test N_{11} che, avendo distribuzione Ipergeometrica, ha funzione di probabilità esplicita. Al contrario, la statistica test χ^2 è basata su una trasformazione del vettore casuale $(N_{11}, N_{12}, N_{21}, N_{22})$ con una funzione di probabilità che non può essere espressa in forma semplice, ma con una distribuzione nota per grandi campioni, ovvero la distribuzione Chi-quadrato. Si deve sottolineare che i due approcci derivano da paradigmi inferenziali differenti. \square

8.5. Riferimenti bibliografici

- Agresti, A. (2019) *An Introduction to Categorical Data Analysis*, terza edizione, Wiley, New York.
- Bretz, F., Hothorn, T. e Westfall, P. (2011) *Multiple Comparisons Using R*, Chapman & Hall/CRC Press, Boca Raton.
- Chernick, M.R. (2008) *Bootstrap Methods*, seconda edizione, Wiley, New York.
- Davison, A.C. e Hinkley, D.V. (1997) *Bootstrap Methods and their Application*, Cambridge University Press, Cambridge.
- Dickhaus, T. (2014) *Simultaneous Statistical Inference*, Springer, New York.
- Dikta, G. e Scheer, M. (2021) *Bootstrap Methods*, Springer, Cham.
- Efron, B. e Tibshirani, R. (1993) *An Introduction to the Bootstrap*, Chapman & Hall, London.
- Fisher, R.A. (1970) *Statistical Methods for Research Workers*, quattordicesima edizione, Oliver & Boyd, London.
- Hájek, J. (1969) *Nonparametric Statistics*, Holden Day, San Francisco.
- Hájek, J. e Šidák, Z. (1967) *Theory of Rank Tests*, Academic Press, New York.

-
- Hettmansperger, T.P. e McKean, J.W. (2011) *Robust Nonparametric Statistical Methods*, seconda edizione, Chapman & Hall/CRC Press, Boca Raton.
- Hollander, M., Wolfe, D.A. e Chicken, E. (2014) *Nonparametric Statistical Methods*, terza edizione, Wiley, New York.
- Kloke, J e McKean, J. (2024) *Nonparametric Statistical Methods Using R*, seconda edizione, CRC Press, Boca Raton.
- Kolassa, J.E. (2020) *An Introduction to Nonparametric Statistics*, CRC Press, Boca Raton.
- Kvam, P., Vidakovic, B. e Kim, S. (2023) *Nonparametric Statistics with Applications to Science and Engineering with R*, seconda edizione, Wiley, New York.
- Lehmann, E.L. e Romano, J.P. (2022) *Testing Statistical Hypothesis*, quarta edizione, Springer, Switzerland.
- Shao, J. e Tu, D. (1995) *The Jackknife and Bootstrap*, Springer, New York.

Pagina intenzionalmente vuota

Capitolo 9

La regressione

9.1. La regressione lineare

Nella sua versione più semplice, ovvero quando si dispone di un regressore e di una variabile di risposta, il modello statistico di regressione analizza la struttura di dipendenza fra le due variabili. Il modello di regressione lineare assume ovviamente un legame lineare fra le variabili. In questo caso, si consideri il modello

$$Y_i = \beta_0 + \beta_1 x_i + \mathcal{E}_i,$$

dove $\mathcal{E}_1, \dots, \mathcal{E}_n$ sono variabili casuali indipendenti (detti errori) tali che $E[\mathcal{E}_i] = 0$ e $\text{Var}[\mathcal{E}_i] = \sigma^2$. La formulazione alternativa del modello di regressione lineare è data dalle relazioni

$$E[Y_i] = \beta_0 + \beta_1 x_i$$

e

$$\text{Var}[Y_i] = \sigma^2.$$

In questo modello di regressione esistono quindi tre parametri.

Le stime di β_0 e β_1 ottenute con il metodo dei minimi quadrati (che coincidono con le stime di massima verosimiglianza assumendo la Normalità degli errori) risultano

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

e

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}.$$

Indicando per semplicità stima e stimatore con lo stesso simbolo (un abuso di notazione che è usualmente adottato, anche se può produrre confusione), gli stimatori $\hat{\beta}_0$ e $\hat{\beta}_1$ sono corretti con varianze

$$\text{Var}[\hat{\beta}_0] = \frac{\sigma^2}{n s_x^2} (s_x^2 + \bar{x}^2)$$

e

$$\text{Var}[\hat{\beta}_1] = \frac{\sigma^2}{n s_x^2}.$$

Questi risultati saranno discussi in generale nel caso della regressione lineare multipla presentata nel prossimo Capitolo. I valori stimati risultano inoltre

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

mentre le quantità $(y_i - \hat{y}_i)$ sono dette residui. Il parametro σ^2 può essere stimato in modo corretto mediante la varianza corretta dei residui, ovvero

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

Le stime di $\text{Var}[\hat{\beta}_0]$ e $\text{Var}[\hat{\beta}_1]$ possono essere infine ottenute sostituendo $\hat{\sigma}^2$ al posto di σ^2 .

La variabilità totale delle osservazioni relative alla variabile di risposta può essere scomposta come

$$\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2,$$

dove la prima componente rappresenta la variabilità degli errori e la seconda componente la variabilità dovuta al modello lineare. Di conseguenza, la quantità

$$r_{xy}^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

rappresenta la percentuale di variabilità spiegata dalla relazione lineare rispetto alla variabilità totale ed è il cosiddetto coefficiente di determinazione lineare. Ovviamente, risulta $r_{xy}^2 \in [0, 1]$ e il valore uno indica la presenza di linearità perfetta. Si dimostra facilmente che $r_{xy}^2 = s_{xy}^2 / (s_x^2 s_y^2)$, ovvero il coefficiente di determinazione lineare è il quadrato del coefficiente di correlazione lineare. Si noti che può esistere una relazione (non lineare) perfetta fra variabile di risposta e regressore per cui $r_{xy}^2 = 0$.

• **Esempio 9.1.1.** Si dispone delle osservazioni di distanze lineari e stradali fra località a Sheffield (in km) (Fonte: Gilchrist, W., 1984, *Statistical Modelling*, Wiley, New York, p.5). La variabile di risposta è la distanza stradale, mentre il regressore è la distanza lineare. I dati sono contenuti nel file `roaddistance.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\roaddistance.txt",
+   header = T)
> attach(d)
```

Le stime dei parametri del modello di regressione lineare vengono ottenute mediante il seguente comando:

```
> lm(Road ~ Linear)
```

Call:

```
lm(formula = Road ~ Linear)
```

Coefficients:

```
(Intercept)      Linear
    0.3791         1.2694
```

Il diagramma di dispersione con retta di regressione stimata viene ottenuto mediante i seguenti comandi:

```
> plot(Linear, Road, xlab = "Linear distance (km)",
+   ylab = "Road distance (km)", main = "Scatter plot")
> abline(lm(Road ~ Linear))
```

I precedenti comandi forniscono il grafico della Figura 9.1.1.

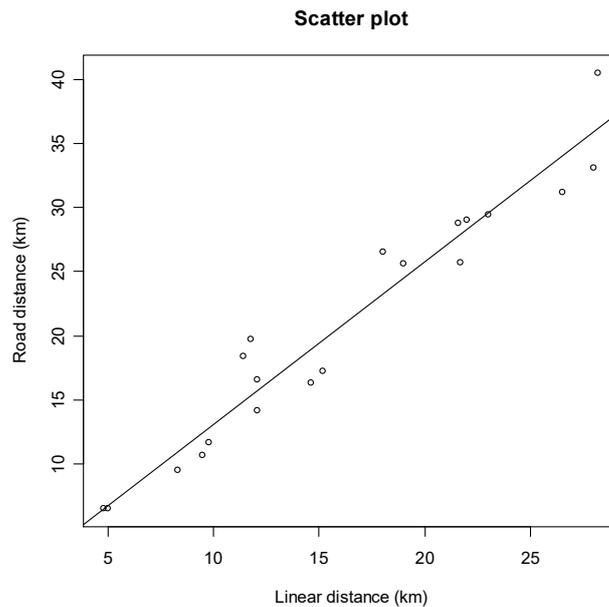


Figura 9.1.1.

Il diagramma di dispersione evidenzia l'adeguatezza del modello lineare. □

Oltre alla costruzione dell'indice di determinazione, i residui permettono di considerare le diagnostiche grafiche per esaminare la validità della relazione lineare. In particolare, il diagramma di Anscombe, che fornisce il diagramma cartesiano dei residui rispetto ai valori stimati, dovrebbe presentare una disposizione casuale dei punti se effettivamente il modello lineare è adeguato. Questo grafico viene anche riportato con le radici dei valori assoluti dei residui standardizzati. Il diagramma quantile-quantile, che fornisce il diagramma cartesiano dei residui standardizzati e ordinati rispetto ai quantili della distribuzione normale standardizzata, dovrebbe avere una disposizione dei punti lungo la bisettrice se l'ipotesi di normalità per gli errori è valida. Il diagramma con le distanze di Cook consente di verificare l'impatto della rimozione di ogni singola osservazione sulle stime dei parametri e quindi l'influenza delle singole osservazioni.

• **Esempio 9.1.2.** Si considerano di nuovo i dati relativi alle distanze stradali dell'Esempio 9.1.1. I diagrammi di Anscombe, il diagramma quantile-quantile e il diagramma con le distanze di Cook vengono ottenuti mediante i seguenti comandi:

```
> par(mfrow = c(2, 2))
> plot(lm(Road ~ Linear), which = c(1:4), add.smooth = F)
> par(mfrow = c(1, 1))
```

I precedenti comandi forniscono il grafico della Figura 9.1.2. I punti che hanno maggiore influenza possono essere evidenziati sul diagramma di dispersione mediante i seguenti comandi:

```
> plot(Linear, Road, xlab = "Linear distance (km)",
+      ylab = "Road distance (km)", main = "Scatter plot")
> abline(lm(Road ~ Linear))
> text(x = Linear[9] + 0.3, y = Road[9], labels = "9", adj=0)
> text(x = Linear[11] + 0.3, y = Road[11], labels = "11", adj=0)
> text(x = Linear[17] - 0.5, y = Road[17] - 1, labels = "17", adj=0)
```

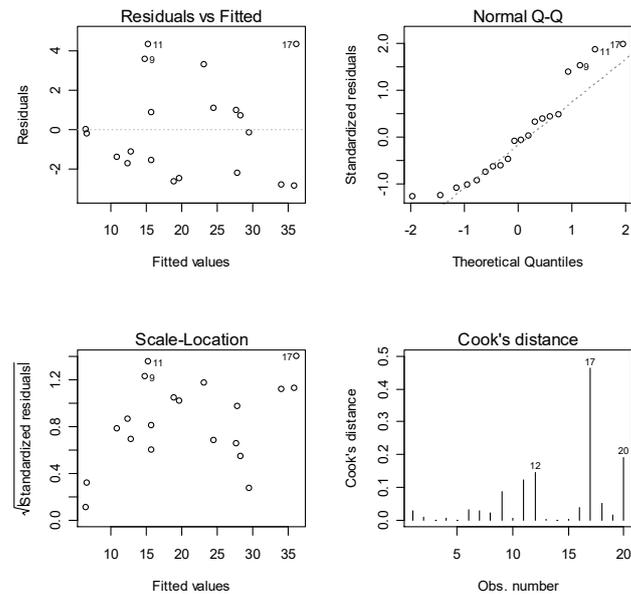


Figura 9.1.2.

I precedenti comandi forniscono il grafico della Figura 9.1.3.

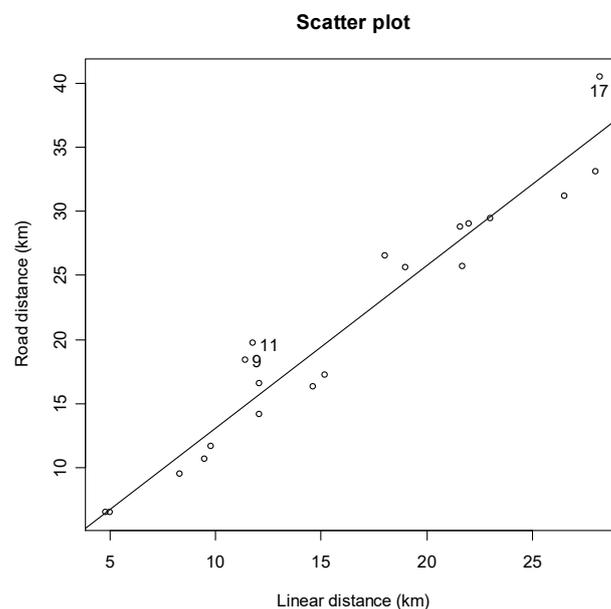
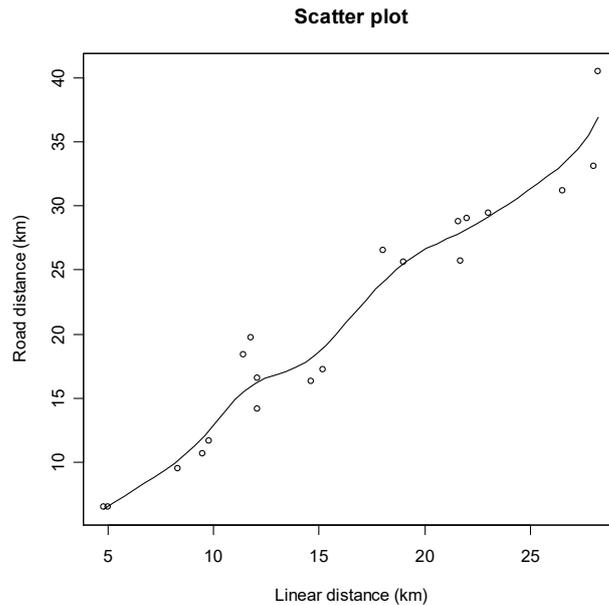


Figura 9.1.3.

Infine, il grafico della stima della funzione di regressione basata sulla regressione lineare locale è ottenuto mediante i seguenti comandi:

