UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

CORSO DI LAUREA IN STATISTICA E TECNOLOGIE INFORMATICHE

TESI DI LAUREA TRIENNALE

Il modello di regressione con variabile risposta Beta

Relatore
Ch.ma Prof.ssa Laura Ventura

Laureando
Eros Magro
Matr. N. 600759 / STI

Anno Accademico 2010 / 2011

Alla mia Famiglia

Indice

Introduzione

In statistica è comune utilizzare un modello di regressione per spiegare una variabile risposta in relazione ad altre variabili concomitanti. Lo scopo di questa tesi è presentare un particolare modello di regressione adatto a trattare variabili risposta definite su un dominio limitato, come in particolare tassi e proporzioni. Il modello in questione è il modello di regressione con variabile risposta Beta, che è stato introdotto da Ferrari e Cribari-Neto nel 2004. La tesi si compone di tre capitoli: nel primo capitolo si presentano le caratteristiche della variabile casuale (v.c.) Beta e del modello di regressione corrispondente, nel secondo capitolo si introduce una versione estesa del modello e si spiega come utilizzare il pacchetto sviluppato per il software statistico R; infine nel terzo capitolo si utilizza quanto esposto per discutere un'applicazione del modello ad un dataset contenente dati reali ralativi ad ascolti televisivi.

Capitolo 1

Il modello di regressione Beta

In questo capitolo viene introdotto il modello di regressione Beta. In particolare si introducono la v.c. Beta, le principali caratteristiche del modello, l'inferenza basata sulla verosimiglianza, e alcuni test e metodi diagnostici per valutare la bontà di adattamento del modello. Quasto modello di regressione è particolarmente adatto per modellare variabili risposta y con dominio (0,1). In realtà, questo modello può essere usato anche per modellare variabili risposta che variano in un intervallo (a, b), dove a e b sono scalari, e $a < b$, modellando la variabile $(y - a)/(b - a)$ al posto della variabile y.

1.1 La variabile casuale Beta

La v.c. Beta è una v.c. avente supporto (0,1) ed è adatta quindi a modellare variabili come tassi e proporzioni. Inoltre, alcuni autori

(cfr. Ferrari e Cribari-Neto, 2004) ritengono che questa v.c. sia la scelta migliore per questo tipo di variabili.

A partire da questa distribuzione è possibile creare un modello statistico simile ad un modello lineare generalizzato, con la differenza che questa distribuzione non appartiene alla famiglia esponenziale. La distribuzione di densità una v.c. Beta è

$$f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \quad 0 < y < 1, \tag{1.1}$$

dove $p > 0$ e $q > 0$ sono opportuni parametri, e $\Gamma(\cdot)$ è la funzione gamma.

La media e la varianza di una v.c. $Y \sim Beta(p,q)$ sono date, rispettivamente, da

$$E(Y) = \frac{p}{p+q}$$

e

$$Var(Y) = \frac{pq}{(p+q)^2(p+q+1)}.$$

Inoltre, al variare di p e q, la forma della (1.1) varia considerevolmente, rendendola molto flessibile. In particolare, si ha (si veda anche la Figura 1.1):

- se $p > 1$, $q > 1$, la (1.1) è campanulare con un'unica moda in $\frac{p-1}{p+q-2}$;
- se $p < 1$, $q < 1$, la (1.1) ha forma ad U con minimo in $\frac{p-1}{p+q-2}$;
- se $p > 1$, $q \leq 1$, la (1.1) è monotona crescente;
- se $p \leq 1$, $q > 1$, la (1.1) è monotona decrescente;
- se $p = q = 1$, si ottiene la distribuzione uniforme U(0,1).

È facile notare, inoltre, che se si scambiano p e q si ottiene l'immagine riflessa della densità attorno all'ascissa 1/2; inoltre, se $p = q$ si ha che la funzione di densità è simmetrica attorno a 1/2.

Al fine della modellazione si userà una diversa parametrizzazione, in modo da avere un parametro per la media, che chiameremo μ, e un parametro di precisione, che chiameremo ϕ. Si pone

$$\mu = \frac{p}{p+q}$$

e

$$\phi = p + q,$$

da cui segue che $p = \mu\phi$ e $q = \phi(1-\mu)$. Da questo segue che

$$E(Y) = \mu$$

e

$$Var(Y) = \frac{V(\mu)}{1+\phi},$$

con $V(\mu) = \mu(1-\mu)$. Si noti che, al diminuire di $Var(Y)$, per μ fissato, aumenta il valore del parametro di precisione ϕ.

Procedendo per sostituzione, dalla (1.1) si ottiene la seguente distribuzione di densità

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}. \qquad (1.2)$$

Nel seguito, si indicherà con $Y \sim Beta(\mu,\phi)$ una v.c. con distribuzione (1.2).

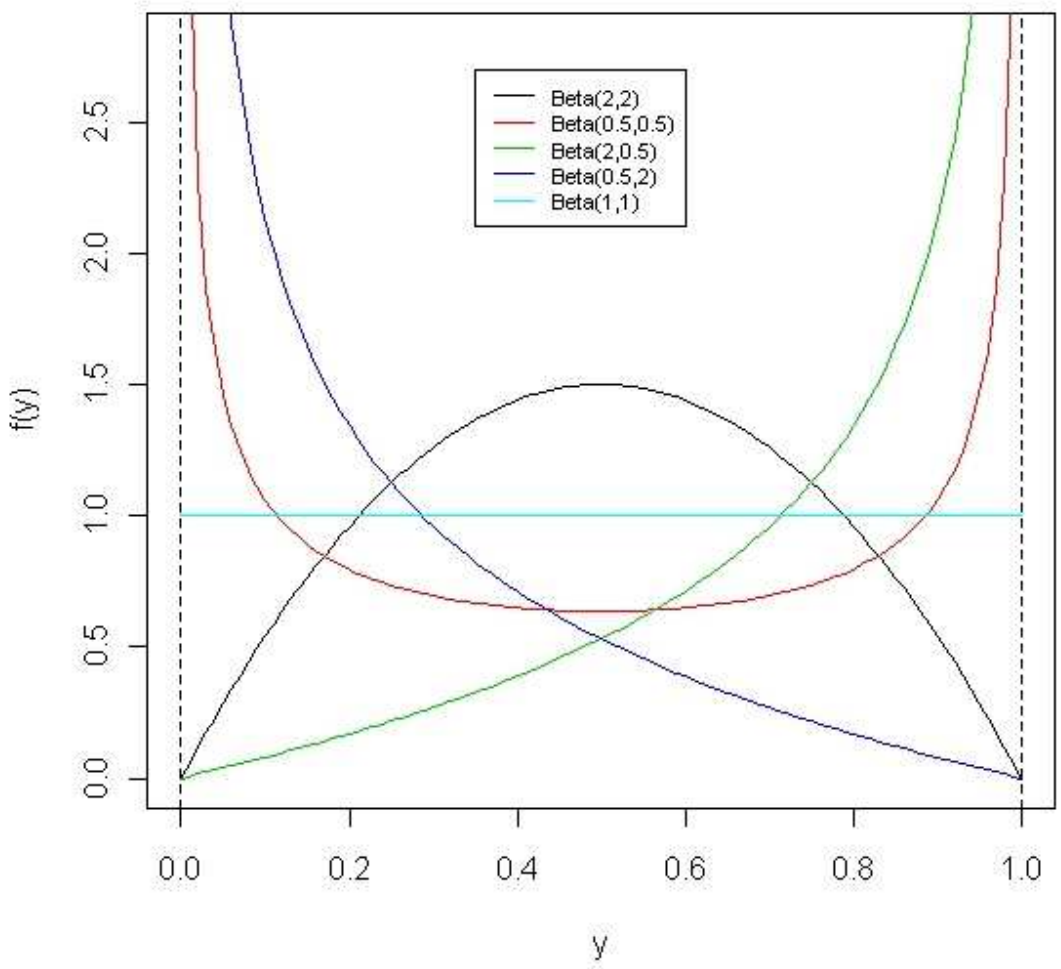

Figura 1.1: La distribuzione Beta per alcuni valori di p e q.

1.2 Il modello di regressione Beta

Siano $y_1, ..., y_n$ realizzazioni indipendenti dalle v.c. $Y_i \sim Beta(\mu_i, \phi)$, di media μ_i e parametro di precisione ϕ ignoti. Il modello di regressione Beta assume che ci sia un predittore lineare per la media, espresso come

$$g(\mu_i) = \sum_{j=1}^{k} x_{ij}\beta_j = \eta_i, \tag{1.3}$$

dove $\beta = (\beta_1, ..., \beta_k)$, con $\beta \in \Re^k$, è un vettore di ignoti parametri di regressione e $x_{i1}, ..., x_{ik}$, con $k < n$, sono k costanti note e fissate. La funzione $g(\cdot)$ è una funzione legame, definita da (0, 1) in $\Re$, strettamente monotona e derivabile due volte. Ci sono diverse possibili scelte per la funzione legame $g(\cdot)$, tra cui le più usate, in analogia con i modelli modelli per variabili binomiali, sono:

- la funzione logit: $g(\mu) = \log \frac{\mu}{1-\mu}$;
- la funzione probit: $g(\mu) = \Phi^{-1}(\mu)$, dove $\Phi(\cdot)$ è la funzione di ripartizione di una normale standard;
- la funzione complementary log-log: $g(\mu) = \log(-\log(1-\mu))$.

Nel caso del legame logit si ha che

$$\mu_i = \frac{e^{x_i^T\beta}}{1 + e^{x_i^T\beta}},$$

con $x_i^T = (x_{i1}, ..., x_{ik})$, $i = 1, ..., n$.

Si supponga che, assumendo la funzione legame logit, il valore dell'j-esimo regressore venga aumentato di una data costante c e che

tutte le altre variabili indipendenti rimangano inalterate. Sia μ^+ la media della Y sotto queste condizioni, e sia μ la media della Y sotto le condizioni di partenza, ovvero con le covariate originali. Allora si ha che

$$e^{c\beta_j} = \frac{\mu^+/(1-\mu^+)}{\mu/(1-\mu)},$$

ovvero si ha che il parametro a cui la covariata è legata moltiplicato per c può essere intrerpretato come il logaritmo tra l'odds-ratio sotto le nuove condizioni e quello sotto le vecchie condizioni.

1.3 La verosimiglianza

La log-verosimiglianza basata sul campione di n osservazioni indipendenti è

$$l(\beta, \phi) = \sum_{t=1}^{n} l_i(\mu_i, \phi), \tag{1.4}$$

dove

$$\begin{aligned} l_i(\mu_i, \phi) &= \log\Gamma(\phi) - \log\Gamma(\mu_i\phi) - \log\Gamma((1-\mu_i)\phi) + \\ &+ (\mu_i\phi - 1)\log y_i + ((1-\mu_i)\phi - 1)\log(1-y_i), \end{aligned} \tag{1.5}$$

con μ_i definito come nella (1.3).

Nel seguito si presenta la funzione score, detta anche funzione di punteggio, data da

$$\nabla l(\beta, \phi) = \begin{bmatrix} l_\beta \\ l_\phi \end{bmatrix},$$

con $l_\beta = \frac{\partial l(\beta,\phi)}{\partial \beta}$ e $l_\phi = \frac{\partial l(\beta,\phi)}{\partial \phi}$. Dalla log-verosimiglianza (1.4) si ottiene

$$\frac{\partial l(\beta,\phi)}{\partial \beta_r} = \sum_{i=1}^{n} \frac{\partial l_i(\mu_i,\phi)}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_r}, \quad r = 1, ..., k. \tag{1.6}$$

Si noti che $\partial\eta_i/\partial\mu_i = g'(\mu_i)$, e quindi $\partial\mu_i/\partial\eta_i = 1/g'(\mu_i)$, $i = 1, ..., n$.

Sia $\psi(x)$ la funzione digamma, definita come la derivata del logaritmo della funzione gamma, ovvero

$$\psi(x) = \frac{\partial \log\Gamma(x)}{\partial x}, \quad x > 0. \tag{1.7}$$

Dalla (1.5) si ottiene

$$\frac{\partial l_i(\mu_i,\phi)}{\partial \mu_i} = \phi\left[\log\left(\frac{y_i}{1-y_i}\right) - (\psi(\mu_i\phi) - \psi((1-\mu_i)\phi))\right]. \tag{1.8}$$

Posto

$$\mu_i^* = \psi(\mu_i\phi) - \psi((1-\mu_i)\phi)$$

e

$$y_i^* = \log\left(\frac{y_i}{1-y_i}\right)$$

si ottiene

$$l_r = \frac{\partial l_i(\beta,\phi)}{\partial \beta_r} = \phi\sum_{i=1}^{n}\frac{(y_i^* - \mu_i^*)x_{ir}}{g'(\mu_i)}, \quad r = 1, ..., k,$$

dove l_r indica l'r-esimo elementi del vettore l_β.

