UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI SCIENZE STATISTICHE

CORSO DI LAUREA SPECIALISTICA IN
STATISTICA E INFORMATICA

MODELLI LINEARI GERARCHICI APPLICATI A DATI LONGITUDINALI DI TELEFONIA MOBILE

RELATORE: prof. Bruno Scarpa

LAUREANDO: Giulio Cordioli

INDICE

INTRODUZIONE

I telefoni cellulari hanno avuto, a partire dagli anni Novanta, un'estrema diffusione. Il loro numero è aumentato al punto tale che in Europa, alla fine dell'anno 2007, c'erano più apparecchi mobili che abitanti. Nel 2009 il 61% della popolazione mondiale ne possedeva almeno uno. Di conseguenza, la priorità delle compagnie di telefonia mobile si è concentrata sulla concorrenza nel mercato e sulla ricerca dei migliori prodotti. Tali aziende, alla continua ricerca di nuovi utenti e di strategie per evitare che i clienti lascino la compagnia per altre concorrenti, registrano le informazioni relative ad essi e al traffico telefonico effettuato da ogni singola Sim. Si trovano quindi a dover elaborare grandi quantità di dati, allo scopo di trarre maggiore profitto e di guadagnare competitività nel settore. L'obiettivo della ricerca è lo sviluppo di modelli statistici che prevedano, sulla base delle informazioni disponibili, la quantità di traffico telefonico in uscita di ogni singola Sim. Per fare ciò si utilizzano diversi modelli statistici e si sceglie infine quello che meglio approssima i dati reali. La variabile di interesse è il numero di chiamate effettuate da ciascuna Sim nell'arco di diciotto mesi. Si deve tener conto del fatto che le singole osservazioni disponibili fanno riferimento ad una particolare scheda telefonica. Si va quindi a verificare se esiste un "fattore Sim", cioè se esistono differenze di traffico tra le singole schede telefoniche o se, al contrario, i valori analizzati possono essere considerati come appartenenti ad un unico gruppo, ignorando il fatto che appartengano a Sim diverse. Si deve tener conto anche che alcune persone utilizzano più schede telefoniche contemporaneamente. In questa situazione l'utente potrebbe preferire l'utilizzo di una rispetto alle altre o, al contrario, effettuare chiamate usandole indifferentemente. In questa indagine si cercherà di esaminare questo fenomeno chiamato "fattore cliente" per provare l'esistenza o meno di variazioni di traffico nelle schede appartenenti alla medesima persona. Da ultimo si cercherà di individuare quali informazioni relative ai singoli clienti possano influenzare la variabile risposta e quali invece risultano non essere significative a tale scopo.

1 PRESENTAZIONE DEI DATI

1.1 Premesse

I dati presi in considerazione in questo studio si riferiscono a una delle aziende leader del settore e fanno riferimento ad alcune caratteristiche di circa 30000 schede telefoniche. Si è analizzato il traffico telefonico in uscita nell'arco temporale di diciotto mesi, tra novembre 2004 ed aprile 2006, tenendo in considerazione anche il fatto che l'utente abbia deciso o meno di continuare ad usare lo stesso gestore di telefonia mobile.

Le variabili disponibili per ciascuna Sim sono di seguito elencate:
Anno di nascita dell'utente: non registrata per circa il 3% degli utenti.
Data di attivazione della SIM: schede attivate tutte nel periodo dal 14/03/2003 al 30/09/2003.
Data di inizio stato della SIM: per il 90% delle Sim la data è quella di attivazione, per il restante 10% si riferisce ad un periodo successivo, compreso dal 14/03/2003 al 01/05/2006.
Marca del telefonino: disponibile solo per il 2,5% delle registrazioni.
Piano tariffario: fattore a 8 livelli.
Provincia di registrazione della SIM: fattore con 102 livelli che corrispondono ad altrettante città italiane, non disponibile per il 2% degli utenti.
Sesso dell'utente: fattore a due livelli: maschi 75%, femmine 22%, non disponibile 3% circa.
Stato della SIM: fattore a 3 livelli: attiva (94%), disattiva (6%) e sospesa (solo 12 Sim). Si riferisce al termine dei diciotto mesi di studio.
Causale stato: specifica per quale motivo la Sim è stata disattivata. Fattore a nove livelli. L'informazione è disponibile e utile solo per le Sim (circa il 6% del totale) non più attive al termine dei diciotto mesi di studio.

Per ogni mese da Novembre 2004 ad aprile 2006 sono state registrate le seguenti informazioni:
Numero di chiamate effettuate verso numeri fissi.
Numero di chiamate effettuate verso cellulari dello stesso operatore.
Numero di chiamate effettuate verso cellulari di altri operatori.
Numero di SMS inviati.
Numero di MMS inviati.

La ricerca viene effettuata sulle sole Sim che al termine dei diciotto mesi sono ancora attive, perché si ritiene che le schede disattivate siano caratterizzate da un diverso comportamento, e che quindi possano deviare l'obiettivo dell'analisi. Il grafico 1.1 confronta il numero medio di chiamate effettuate delle Sim attive con quelle che invece verranno disattivate. Si nota una sostanziale differenza tra i due andamenti ad indicare che, in media, le Sim disattivate hanno effettuato un numero nettamente inferiore di chiamate rispetto a quelle attive.

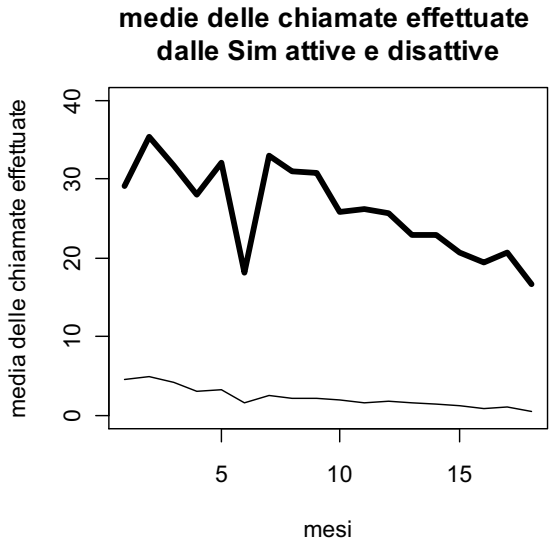

Figura 1.1 Il grafico mostra la media delle chiamate effettuate dalle Sim attive (in grassetto) rispetto alle Sim disattivate al termine dei 18 mesi

1.2 Modifica del dataset

Prima di iniziare l'analisi vera e propria è opportuno compiere alcune modifiche al dataset per poter gestire al meglio le informazioni in nostro possesso.

A tal fine sono state create le seguenti nuove variabili:

Età: alla fine del periodo esaminato (aprile 2006). Sono state create 6 classi di età: dai 18 ai 30 anni, dai 30 ai 40, dai 40 ai 50, dai 50 ai 60, dai 60 ai 70 e over 70. La variabile è un fattore che assume valori da 1 a 6.

Zona: la variabile **Provincia** è stata suddivisa in zone di appartenenza: Nord, Centro, Sud e Isole. In questo modo abbiamo ridotto i livelli del fattore da 102 a 4 cercando di alleggerire l'utilizzo di questa informazione.

ID: ad ogni singola scheda verrà assegnato un numero intero di riconoscimento.