```
> library(sm)
> plot(Linear, Road, xlab = "Linear distance (km)",
+      ylab = "Road distance (km)", main = "Scatter plot")
> sm.regression(Linear, Road, method = "df", add = T)
```

I precedenti comandi forniscono il grafico della Figura 9.1.4.

**Figura 9.1.4.**

Il precedente grafico conferma l'adeguatezza del modello lineare. □

Per quanto riguarda la verifica delle ipotesi, supponendo la normalità degli errori e di conseguenza la normalità di Y_1, \dots, Y_n , la validità del modello lineare viene verificata considerando il sistema di ipotesi $H_0 : \beta_1 = 0$ contro $H_1 : \beta_1 \neq 0$. Questo sistema di ipotesi viene verificato attraverso la statistica test fornita dal rapporto delle verosimiglianze

$$F = (n - 2) \frac{R_{xy}^2}{1 - R_{xy}^2},$$

dove R_{xy} è il rapporto di correlazione campionario. La statistica test F si distribuisce come una variabile casuale di Snedecor $F_{1,n-2}$. Una simile verifica di ipotesi può essere condotta anche su β_0 .

• **Esempio 9.1.3.** Si considerano di nuovo i dati relativi alle distanze stradali dell'Esempio 9.1.1. L'analisi relativa alla verifica delle ipotesi viene implementata mediante il seguente comando:

```
> summary(lm(Road ~ Linear))
Call:
lm(formula = Road ~ Linear)

Residuals:
    Min       1Q   Median       3Q      Max
-2.8231 -1.8604 -0.2011  1.0263  4.3416

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.37908    1.34401   0.282   0.781
Linear       1.26943    0.07617  16.665 2.19e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.436 on 18 degrees of freedom
Multiple R-squared:  0.9391,    Adjusted R-squared:  0.9358
F-statistic: 277.7 on 1 and 18 DF,  p-value: 2.187e-12
```

L'analisi relativa alla verifica delle ipotesi con il modello senza intercetta viene implementata mediante il seguente comando:

```
> summary(lm(Road ~ -1 + Linear))

Call:
lm(formula = Road ~ -1 + Linear)

Residuals:
    Min       1Q   Median       3Q      Max
-2.994 -1.728 -0.097  1.029  4.489

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
Linear  1.28907      0.03012    42.8   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.376 on 19 degrees of freedom
Multiple R-Squared:  0.9897,    Adjusted R-squared:  0.9892
F-statistic: 1832 on 1 and 19 DF,  p-value: < 2.2e-16
```

Sulla base della significatività osservata si deve dunque rifiutare l'ipotesi di base, ovvero si accetta l'esistenza di un forte legame lineare fra le variabili. Inoltre, il modello lineare senza intercetta si adegua quasi perfettamente ai dati. \square

Anche se la relazione fra variabile di risposta e regressore non è lineare, ci si può spesso ricondurre alla linearità mediante trasformazioni monotone opportune. Se G e H sono trasformazioni monotone, il modello di regressione lineare per le osservazioni trasformate risulta

$$G(Y_i) = \beta_0 + \beta_1 H(x_i) + \mathcal{E}_i .$$

In questo caso, la relazione fra la variabile di risposta e il regressore è data quindi dalla funzione $y = G^{-1}(\beta_0 + \beta_1 H(x))$.

• **Esempio 9.1.4.** Si dispone delle osservazioni della percentuale di ricordi nel tempo (in minuti) relativi ad un esperimento psicométrico su un soggetto (Fonte: Mosteller, F., Rourke, R.E.K. e Thomas, G.B. (1970) *Probability with Statistical Applications*, seconda edizione, Addison-Wesley, Reading, p.383). La variabile di risposta è la percentuale di ricordi, mentre il regressore è il tempo. I dati sono contenuti nel file `memory.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\memory.txt", header = T)
> attach(d)
```

Da una analisi del diagramma di dispersione contenuto nella Figura 9.1.5 è evidente che una trasformazione logaritmica sulle osservazioni relative al regressore può essere opportuna. L'analisi relativa al modello di regressione lineare

$$Y_i = \beta_0 + \beta_1 \log x_i + \mathcal{E}_i$$

viene ottenuta mediante il seguente comando:

```
> summary(lm(Memory ~ log(Time)))
```

```

Call:
lm(formula = Memory ~ log(Time))

Residuals:
    Min       1Q   Median       3Q      Max
-0.036077 -0.015330 -0.006415  0.017967  0.037799

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.846415   0.014195   59.63 3.65e-15 ***
log(Time)   -0.079227   0.002416  -32.80 2.53e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02339 on 11 degrees of freedom
Multiple R-Squared:  0.9899,    Adjusted R-squared:  0.989
F-statistic:  1076 on 1 and 11 DF,  p-value: 2.525e-12

```

Il diagramma di dispersione con la relativa funzione di regressione si può ottenere mediante i seguenti comandi:

```

> plot(Time, Memory, xlab = "Time (minutes)",
+       ylab = "Memory retention", main = "Scatter plot")
> lines(seq(1, 11000, 1), predict(lm(Memory ~ log(Time))),
+       data.frame(Time = seq(1, 11000, 1)))

```

I precedenti comandi forniscono il grafico della Figura 9.1.5.

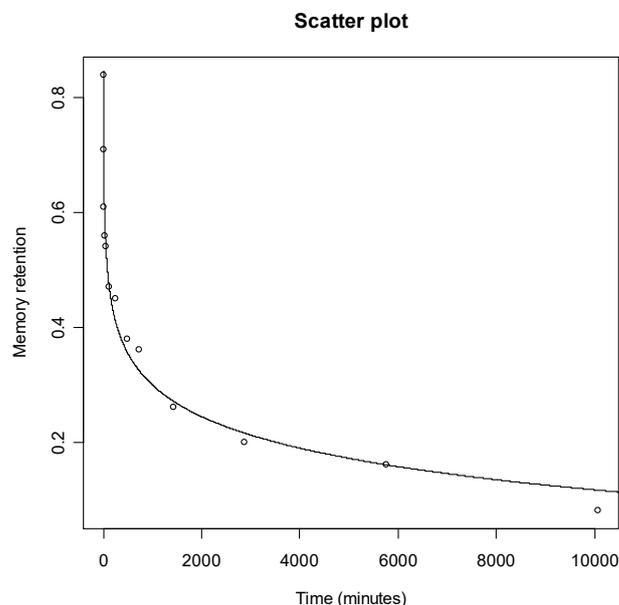


Figura 9.1.5.

Sulla base della significatività osservata si deve dunque rifiutare l'ipotesi di base, ovvero si accetta l'esistenza di un forte legame logaritmico fra le variabili. \square

La struttura inferenziale basata su due campioni casuali indipendenti (di numerosità n_1 e n_2 con $n = n_1 + n_2$) da una variabile casuale Y a due livelli differenti c_1 e c_2 di un fattore, tali che $E[Y_{c_1}] = \mu_1$, $E[Y_{c_2}] = \mu_2$ e $\text{Var}[Y_{c_1}] = \text{Var}[Y_{c_2}] = \sigma^2$, può essere riportata ad un modello di

regressione lineare. In effetti, se Y_1, \dots, Y_n rappresentano le osservazioni relative al campione misto, allora si può scrivere il modello lineare con

$$E[Y_i] = \beta_0 + \beta_1 x_i$$

e

$$\text{Var}[Y_i] = \sigma^2,$$

dove $\beta_0 = \mu_1$, $\beta_1 = \mu_2 - \mu_1$, mentre x_1, \dots, x_n sono i valori assunti da un regressore binario che vale uno se l'osservazione è relativa al secondo livello del fattore e zero altrimenti. L'ipotesi di omogeneità delle medie, supponendo la normalità di Y_1, \dots, Y_n , viene verificata considerando il sistema di ipotesi $H_0 : \beta_1 = 0$ contro $H_1 : \beta_1 \neq 0$. Di conseguenza, le tecniche viste in precedenza possono essere applicate anche in questo caso.

• **Esempio 9.1.5.** Si considera di nuovo i dati relativi alle sfere di acciaio dell'Esempio 8.1.1. Il sistema di ipotesi $H_0 : \mu_1 = \mu_2$ contro $H_1 : \mu_1 \neq \mu_2$ può essere verificato mediante il seguente comando:

```
> summary(lm(Diameter ~ Line))
```

```
Call:
```

```
lm(formula = Diameter ~ Line)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-0.6360 -0.2090  0.0150  0.2915  0.5540
```

```
Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.1940      0.1156  10.327 5.43e-09 ***
LineL2          0.2120      0.1635   1.297  0.211
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.3656 on 18 degrees of freedom
```

```
Multiple R-Squared: 0.08541,    Adjusted R-squared: 0.0346
```

```
F-statistic: 1.681 on 1 and 18 DF,  p-value: 0.2112
```

L'omogeneità delle medie viene ovviamente accettata con lo stesso livello di significatività osservata dell'Esempio 8.1.1. Tuttavia, in questo caso si ha una maggiore informazione sulle due medie. In effetti, la stima del coefficiente β_1 è positiva, anche se non significativa, e dunque la media nel secondo gruppo tende ad essere più elevata rispetto a quella nel primo gruppo. \square

9.2. I modelli lineari generalizzati

Una generalizzazione del modello di regressione lineare può essere ottenuta assumendo una distribuzione non Normale per Y_1, \dots, Y_n e considerando una opportuna funzione legame g tale che

$$g(E[Y_i]) = \beta_0 + \beta_1 x_i.$$

La classe dei modelli lineari generalizzati è basata appunto su questa relazione. Il modello di regressione lineare è un caso particolare di questa classe quando le variabili di risposta hanno una

distribuzione Normale e g è la funzione identità. Il modello lineare generalizzato consente di gestire in modo più appropriato variabili di risposta che sono asimmetriche, discrete o binarie.

Per ogni distribuzione che si assume per la variabile di risposta esiste una funzione legame canonica, ovvero una parametrizzazione naturale del modello. In generale, non è possibile ottenere in forma chiusa le stime di massima verosimiglianza $\hat{\beta}_0$ e $\hat{\beta}_1$ di β_0 e β_1 e delle relative varianze. Quindi si deve ricorrere a procedure numeriche per ottenere queste stime. L'analisi per la validità del modello viene verificata di nuovo mediante il sistema di ipotesi $H_0 : \beta_1 = 0$ contro $H_1 : \beta_1 \neq 0$. Questo sistema di ipotesi è basato sulla devianza, che è una statistica test basata sul metodo del rapporto delle verosimiglianze. La statistica test si distribuisce per grandi campioni come una variabile casuale con distribuzione Chi-quadrato con opportuni gradi di libertà. La devianza si può scomporre in devianza sotto ipotesi di base e devianza residua. Un valore elevato della devianza residua rispetto ai rispettivi gradi di libertà è indice di super dispersione, ovvero della presenza di una variabilità più accentuata delle stime rispetto a quella prevista dal modello lineare generalizzato. Questo fenomeno può essere gestito mediante l'uso di tecniche basate sulla quasi-verosimiglianza per la stima della variabilità degli stimatori di β_0 e β_1 .

Al fine di spiegare la scelta della funzione legame, si consideri una variabile casuale Y che appartiene alla famiglia Esponenziale, ovvero con funzione di densità o probabilità data da

$$f_Y(y; \theta, \phi) = e^{\frac{1}{a(\phi)}(\theta y - b(\theta)) + c(y, \phi)} \mathbf{1}_S(y),$$

dove θ e ϕ sono parametri, mentre a , b e c sono funzioni specifiche e $S \subseteq \mathbb{R}$ è un insieme opportuno che non dipende dai parametri. Inoltre, si assume che b sia una funzione differenziabile al secondo ordine. Se ϕ è un parametro noto, allora la famiglia è detta Esponenziale con parametro canonico θ . Per un dato valore y la log-verosimiglianza risulta

$$l(\theta, \phi; y) = \log f_Y(y; \theta, \phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi).$$

Dunque, si ha

$$\frac{dl}{d\theta} = \frac{y - b'(\theta)}{a(\phi)}, \quad \frac{d^2l}{d\theta^2} = -\frac{b''(\theta)}{a(\phi)},$$

da cui per le proprietà della funzione punteggio risulta

$$E\left[\frac{Y - b'(\theta)}{a(\phi)}\right] = \frac{E[Y] - b'(\theta)}{a(\phi)} = 0$$

e

$$\text{Var}\left[\frac{Y - b'(\theta)}{a(\phi)}\right] = \frac{\text{Var}[Y]}{a(\phi)^2} = \frac{b''(\theta)}{a(\phi)},$$

ovvero

$$E[Y] = b'(\theta)$$

e

$$\text{Var}[Y] = a(\phi)b''(\theta).$$

Si assuma che $\mu = \mu(\theta) = E[Y]$ e $V(\mu) = \text{Var}[Y]$. La funzione inversa $\theta = \mu^{-1}$ fornisce la funzione legame canonica, mentre la funzione $V(\mu)$ è detta funzione di varianza.