In forma matriciale, si può scrivere

$$l_\beta = [l_r] = \frac{\partial l(\beta,\phi)}{\partial \beta} = \phi X^T T(y^* - \mu^*), \tag{1.9}$$

dove X è la matrice del modello $n \times k$, T è una matrice diagonale $n \times n$ definita come $T = \text{diag}\{1/g'(\mu_i), ..., 1/g'(\mu_n)\}$, e $y^* = (y_1^*, ..., y_n^*)$

e $\mu^* = (\mu_1^*, ..., \mu_n^*)$.

Per la funzione di punteggio per ϕ, si ha

$$\begin{aligned}\frac{\partial l_i(\mu_i,\phi)}{\partial\phi} &= \mu_i\left[\log\left(\frac{y_i}{1-y_i}\right) - \psi(\mu_i\phi) + \psi((1-\mu_i)\phi)\right] + \\ &+ \log(1-y_i) - \psi((1-\mu_i)\phi) + \psi(\phi).\end{aligned}$$

Usando la notazione per y^* e μ^*, si ha che

$$l_\phi = \sum_{i=1}^{n}\left[\mu_i(y_i^* - \mu_i^*) + \log(1-y_i) - \psi((1-\mu_i)\phi) + \psi(\phi)\right]. \qquad (1.10)$$

Si considera ora la matrice d'informazione osservata, definita come la matrice delle derivate seconde cambiate di segno. Essa è data da

$$j(\beta,\phi) = \begin{bmatrix} j_{\beta\beta}(\beta,\phi) & j_{\beta\phi}(\beta,\phi) \\ j_{\phi\beta}(\beta,\phi) & j_{\phi\phi}(\beta,\phi) \end{bmatrix} = -\begin{bmatrix} \frac{\partial^2 l(\beta,\phi)}{\partial\beta_1\partial\beta_1} & \cdots & \frac{\partial^2 l(\beta,\phi)}{\partial\beta_1\partial\beta_k} & \frac{\partial^2 l(\beta,\phi)}{\partial\beta_1\partial\phi} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 l(\beta,\phi)}{\partial\beta_k\partial\beta_1} & \cdots & \frac{\partial^2 l(\beta,\phi)}{\partial\beta_k\partial\beta_k} & \frac{\partial^2 l(\beta,\phi)}{\partial\beta_k\partial\phi} \\ \frac{\partial^2 l(\beta,\phi)}{\partial\phi\partial\beta_1} & \cdots & \frac{\partial^2 l(\beta,\phi)}{\partial\phi\partial\beta_k} & \frac{\partial^2 l(\beta,\phi)}{\partial\phi\partial\phi} \end{bmatrix}.$$

Dalla (1.6) si ottiene

$$\frac{\partial^2 l(\beta,\phi)}{\partial\beta_r\partial\beta_s} = \sum_{i=1}^{n}\frac{\partial^2 l_i(\mu_i,\phi)}{\mu_i^2}\frac{1}{g'(\mu_i)^2}x_{ir}x_{is}, \quad r,s = 1,...,k.$$

Derivando ulteriormente la (1.8) rispetto a μ_i si ottiene

$$\frac{\partial^2 l_i(\mu_i,\phi)}{\partial\mu_i^2} = -\phi^2[\psi'(\mu_i\phi) + \psi'(\phi(1-\mu_i))],$$

e ponendo

$$w_i = \phi[\psi'(\mu_i\phi) + \psi'(\phi(1-\mu_i))]\frac{1}{g'(\mu_i)^2},$$

si giunge alle seguente espressione

$$j_{\beta\beta}(\beta,\phi) = \left[-\frac{\partial^2 l(\beta,\phi)}{\partial\beta_r\partial\beta_s}\right] = \phi\left[\sum_{i=1}^{n} w_i x_{ir} x_{is}\right]. \tag{1.11}$$

Quest'ultima quantità rappresenta il blocco (β,β) dell'informazione osservata, che coincide col rispettivo blocco dell'informazione attesa. La (1.11) può essere espressa anche in forma matriciale nel seguente modo

$$i_{\beta\beta}(\beta,\phi) = j_{\beta\beta}(\beta,\phi) = \phi X^T W X,$$

dove $W = \text{diag}\{w_1, ..., w_n\}$ è una matrice diagonale.

Dalla (1.6), si ottiene anche

$$\frac{\partial^2 l(\beta_r,\phi)}{\partial\beta\partial\phi} = \sum_{i=1}^{n}\left\{\left[(y_i^* - \mu_i^*) - \phi\frac{\partial\mu_i^*}{\partial\phi}\right]\frac{1}{g'(\mu_i)}x_{ir}\right\},$$

dove $\frac{\partial\mu_i^*}{\partial\phi} = \phi[\psi(\mu_i\phi)\mu_i - \psi(\phi(1-\mu_i))(1-\mu_i)]$, $i = 1, ..., n$.

Se si pone $c_i = \frac{\partial\mu_i^*}{\partial\phi}\phi$, si ha che l'elemento $j_{\beta_r\phi}(\beta,\phi)$ dell'informazione osservata è

$$j_{\beta_r\phi}(\beta,\phi) = -\sum_{i=1}^{n}[(y_i^* - \mu_i^*) - c_i]\frac{1}{g'(\mu_i)}x_{ir}.$$

Si nota ora che, essendo $E[y^*] = \mu^*$, il corrispondente elemento dell'informazione attesa è

$$i_{\beta_r\phi}(\beta,\phi) = \sum_{i=1}^{n} c_i\frac{1}{g'(\mu_i)}x_{ir}, \quad r = 1, ..., k.$$

Si giunge quindi alla forma matriciale $i_{\beta\phi}(\beta,\phi) = X^T T c$.

Infine, derivando la (1.10) rispetto a ϕ, si trova

$$\frac{\partial^2 l(\beta,\phi)}{\partial \phi^2} = -\sum_{i=1}^{n} \left[\psi'(\mu_i\phi)\mu_i^2 + (1-\mu_i)^2\psi'((1-\mu_i)\phi) - \psi'(\phi)\right].$$

Ponendo $D = \text{diag}\{d_1, ..., d_n\}$, dove il singolo elemento d_i è dato da

$$d_i = \psi'(\mu_i\phi)\mu_i^2 + (1-\mu_i)^2\psi'(\phi(1-\mu_i)) - \psi'(\phi), \quad i = 1, ..., n,$$

si ottiene che

$$i_{\phi\phi}(\beta,\phi) = E\left[\sum_{i=1}^{n} d_i\right] = tr(D).$$

La matrice d'informazione attesa, $i(\beta,\phi)$ risulta quindi

$$i(\beta,\phi) = \begin{bmatrix} \phi X^T W X & X^T T c \\ (X^T T c)^T & tr(D) \end{bmatrix}. \tag{1.12}$$

Si noti che i blocchi fuori diagonale non sono nulli; ovvero, i parametri β e ϕ non sono ortogonali, al contrario di quanto succede nei modelli lineari generalizzati.

1.4 L'inferenza

Le stime di massima verosimigianza per β e per ϕ sono ottenibili dalle equazioni di verosimiglianza, che sono le funzioni di punteggio poste uguali a zero, ovvero $l_\beta = 0$ e $l_\phi = 0$. Nel caso specifico di questo tipo di modelli, queste equazioni non sono risolvibili analicamente e si deve ricorrere a metodi numerici per risolverle come, ad esempio, l'algoritmo di Newton-Raphson (Azzalini, 2008).

La regola di aggiornamento dell'algoritmo di Newton-Raphson al passo s per il vettore $(k+1)$-dimensionale $\theta = (\beta_1, ..., \beta_k, \phi)$ è data da

$$\hat{\theta}_{s+1} = \hat{\theta}_s - \left(\frac{\partial^2 l(\hat{\theta}_s)}{\partial\theta\partial\theta^T}\right)^{-1} \frac{\partial l(\hat{\theta}_s)}{\partial\theta}, \quad s = 0, 1, ..., \tag{1.13}$$

dove $\hat{\theta}_s = (\hat{\beta}_{1s}, ..., \hat{\beta}_{ks}, \hat{\phi}_s)$. È noto che i metodi numerici necessitano di valori inizali per i parametri dai quali far partire la prima iterazione dell'algoritmo, dati da β_0 e ϕ_0. Questi valori possono essere scelti arbitrariamente, ma certe scelte risultano migliori e più leggere computazionalmente di altre. In particolare, Ferrari e Cribari-Neto (2004) suggeriscono la seguente soluzione. Per il vettore β, una buona scelta è data da

$$\beta_0 = (X^T X)^{-1} X^T z, \tag{1.14}$$

dove $z = (g(y_1), ..., g(y_n))^T$. La (1.14) deriva dall'idea di effettuare una regressione lineare utilizzando come variabili risposta le $g(y_1), ..., g(y_n)$, e di simare i parametri β con il metodo dei minimi quadrati.

Per la stima iniziale di ϕ, invece, si ricorda che $Var(Y_i) = \frac{\mu_i(1-\mu_i)}{1+\phi}$, da cui è facile ricavare che $\phi = \frac{\mu_i(1-\mu_i)}{Var(Y_i)} - 1$, $i = 1, ..., n$. Ma utilizzando i primi due termini dello sviluppo di Taylor di $g(y_i)$ in μ_i, si ha

$$Var(g(y_i)) \doteq Var(g(\mu_i) + (y_i - \mu_i)g'(\mu_i)) = Var(y_i)[g'(\mu_i)]^2$$

e quindi che

$$Var(y_i) \doteq \frac{Var(g(y_i))}{[g'(\mu_i)]^2}, \quad i = 1, ..., n.$$

Di conseguenza, il valore iniziale per ϕ è

$$\phi_0 = \frac{1}{n}\sum_{i=1}^{n} \frac{\check{\mu}_i(1-\check{\mu}_i)}{\check{\sigma}_i^2} - 1,$$

dove $\check{\mu}_i$ è ottenuto applicando $g^{-1}(\cdot)$ all'i-esimo valore stimato dal modello di regressione lineare di $g(y_1), ..., g(y_n)$ su X, ovvero

$$\check{\mu}_i = g^{-1}(x_i^T (X^T X)^{-1} X^T z), \quad i = 1, ..., n,$$

mentre

$$\check{\sigma}_i^2 = \frac{\check{e}^T \check{e}}{(n-k)[g^{-1}(\check{\mu}_i)]^2},$$

con $\check{e}$ vettore dei residui empirici della regressione di z su X, ossia $\check{e} = z - X(X^T X)^{-1} X^T z$.

Per determinare la distribuzione asintotica degli stimatori di massima verosimiglianza risulta utile la matrice d'informazione attesa (1.12), in quanto la sua inversa fornisce una stima della matrice di covarianza asintotica di $(\hat{\beta}, \hat{\phi})$ sotto (β, ϕ). Sotto condizioni di regolarità, vale

$$\begin{pmatrix} \hat{\beta} \\ \hat{\phi} \end{pmatrix} \dot{\sim} N_{k+1}\left(\begin{pmatrix} \beta \\ \phi \end{pmatrix}, i(\hat{\beta}, \hat{\phi})^{-1} \right), \tag{1.15}$$

dove $\hat{\beta}$ e $\hat{\phi}$ sono gli stimatori di massima verosimiglianza di β e di ϕ.

È utile ottenere l'espressione dell'inversa della matrice d'informazione attesa. Per le proprietà delle matrici a blocchi, si ha che

$$i(\beta, \phi)^{-1} = \begin{pmatrix} i^{\beta\beta} & i^{\beta\phi} \\ i^{\phi\beta} & i^{\phi\phi} \end{pmatrix}, \tag{1.16}$$

dove

$$i^{\beta\beta} = \frac{1}{\phi}(X^T W X)^{-1} \left(I_k + \frac{X^T T c c^T T^T X (X^T W X)^{-1}}{\gamma \phi} \right),$$

$$i^{\beta\phi} = (i^{\phi\beta})^T = -\frac{1}{\gamma\phi}(X^T W X)^{-1} X^T T c$$

e

$$i^{\phi\phi} = \frac{1}{\gamma},$$

con $\gamma = tr(D) - \phi^{-1} c^T T^T X (X^T W X)^{-1} X^T T c$, e I_k è la matrice identità di dimensione $k \times k$.