CustomerID: ogni utente può essere in possesso di più Sim, questa variabile identifica i singoli clienti. L'azienda è in possesso di questa informazione, che però non è presente nel dataset. Per ovviare a ciò con-

sideriamo appartenenti allo stesso cliente le schede registrate da utenti con la stessa data di nascita, la stessa provincia di appartenenza e lo stesso sesso (così da ricostruire l'informazione non disponibile).

Chiamate: è la somma delle chiamate effettuate da ciascuna Sim verso numeri fissi, cellulari dello stesso operatore o altri, per ogni singolo mese.

Mese: assume valori da 1 a 18 e indica, per ciascuna osservazione, il mese nel quale è stata conteggiato il numero di chiamate effettuato.

Sono state eliminate quelle Sim che hanno come data di attivazione un valore successivo al primo mese di registrazione del traffico, per le quali non si conosceva quindi il traffico dei mesi iniziali. Circa il 5% delle Sim presentano dei valori mancanti per sesso, data di nascita e provincia. Si è deciso di eliminare dal dataset tali elementi per i quali non era possibile calcolare la variabile **customerID**. Sono presenti nel dataset schede telefoniche che sono state utilizzate pochissimo e addirittura alcune che non sono mai state usata. Si è eliminato chi ha effettuato meno di 50 chiamate nei diciotto mesi e chi non ha mai chiamato per almeno sette mesi. Non abbiamo considerato queste Sim, così poco utilizzate, poiché l'analisi che stiamo delineando vuole essere uno strumento che possa essere consultato dai gestori ai fini di un incremento positivo del traffico in uscita. L'azienda secondo noi non può concentrarsi su utenti con parametri di chiamate che sfiorano lo zero, ma dovrebbe tentare di trovare margini di miglioramento in quelli che già contano un sufficiente uso dei loro prodotti. Sono così rimaste 21758 registrazioni. Dal dataset possiamo ora togliere le variabili riferite alla data di nascita, alla data di attivazione e di inizio stato. I dati sono di seguito registrati:

customer							chiamate			SMS			MMS		
ID	**ID**	**età**	**piano**	**prov**	**sesso**	**zona**	**nov.04**	**…**	**apr.06**	**nov.04**	**…**	**apr.06**	**nov.04**	**…**	**apr.06**
1	1	6	E	PO	F	Centro	9	…	0	7	…	0	0	…	0
2	2	6	A	SR	M	Isole	216	…	182	19	…	40	0	…	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
1006	903	3	A	PD	M	Nord	49	…	6	3	…	5	0	…	0
1007	903	3	A	PD	M	Nord	12	…	16	0	…	0	0	…	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
21758	19373	1	A	MI	M	Nord	12	…	0	0	…	0	0	…	0

Tabella 1.1 La tabella mostra i dati registrati in un unico file composto da 21758 righe corrispondenti alle Sim osservate e 61 colonne corrispondenti alle variabili disponibili

id	id1	età	piano	prov	sesso	zona
1	1	6	E	PO	F	Centro
2	2	6	A	SR	M	Isole
…	…	…	…	…	…	…
1006	903	3	A	PD	M	Nord
1007	903	3	A	PD	M	Nord
…	…	…	…	…	…	…
21758	19373	1	A	MI	M	Nord

id	mese	chiamate	SMS	MMS
1	1	9	7	0
1	…	…	…	…
1	18	0	0	0
…	…	…	…	…
21758	1	12	0	0
21758	…	…	…	…
21758	18	0	0	0

Tabella 1.2 (a) Variabili riferite alla Sim e all'utente. 5(b) Variabili riferiti al traffico telefonico.

La tabella 1.1 mostra il dataset composto da un unico data-frame di 21758 righe e 61 colonne. Gli stessi dati possono essere memorizzati entro due distinti dataset (tabella 1.2). Uno contiene le informazioni relative alla Sim e al suo possessore, l'altro menziona il traffico telefonico mensile.

Con i dati disponibili sono stati creati, tramite una selezione casuale, due gruppi distinti. Il primo verrà utilizzato per stimare i modelli e il secondo verrà impiegato per la verifica degli stessi. Il dataset di stima prevede 12243 registrazioni mentre quello di verifica 9515. Sono state tolte dal primo gruppo le osservazioni riferite ad aprile 2006 per poter poi confrontare le previsioni di quest'ultimo mese con i rispettivi valori reali.

1.3 Analisi preliminare

In questo paragrafo tramite alcuni grafici e altri semplici strumenti descrittivi si cerca di capire meglio la natura delle variabili a disposizione.

1.3.1 Variabile CHIAMATE

Chiamate: è la variabile risposta che cercheremo di stimare conoscendo tutte le altre informazioni.
Il grafico mostra un lieve andamento decrescente del numero di chiamate con il passare dei mesi, ed evidenzia un vistoso calo in aprile del 2005, probabilmente dovuto ad errori nella registrazione del traffico in quel periodo.

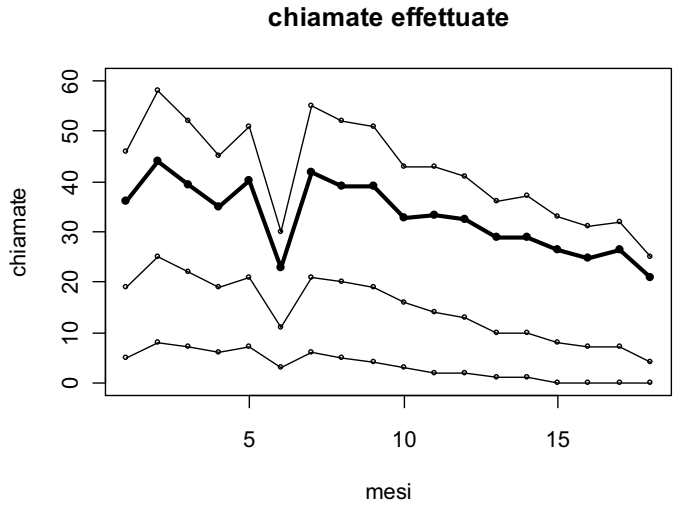

Figura 1.2 75% percentile, la media (in grassetto), la mediana e il 25% percentile della distribuzione delle chiamate nei diciotto mesi

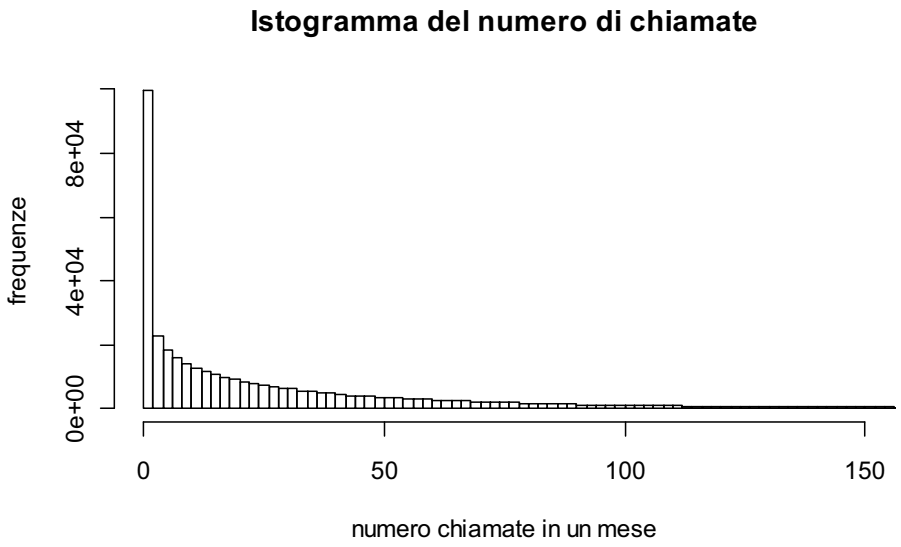

Figura 1.3 Istogramma del numero di chiamate nei vari mesi. Sono esclusi i valori maggiori di 150 (circa il 3%).