• **Esempio 9.2.1.** La distribuzione Normale $N(\mu, \sigma^2)$ fa parte della famiglia Esponenziale, in quanto si può scrivere

$$f_Y(y; \theta, \phi) = e^{\frac{1}{\sigma^2}(y\mu - \mu^2/2) - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)},$$

per cui $\theta = \mu$ e $\phi = \sigma^2$, mentre

$$a(\phi) = \phi,$$

$$b(\theta) = \frac{\mu^2}{2} = \frac{\theta^2}{2}$$

e

$$c(y, \phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2).$$

Dunque, si ha

$$E[Y] = b'(\theta) = \theta.$$

In questo caso, essendo $E[Y] = \mu$, si ha $\mu = \theta$ e $V(\mu) = \sigma^2$, ovvero la funzione legame canonica è la funzione identità, mentre la funzione di varianza è costante. \square

• **Esempio 9.2.2.** La distribuzione di Poisson $P(\lambda)$ fa parte della famiglia Esponenziale, in quanto si può scrivere

$$f_Y(y; \theta, \phi) = e^{y\log(\lambda) - \lambda - \log(y!)} \mathbf{1}_{\mathbb{N}}(y),$$

per cui $\theta = \log(\lambda)$, ovvero $\lambda = e^\theta$, e $\phi = 1$, mentre

$$a(\phi) = 1,$$

$$b(\theta) = \lambda = e^\theta$$

e

$$c(y, \phi) = -\log(y!).$$

Dunque, si ha

$$E[Y] = b'(\theta) = e^\theta.$$

In questo caso, essendo $\mu = E[Y] = \lambda$, si ha $\theta = \log(\lambda)$ e $V(\mu) = \mu$, ovvero la funzione legame canonica è la funzione logaritmo, mentre la funzione di varianza è la funzione identità. \square

• **Esempio 9.2.3.** La distribuzione Binomiale $B(1, p)$ fa parte della famiglia Esponenziale, in quanto si può scrivere

$$f_Y(y; \theta, \phi) = e^{y\log\left(\frac{p}{1-p}\right) + \log(1-p)} \mathbf{1}_{\{0,1\}}(y),$$

per cui $\theta = \log\left(\frac{p}{1-p}\right)$, ovvero $p = \frac{e^\theta}{1+e^\theta}$, e $\phi = 1$, mentre

$$a(\phi) = 1,$$

$$b(\theta) = -\log(1-p) = \log(1+e^\theta)$$

e

$$c(y, \phi) = 0.$$

Dunque, si ha

$$E[Y] = b'(\theta) = \frac{e^\theta}{1 + e^\theta}.$$

In questo caso, essendo $\mu = E[Y] = p$, si ha $\theta = \log\left(\frac{\mu}{1-\mu}\right)$ e $V(\mu) = \mu(1-\mu)$, ovvero la funzione legame canonica è la funzione logit. \square

Quando le variabili di risposta prendono valori sugli interi positivi è conveniente assumere che Y_1, \dots, Y_n siano variabili casuali di Poisson. Questo caso particolare del modello lineare generalizzato costituisce la regressione di Poisson. Tenendo presente l'Esempio 9.2.2, la funzione legame canonica è la funzione logaritmo, ovvero si ha

$$\log(E[Y_i]) = \beta_0 + \beta_1 x_i.$$

• **Esempio 9.2.4.** Si dispone delle osservazioni relative alla lunghezza di pezzi di stoffa (in metri) e del relativo numero di difetti (Fonte: Bissell, A.F., 1972, A negative binomial model with varying element sizes, *Biometrika* **59**, 435-441). La variabile di risposta è il numero di difetti riscontrati, mentre il regressore è la lunghezza. I dati sono contenuti nel file `clothes.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\clothes.txt", header = T)
> attach(d)
```

L'analisi relativa al modello di regressione di Poisson viene ottenuta mediante il seguente comando:

```
> summary(glm(Defects ~ Length, poisson))
```

Call:

```
glm(formula = Defects ~ Length, family = poisson)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.74127	-1.13312	-0.03904	0.66179	3.07446

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.9717506	0.2124693	4.574	4.79e-06 ***
Length	0.0019297	0.0003063	6.300	2.97e-10 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 103.714 on 31 degrees of freedom
 Residual deviance: 61.758 on 30 degrees of freedom
 AIC: 189.06

Number of Fisher Scoring iterations: 4

L'analisi della devianza residua indica la presenza di super dispersione e quindi è conveniente impiegare un metodo di quasi-verosimiglianza, ovvero:

```
> summary(glm(Defects ~ Length, quasipoisson))
```

Call:

```
glm(formula = Defects ~ Length, family = quasipoisson)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.74127	-1.13312	-0.03904	0.66179	3.07446

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	0.9717506	0.3095033	3.140	0.003781	**
Length	0.0019297	0.0004462	4.325	0.000155	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 2.121965)

Null deviance: 103.714 on 31 degrees of freedom

Residual deviance: 61.758 on 30 degrees of freedom

AIC: NA

Number of Fisher Scoring iterations: 4

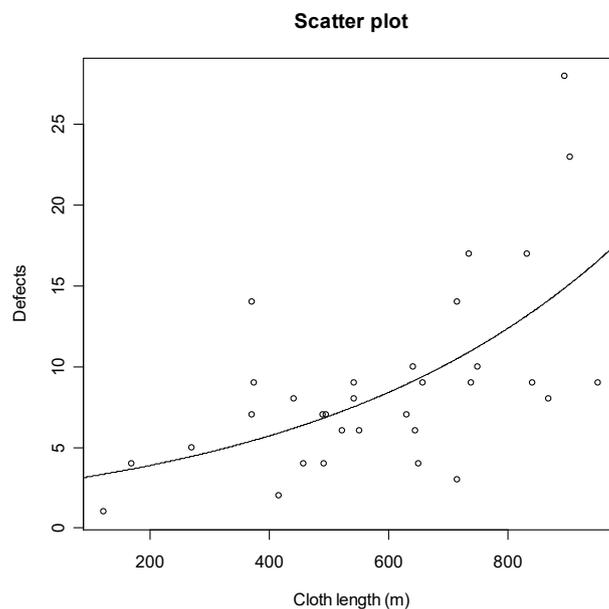


Figura 9.2.1.

La precedente elaborazione indica la validità del modello. Il diagramma di dispersione con la relativa funzione legame stimata viene ottenuto mediante i seguenti comandi:

```
> plot(Length, Defects, xlab = "Cloth length (m)", ylab = "Defects",
+      main = "Scatter plot")
> lines(seq(0, 1000, 1),
+       exp(predict(glm(Defects ~ Length, quasipoisson),
+                     data.frame(Length = seq(0, 1000, 1)))))
```

I precedenti comandi forniscono il grafico della Figura 9.2.1. □

Quando le variabili di risposta sono binarie è conveniente assumere che Y_1, \dots, Y_n siano variabili casuali di Bernoulli. Questo caso particolare del modello lineare generalizzato costituisce la regressione logistica. Dall'Esempio 9.2.3, la funzione legame canonica è la funzione logit, ovvero si ha

$$\log\left(\frac{E[Y_i]}{1 - E[Y_i]}\right) = \beta_0 + \beta_1 x_i.$$

• **Esempio 9.2.5.** Si dispone delle osservazioni relative alla presenza di danneggiamento dei pannelli di protezione e delle temperature (in gradi Fahrenheit) per alcuni voli di shuttle (Fonte: Dalal, S.R., Fowlkes, E.B. e Hoadley, B. (1989) Risk analysis of the space shuttle: pre-challenger prediction of failure, *Journal of the American Statistical Association* **84**, 945-957). La variabile di risposta è la presenza di danneggiamento, mentre il regressore è la temperatura. I dati sono contenuti nel file `shuttle.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\shuttle.txt", header = T)
> attach(d)
```

L'analisi relativa al modello di regressione logistica viene ottenuta mediante il seguente comando:

```
> summary(glm(Failure ~ Temp, binomial))
```

Call:

```
glm(formula = Failure ~ Temp, family = binomial)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.0611	-0.7613	-0.3783	0.4524	2.2175

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	15.0429	7.3786	2.039	0.0415 *
Temp	-0.2322	0.1082	-2.145	0.0320 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 28.267 on 22 degrees of freedom
 Residual deviance: 20.315 on 21 degrees of freedom
 AIC: 24.315

Number of Fisher Scoring iterations: 5

La precedente elaborazione indica una certa adeguatezza del modello. Il diagramma di dispersione con la relativa funzione legame stimata viene ottenuto mediante i seguenti comandi:

```
> plot(Temp, Failure, xlab = "Temperature (°F)", ylab = "Failure",
+       main = "Scatter plot")
> lines(seq(50, 90, 0.1),
+       predict(glm(Failure ~ Temp, binomial),
+       data.frame(Temp = seq(50, 90, 0.1)), type = "response"))
```

I precedenti comandi forniscono il grafico della Figura 9.2.2. □

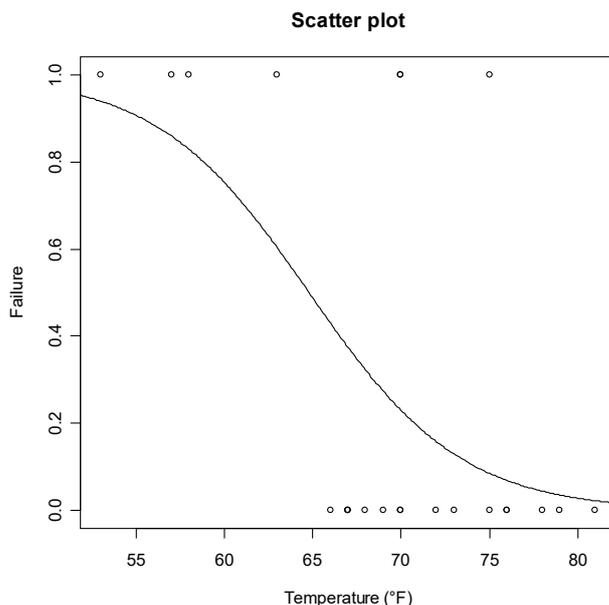


Figura 9.2.2.

La regressione logistica può essere applicata anche quando esistono solo c_1, \dots, c_r livelli distinti del regressore (che potrebbero essere anche i valori centrali di classi opportune) per i due livelli $d_1 = 0$ e $d_2 = 1$ della variabile di risposta e quindi le osservazioni possono essere organizzate in una tabella a doppia entrata con r righe e 2 colonne. In questo caso, la regressione logistica può essere applicata considerando le variabili di risposta $n_{12}/n_{1+}, \dots, n_{r2}/n_{r+}$, ovvero le proporzioni della variabile di risposta per ogni livello del regressore, ottenendo il modello

$$\log\left(\frac{E[n_{j2}/n_{j+}]}{1 - E[n_{j2}/n_{j+}]}\right) = \beta_0 + \beta_1 c_j.$$

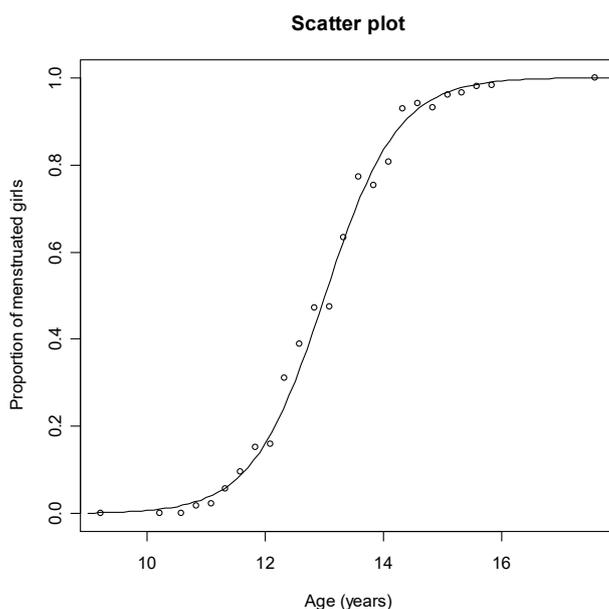


Figura 9.2.3.

• **Esempio 9.2.6.** Si dispone delle osservazioni relative alla percentuale di un gruppo di adolescenti polacche con menarca per vari livelli d'età (Fonte: Morgan, B.J.T. (1989) *Analysis of Quantal Response Data*, Chapman and Hall, London, p.7). La variabile di risposta è la percentuale di un gruppo di adolescenti polacche che hanno avuto il menarca, mentre il regressore è l'età. I dati sono contenuti nel file `menarche.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\menarche.txt", header = T)
> attach(d)
```