A partire dalla (1.15) è possibile ricavare anche stime intervallari approssimate per i parametri. Posto $z_{1-\alpha/2}$ il quantile $1 - \frac{\alpha}{2}$ della $N(0,1)$, con $0 < \alpha < 1/2$, e posto $i^{rr}(\hat{\beta}, \hat{\phi})$ l'r-esima componente della diagonale dell'inversa della matrice dell'informazione attesa (1.16) valutata in $(\hat{\beta}, \hat{\phi})$ con $r = 1, ..., k+1$, si ha che

$$\left[\ \hat{\beta}_r - z_{1-\alpha/2}(i^{rr}(\hat{\beta}, \hat{\phi}))^{1/2}, \ \hat{\beta}_r + z_{1-\alpha/2}(i^{rr}(\hat{\beta}, \hat{\phi}))^{1/2} \ \right]$$

e

$$\left[\ \hat{\phi} - z_{1-\alpha/2}(i^{(k+1)(k+1)}(\hat{\beta}, \hat{\phi}))^{1/2}, \ \hat{\phi} + z_{1-\alpha/2}(i^{(k+1)(k+1)}(\hat{\beta}, \hat{\phi}))^{1/2} \ \right]$$

rappresentano, rispettivamente, intervalli di confidenza per β_r e per ϕ di livello approssimato $1 - \alpha$.

Un intervallo di confidenza di livello approssimato $1 - \alpha$ per la media μ, è dato da

$$[g^{-1}(\hat{\eta} - z_{1-\alpha/2} se(\hat{\eta})), \ g^{-1}(\hat{\eta} + z_{1-\alpha/2} se(\hat{\eta}))],$$

dove $\hat{\eta} = x^T \hat{\beta}$ e $se(\hat{\eta}) = \sqrt{x^T i^{\beta\beta}(\hat{\beta}, \hat{\phi}) x}$, con x vettore di covariate fissato e $i^{\beta\beta}(\hat{\beta}, \hat{\phi})$ componente (β, β) dell'inversa dell'informazione di Fisher valutata con in $(\hat{\beta}, \hat{\phi})$. Si noti che questo intervallo di confidenza

è valido solo se la funzione di legame è strettamente crescente.

Supponiamo ora di essere interessati ad effettuare una verifica d'ipotesi del tipo

$$H_0 : \beta_1 = \beta_1^{(0)} \; vs \; H_1 : \beta_1 \neq \beta_1^{(0)},$$

con $\beta_1 = (\beta_1, ..., \beta_m)^T$, $\beta_1^{(0)} = (\beta_1^{(0)}, ..., \beta_m^{(0)})^T$, per $m \leq k$, e $\beta_1^{(0)}$ vettore di costanti note e fissate. Per tale problema, si può far riferimento alla statistica test log-rapporto di verosimiglianza (Pace e Salvan, 2001), data da

$$W_{lr} = 2(l(\hat{\beta}, \hat{\phi}) - l(\tilde{\beta}, \tilde{\phi})),$$

dove $(\tilde{\beta}, \tilde{\phi})$ sono le stime di massima verosimiglianza di (β, ϕ) sotto l'ipotesi nulla H_0. Sotto condizioni di regolarità e sotto H_0, si ha che $W_{lr} \xrightarrow{d} \chi^2_m$. L'ipotesi nulla viene rifiuta per valori alti della statistica W_{lr}.

Un altro test che si può utilizzare è il test alla Wald, dato da

$$W_w = (\hat{\beta}_1 - \beta_1^{(0)})^T (i_{11}^{\beta\beta}(\hat{\beta}, \hat{\phi}))^{-1} (\hat{\beta}_1 - \beta_1^{(0)}),$$

dove $i_{11}^{\beta\beta}(\hat{\beta}, \hat{\phi})$ equivale a $i_{11}^{\beta\beta}$ valutata nelle stime di massima verosimiglianza. Con $i_{11}^{\beta\beta}$ si intende il blocco $i^{\beta\beta}$ privato delle righe e delle colonne in cui compaiono gli elementi diagonali legati ai parametri non testati. Il test W_w è asintoticamente equivalente al test W_{lr}. In particolare, per testare se il j-esimo parametro di regressione β_j $(j = 1, ..., k)$ è significativo, si può utilizzare la statistica test di Wald $\hat{\beta}_j / \sqrt{i_{jj}^{\beta\beta}(\hat{\beta}, \hat{\phi})}$, che si distribuisce asintoticamente come una normale standard.

I test W_{lr} e W_w possono essere anche utilizzati per confrontare modelli annidati.

1.5 Bontà di adattamento

Una volta stimato un modello, è importante effettuare un'analisi diagnostica per valutare la bontà dell'adattamento.

Una misura globale della varianza spiegata, e quindi dell'adattamento del modello ai dati, può essere ottenuta calcolando l'indice pseudo-R^2 (Ferrari e Cribari-Neto, 2004), denotato con R^2_p. Posto $\hat{\eta} = (\hat{\eta}_1, ..., \hat{\eta}_n)$ e $g(y) = (g(y_1), ..., g(y_n))$, l'indice pseudo-$R^2$ è definito come il quadrato del coefficiente di correlazione calcolato tra $\hat{\eta}$ e $g(y)$, per cui $0 \leq R^2_p \leq 1$. Per quanto riguarda l'interpretazione di tale indice si può dire che è analoga a quella dell'R^2 per i modelli lineari normali. Un'ultima cosa da notare sull'indice pseudo-R^2 è che, in caso di perfetto accordo tra $\hat{\eta}$ e $g(y)$, che equivale ad un accordo perfetto tra $\hat{\mu}$ e y, esso assume il valore 1.

Un'altra valutazione del modello può essere ottenuta a partire dai residui standardizzati, o di Pearson, dati da

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{var}(Y_i)}}, \quad i = 1, ..., n,$$

dove $\hat{\mu}_i = g^{-1}(x_i^T \hat{\beta})$ e $\hat{var}(Y_i) = [\hat{\mu}_i(1 - \hat{\mu}_i)]/(1 + \hat{\phi})$. Questi residui possono essere utilizzati per costruire diagrammi di dispersione che li mettono a confronto con gli indici delle osservazioni i, formando le coppie di punti (i, r_i), oppure con i valori $\hat{\eta}_i$, formando le coppie $(\hat{\eta}_i, r_i)$, $i = 1, ..., n$. La presenza in questi grafici di andamenti sistematici indica che il modello adottato non è adeguato.

Un altro tipo di residui che si possono considerare sono i residui di

devianza, definiti come

$$r_i^d = \text{sgn}(y_i - \hat{\mu}_i)\sqrt{2(l_i(\ddot{\mu}_i, \hat{\phi}) - l_i(\hat{\mu}_i, \hat{\phi}))}, \quad i = 1, ..., n,$$

dove $\ddot{\mu}_i$ è il valore di μ_i che risolve l'equazione $\partial l_i/\partial \mu_i = 0$, ossia $\phi(y_i^* - \mu_i^*) = 0$. Questi residui derivano dalla quantità

$$D(y; \hat{\mu}, \hat{\phi}) = \sum_{i=1}^{n} 2(l_i(\ddot{\mu}_i, \hat{\phi}) - l_i(\hat{\mu}_i, \hat{\phi})),$$

che è detta devianza del modello. È facile notare che la relazione che lega le quantità r_i^d e $D(y; \hat{\mu}, \hat{\phi})$ è data da $D(y; \hat{\mu}, \hat{\phi}) = \sum_{i=1}^{n}(r_i^d)^2$. Si ha quindi che più il valore di r_i^d è grande, più l'i-esima osservazione contribuisce alla devianza del modello, e viceversa. I residui di devianza vengono analizzati con gli stessi grafici dei residui standard e ci si aspetta, come nei primi, che se il modello è buono non ci siano andamenti sistematici.

L'ultima misura diagnostica che può essere considerata è la distanza di Cook (Ferrari e Cribari-Neto, 2004). Essa misura l'influenza di una singola osservazione sulle stime dei parametri di regressione, nel momento in cui viene tolta dal singolo processo di stima. Nel caso in cui la distanza di Cook assume valori elevati (solitamente maggiori di 1) si può affermare che l'osservazione è molto influente, e si può quindi scegliere di ignorarla, se si ritiene che essa alteri in maniera scorretta le stime. La distanza di Cook è definita come

$$Cook_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T X^T W X (\hat{\beta} - \hat{\beta}_{(i)})}{k}, \quad i = 1, ..., n,$$

dove $\hat{\beta}_{(i)}$ è la stima di massima verosimiglianza del parametro β effet-

tuata senza l'i-esima osservazione. Si noti che tale quantità rappresenta una distanza tra $\hat{\beta}_{(i)}$ e $\hat{\beta}$. Infine, si osserva che, per evitare di dover stimare il modello per $n+1$ volte, e quindi affrontare un algoritmo computazionalmente pesante, si può usare una comoda approssimazione della distanza di Cook, data da

$$C_i = \frac{h_{ii} - r_i^2}{k(1-h_{ii})^2},$$

dove h_{ii} indica l'i-esimo elemento della diagonale della matrice di proiezione $H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$ ed r_i indica l'i-esimo residuo standard. Per vedere rapidamente quali valori si possono eliminare, si consiglia di rappresentare tutti i valori su un diagramma a bastoncini.

1.6 Conclusioni

In questo capitolo si è introdotto il modello di regressione Beta, molto flessibile per modellare variabili che variano in intervalli fissati. Sono state introdotte la verosimiglianza e le sue quantità, le distribuzioni asintotiche per i parametri, sono stati presentati intervalli di confidenza approssimati per i parametri e la media, statistiche test, indici per la varianza spiegata del modello, residui e distanze di Cook.

Nel prossimo capitolo si introdurrà un pacchetto R per adattare modelli di regressione Beta con parametro ϕ costante, ma anche per situazioni con parametro ϕ che varia da un'osservazione all'altra e si introdurrà brevemente questa estensione del modello trattato.

Capitolo 2

Adattamento in R

Il pacchetto statistico R da utilizzare per adattare un modello di regressione Beta è il pacchetto `betareg`. Esso in realtà implementa un'estensione del modello trattato in precedenza, che verrà brevemente introdotto. Si passerà poi a descrivere il funzionamento dei comandi R, e si utilizzerà un dataset contenuto nella libreria per chiarire meglio come lavorare con tale pacchetto.

2.1 Estensione al caso di eteroschedasticità

Un'estensione del modello introdotto nel capitolo precedente è implementata nel pacchetto R `betareg`, e si tratta del modello di regressione beta con dispersione variabile. In tale modello, il parametro di precisione ϕ non è assunto costante per tutte le osservazioni, ma varia da un'osservazione all'altra.

Si considerino n osservazioni indipendenti tali che $Y_i \sim Beta(\mu_i, \phi_i)$, con $i = 1, ..., n$, e si assumano due predittori lineari, uno per la media

e uno per la precisione, definiti come segue

$$g_1(\mu_i) = \eta_{1i} = x_i^T \beta,$$

$$g_2(\phi_i) = \eta_{2i} = z_i^T \zeta,$$

dove $\beta = (\beta_1, ..., \beta_k)$ e $\zeta = (\zeta_1, ..., \zeta_h)$, con $k + h < n$, sono i coefficienti di regressione e x_i e z_i sono i vettori dei regressori. Analogamente al capitolo precedente, gli elementi dei vettori β e ζ sono stimati con la massima verosimiglianza, sostituendo ϕ_i a ϕ nella (1.5). Si noti che se $g_2(\cdot)$ è la funzione identità, $\zeta = \zeta_1$ e $z_i = 1$, si ha $\phi_i = \zeta_1$, ovvero si torna ad avere il modello esposto nel capitolo precedente. Per quanto riguarda i residui si possono utilizzare i residui di Pearson avendo l'accortezza di sostituire $\hat{\phi}$ con $\hat{\phi}_i = g_2^{-1}(z_i^T \hat{\zeta})$. Un procedimento analogo può essere usato per i residui di devianza e per gli intervalli di confidenza. Un'altro modo per calcolare i residui, implementato dal pacchetto e che risulta avere migliori proprietà rispetto ai residui di Pearson e di devianaza (Cribari-Neto e Zeileis, 2010), è quello di utilizzare i residui standard pesati dati da

$$r_i^{sw2} = \frac{y_i^* - \hat{\mu}_i^*}{\sqrt{\hat{v}_i(1 - h_{ii})}},$$

dove $\hat{v}_i = \psi'(\hat{\mu}_i \hat{\phi}_i) + \psi'((1 - \hat{\mu}_i)\hat{\phi}_i)$, $i = 1, ..., n$.