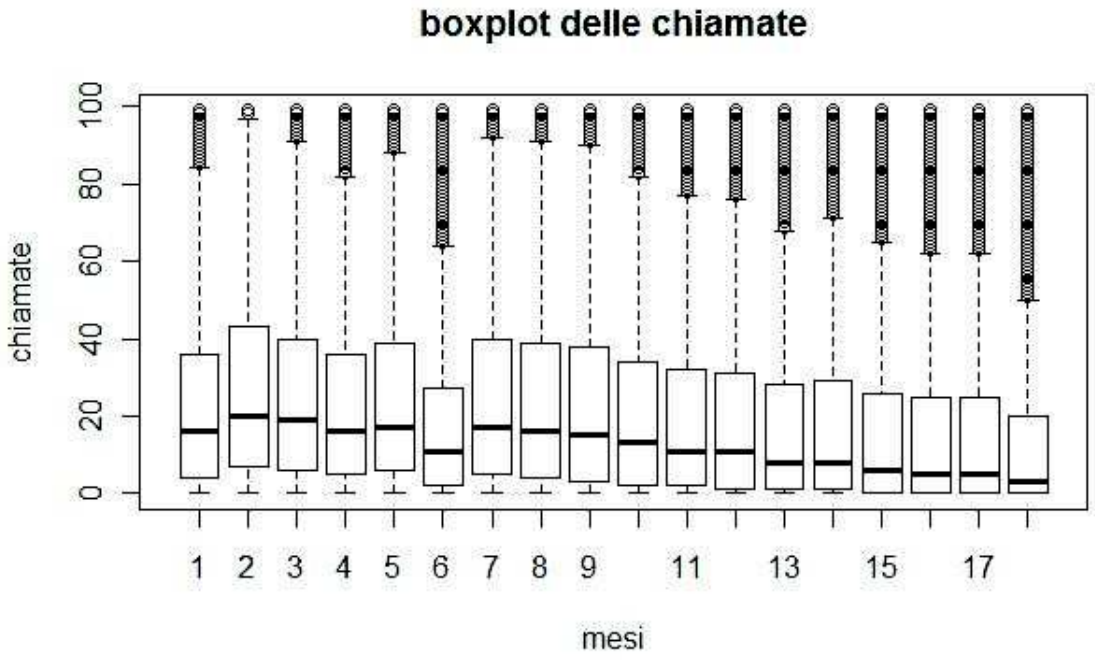

Figura 1.4 Box-plot delle chiamate nei diciotto mesi (sono esclusi i valori maggiori di 100)

Le figure 1.3 e 1.4 mostrano la distribuzione del numero di chiamate nei vari mesi disponibili. E' evidente una forte asimmetria positiva, dato che il numero di chiamate è infatti per il 92% minore di 100 e per il 50% minore di 15.

minimo	1st quartile	mediana	media	3rd quartile	massimo
0	2	15	32.89	42	2693

Tabella 1.3 Numero di chiamate di tutti gli utenti in tutti i 18 mesi

1.3.2 Variabile SMS e MMS

Osserviamo ora le distribuzioni del numero di **SMS** e **MMS** inviati dagli utenti nel periodo di osservazione.

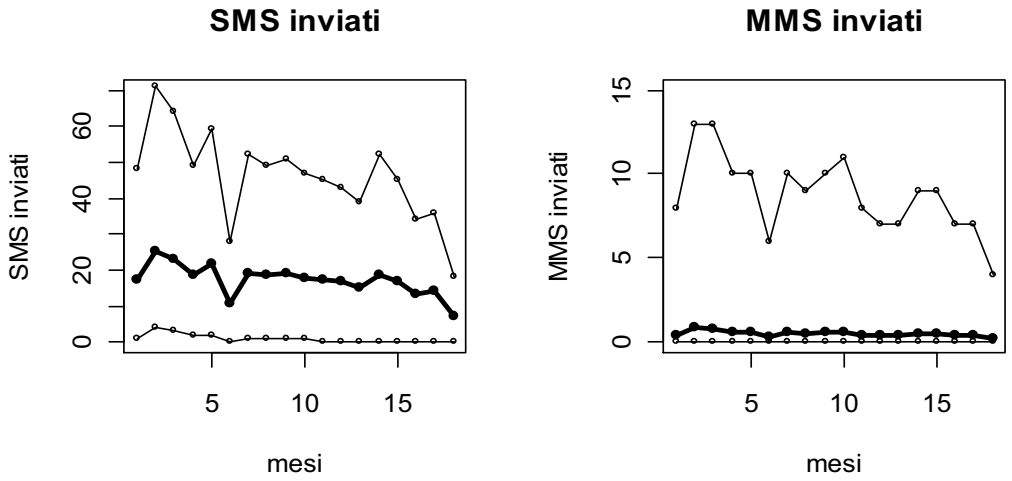

Figura 1.5(a) 90° percentile, media e mediana degli SMS inviati nei diciotto mesi. 1.5(b) 99° percentile, media e mediana degli MMS inviati nei diciotto mesi

minimo	1st quartile	mediana	media	3rd quartile	massimo
0	0	1	17.25	11	4161

Tabella 1.4 Numero di SMS inviati

Minimo	1st quartile	mediana	media	3rd quartile	massimo
0	0	0	0.49	0	334

Tabella 1.5 Numero di MMS inviati.

I grafici e le tabelle descrivono brevemente queste due variabili; la particolarità che le accomuna è che gran parte degli utenti non inviano né Sms né Mms, infatti si nota che circa la metà di essi nel corso dei diciotto mesi non invia alcun messaggio, mentre l'87% non invia alcun MMS.

1.3.3 Variabile ETA'

Per poter possedere una Sim è necessario essere maggiorenni, di conseguenza l'età degli utenti sarà necessariamente maggiore di 18. La variabile **età** è stata suddivisa in classi di 10 anni:

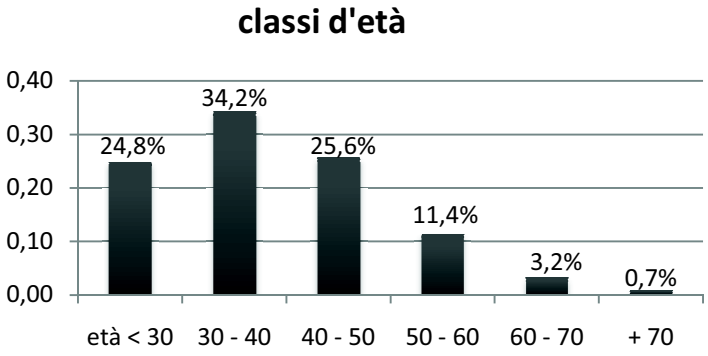

Figura 1.6 Percentuali di utenti nelle varie classi di età

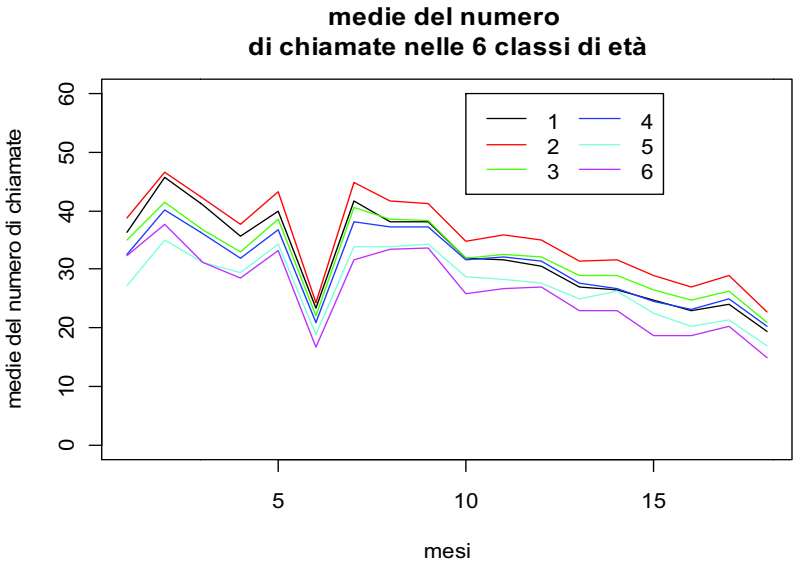

Figura 1.7 Numero di chiamate mediano e medio per ogni mese per le varie classi di età

Gli utenti tra i 30 e i 40 anni registrano il maggior numero di chiamate, con l'aumento dell'età si evidenzia un calo dell'uso del cellulare.

1.3.4 Variabile ZONA

Vediamo ora come sono distribuite sul territorio gli utenti della nostra azienda.

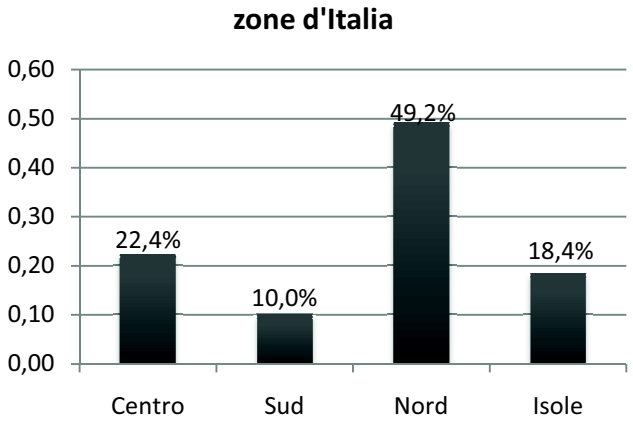

Figura 1.8 Percentuali delle utenti nelle 4 zone demografiche.

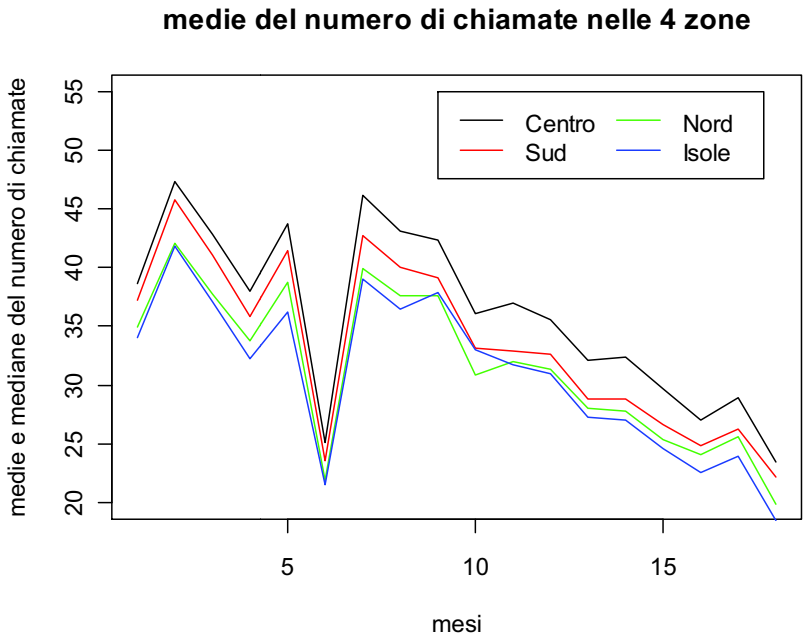

Figura 1.9 Medie delle chiamate nei 18 mesi delle 4 zone geografiche.

Cerchiamo di valutare le possibili differenze per quanto riguarda l'uso del cellulare tra la varie zone d'Italia.

	Minimo	1st quartile	mediana	media	3rd quartile	massimo
Centro	0	3	17	36.09	48	780
Sud	0	2	13	30.86	39	726
Nord	0	3	15	31.62	41	947
Isole	0	2	13	33.49	41	2693

Tabella 1.6 Caratteristiche della variabile Zona nelle sue 4 modalità: Centro, Sud, Nord e Isole

I grafici delineano in tutte e quattro le regioni un'uguale tendenza verso il basso, con un picco negativo nel mese di aprile del 2005. Il centro risulta essere la zona d'Italia in cui si effettuano, mediamente, più chiamate.

1.3.5 Variabile SESSO

gli utenti sono per il 23% femmine mentre il restante 77% sono maschi.

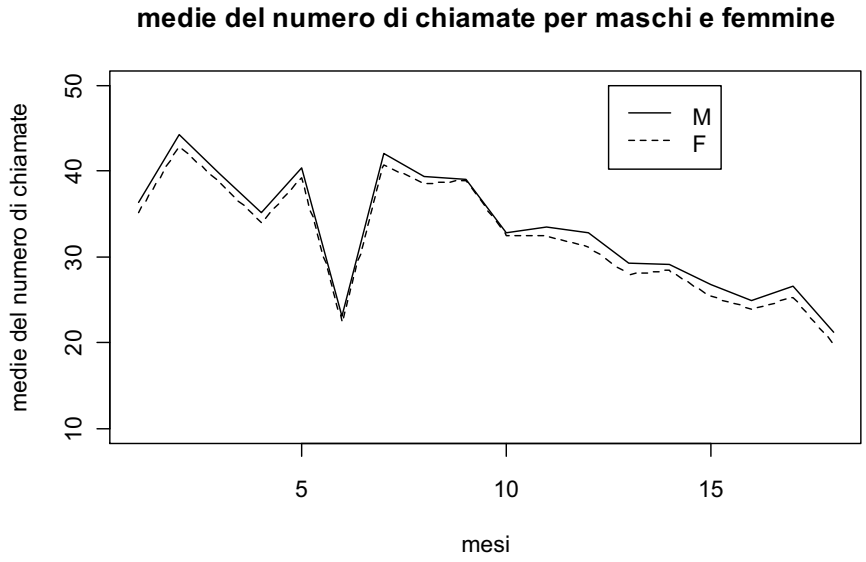

Figura 1.10 Medie dei maschi e delle femmine nei diciotto mesi

I grafici non evidenziano particolari differenze tra maschi e femmine nell'uso del telefono.

1.3.6 Variabile CUSTOMERID

Abbiamo nel dataset 21758 Sim che appartengono a 19373 utenti.

numero Sim	**1**	**2**	**3**	**4**	**6**	**8**
frequenze	17100	21719	82	10	1	1

Tabella 1.7 frequenze del numero di Sim per utente

Solo il 12% degli utenti possiede più di una scheda telefonica.