L'analisi relativa al modello di regressione logistica viene ottenuta mediante il seguente comando:

```
> Proportion <- cbind(Menarche, Total - Menarche)
> summary(glm(Proportion ~ Age, binomial))
```

Call:

```
glm(formula = Proportion ~ Age, family = binomial)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.0363	-0.9953	-0.4900	0.7780	1.3675

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-21.22639	0.77068	-27.54	<2e-16 ***
Age	1.63197	0.05895	27.68	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	3693.884	on 24	degrees of freedom
Residual deviance:	26.703	on 23	degrees of freedom
AIC:	114.76		

Number of Fisher Scoring iterations: 4

La precedente elaborazione indica la validità del modello. Il diagramma di dispersione con la relativa funzione legame stimata viene ottenuto mediante i seguenti comandi:

```
> plot(Age, Menarche / Total, xlab = "Age (years)",
+      ylab = "Proportion of menstruated girls",
+      main = "Scatter plot")
> lines(seq(9, 18, 0.1),
+       predict(glm(Proportion ~ Age, binomial),
+       data.frame(Age = seq(9, 18, 0.1)), type = "response"))
```

I precedenti comandi forniscono il grafico della Figura 9.2.3. □

9.3. Riferimenti bibliografici

Agresti, A. (2013) *Categorical Data Analysis*, terza edizione, Wiley, New York.

Agresti, A. (2015) *Foundations of Linear and Generalized Linear Models*, Wiley, New York.

Agresti, A. (2019) *An Introduction to Categorical Data Analysis*, terza edizione, Wiley, New York.

- Atkinson, A. e Riani, M. (2000) *Robust Diagnostic Regression Analysis*, Springer-Verlag, New York.
- Belsley, D.A., Kuh, E. e Welsch, R.E. (1980) *Regression Diagnostics*, Wiley, New York.
- Chatterjee, S. e Simonoff, J.S. (2013) *Handbook of Regression Analysis*, Wiley, New York.
- Cook, R.D. e Weisberg, S. (1999) *Applied Regression Including Computing and Graphics*, Wiley, New York.
- Christensen, R. (2025) *Log-Linear Models and Logistic Regression*, terza edizione, Springer, Cham.
- Dunn, P.K. e Smyth, G.K. (2018) *Generalized Linear Models with Examples in R*, Springer, New York.
- Gelman, A., Hill, J. e Vehtari, A. (2021) *Regression and Other Stories*, Cambridge University Press, Cambridge.
- Hadi, A.S. e Chatterjee, S. (2024) *Regression Analysis by Example Using R*, sesta edizione, , Wiley, New York.
- James, G., Witten, D., Hastie T. e Tibshirani, R. (2013) *An Introduction to Statistical Learning*, Springer, New York.
- McCullagh, P. e Nelder, J.A. (1989) *Generalized Linear Models*, seconda edizione, Chapman and Hall, London.
- Seber, G.A.F. (2015) *The Linear Model and Hypothesis*, Springer, New York.
- Sheather, S.J. (2009) *A Modern Approach to Regression with R*, Springer, New York.
- Simonoff, J.S. (2003) *Analyzing Categorical Data*, Springer, New York.
- Stasinopoulos, M.D., Kneib, T., Klein, N., Mayr, A. e Heller, G.Z. (2024) *Generalized Additive Models for Location, Scale and Shape*, Cambridge University Press, Cambridge.
- Xu, J. (2023) *Modern Applied Regressions*, CRC Press, Boca Raton.
- Weisberg, S. (2014) *Applied Linear Regression*, Wiley, New York.

Capitolo 10

La regressione multipla

10.1. La regressione lineare multipla

Il modello di regressione multipla analizza la struttura di dipendenza fra un insieme di regressori e una variabile di risposta. Nella versione più semplice, il modello di regressione multipla assume un legame lineare fra le variabili e questo modello include come caso particolare anche l'analisi della varianza. Quando la variabile di risposta dipende da p regressori si consideri dunque il modello

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \mathcal{E}_i,$$

dove $\mathcal{E}_1, \dots, \mathcal{E}_n$ sono variabili casuali indipendenti tali che $E[\mathcal{E}_i] = 0$ e $\text{Var}[\mathcal{E}_i] = \sigma^2$. La formulazione alternativa del modello di regressione lineare multipla è data quindi dalle relazioni

$$E[Y_i] = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

e

$$\text{Var}[Y_i] = \sigma^2.$$

In questo modello esistono dunque $(p + 2)$ parametri.

Sia $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^T$ il vettore dei parametri di ordine $(p + 1)$ e sia $\mathbf{X} = (x_{ij})$ la matrice di ordine $(n \times (p + 1))$ delle osservazioni relative ai regressori, in cui la prima colonna è composta da unità (ovvero, si è aggiunto in pratica un ulteriore regressore che assume valori pari all'unità). Inoltre, sia $\mathbf{y} = (y_1, \dots, y_n)^T$ il vettore delle determinazioni del vettore casuale $\mathbf{Y} = (Y_1, \dots, Y_n)^T$. Infine, sia $\boldsymbol{\mathcal{E}} = (\mathcal{E}_1, \dots, \mathcal{E}_n)^T$ con vettore medio $E[\boldsymbol{\mathcal{E}}] = \mathbf{0}$ e matrice di varianza-covarianza $\text{Var}[\boldsymbol{\mathcal{E}}] = \sigma^2 \mathbf{I}_n$. In questo caso, il modello di regressione lineare multipla può essere espresso in notazione matriciale come

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mathcal{E}},$$

o alternativamente mediante le due relazioni

$$E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$$

e

$$\text{Var}[\mathbf{Y}] = \sigma^2 \mathbf{I}_n.$$

Al fine di stimare il vettore dei parametri $\boldsymbol{\beta}$, il metodo dei minimi quadrati consiste nel minimizzare la somma degli scarti al quadrato dei valori osservati dai valori teorici della variabile di risposta, ovvero nel minimizzare la forma quadratica

$$\varphi(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}^T \mathbf{y} - 2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y} + \boldsymbol{\beta}^T \mathbf{V} \boldsymbol{\beta},$$

dove la matrice $\mathbf{V} = \mathbf{X}^T \mathbf{X}$ è evidentemente definita positiva. A questo fine, si ha

$$\frac{\partial \varphi}{\partial \boldsymbol{\beta}^T} = -2\mathbf{X}^T \mathbf{y} + 2\mathbf{V}\boldsymbol{\beta}.$$

Uguagliando questa quantità a $\mathbf{0}$, si ottiene l'equazione vettoriale

$$\mathbf{V}\boldsymbol{\beta} = \mathbf{X}^T \mathbf{y}.$$

Se si suppone che il rango di \mathbf{X} sia $p + 1 < n$, ovvero che non vi siano relazioni lineari fra le colonne di \mathbf{X} e che la numerosità campionaria sia maggiore del numero dei regressori (per evitare relazioni esatte fra i regressori), si può invertire la matrice \mathbf{V} , ottenendo in questo modo la stima di $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = \mathbf{V}^{-1} \mathbf{X}^T \mathbf{y}.$$

La matrice hessiana

$$\frac{\partial^2 \varphi}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} = 2\mathbf{V}$$

è definita positiva, per cui φ è una funzione convessa e quindi $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)^T$ è effettivamente un punto di minimo. Con un abuso in notazione comune in letteratura, lo stimatore $\hat{\boldsymbol{\beta}} = \mathbf{V}^{-1} \mathbf{X}^T \mathbf{Y}$ viene indicato con lo stesso simbolo della stima.

Si può verificare che lo stimatore $\hat{\boldsymbol{\beta}}$ è corretto e la relativa matrice di varianza-covarianza è data da $\text{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2 \mathbf{V}^{-1}$. Se il vettore dei valori stimati è denotato con $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, la somma dei quadrati dei residui è data dalla forma quadratica

$$S = (\mathbf{Y} - \hat{\mathbf{Y}})^T (\mathbf{Y} - \hat{\mathbf{Y}}).$$

Dal momento che si ha $E[S] = (n - p - 1)\sigma^2$, uno stimatore corretto di σ^2 è dato da

$$\hat{\sigma}^2 = \frac{S}{n - p - 1}.$$

Indicando con s la realizzazione di S , la stima di $\text{Var}[\hat{\boldsymbol{\beta}}]$ è dunque data da

$$\widehat{\text{Var}}[\hat{\boldsymbol{\beta}}] = \frac{s}{n - p - 1} \mathbf{V}^{-1}.$$

Anche per il modello di regressione multipla il coefficiente di determinazione viene definito come

$$r_{xy}^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

dove $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip}$. L'interpretazione di r_{xy}^2 è analoga a quella dello stesso indice considerato nel modello lineare con un singolo regressore. Infine, se si suppone che $\boldsymbol{\mathcal{E}}$ abbia distribuzione Normale $N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, essendo \mathbf{Y} e $\hat{\boldsymbol{\beta}}$ trasformate lineari di $\boldsymbol{\mathcal{E}}$, allora \mathbf{Y} ha distribuzione Normale $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$, mentre $\hat{\boldsymbol{\beta}}$ ha distribuzione Normale $N_{p+1}(\boldsymbol{\beta}, \sigma^2 \mathbf{V}^{-1})$. Inoltre, si può verificare che S/σ^2 ha distribuzione Chi-quadrato χ_{n-p-1}^2 .

In alternativa al metodo dei minimi quadrati, sotto le assunzioni fatte per la distribuzione del vettore casuale \mathbf{Y} , si può considerare il metodo della massima verosimiglianza. Posto per semplicità di notazione $v = \sigma^2$, si ottiene la seguente funzione di verosimiglianza

$$L(\boldsymbol{\beta}, v) = c e^{-\frac{1}{2v}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})},$$

da cui si ha la seguente funzione di log-verosimiglianza

$$l(\boldsymbol{\beta}, v) = \log c - \frac{n}{2} \log v - \frac{1}{2v} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Derivando opportunamente la log-verosimiglianza si hanno le equazioni

$$\frac{\partial l}{\partial \boldsymbol{\beta}^\top} = \frac{1}{v} \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$$

e

$$\frac{\partial l}{\partial v} = -\frac{n}{2v} + \frac{1}{2v^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0.$$

Dalle precedenti relazioni si ottengono le stime di massima verosimiglianza date da

$$\widehat{\boldsymbol{\beta}} = \mathbf{V}^{-1} \mathbf{X}^\top \mathbf{y}$$

e

$$\widehat{v} = \frac{1}{n} (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \frac{s}{n}.$$

Dunque, lo stimatore di $\boldsymbol{\beta}$ coincide con quello ottenuto con il metodo dei minimi quadrati, mentre lo stimatore di σ^2 è distorto. Inoltre, la matrice hessiana è data da

$$\begin{pmatrix} \frac{\partial^2 l}{\partial \boldsymbol{\beta}^\top \partial \boldsymbol{\beta}} & \frac{\partial^2 l}{\partial \boldsymbol{\beta}^\top \partial v} \\ \frac{\partial^2 l}{\partial \boldsymbol{\beta}^\top \partial v} & \frac{\partial^2 l}{\partial v^2} \end{pmatrix} = -\frac{1}{v} \begin{pmatrix} \mathbf{V} & \mathbf{0} \\ \mathbf{0} & \frac{n}{2v} \end{pmatrix},$$

che è definita negativa e quindi l è una funzione concava sullo spazio parametrico. Dunque, le stime ottenute sono effettivamente dei punti di massimo. La matrice di varianza-covarianza per grandi campioni degli stimatori di massima verosimiglianza si ottiene mediante la matrice d'informazione di Fisher $\mathbf{I}(\boldsymbol{\beta}, v)$, che coincide con la precedente matrice hessiana. Dunque, la matrice di varianza-covarianza per grandi campioni risulta

$$\text{Var}[\widehat{\boldsymbol{\beta}}, \widehat{v}] = -\mathbf{I}(\boldsymbol{\beta}, v)^{-1} = v \begin{pmatrix} \mathbf{V}^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2v}{n} \end{pmatrix},$$

che può essere stimata sostituendo \widehat{v} al posto di v . Per le proprietà del metodo della massima verosimiglianza gli stimatori $\widehat{\boldsymbol{\beta}}$ e \widehat{v} sono asintoticamente corretti e con distribuzione Normale per grandi campioni. Le distribuzioni per piccoli campioni sono analoghe a quelle evidenziate in precedenza per il metodo dei minimi quadrati.

Per quanto riguarda la verifica delle ipotesi, supponendo la Normalità del vettore casuale \mathbf{Y} , la validità del modello lineare viene basata sul sistema di ipotesi $H_0 : \beta_1 = \dots = \beta_p = 0$ contro $H_1 : \beta_j \neq 0$ per qualche j . Questo sistema di ipotesi viene verificato attraverso la statistica test fornita dal rapporto delle verosimiglianze

$$F = \frac{n-p}{p-1} \frac{R_{xy}^2}{1-R_{xy}^2},$$

che ha distribuzione di Snedecor $F_{p-1, n-p}$. La verifica di ipotesi su uno specifico parametro β_j , ovvero $H_0 : \beta_j = 0$ contro $H_1 : \beta_j \neq 0$, viene effettuata con la statistica test del rapporto delle verosimiglianze

$$T = \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{v_{jj}}},$$

dove v_{jj} è il j -esimo elemento sulla diagonale della matrice \mathbf{V}^{-1} , e che si distribuisce come $T \sim t_{n-p}$.

Quando si è verificato che alcuni parametri del modello sono nulli, si cerca di costruire un modello più semplice (ovvero con un numero minore di regressori) di quello iniziale. Se si eliminano alcuni parametri, il modello di regressione diventa più semplice, ma la sua adeguatezza ai dati diminuisce. In generale, quando si ipotizza un modello con k parametri $\theta_1, \dots, \theta_k$, il criterio di Akaike è dato da

$$\text{AIC} = -2(\log L(\hat{\theta}_1, \dots, \hat{\theta}_k) - k)$$

dove $L(\theta_1, \dots, \theta_k)$ è la funzione di verosimiglianza relativa al modello. Il criterio di Akaike è in effetti il massimo della log-verosimiglianza penalizzato dal numero di parametri presenti nel modello stesso (il tutto moltiplicato da una costante negativa), ovvero un indice finalizzato a valutare il compromesso fra adeguatezza e semplicità del modello. Quando si deve dunque scegliere fra più modelli, il modello che si preferisce è quindi quello che fornisce il valore minimo del criterio di Akaike.

• **Esempio 10.1.1.** Si dispone delle osservazioni del numero di specie di uccelli in “isole” di vegetazione nel nord delle Ande (Fonte: Vuilleumier, F., 1970, Insular biogeography in continental regions. I. The northern Andes of South America, *American Naturalist* **104**, 373-388). I regressori sono l'area dell'isola di vegetazione (in migliaia di km quadrati), la sua elevazione (in km), la sua distanza dall'equatore (in km) e la sua distanza dall'isola di vegetazione più vicina (in km). I dati sono contenuti nel file `paramo.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\paramo.txt", header = T)
> attach(d)
```