2.2 La funzione `betareg` e le funzioni collegate

All'apertura di una sessione di lavoro di R, affinché si possa utilizzare il pacchetto `betareg` è necessario digitare il comando `library(betareg)`. All'interno di questa libreria, si trova la funzione `betareg()` che funziona in modo simile alla funzione `glm()` per adattare i modelli lineari generalizzati. Gli argomenti della funzione `betareg()` sono

```
betareg(formula, data, subset, na.action, weights, offset,
        link = "logit", link.phi = NULL,
        control = betareg.control(...),
        model = TRUE, y = TRUE, x = FALSE, ...).
```

Se la `formula` è del tipo `y` $\sim$ `x1 + x2`, il modello stimato assume che vi sia omoschedasticità; ossia, è un modello del tipo esposto nel Capitolo 1. Per adattare invece un modello con eteroschedasticità, `formula` deve essere del tipo `y` $\sim$ `x1 + x2 | z1 + z2 + z3`. La funzione legame $g_2(\cdot)$, rappresentata da `link.phi`, di default è la funzione logaritmo se in `formula` è presente il catattere `|`. Si deve quindi prestare attenzione, in quanto, per esempio le formule `y` $\sim$ `x1 + x2` e `y` $\sim$ `x1 + x2 | 1` usano diverse parametrizzazioni per ϕ_i: nella prima $\phi_i = \zeta_1$, nella seconda $\log(\phi_i) = \zeta_1$. Le funzioni legame implementate per la media sono le stesse che vi sono nella famiglia `binomial` della funzione `glm()`; ovvero: `logit` (default), `probit`, `cloglog`, `cauchit`, `log` e `loglog`. Quelle implementate per la precisione, invece, sono: `identity` (default, se non sono specificati i regressori per la precisione), `log` (default negli altri casi) e `loglog`.

All'argomento `control` invece bisogna passare una lista che contiene varie informazioni per stimare il modello con l'utilizzo della funzione `betareg()`. Questa lista è restituita dalla funzione `betareg.control()`, alla quale si può passare un parametro stringa all'argomento `method`, mediante il quale è possibile specificare l'algoritmo da utilizzare per ricavare le stime di massima verosimiglianza. Di default `method` è settato a BFGS, un algoritmo con ottime prestazioni.

Segue ora un elenco di utili funzioni applicabili ad un oggetto di tipo `betareg`, seguite da una breve descrizione:

- `print()`: stampa a video i coefficienti stimati;
- `summary()`: stampa a video varie quantità di regressione;
- `coef()`: ritorna i coefficienti di regressione in un vettore (media e precisione);
- `vcov()`: ritorna la stima della matrice di covarianza degli stimatori;
- `predict()`: ritorna previsioni di medie μ, predittori lineari η_1, parametri di precisione ϕ o varianze $\mu(1-\mu)/(1+\phi)$ per nuovi valori;
- `fitted()`: ritorna le medie stimate per i dati osservati;
- `residuals()`: estrae residui (di risposta $(y_i - \hat{\mu}_i)$, di Pearson, di devianza, o residui standard pesati), di default estrae i residui standard pesati;
- `model.matrix()`: ritorna la matrice X del modello;

- `model.frame()`: ritorna il frame originale utilizzato nel modello;
- `logLik()`: ritorna il valore stimato della log-verosimiglianza $l(\beta, \phi)$;
- `plot()`: crea grafici diagnostici;
- `hatvalues()`: ritorna gli elementi diagonali della matrice di proiezione H;
- `cooks.distance()`: ritorna le distanze di Cook;
- `coeftest()`: applica il test di Wald per testare la significatività dei coefficienti di regressione;
- `waldtest()`: effettua un test alla Wald per modelli annidati;
- `lrtest()`: effettua un test log-rapporto di verosimiglianza per modelli annidati;
- `AIC()`: calcola l'omonimo indice.

2.3 Un esempio illustrativo

Si considera, a scopo illustrativo, il dataset discusso da Prater (1956) sulla benzina, presente nel pacchetto `betareg`. Esso contiene $n = 32$ osservazioni su dati riguardanti la quota di petrolio greggio convertito in benzina dopo la distillazione e frazionamento. In dettaglio, le variabili rilevate sono:

- `yield`: percentuale di petrolio greggio convertito in benzina dopo la distillazione e frazionamento;

- `gravity`: la gravità del petrolio greggio (°API);
- `pressure`: la pressione del petrolio greggio (lbs/in^2);
- `temp10`: la temperatura (°F) alla quale il 10% del petrolio greggio si vaporizza;
- `temp`: la temperatura (°F) alla quale tutto il petrolio greggio si vaporizza;
- `batch`: fattore a 10 livelli che indica un gruppo unico di condizioni `gravity`, `pressure` e `temp10`.

Si inizia caricando il dataset, e adattando un modello di regressione Beta in cui si presume che non vi sia eteroschedasticità.

```
> data("GasolineYield", package="betareg")
> m1 <- betareg(yield~batch+temp, data=GasolineYield)
> summary(m1)

Call:
betareg(formula = yield ~ batch + temp, data = GasolineYield)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-2.8750 -0.8149  0.1601  0.8384  2.0483

Coefficients (mean model with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.1595710  0.1823247 -33.784  < 2e-16 ***
```

```
batch1       1.7277289  0.1012294  17.067  < 2e-16 ***
batch2       1.3225969  0.1179021  11.218  < 2e-16 ***
batch3       1.5723099  0.1161045  13.542  < 2e-16 ***
batch4       1.0597141  0.1023598  10.353  < 2e-16 ***
batch5       1.1337518  0.1035232  10.952  < 2e-16 ***
batch6       1.0401618  0.1060365   9.809  < 2e-16 ***
batch7       0.5436922  0.1091275   4.982 6.29e-07 ***
batch8       0.4959007  0.1089257   4.553 5.30e-06 ***
batch9       0.3857929  0.1185933   3.253  0.00114 **
temp         0.0109669  0.0004126  26.577  < 2e-16 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)    440.3      110.0   4.002 6.29e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood:  84.8 on 12 Df
Pseudo R-squared: 0.9617
Number of iterations in BFGS optimization: 51
```

Nell'output ci sono alcune statistiche di sintesi dei residui pesati standard, le stime di massima verosimiglianza del vettore β e del parametro ϕ (con relativi standard error) e i valori della statistica test di Wald per testare la significatività di ogni coefficiente, con relativo di p-value. Sono quindi riportati il valore della log-verosimiglianza massimizzata con i relativi gradi di libertà $(k + h)$, il valore del-

l'indice pseudo-R^2 e il numero di iterazioni che sono state necessarie per stimare i parametri con il metodo BFGS.

Si nota che il modello ha tutti i coefficienti significativi e un indice pseudo-R^2 elevato (96,17%).

Per una verifica della bontà di adattamento del modello, si esegue una analisi grafica dei residui (Figura 2.1):

```
> par(mfrow=c(3,2))
> plot(m1, which=c(1,4))
> plot(m1,which=c(1,4),type="pearson")
> plot(m1,which=c(1,4),type="deviance")
```

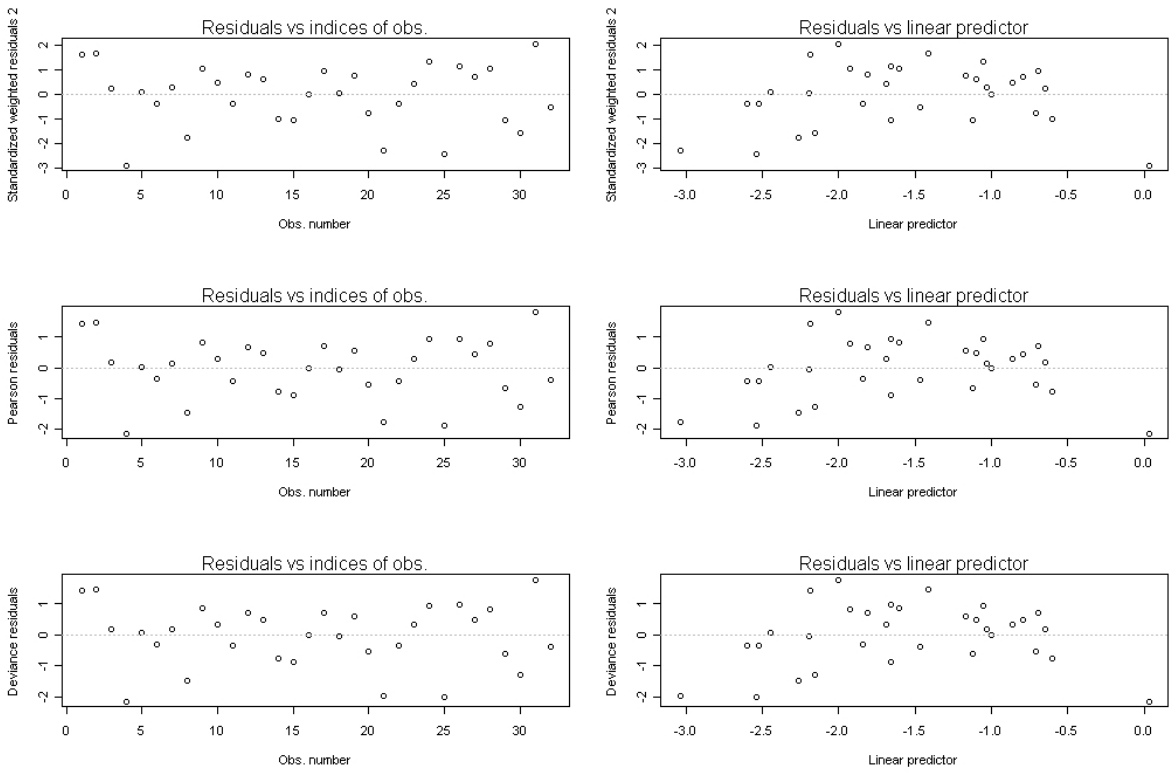

Figura 2.1: Grafici dei residui per il modello `m1`.

È possibile anche considerare una rappresentazione delle distanze di Cook (Figura 2.2):

```
> plot(m1, which=2, sub.caption="")
```

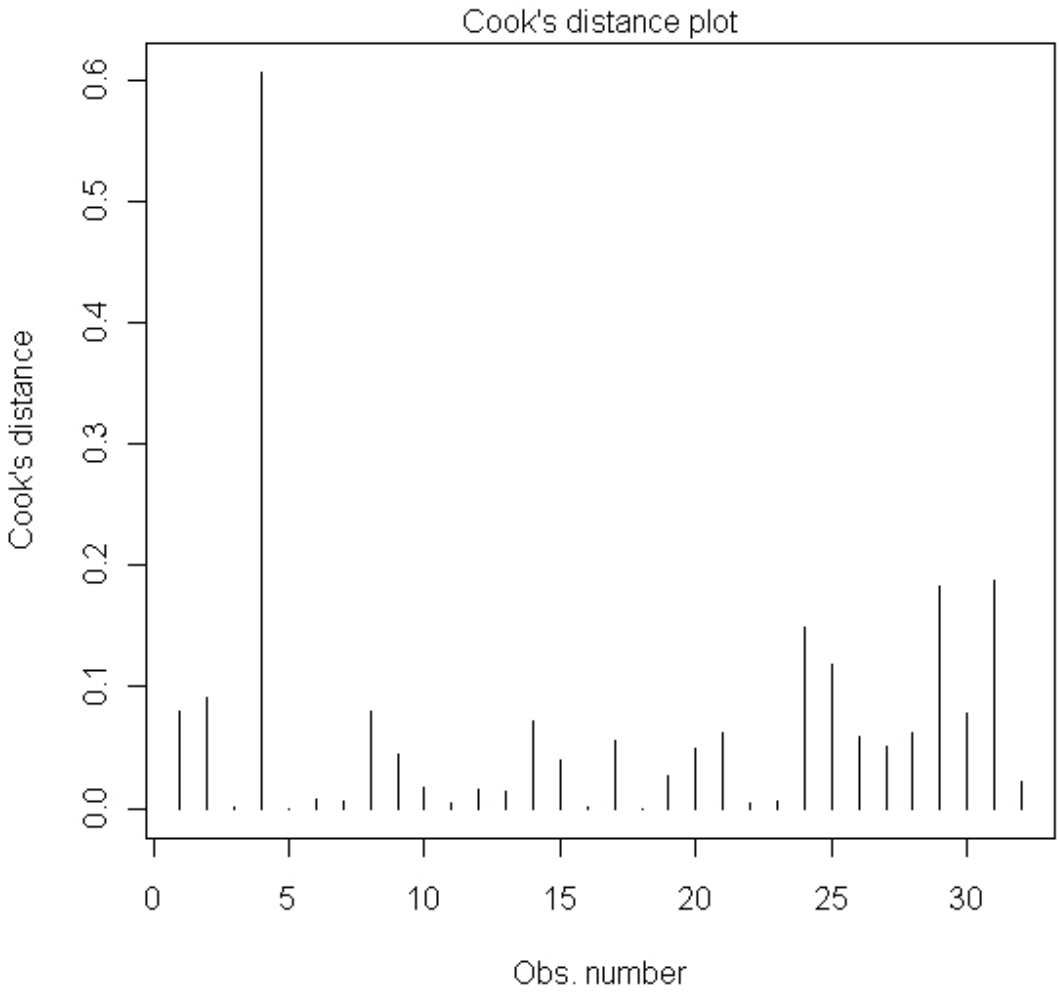

Figura 2.2: Distanze di Cook per il modello `m1`.