1.3.7 Variabile PIANO TARIFFARIO

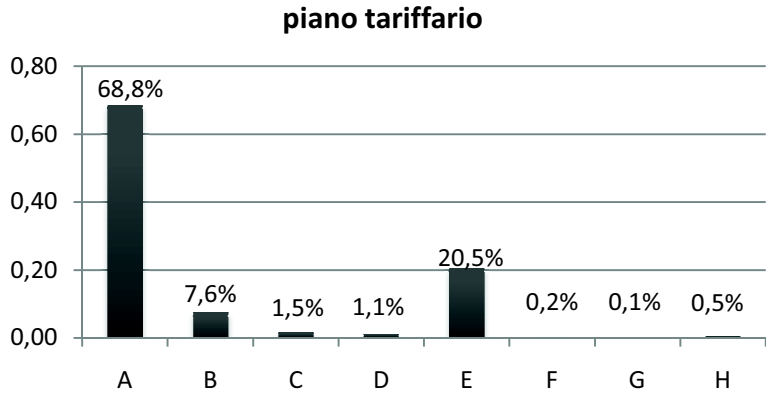

Figura 1.11 Percentuali dei piani tariffari delle Sim

Come emerge dal grafico 1.11, A è il piano tariffario dominante, seguono rispettivamente E e B, i restanti piani hanno una percentuale minore del 3%.

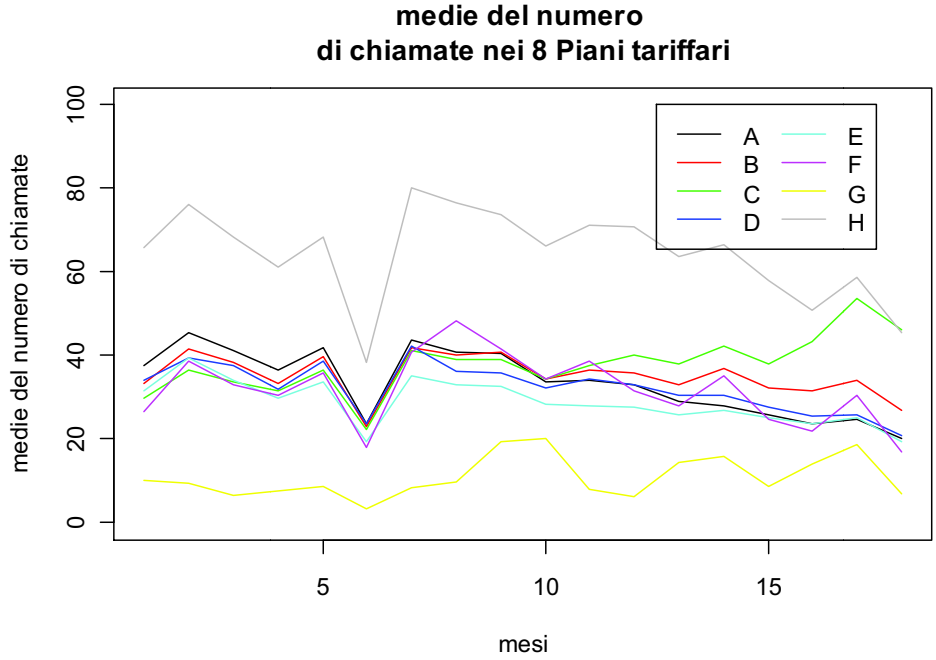

Figura 1.12 Medie del numero di chiamate degli otto piani tariffari per ciascun mese

Analizzando le medie del numero di chiamate per ogni piano, l'H risulta essere quello con più chiamate effettuate, mentre il piano G quello meno utilizzato. Tutti gli altri hanno valori simili fatta eccezione per il C che, a partire dal dodicesimo mese, ha un andamento verso l'alto ad indicare un aumento delle chiamate.

2 REGRESSIONE LINEARE

2.1 Regressione lineare semplice

La regressione lineare è un metodo di stima del valore atteso condizionato di una variabile dipendente Y, dati i valori di altre variabili indipendenti $X_1, \dots, X_k$. La regressione può essere usata per individuare eventuali relazioni tra variabili e soprattutto per fare previsioni della variabile risposta conoscendo i valori dei predittori. La variabile **chiamate**, la quantità da prevedere con i modelli, assume valori tra 0 e 2693. Dato il largo campo di variazione della variabile è ragionevole utilizzare una semplice regressione lineare del tipo:

$$y_t = X_t\beta + \varepsilon_t = \beta_1 X_{t1} + \beta_2 X_{t2} + \cdots + \beta_k X_{tk} + \varepsilon_t \qquad \textit{per t = 1, … , n}$$

dove gli errori ε_t sono indipendenti e distribuiti normalmente con media 0 e deviazione standard σ. Un scrittura equivalente è la seguente: $y_t \sim \mathcal{N}(X_t\beta, \sigma^2)$ dove X è una matrice n x k. Nel nostro dataset, y corrisponde al numero di chiamate effettuate al mese, k = 8 sono i predittori nel vettore X_t (la t-esima riga della matrice X) e n = 208131 le osservazioni totali.

Gli input del modello sono:

X_{t1} un termine costante uguale a uno per tutte le osservazioni (intercetta).

X_{t2} l'età dell'utente a cui è registrata la Sim.

X_{t3} il piano tariffario della scheda telefonica.

X_{t4} la zona d'Italia dove risiede l'utente.

X_{t5} il sesso .

X_{t6} il mese di registrazione (valori da 1 a 18).

X_{t7} il numero di Sms inviati quel mese.

X_{t8}, il numero di Mms inviati quel mese.

Il parametro σ rappresenta la variabilità con la quale i valori predetti dal modello si scostano dai valori reali. Per ogni variabile categoriale, nel nostro caso età, piano tariffario, sesso e zona geografica, è necessario utilizzare γ-1 variabili dummy (dove γ è il numero dei possibili valori della variabile). Una variabile dummy assume solo valori zero e uno. Uno se la variabile categoriale del soggetto assume un preciso livello e zero in tutti gli altri casi. Ad esempio, per la variabile zona, γ è uguale a 4 e quindi utilizzeremo 3 variabili dummy, di cui la prima vale 1 se la Sim è registrata nel sud Italia e 0 negli altri casi. La seconda varrà 1 se, invece, la scheda appartiene a qualche regione del Nord e 0 negli altri casi, la terza varrà 1 solo se la scheda è regi-

strata in Sicilia o Sardegna. Se tutte e tre le variabili assumono valore zero significa che la scheda telefonica è stata registrata in centro Italia.

Il modello viene stimato tramite il metodo dei minimi quadrati implementato dalla funzione lm() di R.

	stima	**st. error**		**stima**	**st. error**
Intercetta	39.468	0.377	piano F	-7.395	3.018
età 30-40	5.315	0.287	piano G	-25.690	6.030
età 40-50	1.243	0.309	piano H	26.969	1.603
età 50-60	0.427	0.390	sud	-4.140	0.409
età 60-70	1.551	0.674	nord	-5.245	0.276
eta +70	-2.542	1.232	isole	-0.624	0.343
piano B	-0.228	0.417	femmine	-1.446	0.259
piano C	-2.066	0.891	mese	-0.830	0.022
piano D	0.643	1.032	SMS	0.204	0.002
piano E	-4.861	0.275	MMS	0.492	0.030

Tabella 2.1 Stime e standard error per i coefficienti del modello lineare

Il modello è composto da 20 variabili esplicative; per ognuna è riportato la stima del rispettivo parametro e il suo errore standard. Avere poche osservazioni di una variabile può generare rilevanti standard error, ad esempio il piano G viene stimato solo su 17 osservazioni perché nel gruppo di stima è presente una sola Sim con quel particolare piano tariffario.