La matrice dei diagrammi di dispersione viene ottenuta mediante il seguente comando:

```
> pairs(d, main = "Scatter-plot matrix")
```

Il precedente comando fornisce il grafico della Figura 10.1.1.

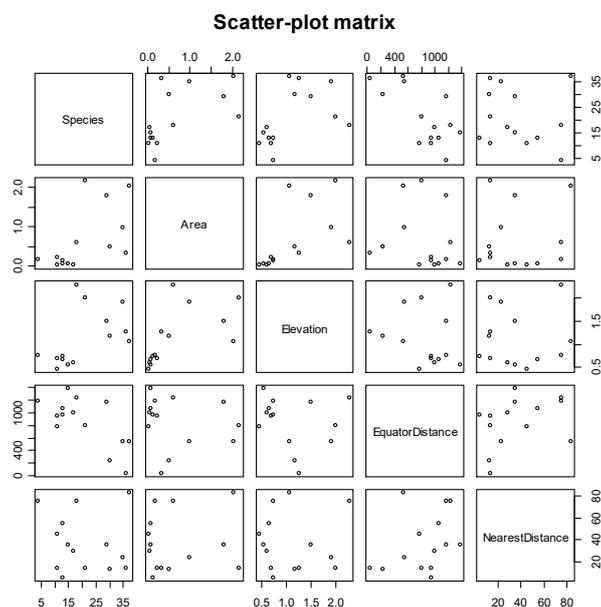


Figura 10.1.1.

L'analisi del modello lineare con tutti i regressori viene implementata mediante il seguente comando:

```
> summary(lm(Species ~ Area + Elevation +
+ EquatorDistance + NearestDistance))
```

Call:

```
lm(formula = Species ~ Area + Elevation +
EquatorDistance + NearestDistance)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-10.66596	-3.40900	0.08345	3.55920	8.23565

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	27.889386	6.181843	4.511	0.00146	**
Area	5.153864	3.098074	1.664	0.13056	
Elevation	3.075136	4.000326	0.769	0.46175	
EquatorDistance	-0.017216	0.005243	-3.284	0.00947	**
NearestDistance	0.016591	0.077573	0.214	0.83541	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.705 on 9 degrees of freedom

Multiple R-Squared: 0.7301, Adjusted R-squared: 0.6101

F-statistic: 6.085 on 4 and 9 DF, p-value: 0.01182

La semplificazione automatica del modello mediante il criterio AIC viene implementata mediante il seguente comando:

```
> summary(step(lm(Species ~ Area + Elevation + EquatorDistance +
+ NearestDistance)))
```

Start: AIC=57.09

Species ~ Area + Elevation + EquatorDistance + NearestDistance

	Df	Sum of Sq	RSS	AIC
- NearestDistance	1	2.06	406.65	55.16
- Elevation	1	26.57	431.15	55.98
<none>			404.59	57.09
- Area	1	124.41	529.00	58.85
- EquatorDistance	1	484.71	889.30	66.12

Step: AIC=55.16

Species ~ Area + Elevation + EquatorDistance

	Df	Sum of Sq	RSS	AIC
- Elevation	1	26.06	432.71	54.03
<none>			406.65	55.16
- Area	1	133.51	540.15	57.14
- EquatorDistance	1	537.39	944.04	64.96

Step: AIC=54.03

Species ~ Area + EquatorDistance

```

              Df Sum of Sq    RSS    AIC
<none>                432.71  54.03
- Area                1   342.64  775.35  60.20
- EquatorDistance    1   557.23  989.94  63.62

```

```

Call:
lm(formula = Species ~ Area + EquatorDistance)

```

```

Residuals:
    Min       1Q   Median       3Q      Max
-10.637  -4.396   0.899   4.084   7.273

```

```

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.797969   4.648155   6.626 3.73e-05 ***
Area          6.683038   2.264403   2.951  0.01318 *
EquatorDistance -0.017057  0.004532  -3.764  0.00313 **
---

```

```

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Residual standard error: 6.272 on 11 degrees of freedom
Multiple R-squared:  0.7113,    Adjusted R-squared:  0.6588
F-statistic: 13.55 on 2 and 11 DF,  p-value: 0.001077

```

I diagrammi di Anscombe, il diagramma quantile-quantile e il diagramma con le distanze di Cook del modello ridotto vengono ottenuti mediante i seguenti comandi:

```

> par(mfrow = c(2, 2))
> plot(lm(Species ~ Area + EquatorDistance),
+      which = c(1:4), add.smooth = FALSE)
> par(mfrow = c(1, 1))

```

I precedenti comandi forniscono il grafico della Figura 10.1.2. □

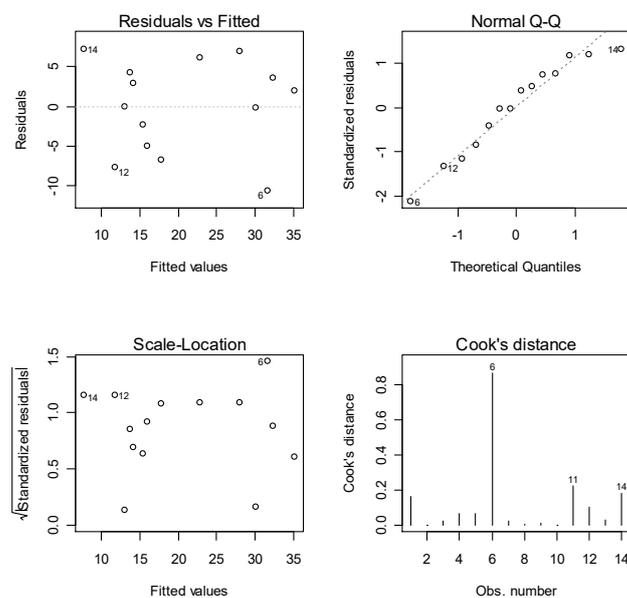


Figura 10.1.2.

Al fine di migliorare l'adeguatezza del modello iniziale di regressione multipla, può essere utile considerare interazioni o ulteriori trasformate dei regressori originali. Ad esempio, un modello iniziale con due regressori

$$E[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

potrebbe essere esteso considerando l'interazione fra le variabili

$$E[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}$$

e introducendo di fatto un nuovo regressore (dato dal prodotto dei due regressori originali). Il modello iniziale potrebbe essere ulteriormente esteso considerando effetti non lineari dei regressori come nel seguente modello che introduce dipendenze quadratiche

$$E[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i2}^2$$

considerando in effetti due nuovi regressori (dati dai quadrati dei due regressori originali). Evidentemente, interazioni e dipendenze non lineari potrebbero essere introdotte congiuntamente

$$E[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \beta_4 x_{i1}^2 + \beta_5 x_{i2}^2$$

considerando tre nuovi regressori. I modelli estesi costruiti in questo modo sono comunque lineari nei parametri e possono essere trattati come visto in precedenza. Tuttavia, si osservi che i nuovi modelli perdono la semplicità del modello originale e possono introdurre difficoltà di interpretazione dei parametri. Modelli più complessi devono dunque essere adottati con cautela.

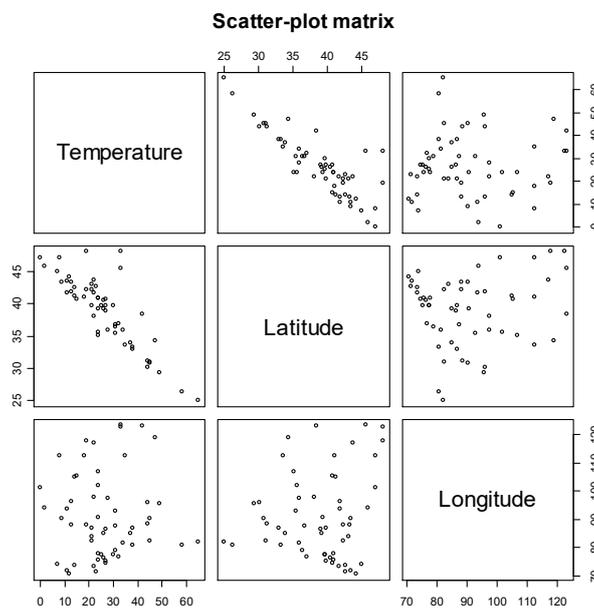


Figura 10.1.3.

• **Esempio 10.1.2.** Si dispone delle osservazioni delle temperature medie del mese di gennaio (in °F) nel periodo 1931-1960 in alcune città degli Stati Uniti (Fonte: Peixoto, J.L., 1990, A property of well-formulated polynomial regression models, *American Statistician* **44**, 26-30). I regressori sono la latitudine e la longitudine della città. I dati sono contenuti nel file `temperature.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\temperature.txt",
+   header = T)
> attach(d)
```

La matrice dei diagrammi di dispersione viene ottenuta mediante il seguente comando:

```
> pairs(d, main = "Scatter-plot matrix")
```

Il precedente comando fornisce il grafico della Figura 10.1.3. L'analisi del modello lineare con i regressori originali viene implementata mediante il seguente comando:

```
> summary(lm(Temperature ~ Latitude + Longitude))
```

Call:

```
lm(formula = Temperature ~ Latitude + Longitude)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-12.9983	-3.8957	0.5577	3.7330	22.0113

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	98.64523	8.32708	11.846	<2e-16	***
Latitude	-2.16355	0.17570	-12.314	<2e-16	***
Longitude	0.13396	0.06314	2.122	0.0386	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.935 on 53 degrees of freedom

Multiple R-Squared: 0.7411, Adjusted R-squared: 0.7314

F-statistic: 75.88 on 2 and 53 DF, p-value: 2.792e-16

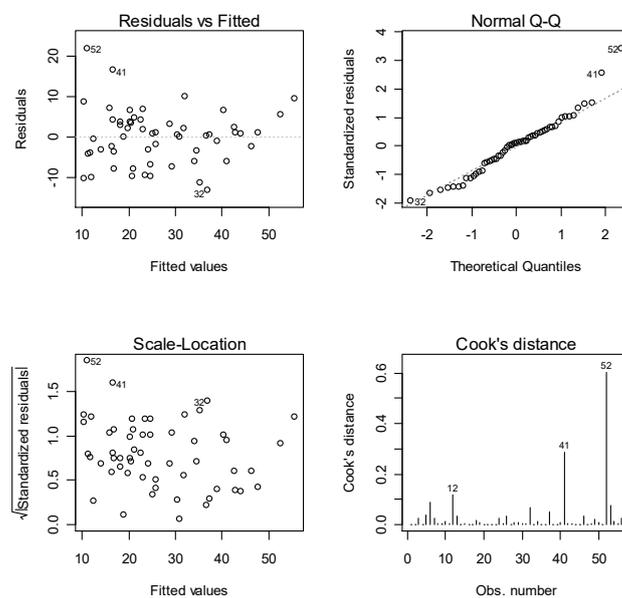


Figura 10.1.4.

I diagrammi di Anscombe, il diagramma quantile-quantile e il diagramma con le distanze di Cook del modello vengono ottenuti mediante i seguenti comandi:

```
> par(mfrow = c(2, 2))
> plot(lm(Temperature ~ Latitude + Longitude), which = c(1:4),
+      add.smooth = FALSE)
> par(mfrow = c(1, 1))
```

I precedenti comandi forniscono il grafico della Figura 10.1.4. L'analisi del modello lineare con i regressori originali senza l'osservazione anomala viene implementata mediante il seguente comando:

```
> summary(lm(Temperature ~ Latitude + Longitude,
+ subset = (1:length(Temperature) != 52)))
```

Call:

```
lm(formula = Temperature ~ Latitude + Longitude,
subset = (1:length(Temperature) != 52))
```

Residuals:

	Min	1Q	Median	3Q	Max
	-12.092772	-3.678680	0.001197	3.505167	19.667543

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	108.30043	7.84331	13.808	<2e-16 ***
Latitude	-2.28584	0.15992	-14.294	<2e-16 ***
Longitude	0.07522	0.05837	1.289	0.203

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.184 on 52 degrees of freedom

Multiple R-squared: 0.7971, Adjusted R-squared: 0.7893

F-statistic: 102.2 on 2 and 52 DF, p-value: < 2.2e-16

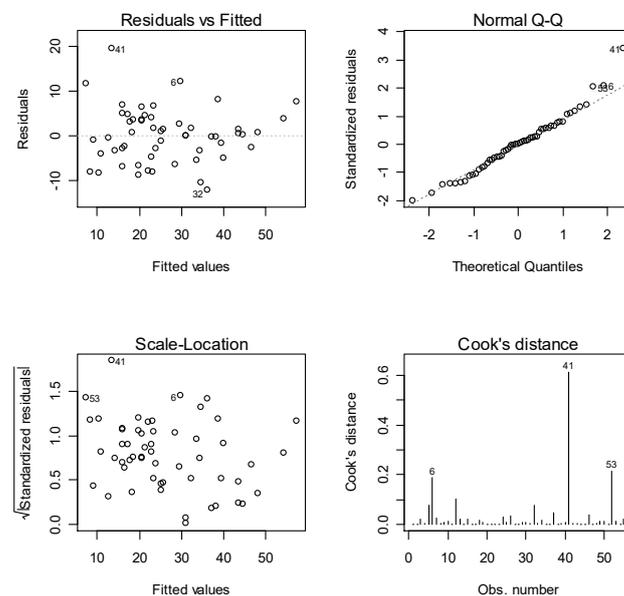


Figura 10.1.5.

I diagrammi di Anscombe, il diagramma quantile-quantile e il diagramma con le distanze di Cook del modello con i regressori originali e senza l'osservazione anomala sono ottenuti mediante i seguenti comandi:

```
> par(mfrow = c(2, 2))
> plot(lm(Temperature ~ Latitude + Longitude,
+ subset = (1:length(Temperature) != 52)),
+ which = c(1:4), add.smooth = FALSE)
> par(mfrow = c(1, 1))
```

I precedenti comandi forniscono il grafico della Figura 10.1.5. L'analisi del modello lineare con i regressori originali e l'interazione viene implementata mediante il seguente comando:

```
> summary(lm(Temperature ~ Latitude * Longitude))

Call:
lm(formula = Temperature ~ Latitude * Longitude)

Residuals:
    Min       1Q   Median       3Q      Max
-11.6738  -2.8165  -0.1268   3.4107  15.0605

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    259.48952    44.71515     5.803 3.93e-07 ***
Latitude       -6.07039     1.08235    -5.609 7.94e-07 ***
Longitude      -1.61025     0.48139    -3.345 0.001533 **
Latitude:Longitude  0.04220     0.01156     3.649 0.000611 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.247 on 52 degrees of freedom
Multiple R-Squared:  0.7939,    Adjusted R-squared:  0.782
F-statistic: 66.77 on 3 and 52 DF,  p-value: < 2.2e-16
```

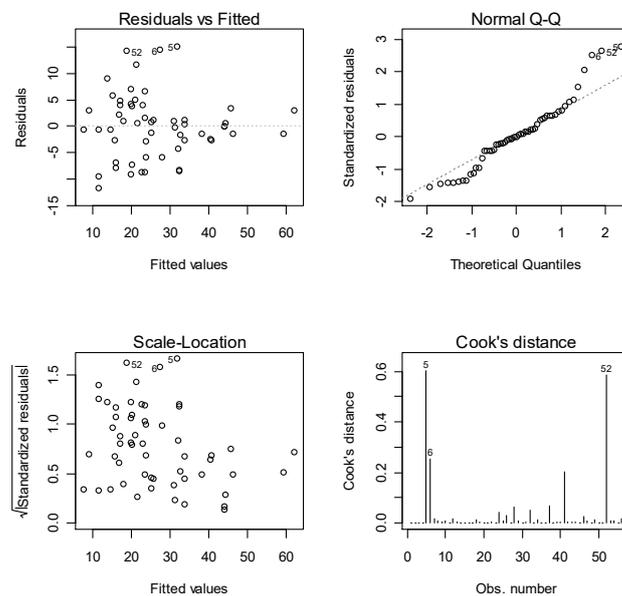


Figura 10.1.6.

I diagrammi di Anscombe, il diagramma quantile-quantile e il diagramma con le distanze di Cook del modello con interazione sono ottenuti mediante i seguenti comandi:

```
> par(mfrow = c(2, 2))
> plot(lm(Temperature ~ Latitude * Longitude), which = c(1:4),
+      add.smooth = FALSE)
> par(mfrow = c(1, 1))
```

I precedenti comandi forniscono il grafico della Figura 10.1.6. L'analisi del modello lineare con i regressori originali, l'interazione e una dipendenza cubica dalla longitudine viene implementata mediante il seguente comando:

```
> summary(lm(Temperature ~ Latitude * Longitude + I(Longitude^3)))
```

Call:

```
lm(formula = Temperature ~ Latitude * Longitude + I(Longitude^3))
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-7.30440 -2.85850  0.04342  2.49406  9.06776
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.908e+02  3.151e+01  12.402 < 2e-16 ***
Latitude      -5.891e+00  6.773e-01  -8.698 1.20e-11 ***
Longitude     -3.632e+00  3.750e-01  -9.687 3.77e-13 ***
I(Longitude^3)  8.064e-05  8.912e-06   9.050 3.47e-12 ***
Latitude:Longitude  3.656e-02  7.261e-03   5.035 6.34e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 3.908 on 51 degrees of freedom
Multiple R-Squared:  0.9209,    Adjusted R-squared:  0.9147
F-statistic: 148.5 on 4 and 51 DF,  p-value: < 2.2e-16
```

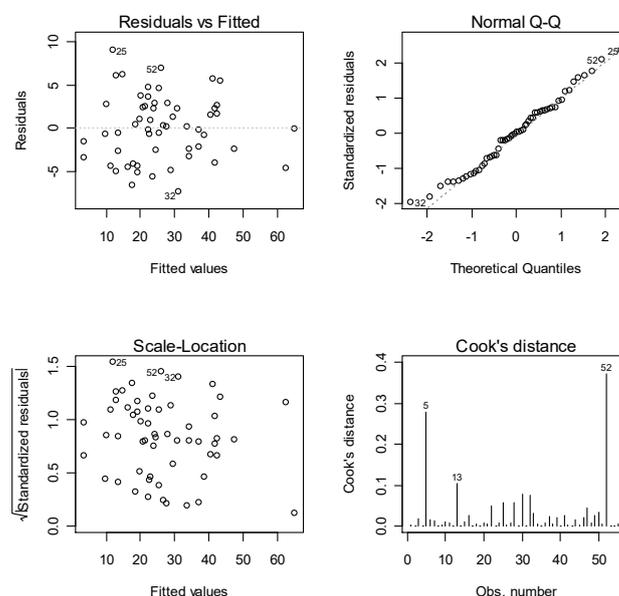


Figura 10.1.7.

I diagrammi di Anscombe, il diagramma quantile-quantile e il diagramma con le distanze di Cook del modello con interazione e dipendenza cubica sono ottenuti mediante i seguenti comandi:

```
> par(mfrow = c(2, 2))
> plot(lm(Temperature ~ Latitude * Longitude + I(Longitude^3)),
+      which = c(1:4), add.smooth = FALSE)
> par(mfrow = c(1, 1))
```

I precedenti comandi forniscono il grafico della Figura 10.1.7. Il modello esteso con interazione e dipendenza cubica dalla longitudine non può essere semplificato:

```
> summary(step(lm(Temperature ~ Latitude * Longitude +
+ I(Longitude^3)))
Start:  AIC=157.41
Temperature ~ Latitude * Longitude + I(Longitude^3)

              Df Sum of Sq      RSS      AIC
<none>                778.71  157.41
- Latitude:Longitude   1    387.05 1165.76  178.00
- I(Longitude^3)       1   1250.42 2029.12  209.04

Call:
lm(formula = Temperature ~ Latitude * Longitude + I(Longitude^3))

Residuals:
    Min       1Q   Median       3Q      Max
-7.30440 -2.85850  0.04342  2.49406  9.06776

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.908e+02  3.151e+01  12.402 < 2e-16 ***
Latitude      -5.891e+00  6.773e-01  -8.698 1.20e-11 ***
Longitude     -3.632e+00  3.750e-01  -9.687 3.77e-13 ***
I(Longitude^3)  8.064e-05  8.912e-06   9.050 3.47e-12 ***
Latitude:Longitude  3.656e-02  7.261e-03   5.035 6.34e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.908 on 51 degrees of freedom
Multiple R-Squared:  0.9209,    Adjusted R-squared:  0.9147
F-statistic: 148.5 on 4 and 51 DF,  p-value: < 2.2e-16
```

Anche se questo modello fornisce il migliore adeguamento, si tenga presente che il significato dei regressori non è facilmente interpretabile. \square

La struttura inferenziale dell'analisi della varianza basata su r campioni casuali indipendenti, ciascuno di numerosità n_j e tali che $\sum_{j=1}^r n_j = n$, da una variabile casuale Y a r livelli c_1, \dots, c_r di un fattore, con $E[Y_{c_j}] = \mu_j$ e $\text{Var}[Y_{c_j}] = \sigma^2$, può essere riportata ad un modello di regressione lineare. Se Y_1, \dots, Y_n sono le osservazioni relative al campione misto, allora si può scrivere il modello lineare

$$E[Y_i] = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{r-1} x_{i,r-1}$$

e

$$\text{Var}[Y_i] = \sigma^2,$$

dove $\beta_0 = \mu_1$ e $\beta_j = \mu_{j+1} - \mu_1$, mentre x_{ij} è il valore assunto sull' i -esima unità da un regressore binario che vale uno se l'osservazione è relativa al $(j+1)$ -esimo livello del fattore e zero altrimenti. L'ipotesi di omogeneità delle medie, supponendo la normalità di Y_1, \dots, Y_n , viene verificata considerando il sistema di ipotesi $H_0 : \beta_1 = \dots = \beta_{r-1} = 0$ contro $H_1 : \beta_j \neq 0$ per qualche j . Di conseguenza, le tecniche viste in precedenza possono essere applicate anche in questo caso. Il

vantaggio di questo approccio è quello di individuare da quale livello del fattore dipende l'eventuale rifiuto dell'ipotesi di base.

• **Esempio 10.1.3.** Si considera di nuovo i dati relativi alla velocità della luce dell'Esempio 8.3.1. L'ipotesi di omogeneità delle medie può essere verificato mediante il seguente comando:

```
> summary.lm(lm(Speed ~ Trial))

Call:
lm(formula = Speed ~ Trial)

Residuals:
    Min       1Q   Median       3Q      Max
-259.00  -42.62    2.25   41.75  161.00

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    909.00     16.60   54.762 < 2e-16 ***
TrialT2        -53.00     23.47   -2.258  0.026251 *
TrialT3        -64.00     23.47   -2.726  0.007627 **
TrialT4       -88.50     23.47   -3.770  0.000283 ***
TrialT5       -77.50     23.47   -3.301  0.001356 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 74.23 on 95 degrees of freedom
Multiple R-Squared:  0.1529,    Adjusted R-squared:  0.1173
F-statistic: 4.288 on 4 and 95 DF,  p-value: 0.003114
```

L'omogeneità delle medie viene ovviamente respinta con lo stesso livello di significatività osservata dell'Esempio 8.3.1. Analizzando le stime dei coefficienti, si nota che il rifiuto è principalmente dovuto ai coefficienti β_3 e β_4 le cui stime sono elevate in valore assoluto e con segno negativo, ovvero il rifiuto risulta principalmente da $\mu_1 > \mu_4$ e $\mu_1 > \mu_5$, analogamente all'Esempio 8.3.1. \square

L'analisi della varianza può essere estesa anche al caso che si abbiano due o più fattori, per cui si ha la cosiddetta analisi della varianza a due o più criteri. Anche in questo caso l'analisi può essere riportata ad un modello di regressione lineare. Per descrivere questo modello, si consideri per semplicità una analisi della varianza a due criteri, dove ogni fattore assume due livelli. Dunque, si hanno 4 campioni, ovvero un campione per ogni combinazione di livelli. Assumendo la normalità delle osservazioni, se Y_1, \dots, Y_n sono le osservazioni relative al campione misto, allora si può scrivere il modello lineare

$$E[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}$$

e

$$\text{Var}[Y_i] = \sigma^2,$$

dove x_{i1} è il valore assunto sull' i -esima unità da un regressore binario che vale uno se l'osservazione è relativa al secondo livello del primo fattore e zero altrimenti, mentre x_{i2} è il valore assunto sull' i -esima unità da un regressore binario che vale uno se l'osservazione è relativa al secondo livello del secondo fattore e zero altrimenti. Evidentemente, in questo caso la verifica d'ipotesi risulta più complessa. In effetti, si può voler verificare l'effetto marginale del primo fattore (ovvero l'ipotesi $H_0 : \beta_1 = 0$), l'effetto marginale del secondo fattore (ovvero l'ipotesi $H_0 : \beta_2 = 0$) o l'effetto dell'interazione dei due fattori (ovvero l'ipotesi $H_0 : \beta_3 = 0$). In una maniera simile anche se con crescente complessità in

notazione, il caso generale dell'analisi della varianza a più criteri può essere riportata ad un modello di regressione. Il vantaggio di questi approcci è quello di individuare da quale combinazioni di livelli dei fattori dipende la media delle osservazioni.

• **Esempio 10.1.4.** Si dispone delle osservazioni dei pesi di topi (in grammi) con quattro differenti genotipi per la genitrice e quattro differenti genotipi per la nidiata (Fonte: Scheffé, H., 1959, *Analysis of Variance*, Wiley, New York, p.140). I dati sono contenuti nel file `foster.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\foster.txt", header = T)
> attach(d)
```