Osservando la Figura 2.2 si nota che la quarta osservazione ha una distanza di Cook elevata. È possibile ristimare il modello senza di essa, con il comando update:

```
> m1.4 <- update(m1, subset = -4)
```

Mettendo a confronto i parametri di precisione dei due modelli si nota che la precisione del secondo è aumentata:

```
> coef(m1, model="precision")
   (phi)
440.2783

> coef(m1.4, model="precision")
   (phi)
577.7907
```

L'eliminazione della quarta osservazione porta quindi a una diminuzione della varianza.

Aggiungiamo ora al modello `m1` il regressore `temp` per il parametro di precisione, utilizzando la funzione legame $g_2(\cdot)$ di tipo logaritmico:

```
> m2 <- betareg(yield~batch+temp | temp, data=GasolineYield)
> summary(m2)

Call:
betareg(formula = yield ~ batch + temp | temp,
                  data = GasolineYield)
```

```
Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-2.5399 -0.7792 -0.1167  0.8621  2.9419

Coefficients (mean model with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.9232377  0.1835262 -32.275  < 2e-16 ***
batch1       1.6019882  0.0638563  25.087  < 2e-16 ***
batch2       1.2972664  0.0991001  13.090  < 2e-16 ***
batch3       1.5653384  0.0997392  15.694  < 2e-16 ***
batch4       1.0300721  0.0632884  16.276  < 2e-16 ***
batch5       1.1541631  0.0656428  17.582  < 2e-16 ***
batch6       1.0194450  0.0663511  15.364  < 2e-16 ***
batch7       0.6222589  0.0656326   9.481  < 2e-16 ***
batch8       0.5645827  0.0601848   9.381  < 2e-16 ***
batch9       0.3594386  0.0671407   5.354 8.63e-08 ***
temp         0.0103595  0.0004362  23.751  < 2e-16 ***

Phi coefficients (precision model with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.364110   1.225781   1.113    0.266
temp        0.014570   0.003618   4.027 5.65e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: 86.98 on 13 Df
Pseudo R-squared: 0.9519
Number of iterations in BFGS optimization: 33
```

Il modello `m2` può essere confrontato col modello `m1` utilizzando il test log-rapporto di verosimiglianza e il test alla Wald, usando le opportune funzioni della libreria `lmtest`:

```
> library(lmtest)

> lrtest(m1,m2)
Likelihood ratio test

Model 1: yield ~ batch + temp
Model 2: yield ~ batch + temp | temp
  #Df LogLik Df Chisq Pr(>Chisq)
1  12 84.798
2  13 86.977  1 4.359    0.03681 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> waldtest(m1,m2)
Wald test

Model 1: yield ~ batch + temp
Model 2: yield ~ batch + temp | temp
  Res.Df Df  Chisq Pr(>Chisq)
1     20
2     19  1 632.92  < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Entrambi i test sono a favore del modello `m2` (p-value$<$ 0.05). Si noti che questi test possono essere interpretati come test per testare l'ipotesi nulla di ϕ costante contro un'alternativa specifica di dispersione.

Si stima ora un modello con omoschedasticità, ma con un legame probit per la media:

```
> m1.probit <- betareg(yield~batch+temp, link="probit",
                       data=GasolineYield)
```

e si confronta il suo indice pseudo-R^2 con quello del modello `m1`.

```
> summary(m1)$pseudo.r.squared
[1] 0.9617312


> summary(m1.probit)$pseudo.r.squared
[1] 0.9754757
```

Si nota che il modello `m1` e il modello `m1.probit` sono molto simili in termini di pseudo-R^2. I due modelli possono essere confrontati anche tramite l'indice di Akaike:

```
> AIC(m1,m1.probit)
           df       AIC
m1         12 -145.5951
m1.probit 12 -155.6575
```

Anche utilizzando l'indice AIC non si nota una grande differenza tra `m1` e `m1.probit`. D'altro, come nei modelli lineari generalizzati, la scelta della funzione logit o probit come funzione legame non influenza molto le conlusioni inferenziali.

2.4 Conclusioni

In questo capitolo si è brevemente introdotta l'estensione al caso di eteroschedasticità del modello di regressione Beta, illustrando il pacchetto R `betareg` necessario a stimare questo tipo di modelli, anche attraverso un esempio illustrativo. Nel capitolo seguente si utilizzerà quanto esposto fin ora per affrontare un'applicazione a dati reali.

Capitolo 3

Applicazione a dati reali

In questo capitolo viene discussa un'applicazione del modello di regressione Beta ad un dataset contenente dati relativi ad ascolti televisivi. Le analisi che sono presentate nel seguito sono state condotte utilizzando il software statistico R e la libreria `betareg` presentata nel capitolo precedente.

3.1 I dati

Il dataset (si veda ad esempio Giudici, 2005, Capitolo 12) contiene osservazioni relative a variabili associate a dati sugli ascolti televisivi rilevati dalla società Auditel. Auditel individua un campione stratificato di famiglie di telespettatori rappresentante le diverse caratteristiche geografiche, demografiche e socioculturali dell'intera nazione, costituita da individui con età maggiore di 4 anni secondo fonti ISTAT. I soggetti inseriti in tale campione, grazie a interviste personali e secondo criteri di casualità e di stratificazione, ruotano del 20% ogni anno. Al termine del campionamento, su ogni televisore presente nelle abitazioni degli

individui prescelti viene applicato un *meter*, ovvero un apparecchio predisposto per raccogliere automaticamente, ogni giorno, minuto per minuto, tutti i canali visualizzati da ogni televisore in funzione.

Il dataset in questione fa riferimento alla fascia *prime time*, ovvero alla fascia oraria compresa tra le 20.30 e le 22.30, e alle principali reti televisive nazionali (Rai1, Rai2, Rai3, Rete4, Canale5, Italia1). Le $n = 366$ osservazioni sono state rilevate, giorno per giorno, per il periodo tra il 29/11/1995 e il 28/11/1996. In dettaglio, le variabili rilevate sono:

- `data`: data in cui è stata effettuata la rilevazione;
- `giorno`: fattore a 7 livelli che indica il giorno della settimana (rappesentato con un numero tra 1 e 7 per i giorni da lunedì a domenica);
- `mese`: fattore a 12 livelli che indica il mese dell'anno (rappesentato con un numero tra 1 e 12 per i mesi tra Gennaio e Dicembre);
- `totale`: totale ascolto[1];
- `p.rai1`, `p.rai2`, `p.rai3`, `p.rete4`, `p.can5`, `p.ita1`: *share*[2], con dominio (0, 100), relativo rispettivamente alle reti televisive Rai1, Rai2, Rai3, Rete4, Canale5, Italia1;
- `g.rai1`, `g.rai2`, `g.rai3`, `g.rete4`, `g.can5`, `g.ita1`: fattore a 5 livelli che indica il tipo di programma (cultura, film,

[1]Somma dell'ascolto medio di tutte le emittenti rilevate, dove con ascolto medio si intende una media calcolata mediante il rapporto tra la somma di tutti i telespettatori presenti in ciascun minuto, in un certo intervallo di tempo, e il numero di minuti dell'intervallo stesso.

[2]Rapporto percentuale tra l'ascolto medio di una emittente, in un intervallo di tempo stabilito, e il totale ascolto nello stesso intervallo.

sport, telefilm, varietà) trasmesso rispettivamente dalle reti televisive Rai1, Rai2, Rai3, Rete4, Canale5, Italia1.

Lo scopo di questo capitolo è modellare lo *share* relativo a Rai1 in funzione delle altre variabili, utilizzando la libreria `betareg`. Poiché lo *share* varia tra 0 e 100, si assegna alla variabile `p.rai1` il valore `p.rai1/100`, e si mantiene questa trasformazione per tutto il seguito del presente capitolo.

3.2 Analisi esplorative

L'analisi esplorativa che segue si prefigge di dare una prima idea delle informazioni presenti nei dati, concentrandosi sulle informazioni inerenti alla rete Rai1. Si inizia osservando la distribuzione della variabile `p.rai1` (Figura 3.1), le cui statistiche di sintesi sono di seguito riportate:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0920  0.1930  0.2170  0.2379  0.2600  0.7250
```

Si nota che lo *share* medio di Rai1 è approssimativamente del 23.8%. Inoltre, si nota che la distibuzione presenta una asimmetria positiva, dovuta alla presenza di un buon numero di picchi di ascolto. Una conferma è data dall'indice di asimmetria di Pearson che, calcolato su `p.rai1`, assume approssimativamente valore 2.01. Si nota anche la presenza di un picco di ascolto negativo dovuto ad un film, trasmesso il mercoledì 03/12/1996, che ha totalizzato un *share* del 9.2% su un totale ascolto di 30597.

Si analizza ora `g.rai1`, ovvero il tipo di programma trasmesso dalla rete (Figura 3.2, grafico a sinistra), e si nota che i più frequenti

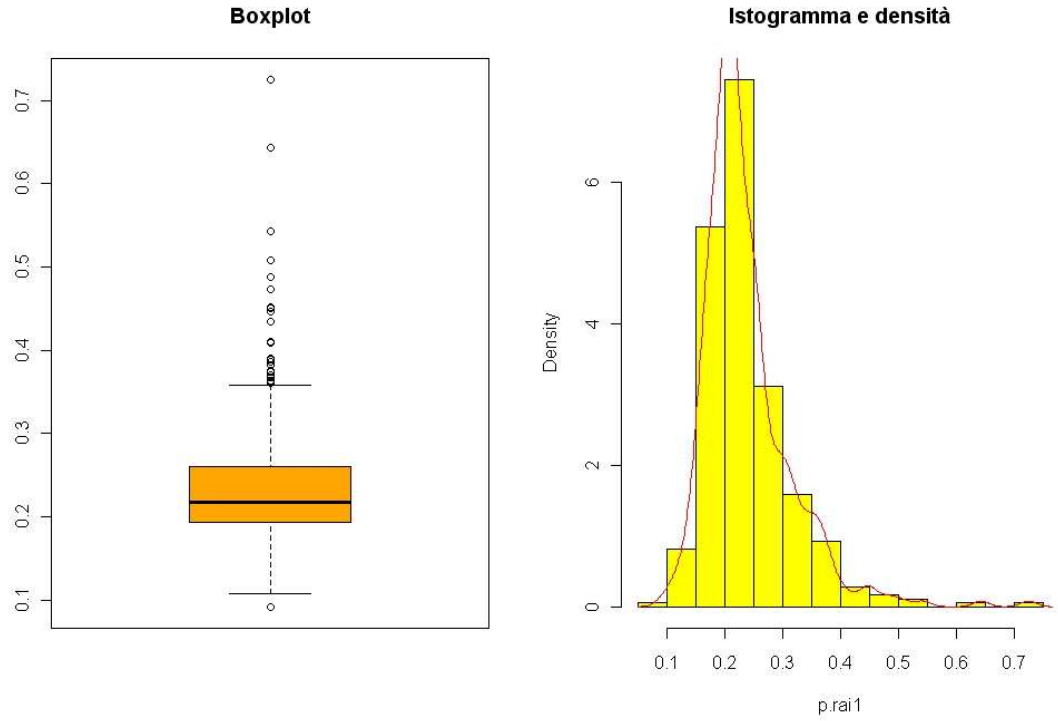

Figura 3.1: Boxplot (sinistra) e istogramma (destra) della variabile `p.rai1`.

sono i varietà (35%), seguiti dai film (28%) e dai telefim (19%), ed infine i meno trasmessi sono i programmi di cultura e di sport (9%).