L'intercetta pari a 39.468 corrisponde al numero medio di chiamate effettuate dalle Sim quanto tutte le altre variabili assumono un valore nullo. I coefficienti con stime negative porteranno una diminuzione della stima $\hat{y}_t$. Ad esempio le femmine effettuano, in media, quasi una chiamata e mezza in meno al mese dei maschi, mentre per ogni mese trascorso si ha una diminuzione media di 0.83 chiamate. I coefficienti con stime positive portano, invece, un aumento della stima di $\hat{y}_t$. Coloro che hanno piano tariffario H effettuano, in media, quasi 27 chiamate in più al mese rispetto a coloro che hanno piano A. Si stima che coloro che hanno il valore più alto di chiamate sono gli uomini del centro Italia di età compresa tra i 30 e i 40 anni, con piano tariffario H. Un altro dato interessante risulta essere la relazione tra numero di chiamate e messaggi inviati. Le variabili sono direttamente proporzionali, quando aumenta il numero di sms o mms aumentano anche le telefonate

La deviazione standard residua $\hat{\sigma} = \sqrt{\sum_{t=1}^{n} {r_t}^2/n}$ riassume la scala dei residui e nel nostro esempio vale 49.688. La deviazione standard è una misura della distanza media di ogni osservazione dal rispettivo valore stimato dal modello. Un altro indicatore della bontà del modello è R^2, la frazione di varianza spiegata dal modello. Nel nostro caso R^2 è circa del 7%.

Assunzioni della regressione

1) **Validità**. I dati che si analizzano devono essere tali da poter rispondere alle domande poste dalla stessa analisi. Significa che la variabile risposta deve riflettere accuratamente il fenomeno d'interesse, il modello deve includere tutti i predittori rilevanti, e essere utilizzato quando esso è applicabile. Nel nostro studio tale assunzione può ritenersi valida in quanto la variabile risposta descrive il traffico telefonico delle Sim e vengono utilizzate nel modello tutte le variabili disponibili del dataset originale.

2) **Additività e linearità.** Il più importante vincolo matematico del modello di regressione è che le sue componenti deterministiche sono una funzione lineare dei coefficienti: $y_i = \beta_1 x_1 + \beta_2 x_2 + \ldots$.

Nel nostro esempio ogni variabile utilizzata contribuisce, tramite il proprio valore e quello del suo coefficiente β, alla stima di y_t, tramite una funzione lineare.

3) **Indipendenza degli errori.** L'ipotesi dell'indipendenza degli errori è spesso violata quando i dati sono raccolti nel corso del tempo. I residui di un certo tempo *t* tendono ad essere simili ai residui che avvengono in tempi contigui. In questo caso si dice che tra i residui vi è ***autocorrelazione,*** come mostrato in figura 2.1

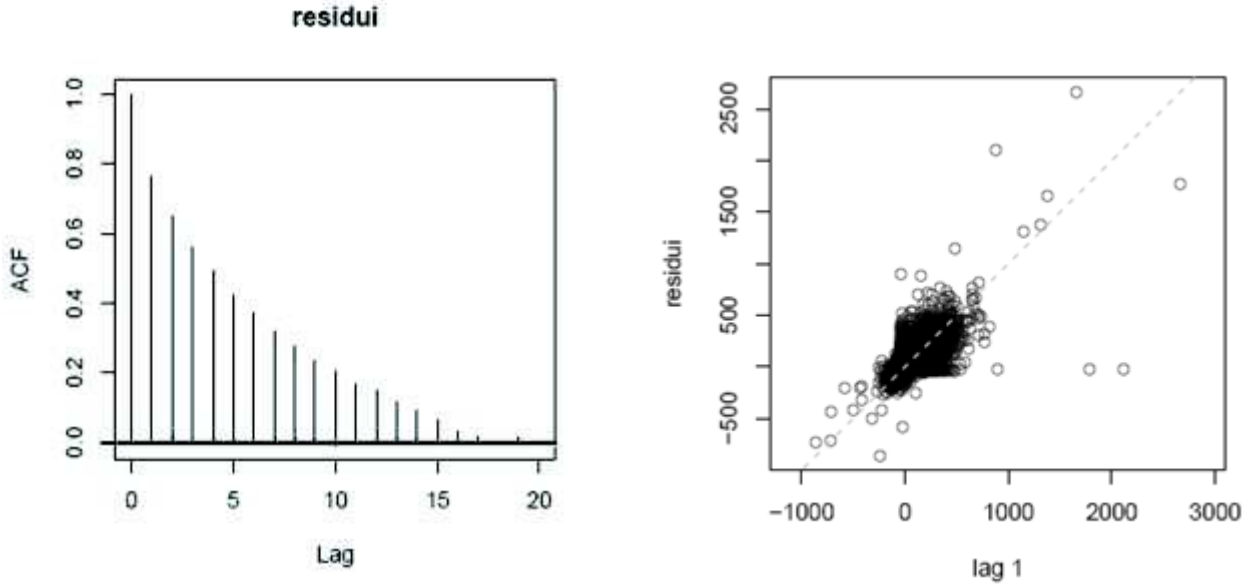

Figura 2.1 (a) acf dei residui (b) lag.plot dei residui del modello lineare

4) **Omoschedasticità degli errori**. Se la varianza degli errori di regressione è diversa, le stime sono più efficienti usando i minimi quadrati pesati, con pesi inversamente proporzionati alla propria varianza. Aver diverse varianze non è, comunque, uno degli aspetti più importanti della regressione.

5) **Normalità degli errori**. L'assunto richiede la normalità degli errori. La diagnostica della normalità dei residui non è comunque così importante. Nel nostro modello lineare i residui non sono normali, come si nota dal grafico 2.3(b).