Il seguente comando permette di ottenere le medie per tutte le combinazioni dei fattori:

```
> tapply(Weight, list(Mother, Litter), mean)
      A      B      I      J
A 63.680 52.325 47.10000 54.35000
B 52.400 60.640 64.36667 56.10000
I 54.125 53.925 51.60000 54.53333
J 48.960 45.900 49.43333 49.06000
```

L'analisi della varianza a due criteri può essere implementata mediante il seguente comando:

```
> summary(aov(Weight ~ Litter * Mother))
              Df Sum Sq Mean Sq F value    Pr(>F)
Litter          3   60.16    20.05   0.3697  0.775221
Mother          3  775.08   258.36   4.7632  0.005736 **
Litter:Mother    9  824.07    91.56   1.6881  0.120053
Residuals      45 2440.82    54.24
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Il seguente comando che semplifica il modello di regressione è equivalente alla precedente analisi della varianza:

```
> summary(step(lm(Weight ~ Litter * Mother)))
Start:  AIC=257.04
Weight ~ Litter * Mother

              Df Sum of Sq    RSS    AIC
- Litter:Mother  9      824.1 3264.9  256.8
<none>                2440.8  257.0

Step:  AIC=256.79
Weight ~ Litter + Mother

              Df Sum of Sq    RSS    AIC
- Litter      3      63.6 3328.5  252.0
<none>                3264.9  256.8
- Mother      3     775.1 4040.0  263.8

Step:  AIC=251.96
Weight ~ Mother
```

```

              Df Sum of Sq    RSS    AIC
<none>                3328.5  252.0
- Mother    3         771.6 4100.1  258.7

Call:
lm(formula = Weight ~ Mother)

Residuals:
    Min     1Q   Median     3Q     Max
-19.10  -5.90   1.50   5.32  12.80

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    55.400     1.910  28.999  <2e-16 ***
MotherB         3.300     2.797   1.180  0.2429
MotherI        -2.038     2.702  -0.754  0.4539
MotherJ        -6.720     2.746  -2.447  0.0175 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.642 on 57 degrees of freedom
Multiple R-Squared:  0.1882,    Adjusted R-squared:  0.1455
F-statistic: 4.405 on 3 and 57 DF,  p-value: 0.007433

```

Dunque, si rifiuta l'ipotesi di base ed il rifiuto è dovuto al genotipo della genitrice. L'analisi con il modello di regressione evidenzia in particolare che il primo genotipo della genitrice tende a produrre nidiate di tipo maggiore rispetto all'ultimo genotipo. \square

10.2. I modelli additivi generalizzati

In molti casi risulta difficile specificare la struttura di dipendenza della variabile di risposta dai regressori e questo fatto preclude la costruzione di un modello di regressione lineare (o comunque riconducibile a tale). Un approccio “distribution-free” alla regressione multipla può essere ottenuto considerando il modello additivo generalizzato con

$$E[Y_i] = \beta_0 + m_1(x_{i1}) + \dots + m_p(x_{ip})$$

e

$$\text{Var}[Y_i] = \sigma^2,$$

dove m_1, \dots, m_p sono funzioni non note. Dunque, il modello additivo generalizzato suppone che la variabile di risposta sia dipendente da una somma di trasformazioni non note dei regressori. Le funzioni m_1, \dots, m_p e la quantità β_0 vengono stimate con una procedura simile al caso della regressione lineare locale, ottenendo i valori stimati

$$\hat{y}_i = \hat{\beta}_0 + \hat{m}_1(x_{i1}) + \dots + \hat{m}_p(x_{ip}).$$

L'adeguatezza del modello può essere valutata mediante la devianza che è la somma dei quadrati dei residui dei valori osservati da quelli stimati, ovvero

$$D = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

• **Esempio 10.2.1.** Si considera di nuovo i dati relativi alle temperature delle città degli Stati Uniti. La stima del modello viene ottenuta richiamando la libreria `mgcv`. L'analisi del modello additivo generalizzato viene implementata come segue:

```
> library(mgcv)
> summary(gam(Temperature ~ s(Latitude) + s(Longitude)))

Family: gaussian
Link function: identity

Formula:
Temperature ~ s(Latitude) + s(Longitude)

Parametric coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.5179      0.3487   76.04  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
              edf Est.rank      F  p-value
s(Latitude)   6.579         9 119.85 < 2e-16 ***
s(Longitude)  4.867         9  31.31 3.46e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.962   Deviance explained =   97%
GCV score = 8.7557   Scale est. = 6.8098     n = 56
```

I grafici delle funzioni $\hat{m}_1(x_1)$ e $\hat{m}_2(x_2)$ possono essere ottenuti mediante il comando

```
> plot(gam(Temperature ~ s(Latitude) + s(Longitude)))
```

Il precedente comando fornisce i grafici della Figura 10.2.1 e della Figura 10.2.2.

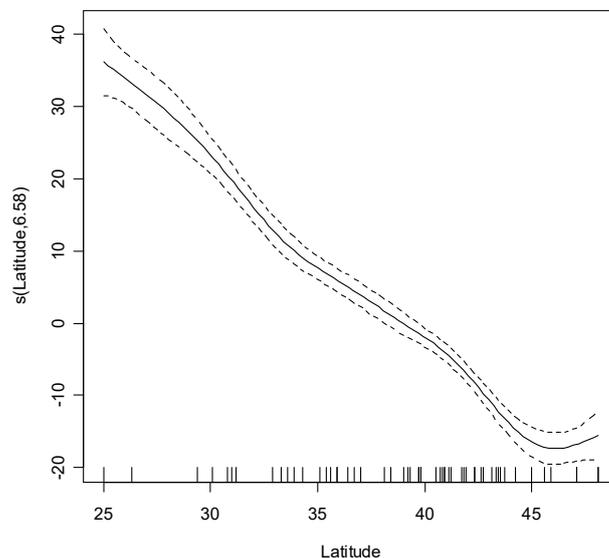


Figura 10.2.1.

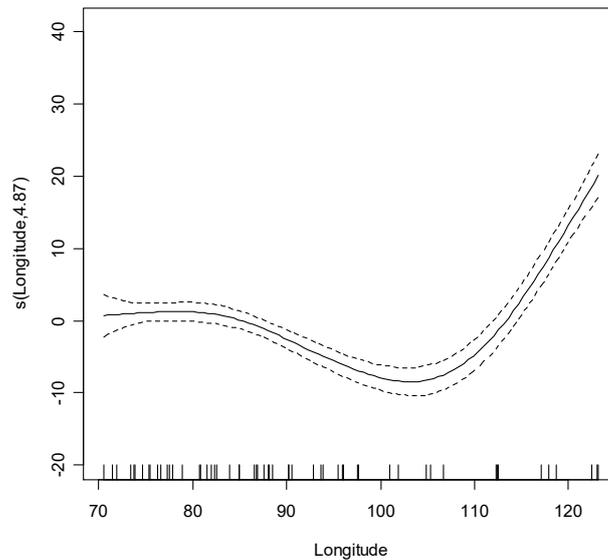


Figura 10.2.2.

Dai precedenti grafici si osserva che l'effetto della latitudine è sostanzialmente lineare quando lo spostamento avviene verso nord, mentre l'effetto della longitudine è curvilineo quando lo spostamento avviene da est verso ovest. □

10.3. I modelli lineari generalizzati multipli

Il modello lineare generalizzato multiplo viene ottenuto assumendo una distribuzione non Normale per Y_1, \dots, Y_n e considerando una opportuna funzione legame g tale che

$$g(E[Y_i]) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} .$$

Il modello di regressione lineare multiplo è un caso particolare di questa classe quando le variabili di risposta hanno distribuzione Normale e g è la funzione identità. In modo simile al caso univariato, per ogni distribuzione che si assume per la variabile di risposta esiste una funzione legame canonica, ovvero una parametrizzazione naturale del modello.

Quando si assume che Y_1, \dots, Y_n siano variabili casuali di Poisson, si ottiene la regressione di Poisson multipla e la funzione legame canonica è la funzione logaritmo, ovvero si ha

$$\log(E[Y_i]) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} .$$

Inoltre, se si assume che Y_1, \dots, Y_n siano variabili casuali di Bernoulli, si ha la regressione logistica multipla e la funzione legame canonica è la funzione logit, ovvero si ha

$$\log\left(\frac{E[Y_i]}{1 - E[Y_i]}\right) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} .$$

In modo analogo al caso univariato, se si dispone delle frequenze relative ai vari livelli dei regressori, la regressione logistica può essere applicata considerando le proporzioni della variabile di risposta per ogni livello dei regressori.

I modelli additivi possono essere adattati anche ai modelli lineari generalizzati introducendo una relazione del tipo

$$g(E[Y_i]) = \beta_0 + m_1(x_{i1}) + \dots + m_p(x_{ip}),$$

dove m_1, \dots, m_p sono funzioni non note. Questo tipo di modello risulta molto flessibile e particolarmente adatto per le applicazioni reali.

• **Esempio 10.3.1.** Si dispone dei dati relativi ai fallimenti di industrie degli Stati Uniti che hanno operato nelle telecomunicazioni fra il 2000 e il 2002 (Fonte: Simonoff, J.S., 2003, *Analyzing Categorical Data*, Springer, New York, p.381). I regressori sono il capitale da lavoro rapportato all'attivo totale, il guadagno rapportato all'attivo totale, il guadagno al netto di interessi e tasse rapportato all'attivo totale, le vendite rapportate all'attivo totale e il valore dell'azienda rapportato alla passività. I dati sono contenuti nel file `bankruptcy.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\bankruptcy.txt", header = T)
> attach(d)
```

La matrice dei diagrammi di dispersione viene ottenuta mediante il seguente comando:

```
> pairs(d[,2:7], main = "Scatter-plot matrix")
```

Il precedente comando fornisce il grafico della Figura 10.3.1.

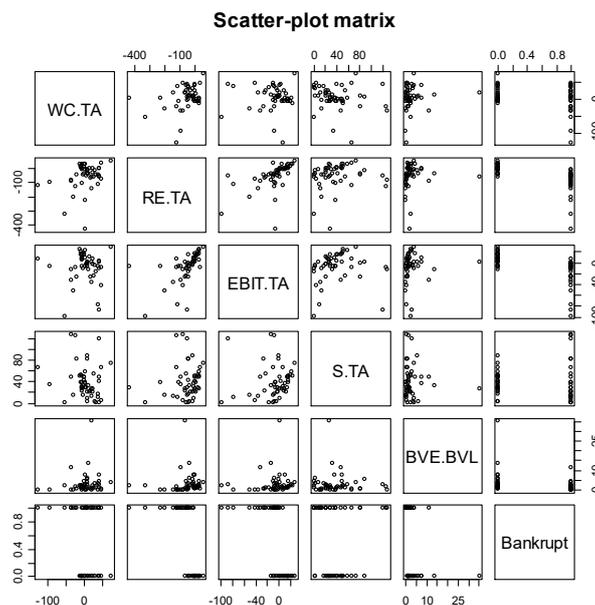


Figura 10.3.1.