Considerando la distribuzione dello *share* condizionatamente al tipo di programma trasmesso (Figura 3.2, grafico a destra) si conclude che i valori più alti dello *share* di Rai1 si raggiungono con sport e varietà. Inoltre si nota che per i programmi sportivi e per i varietà vi è una maggiore variabilità. Si noti, infine, la presenza di due valori particolarmente elevati rilevati con la trasmissione di programmi sportivi.

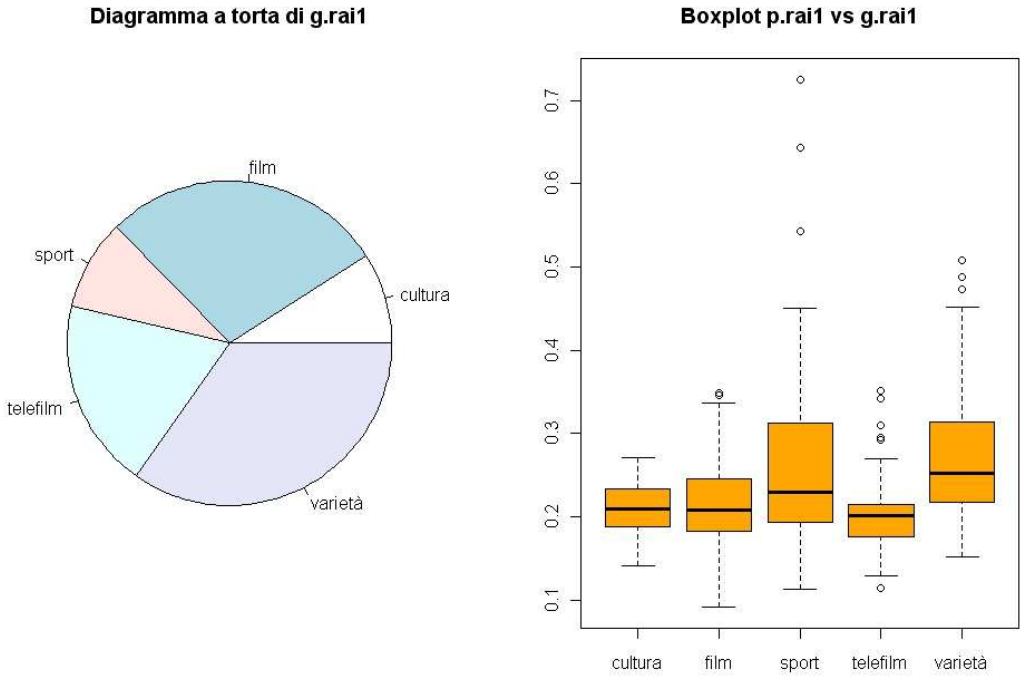

Figura 3.2: Diagramma a torta di `g.rai1` (sinistra) e boxplot di `p.rai1`(destra) *versus* `g.rai1`.

Osservando come varia lo *share* condizionatamente al giorno della settimana (Figura 3.3, sinistra) si nota che esso risulta maggiore la domenica. Per quanto riguarda la distribuzione condizionata al mese (Figura 3.3, grafico di destra) si nota una grande variabilità nel mese di Febbraio.

Risulta evidente la presenza di due osservazioni che presentano valori molto elevati per la variabile `p.rai1`, ossia

- l'osservazione numero 204 effettuata il giovedì 19/06/1996 relativa ad un programma sportivo, con uno *share* del 72.5% su un totale ascolto di 28800;

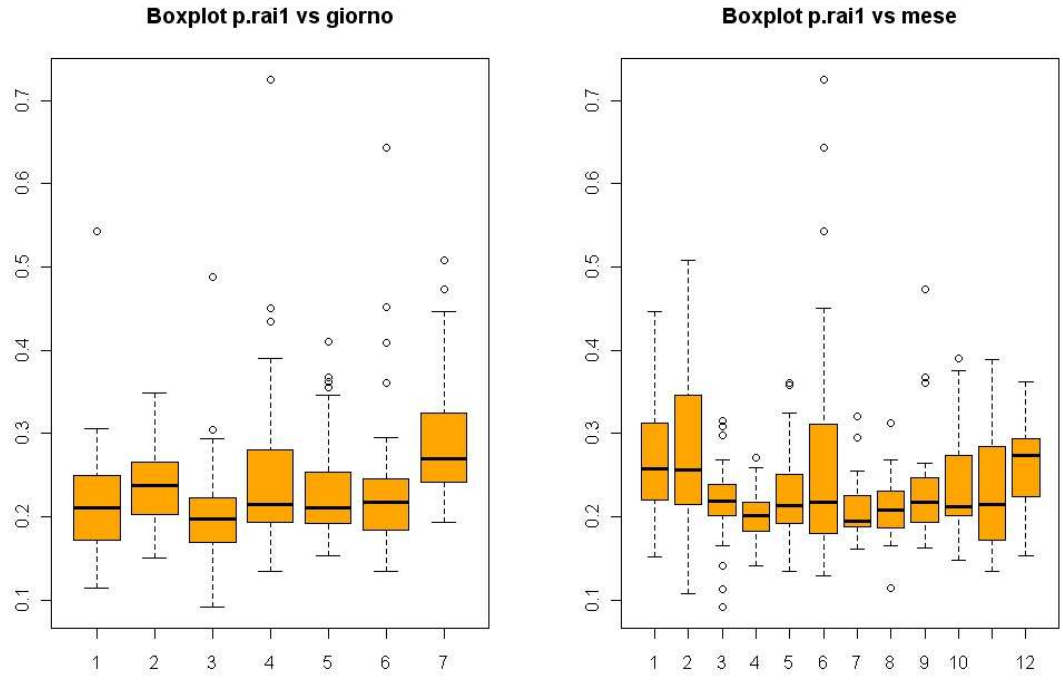

Figura 3.3: Boxplot di `p.rai1` *versus* `giorno` (sinistra) e `mese` (destra).

- l'osservazione numero 199 effettuata il sabato 14/06/1996 relativa ad un programma sportivo, con uno *share* del 64.3% su un totale ascolto di 26584.

Entrambe le osservazioni riguardano due importanti partite di calcio giocate dall'Italia. Nelle analisi successive, ritenendo tali osservazioni delle eccezioni, non le si prenderà in considerazione.

Si osserva ora la distribuzione della variabile `totale` (Figura 3.4), le cui statistiche di sintesi sono:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  11660   22140   25860   24590   28030   31200
```

Si nota la presenza di una asimmetria negativa, confermata da un indice di asimmetria di Pearson di circa -0.91.

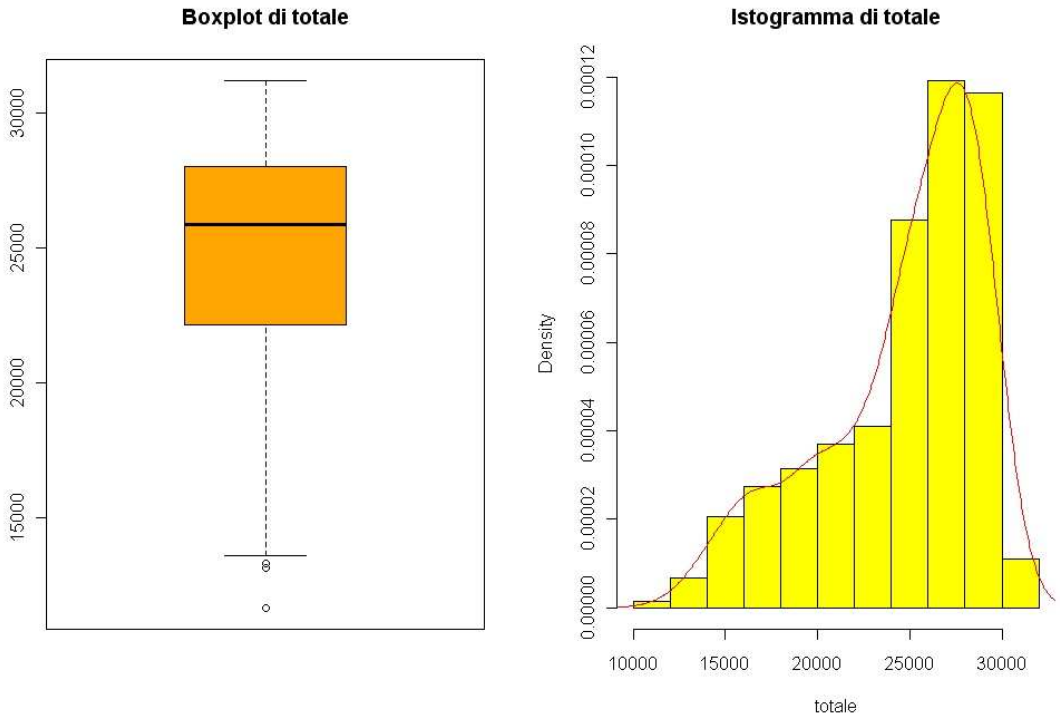

Figura 3.4: Boxplot (sinistra) e istogramma (destra) della variabile `totale`.

Nei mesi estivi (Giugno, Luglio e Agosto) il totale ascolto risulta più basso che negli altri periodi dell'anno (Figura 3.5, grafico di destra). Il totale ascolti inoltre risulta più basso il lunedì e la domenica (Figura 3.5, grafico di sinistra).

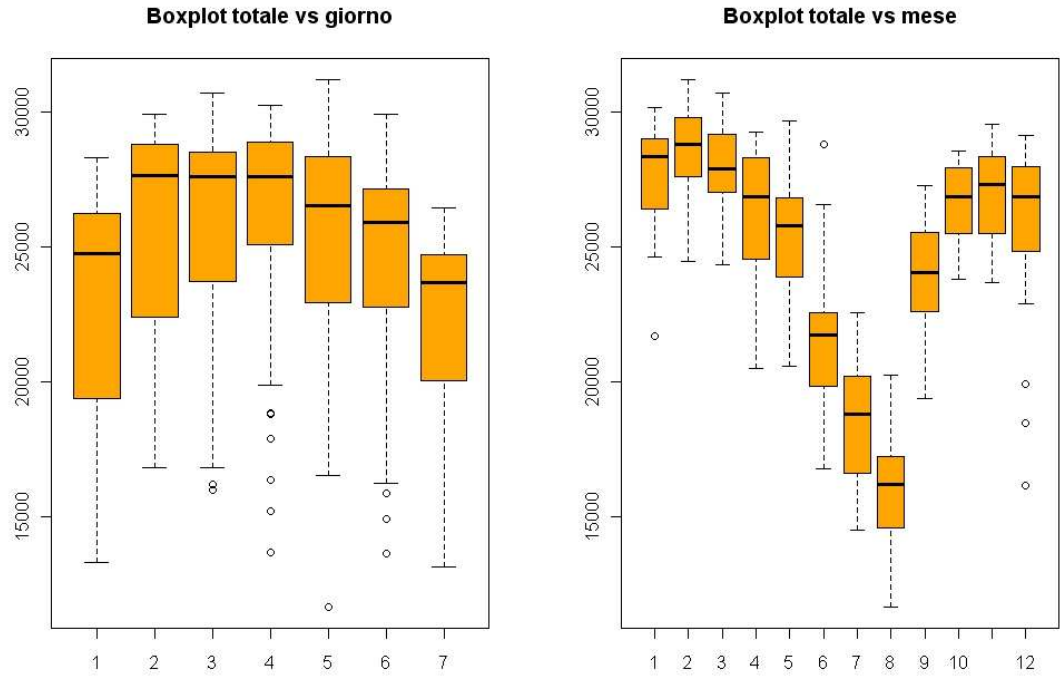

Figura 3.5: Boxplot di `totale` *versus* `giorno` (sinistra) e `mese` (destra).

3.3 Adattamento del modello

In questo paragrafo si presenta un'analisi con un modello di regressione Beta, considerando come dataset di riferimento quello privato delle osservazioni 204 e 199. Le variabili dello *share* e del tipo di programma trasmesso relative alle reti diverse da Rai1 non verranno prese in considerazione.

3.3.1 Modello con ϕ costante

Si inizia considerando un modello con parametro di precisione costante e un predittore lineare per la media che include l'intercetta ed il regressore `totale`. La funzione legame scelta è la funzione logit. La stima del modello fornisce:

```
Modello b1:
betareg(formula = p.rai1 ~ totale)
-----------------------------------------------------
Coefficients (mean model with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.368e+00  1.092e-01 -12.525   <2e-16 ***
totale       7.853e-06  4.365e-06   1.799    0.072 .
-----------------------------------------------------
Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)   41.536      3.057   13.59   <2e-16 ***
-----------------------------------------------------
Log-likelihood: 481.5 on 3 Df
Pseudo R-squared: 0.009807
-----------------------------------------------------
```

L'intercetta risulta significativa (p-value $< 2 \cdot 10^{-16}$), mentre il coefficiente angolare di `totale` risulta significativo al 10% (p-value $= 0.072$). Confrontando il modello `b1` con quello avente solo l'intercetta utilizzando il test log-rapporto di verosimiglianza (p-value $= 0.06867$) e il test alla Wald (p-value $= 0.07199$), si accetta il modello `b1` con una significatività del 10%. Si noti che il coefficiente stimato per `totale` è

maggiore di zero, questo significa che all'aumentare dell'ascolto totale, secondo il modello `b1`, corrisponde un aumento dello *share*. L'indice pseudo-R^2 del modello `b1` risulta essere molto basso ($R^2_p = 0.009807$), e pertanto il modello corrente non è adeguato. Si stima ora il modello con l'aggiunta del regressore `g.rai1`, ovvero tenendo conto del tipo di programma trasmesso.

```
Modello b2:
betareg(formula = p.rai1 ~ totale + g.rai1)
-----------------------------------------------------
Coefficients (mean model with logit link):
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.569e+00  1.161e-01 -13.509  < 2e-16 ***
totale          1.049e-05  4.026e-06   2.606  0.00916 **
g.rai1film      3.955e-02  6.757e-02   0.585  0.55837
g.rai1sport     2.119e-01  8.244e-02   2.571  0.01015 *
g.rai1telefilm -4.520e-02  7.194e-02  -0.628  0.52981
g.rai1varietà   3.131e-01  6.522e-02   4.800 1.59e-06 ***
-----------------------------------------------------
Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)   50.091      3.692   13.57   <2e-16 ***
-----------------------------------------------------
Log-likelihood: 515.3 on 7 Df
Pseudo R-squared: 0.1864
-----------------------------------------------------
```

Si ha un aumento dell'indice pseudo-R^2 ($R^2_p = 0.1864$). Si nota inoltre che i coefficienti di `g.rai1film` (p-value $= 0.55837$) e

`g.rai1telefilm` (p-value $= 0.52981$) non risultano significativi. Confrontando i modelli `b1` e `b2` con il test log-rapporto di verosimiglianza (p-value $= 7,33 \cdot 10^{-14}$) si accetta il modello `b2` e, quindi, `g.rai1` nel complesso risulta significativo. In termini di indice di Akaike per `b1` (AIC = -957.0225) e per `b2` (AIC = -1016.61), si è a favore del modello `b2`. Si continua aggiungendo a `b2` il regressore `giorno`.

```
Modello b3:
betareg(formula = p.rai1 ~ totale + g.rai1 + giorno)
-------------------------------------------------------
Coefficients (mean model with logit link):
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.624e+00  1.204e-01 -13.497  < 2e-16 ***
totale          1.626e-05  3.991e-06   4.074 4.62e-05 ***
g.rai1film     -1.318e-01  7.285e-02  -1.810  0.07037 .
g.rai1sport     1.691e-01  8.095e-02   2.089  0.03670 *
g.rai1telefilm -1.934e-01  7.895e-02  -2.449  0.01432 *
g.rai1varietà   2.092e-01  7.086e-02   2.953  0.00315 **
giorno2         1.819e-01  6.984e-02   2.604  0.00920 **
giorno3        -2.866e-01  6.742e-02  -4.252 2.12e-05 ***
giorno4         1.216e-01  6.134e-02   1.982  0.04746 *
giorno5        -3.940e-02  6.158e-02  -0.640  0.52227
giorno6         1.112e-02  6.859e-02   0.162  0.87125
giorno7         1.875e-01  6.330e-02   2.961  0.00306 **
-------------------------------------------------------
Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)   61.618      4.548   13.55   <2e-16 ***
```

```
-----------------------------------------------------
Log-likelihood: 552.6 on 13 Df
Pseudo R-squared: 0.3386
-----------------------------------------------------
```

Osservando i risultati prodotti dal modello `b3` si nota un miglioramento in termini di pseudo-R^2, che assume valore 33.86%. Si nota inoltre che i coefficienti di `giorno5` e di `giorno6` della variabile `giorno` non risultano significativi. Il test log-rapporto di verosimiglianza per confontare il modello `b3` con il modello `b2` è a favore del modello `b3` (p-value $= 4,642 \cdot 10^{-14}$), e quindi la variabile `giorno` nel complesso è significativa. Anche l'indice di Akaike risulta a favore del modello `b3` (AIC = -1079.209) se paragonato al modello `b2` (AIC = -1016.610). Si prova a migliorare il modello `b3` introducendo il regressore `mese`.

```
Modello b4:
betareg(formula = p.rai1 ~ totale + g.rai1 + giorno + mese)
-----------------------------------------------------
Coefficients (mean model with logit link):
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.378e+00  2.883e-01  -4.780 1.75e-06 ***
totale          1.140e-05  1.063e-05   1.072 0.283631
g.rai1film     -1.929e-01  7.333e-02  -2.630 0.008533 **
g.rai1sport     1.160e-01  8.145e-02   1.425 0.154238
g.rai1telefilm -2.450e-01  8.032e-02  -3.051 0.002284 **
g.rai1varietà   1.253e-01  7.261e-02   1.726 0.084349 .
giorno2         1.961e-01  7.551e-02   2.597 0.009408 **
giorno3        -2.769e-01  7.286e-02  -3.801 0.000144 ***
giorno4         1.359e-01  6.678e-02   2.036 0.041777 *
```

```
giorno5        -1.675e-02  6.522e-02  -0.257 0.797296
giorno6        -5.796e-03  6.871e-02  -0.084 0.932784
giorno7         2.044e-01  6.276e-02   3.256 0.001130 **
mese2           2.341e-02  7.168e-02   0.327 0.744012
mese3          -1.733e-01  7.572e-02  -2.289 0.022105 *
mese4          -2.197e-01  7.661e-02  -2.868 0.004136 **
mese5          -1.145e-01  7.677e-02  -1.492 0.135716
mese6          -2.340e-02  9.992e-02  -0.234 0.814810
mese7          -1.443e-01  1.219e-01  -1.183 0.236693
mese8          -1.064e-01  1.409e-01  -0.755 0.450358
mese9          -1.271e-01  8.313e-02  -1.529 0.126290
mese10          1.279e-02  7.455e-02   0.172 0.863758
mese11         -6.661e-02  7.404e-02  -0.900 0.368329
mese12          2.273e-02  7.216e-02   0.315 0.752725
---------------------------------------------------
Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)    65.83       4.86   13.54   <2e-16 ***
---------------------------------------------------
Log-likelihood: 564.5 on 24 Df
Pseudo R-squared: 0.3791
---------------------------------------------------
```

Si nota un aumento dell'indice pseudo-R^2 a 37.91%. Si nota inoltre che solo due coefficienti relativi alle modalità della variabile `mese` sono significativi, e che il coefficiente di `totale` non è significativo (p-value = 0.283631). Il test log-rapporto di verosimiglianza per confontare il modello `b4` con il modello `b3` è a favore del modello più

complicato con significatività del 5% (p-value = 0.01356), ovvero la variabile `mese` nel complesso è significativa. Anche l'indice di Akaike risulta a favore del modello `b4` (AIC = -1081.019) se paragonato al modello `b3` (AIC = -1079.209). Si continua togliendo a `b4` il regressore `totale`, ottenendo:

```
Modello b5:
betareg(formula = p.rai1 ~ g.rai1 + giorno + mese)
-------------------------------------------------------
Coefficients (mean model with logit link):
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.084708   0.093618 -11.587  < 2e-16 ***
g.rai1film     -0.197729   0.073287  -2.698 0.006976 **
g.rai1sport     0.116607   0.081527   1.430 0.152635
g.rai1telefilm -0.242314   0.080391  -3.014 0.002577 **
g.rai1varietà   0.125753   0.072711   1.730 0.083719 .
giorno2         0.234772   0.065694   3.574 0.000352 ***
giorno3        -0.240085   0.064489  -3.723 0.000197 ***
giorno4         0.170913   0.058159   2.939 0.003296 **
giorno5         0.013058   0.058986   0.221 0.824797
giorno6         0.012460   0.066512   0.187 0.851396
giorno7         0.195214   0.062295   3.134 0.001726 **
mese2           0.032461   0.071184   0.456 0.648375
mese3          -0.167910   0.075640  -2.220 0.026429 *
mese4          -0.235864   0.075250  -3.134 0.001722 **
mese5          -0.139666   0.073044  -1.912 0.055866 .
mese6          -0.094874   0.075755  -1.252 0.210433
mese7          -0.249312   0.074152  -3.362 0.000773 ***
```