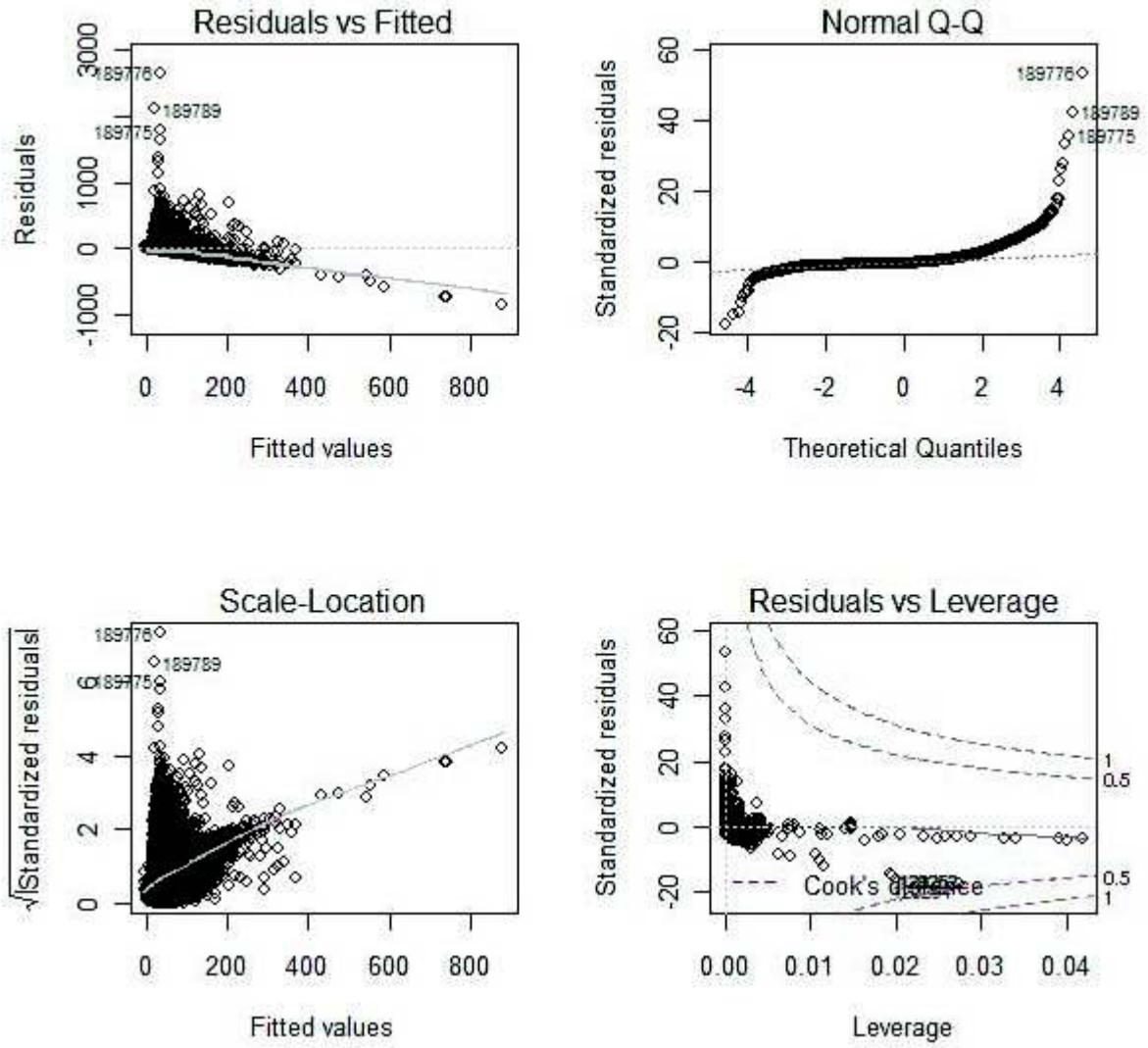

Figura 2.3 grafici diagnostici della bontà della stima del modello `plot()`*: grafici diagnostici della bontà della stima del modello:*

- *Grafico dei residui contro i valori teorici: può rivelare la presenza di una residua dipendenza sistematica non individuata dal modello lineare stimato. In un buon modello questo grafico dovrebbe apparire come completamente accidentale;*
- *Normal q-q plot dei residui standardizzati: verifica visuale dell'assunzione di normalità della componente erratica del modello lineare. Quanto più i punti che rappresentano i residui ordinati giacciono in prossimità della linea q-q, tanto più plausibile è detta assunzione;*
- *Grafico delle radici quadrate dei residui standardizzati contro i valori teorici: utile nell'individuazione di osservazioni anomale e per visualizzare strutture di dipendenza residue non individuate dal modello stimato;*
- *Grafico delle distanze di Cook: misure dell'influenza di ciascun individuo sulla stima dei parametri del modello.*

I grafici in figura 2.3 servono per verificare la bontà del modello. Valutando il loro esito si ottiene un modello lineare non adatto alla stima della variabile chiamate. Questa conclusione è confermata anche da un R^2 molto basso e un'elevata deviazione standard.

Un buon metodo di verifica delle assunzioni è disegnare dei grafici con i residui verso i valori predetti $X_i\hat{\beta}$ o semplicemente predittori individuali x_i.

La figura 2.4 mostra i residui verso il numero di sms e mms inviati. Dai grafici si nota che gli errori più elevati si hanno per Sim che non inviano sms né mms. Questo perché le variabili **SMS** e **MMS** sono correlate positivamente con le chiamate effettuate, ma per alcune Sim si ha un gran numero di telefonate pur non inviando alcun tipo di messaggio.

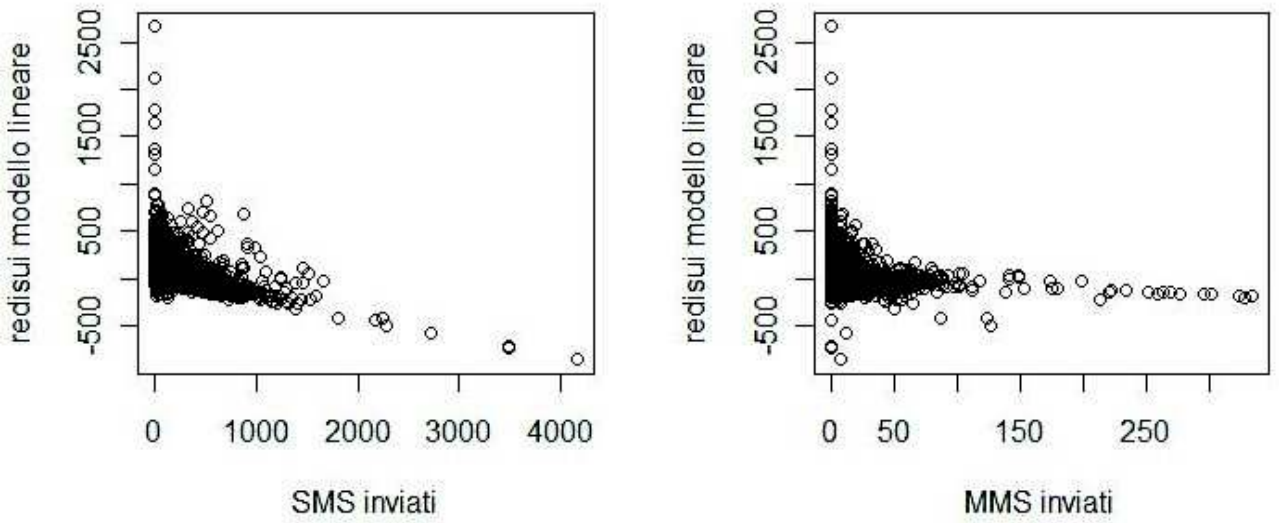

Figura 2.4 (a) residui verso SMS inviati (b) residui verso MMS inviati

Si è inoltre stimato il numero di chiamate effettuate da ogni singola Sim nel mese di aprile 2006, per poi confrontarli con i valori reali. La deviazione standard residua $\hat{\sigma}$ vale 37.680. Questa quantità servirà per confrontare questo modello con altri che svilupperemo in seguito.

Il modo migliore per validare un modello, in qualsiasi ambito, è di confrontare i valori stimati dal modello con i dati reali. Utilizzeremo, nel capitolo 6, il campione di verifica per fare questa operazione.