L'analisi del modello di regressione logistica viene effettuata sui primi quattro regressori, dato che il quinto non sembra influenzare la probabilità di fallimento. La semplificazione automatica di questo modello mediante il criterio AIC viene implementata mediante il seguente comando:

```
> summary(step(glm(Bankrupt ~ WC.TA + RE.TA + EBIT.TA +
+ S.TA, binomial)))
```

```
Start: AIC=32.92
Bankrupt ~ WC.TA + RE.TA + EBIT.TA + S.TA
```

	Df	Deviance	AIC
- S.TA	1	22.968	30.968
- WC.TA	1	24.493	32.493
<none>		22.923	32.923
- RE.TA	1	26.347	34.347
- EBIT.TA	1	28.529	36.529

```
Step: AIC=30.97
Bankrupt ~ WC.TA + RE.TA + EBIT.TA
```

	Df	Deviance	AIC
- WC.TA	1	24.532	30.532
<none>		22.968	30.968
- RE.TA	1	26.407	32.407
- EBIT.TA	1	28.594	34.594

```
Step: AIC=30.53
Bankrupt ~ RE.TA + EBIT.TA
```

	Df	Deviance	AIC
<none>		24.532	30.532
- EBIT.TA	1	28.697	32.697
- RE.TA	1	35.979	39.979

```
Call:
glm(formula = Bankrupt ~ RE.TA + EBIT.TA, family = binomial)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.66450  -0.33579  -0.01778   0.28999   1.97243
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.95801    0.81637  -2.398   0.0165 *
RE.TA        -0.04405    0.01989  -2.214   0.0268 *
EBIT.TA      -0.10761    0.06445  -1.670   0.0950 .
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 69.315  on 49  degrees of freedom
Residual deviance: 24.532  on 47  degrees of freedom
AIC: 30.532
```

```
Number of Fisher Scoring iterations: 7
```

Un modello più semplice basato sul solo primo regressore potrebbe essere ottenuto notando che esiste un valore anomalo, ovvero la prima osservazione. L'analisi di questo modello viene implementata come segue:

```
> summary(glm(Bankrupt ~ RE.TA, binomial,
+ subset = (1:length(Bankrupt) != 1)))
```

```

Call:
glm(formula = Bankrupt ~ RE.TA, family = binomial,
subset = (1:length(Bankrupt) != 1))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.07214  -0.30405  -0.03211   0.24542   2.07304

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.05107    1.08850  -2.803  0.00506 **
RE.TA       -0.08277    0.02591  -3.194  0.00140 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 67.908  on 48  degrees of freedom
Residual deviance: 24.256  on 47  degrees of freedom
AIC: 28.256

Number of Fisher Scoring iterations: 8

```

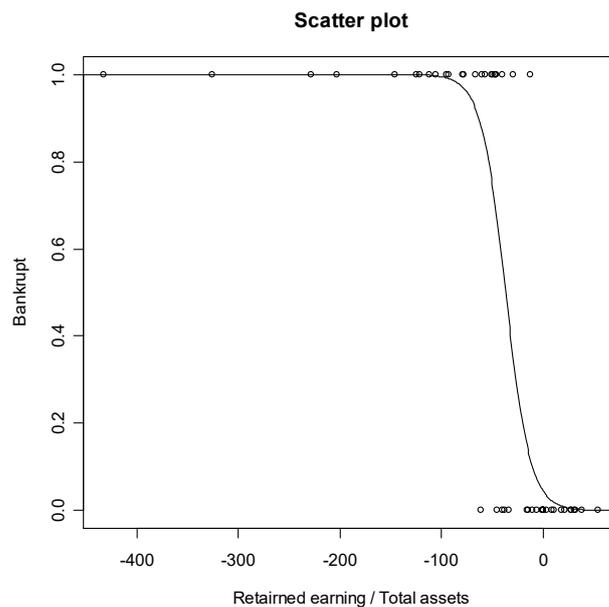


Figura 10.3.2.

Nell'ultimo modello, il diagramma di dispersione con la relativa funzione legame stimata viene ottenuto mediante i seguenti comandi:

```

> Bankrupt.h <- Bankrupt[2:50]
> RE.TA.h <- RE.TA[2:50]
> plot(RE.TA.h, Bankrupt.h,
+      xlab = "Retained earning / Total assets",
+      ylab = "Bankrupt", main = "Scatter plot")
> lines(seq(-500, 100, 1),
+       predict(glm(Bankrupt.h ~ RE.TA.h, binomial),
+               data.frame(RE.TA.h = seq(-500, 100, 1)), type = "response"))

```

I precedenti comandi forniscono il grafico della Figura 10.3.2. Dall'analisi si questo grafico si osserva che la probabilità di fallimento sale drasticamente da 0 a 1 quando il guadagno rapportato all'attivo totale tende a zero, un risultato facilmente intuibile. \square

I modelli lineari generalizzati possono essere applicati anche all'analisi delle tabelle a due o più entrate. Questi modelli sono detti log-lineari. Per descrivere questo modello si consideri per semplicità una tabella a doppia entrata, dove ogni fattore assume due livelli. Supponendo che le quattro frequenze osservate N_{ij} siano delle variabili casuali di Poisson (e dunque si considera in effetti lo schema probabilistico di Poisson descritto nell'Esempio 8.4.6), il modello log-lineare è dato da

$$\log(E[N_{ij}]) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2,$$

dove x_1 è un regressore binario che vale uno se la frequenza è relativa al secondo livello del primo fattore e zero altrimenti, mentre x_2 è un regressore binario che vale uno se la frequenza è relativa al secondo livello del secondo fattore e zero altrimenti. La funzione legame canonica è evidentemente la funzione logaritmo per le assunzioni fatte. In questo caso, si può voler verificare l'effetto marginale del primo fattore (ovvero l'ipotesi $H_0 : \beta_1 = 0$), l'effetto marginale del secondo fattore (ovvero l'ipotesi $H_0 : \beta_2 = 0$) o l'effetto dell'interazione dei due fattori (ovvero l'ipotesi $H_0 : \beta_3 = 0$). In una maniera simile anche se con complessità in notazione, il caso generale dell'analisi delle tabelle a più entrate può essere riportata ad un modello log-lineare.

• **Esempio 10.3.2.** Durante uno studio sull'uso di droghe da parte degli studenti nell'università di Dayton (Ohio) nel 1992 sono stati considerati un gruppo di studenti, ognuno dei quali è stato classificato per l'uso o meno di alcolici, di sigarette e di marijuana (Agresti, A., 1990, *Categorical data analysis*, prima edizione, Wiley, New York, p.152). I dati sono contenuti nel file `drug.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\drug.txt", header = T)
> attach(d)
```

La tabella a più entrate viene ottenuta mediante il comando:

```
> xtabs(Count ~ Alcohol + Cigarette + Marijuana)
, , Marijuana = No
```

	Cigarette	
Alcohol	No	Yes
No	279	43
Yes	456	538

```
, , Marijuana = Yes
```

	Cigarette	
Alcohol	No	Yes
No	2	3
Yes	44	911

Il diagramma a nastri condizionato viene ottenuto mediante il comando:

```
> library(lattice)
> barchart(xtabs(Count ~ Alcohol + Cigarette + Marijuana),
+         ylab = "Alcohol use",
+         auto.key = list(title = "Marijuana use", cex = 0.8),
+         strip = strip.custom(strip.names = T, strip.levels = T))
```

I precedenti comandi forniscono il grafico della Figura 10.3.3.

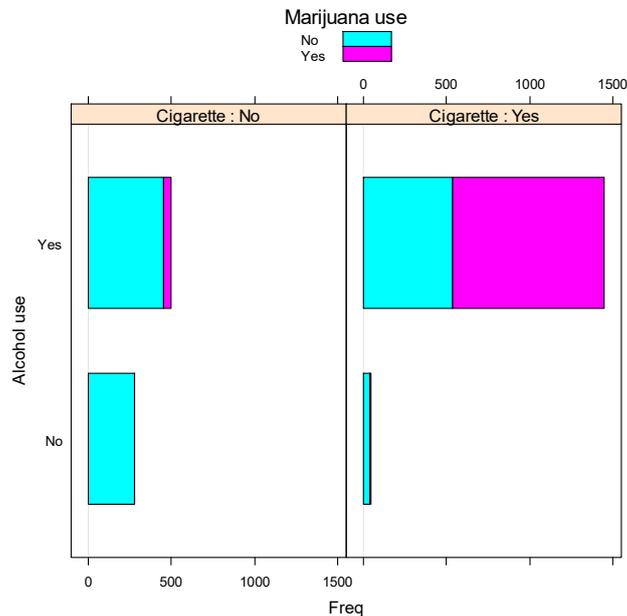


Figura 10.3.3.

Il comando `chisq.test` fornisce l'implementazione del test χ^2 per l'indipendenza:

```
> chisq.test(xtabs(Count ~ Alcohol + Cigarette + Marijuana))
```

Chi-squared test for given probabilities

```
data: xtabs(Count ~ Alcohol + Cigarette + Marijuana)
X-squared = 2676.337, df = 7, p-value < 2.2e-16
```

Una volta rifiutata l'indipendenza, il modello log-lineare può essere analizzato mediante il seguente comando:

```
> summary(step(glm(Count ~ Alcohol * Cigarette * Marijuana,
+ poisson)))
```

Start: AIC=65.04

Count ~ Alcohol * Cigarette * Marijuana

	Df	Deviance	AIC
- Alcohol:Cigarette:Marijuana	1	0.374	63.417
<none>		-2.92e-13	65.043

Step: AIC=63.42

Count ~ Alcohol + Cigarette + Marijuana + Alcohol:Cigarette +
Alcohol:Marijuana + Cigarette:Marijuana

	Df	Deviance	AIC
<none>		0.37	63.42
- Alcohol:Marijuana	1	92.02	153.06
- Alcohol:Cigarette	1	187.75	248.80
- Cigarette:Marijuana	1	497.37	558.41

```

Call:
glm(formula = Count ~ Alcohol + Cigarette + Marijuana +
Alcohol:Cigarette + Alcohol:Marijuana + Cigarette:Marijuana,
family = poisson)

Deviance Residuals:
    1         2         3         4         5         6         7
 0.02044 -0.02658 -0.09256  0.02890 -0.33428  0.09452  0.49134
    8
-0.03690

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      5.63342    0.05970   94.361 < 2e-16 ***
AlcoholYes       0.48772    0.07577    6.437 1.22e-10 ***
CigaretteYes    -1.88667    0.16270  -11.596 < 2e-16 ***
MarijuanaYes    -5.30904    0.47520  -11.172 < 2e-16 ***
AlcoholYes:CigaretteYes  2.05453    0.17406   11.803 < 2e-16 ***
AlcoholYes:MarijuanaYes  2.98601    0.46468    6.426 1.31e-10 ***
CigaretteYes:MarijuanaYes 2.84789    0.16384   17.382 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 2851.46098 on 7 degrees of freedom
Residual deviance: 0.37399 on 1 degrees of freedom
AIC: 63.417

Number of Fisher Scoring iterations: 4

```

L'analisi del modello porta dunque a concludere che vi è una forte dipendenza fra le coppie di sostanze, mentre non esiste una dipendenza fra l'uso contemporaneo delle tre sostanze. \square

10.4. Riferimenti bibliografici

- Agresti, A. (2013) *Categorical Data Analysis*, terza edizione, Wiley, New York.
- Agresti, A. (2015) *Foundations of Linear and Generalized Linear Models*, Wiley, New York.
- Atkinson, A.C., Riani, M. e Cerioli, A. (2004) *Exploring Multivariate Data with the Forward Search*, Springer, New York.
- Bretz, F., Hothorn, T. e Westfall, P. (2011) *Multiple Comparisons using R*, Chapman & Hall/CRC Press, Boca Raton.
- Dickhaus, T. (2014) *Simultaneous Statistical Inference*, Springer, New York.
- Dunn, P.K. e Smyth, G.K. (2018) *Generalized Linear Models with Examples in R*, Springer, New York.
- Everitt, B. e Hothorn, T. (2011) *An Introduction to Applied Multivariate Analysis with R*, Springer, New York.
- Fieller, N. (2016) *Basics of Matrix Algebra for Statistics with R*, CRC Press, Boca Raton.
- Friendly, M. e Meyer, D. (2019) *Discrete Data Analysis with R*, Chapman and Hall, London.
- Hastie, T., Tibshirani, R. e Friedman, J.H. (2009) *The Elements of Statistical Learning*, seconda edizione, Springer, New York.

- Hastie, T., Tibshirani, R. e Wainwright, M. (2015) *Statistical Learning with Sparsity*, Chapman and Hall/CRC Press, Boca Raton.
- Harezlak, J., Ruppert, D. e Wand, M.P. (2018) *Semiparametric Regression with R*, Springer, New York.
- James, G., Witten, D., Hastie, T. e Tibshirani, R. (2014) *An Introduction to Statistical Learning*, Springer, New York.
- Klemelä, J. (2014) *Multivariate Nonparametric Regression and Visualization*, Wiley, New York.
- McCullagh, P. e Nelder, J.A. (1989) *Generalized Linear Models*, seconda edizione, Chapman and Hall, London.
- Simonoff, J.S. (2003) *Analyzing Categorical Data*, Springer, New York.
- Seber, G.A.F. (2015) *The Linear Model and Hypothesis*, Springer, New York.
- Stasinopoulos, M.D., Rigby, R.A., Heller, G.Z., Voudouris, V. e De Bastiani, F. (2017) *Flexible Regression and Smoothing*, CRC Press, Boca Raton.
- Weisberg, S. (2014) *Applied Linear Regression*, Wiley, New York.
- Wood, S.N. (2017) *Generalized Additive Models*, seconda edizione, Chapman and Hall, London.
- Xu, J. (2023) *Modern Applied Regressions*, CRC Press, Boca Raton.