```
mese8          -0.235347   0.074298  -3.168 0.001537 **
mese9          -0.167967   0.073850  -2.274 0.022940 *
mese10          0.001025   0.073904   0.014 0.988932
mese11         -0.073461   0.073903  -0.994 0.320213
mese12          0.007401   0.070756   0.105 0.916697
---------------------------------------------------
Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)   65.618      4.845   13.54   <2e-16 ***
---------------------------------------------------
Log-likelihood: 563.9 on 23 Df
Pseudo R-squared: 0.3779
---------------------------------------------------
```

Vi è un trascurabile calo dell'indice pseudo-R^2 a 37.79%. L'indice di Akaike risulta essere leggermente minore del modello b4 (AIC = -1081.893). Il test log-rapporto di verosimiglianza per confontare il modello `b5` con il modello `b4` è a favore del modello più semplice (p-value = 0.2884). Il modello `b5` è quindi preferibile al modello `b4`.

Si effettua ora un'analisi dei residui basata sui residui pesati standard per il modello `b5`.

Osservando la Figura 3.6 si nota che nel grafico di sinistra non emergono andamenti sistematici o stagionali, mentre nel grafico di destra si nota che i residui tendono ad avere una variabilità maggiore all'aumentare del predittore lineare. Inoltre dalla Figura 3.7 si nota che i residui relativi a osservazioni effettuate al trasmettere di programmi sportivi e varietà hanno una variabilità diversa (maggiore) da quella dei residui relativi a osservazioni effettuate al trasmettere delle altre

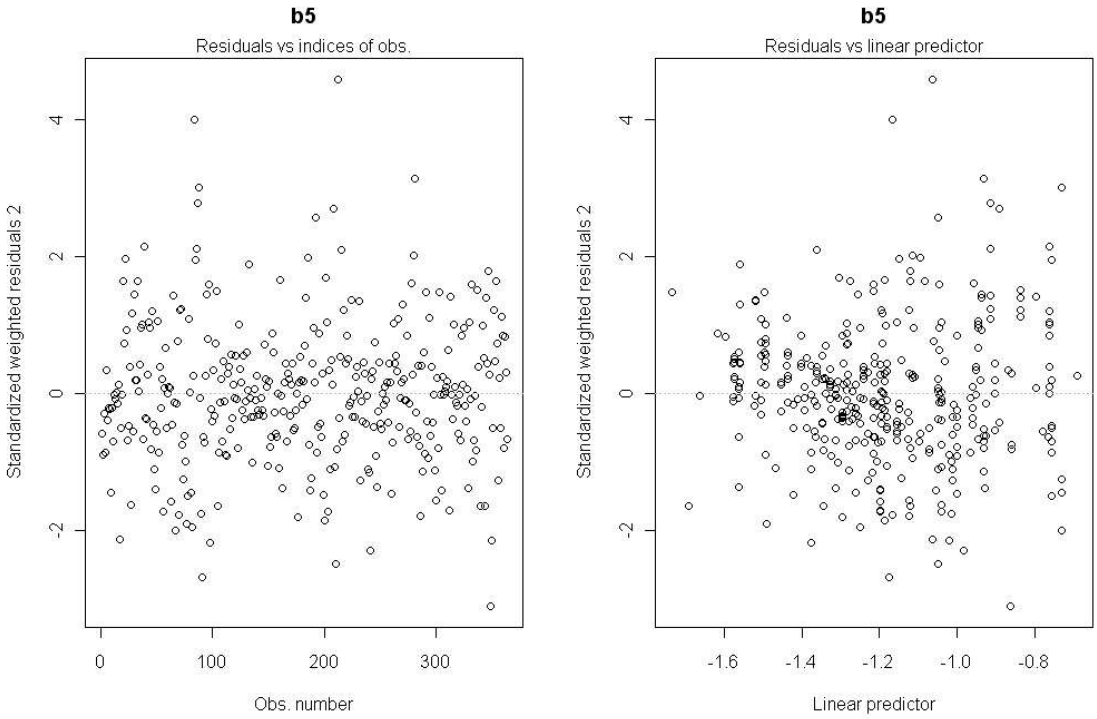

Figura 3.6: Grafici dei residui standard pesati del modello `b5` (sinistra) e *versus* η_i (destra).

tipologie di programmi. È quindi lecito presumere che vi sia un legame tra la variabile `g.rai1` e il parametro di precisione che non può essere colto dal modello `b5` a causa dell'assunto di ϕ costante.

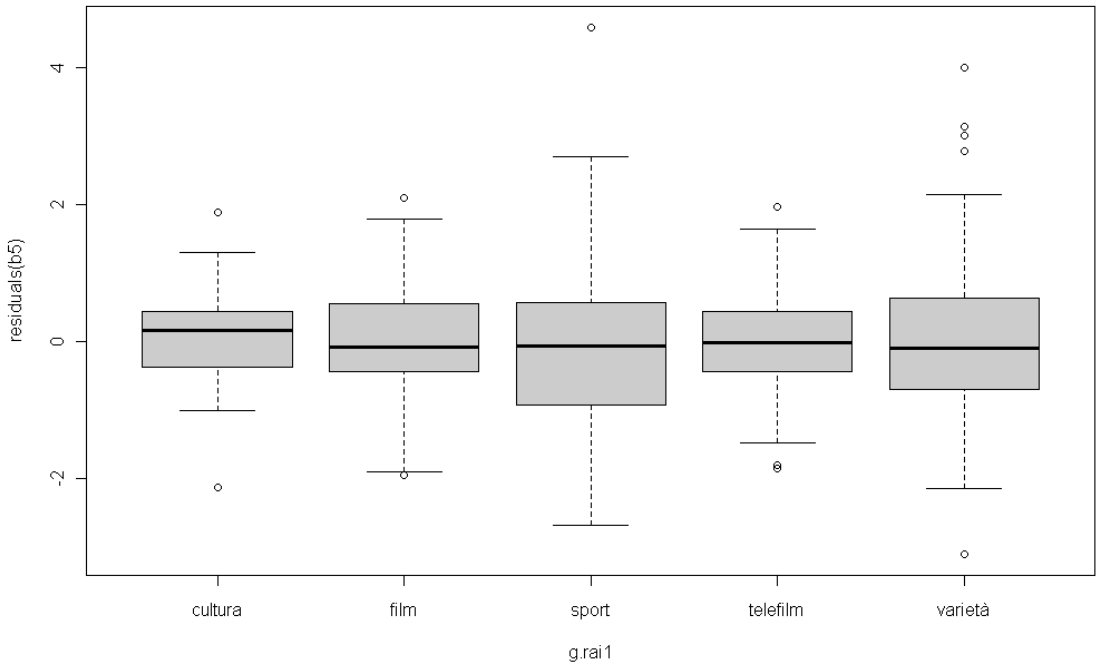

Figura 3.7: Boxplot dei residui standard pesati del modello `b5` *versus* `g.rai1`.

3.3.2 Modello con ϕ variabile

In virtù delle considerazioni fatte precedentemente nell'analisi dei residui del modello `b5` si adatta un modello con parametro di precisione variabile. Nel modello si mantiene lo stesso predittore lineare per la media

con funzione legame logit, mentre per il parametro di precisione si ha

$$\phi_i = \eta_{2i} = \zeta_1 + \zeta_2 z_i, \quad i = 1, ..., 364,$$

dove z_i è una variabile che assume valore 1 se l'osservazione i-esima è stata rilevata in corrispondenza di trasmissioni sportive o varietà di Rai1 e 0 altrimenti. Nell'output di R, la variabile z_i verrà indicata con `p.sv`. La stima del modello fornisce:

```
Modello z1:
betareg(formula = p.rai1 ~ g.rai1 + giorno + mese
                         | p.sv, link.phi = "identity")
-------------------------------------------------------
Coefficients (mean model with logit link):
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.112279   0.085025 -13.082  < 2e-16 ***
g.rai1film     -0.201419   0.060340  -3.338 0.000844 ***
g.rai1sport     0.138707   0.081894   1.694 0.090314 .
g.rai1telefilm -0.240427   0.067840  -3.544 0.000394 ***
g.rai1varietà   0.154961   0.066541   2.329 0.019868 *
giorno2         0.246445   0.056098   4.393 1.12e-05 ***
giorno3        -0.224842   0.063991  -3.514 0.000442 ***
giorno4         0.160601   0.050976   3.151 0.001630 **
giorno5         0.001071   0.056324   0.019 0.984831
giorno6        -0.016340   0.060628  -0.270 0.787536
giorno7         0.188627   0.067005   2.815 0.004876 **
mese2          -0.002833   0.071279  -0.040 0.968295
mese3          -0.108370   0.068935  -1.572 0.115933
mese4          -0.199030   0.069343  -2.870 0.004102 **
```

```
mese5          -0.130036   0.069029  -1.884 0.059594 .
mese6          -0.131414   0.074495  -1.764 0.077722 .
mese7          -0.175919   0.069275  -2.539 0.011103 *
mese8          -0.188220   0.069852  -2.695 0.007049 **
mese9          -0.138661   0.072994  -1.900 0.057482 .
mese10          0.019669   0.068480   0.287 0.773947
mese11         -0.054286   0.069723  -0.779 0.436219
mese12          0.014306   0.069328   0.206 0.836515
---------------------------------------------------
Phi coefficients (precision model with identity link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   110.92      10.92  10.153  < 2e-16 ***
p.sv          -68.87      11.89  -5.794 6.86e-09 ***
---------------------------------------------------
Log-likelihood: 582.7 on 24 Df
Pseudo R-squared: 0.3706
---------------------------------------------------
```

Si nota un'indice pseudo-R^2 del 37.06%, leggermente minore di quello del modello `b5`. In compenso si rivela un indice di Akaike di -1117.427, inferiore di quello del modello `b5` (AIC = -1081.893). Inoltre confrontando i modelli `b5` e `z1` con il test log-rapporto di verosimiglianza si predilige il modello `z1` (p-value $= 8.979 \cdot 10^{-10}$). Si nota inoltre che il coefficiente di `p.sv` risulta significativo (p-value $= 6.86 \cdot 10^{-9}$). Un'osservazione importante da fare riguardo il coefficiente di `p.sv` riguarda il segno, che risulta essere negativo. Questo significa che secondo il modello `z1` al trasmettere su Rai1 di trasmissioni sportive e varietà corrisponde un calo del parametro di precisione.

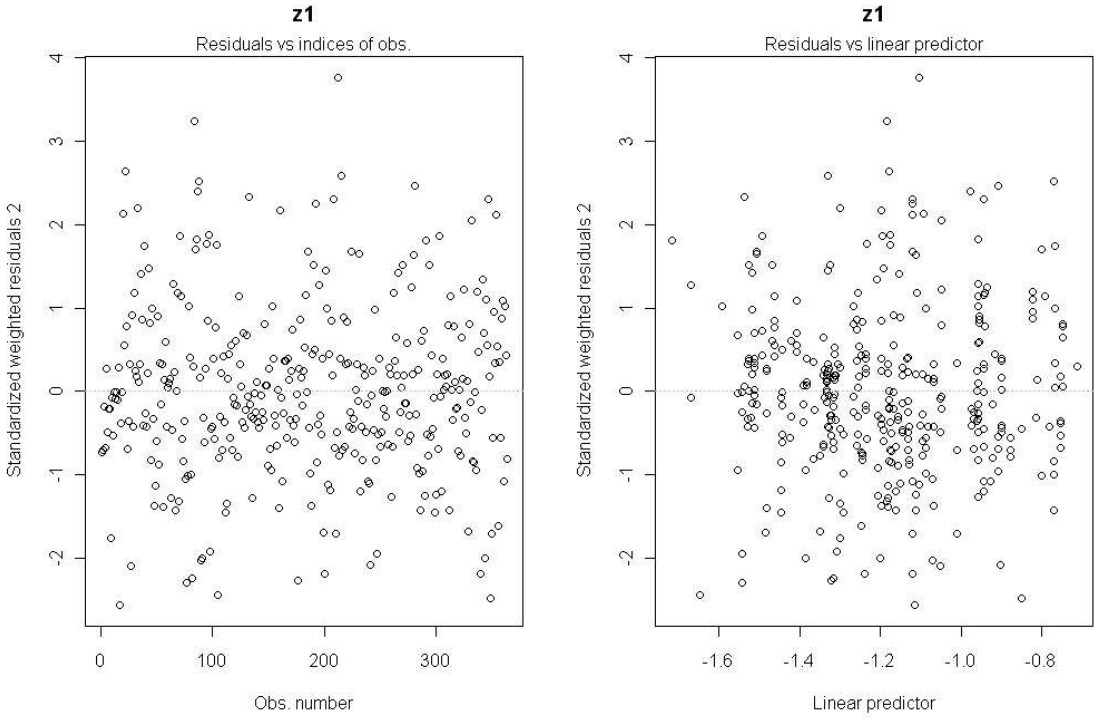

Figura 3.8: Grafici dei residui standard pesati del modello `z1` (sinistra) e *versus* η_{1i} (destra).

Si osservino ora nella Figura 3.8 i grafici diagnostici prodotti coi residui standard pesati. Si nota che l'assenza di andamenti sistematici in entrambi i grafici, e in particolare all'aumentare del predittore lineare (grafico di destra) non si rileva più un aumento della variabilità dei residui.

Sulla base di quanto esposto si conferma `z1` come modello definitivo, nonostante presenti un'indice pseudo-R^2 non molto elevato.

3.4 Conclusoni

In questo capitolo si è affrontato il problema di adattare il modello di regressione Beta ad un dataset contenente dati di ascolti televisivi per prevedere lo *share* di Rai1. Si è presentata la natura dei dati e si è effettuata un'analisi esplorativa, che ha portato all'eliminazione di due osservazioni anomale dovute a due importanti partite di calcio giocate dall'Italia. In seguito, si è adattato prima un modello con ϕ costante, e poi con ϕ variabile, preferendo infine quest'ultimo. Dall'analisi risulta che il totale ascolto non influenza lo *share* e che la trasmissione sulla rete di varietà e trasmissioni sportive comporta una maggiore variabilità delle percentuali di *share*.

Bibliografia

[1] AZZALINI, A. (2008). *Inferenza Statistica - Una presentazione basata sul concetto di verosimiglianza*. Seconda Edizione. Springer.

[2] CRIBARI-NETO, F. e ZEILEIS A. (2010). Beta Regression in R. *Journal of Statistical Software*. **34**(2).

[3] FERRARI SPL. e CRIBARI-NETO F. (2004). Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistic*. **31**(7). 799-815.

[4] GIUDICI P. (2005). *Data Mining - Metodi informatici, statistici e applicazioni*. Seconda Edizione. McGraw-Hill.

[5] ROSS SM. (2007). *Calcolo delle probabilità*. Seconda Edizione. Apogeo.

[6] PACE, L. e SALVAN A. (2001). *Introduzione alla Statistica II - Inferenza, verosimiglianza, modelli*. Cedem.