2.2 Regressione lineare con trasformata radice quadrata della variabile risposta

Si prova qui di seguito a stimare una trasformata della variabile **chiamate** con l'obiettivo di ottenere un modello di previsione migliore del precedente. Una possibile soluzione è di utilizzare la radice quadrata del numero di chiamate effettuate. Si avrà una regressione del tipo:

$$\sqrt{y_t} = \boldsymbol{X_t\beta} + \boldsymbol{\varepsilon_t} = \boldsymbol{\beta_1 X_{t1}} + \boldsymbol{\beta_2 X_{t2}} + \cdots + \boldsymbol{\beta_k X_{tk}} + \boldsymbol{\varepsilon_t} \qquad \textit{per t = 1, … , n}$$

dove gli errori ε_t sono indipendenti e normalmente distribuiti con media 0 e deviazione standard ε.

Le stime dei coefficienti sono riassunti dalla tabella 2.2:

	stima	**st. error**		**stima**	**st. error**
Intercetta	5.152	0.026	piano F	-0.436	0.211
età 30-40	0.403	0.020	piano G	-2.164	0.422
età 40-50	0.129	0.022	piano H	2.572	0.112
età 50-60	0.072	0.027	sud	-0.351	0.029
età 60-70	0.170	0.047	nord	-0.331	0.019
eta +70	-0.163	0.086	isole	-0.175	0.024
piano B	0.311	0.029	femmine	-0.070	0.018
piano C	0.016	0.062	mese	-0.101	0.002
piano D	0.083	0.072	SMS	0.016	0.001
piano E	-0.079	0.019	MMS	0.041	0.002

Tabella 2.2 Stime e standard error per i coefficienti del modello con trasformata radice quadrata della variabile risposta ***chiamate***

La deviazione standard residua $\hat{\sigma}$ vale ora 55.431.

Analizzando i grafici della figura 2.2 si raggiungono le medesime conclusioni del modello precedente. La deviazione standard residua, ottenuta confrontando le stime per il mese di aprile 2006 con i veri valori, $\hat{\sigma}$ vale 38.680. $\hat{\sigma}$ è aumentato rispetto al modello precedente. Non sembra quindi che la trasformata della radice quadrata per la variabile risposta **chiamate** possa portare alcun miglioramento al modello.

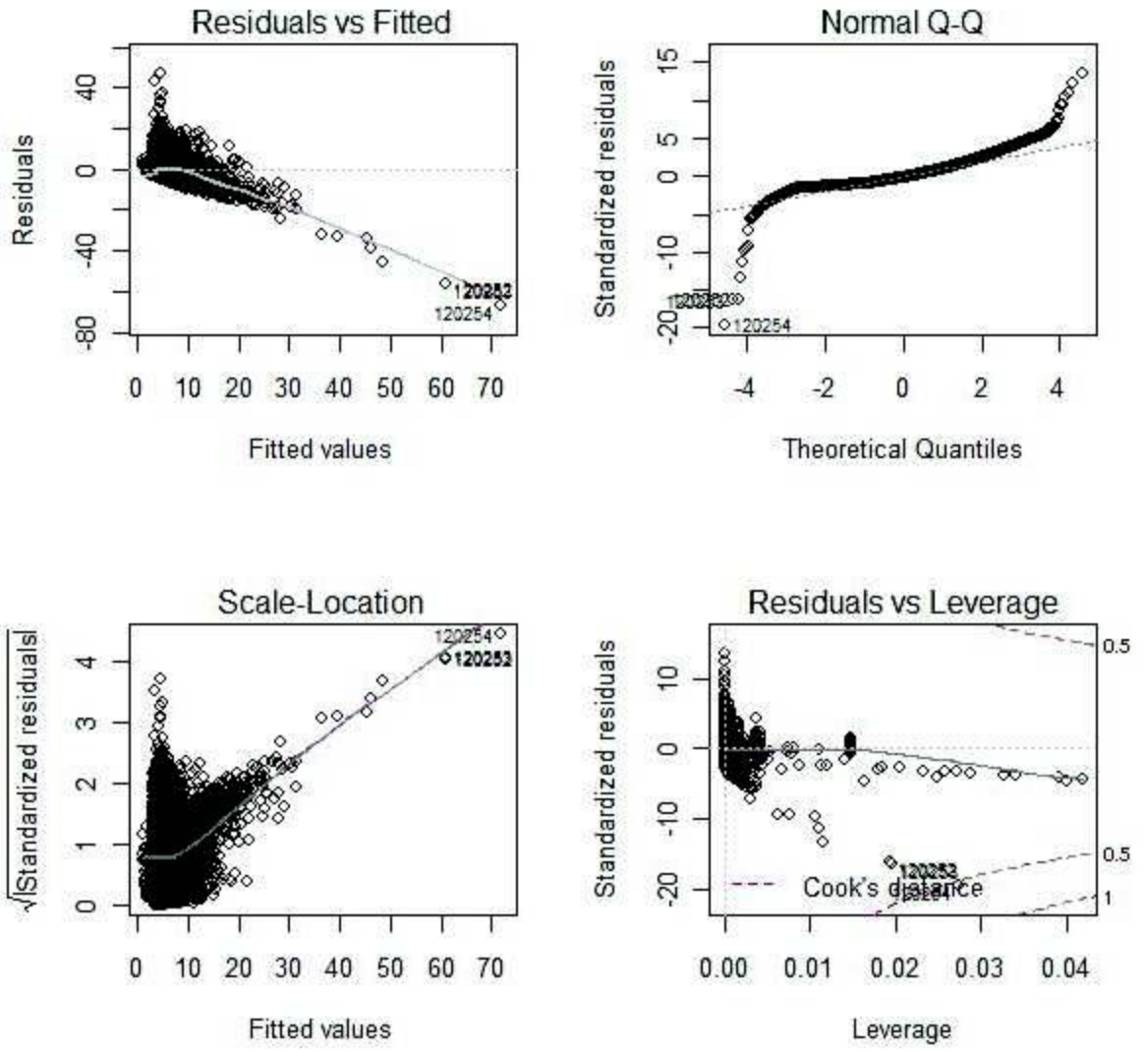

Figura 2.3 grafici diagnostici della bontà della stima del modello $plot$

La diagnostica dei residui non evidenzia miglioramenti significativi nell'utilizzo della trasformata radice quadrata della variabile risposta.

3 MODELLI LINEARI GERARCHICI CON INTERCETTA VARIABILE

3.1 Introduzione

Quando si devono analizzare dati strutturati in gruppi, i modelli gerarchici rappresentano una generalizzazione della regressione lineare, dove l'intercetta e/o altri coefficienti possono variare nei gruppi. Con dati raggruppati, una regressione che include indicatori per gruppi è chiamata *modello ad intercetta variabile.* Esso può essere interpretato come un unico modello ma con tante differenti intercette quante sono i gruppi. La figura 3.1(a) mostra un esempio di un modello con un solo predittore x e indicatori per J = 5 gruppi. Il modello può essere scritto come una regressione con 6 coefficienti, oppure come una regressione con 2 predittori (termine costante e x) ma con l'intercetta che varia nei gruppi:

modello ad intercetta variabile: $y_i = \alpha_{j[i]} + \beta x_i + \epsilon_i.$

La figura 3.1(b) mostra un modello con intercetta fissa e il coefficiente per x che varia per gruppo:

modello con coefficiente variabile: $y_i = \alpha + \beta_{j[i]} x_i + \epsilon_i.$

La figura 3.1(c) mostra un modello con intercetta e coefficiente β che variano per gruppo:

modello ad intercetta e coefficiente variabile: $y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i.$

I coefficienti variabili sono le interazioni tra il predittore x e gli indicatori per i gruppi.

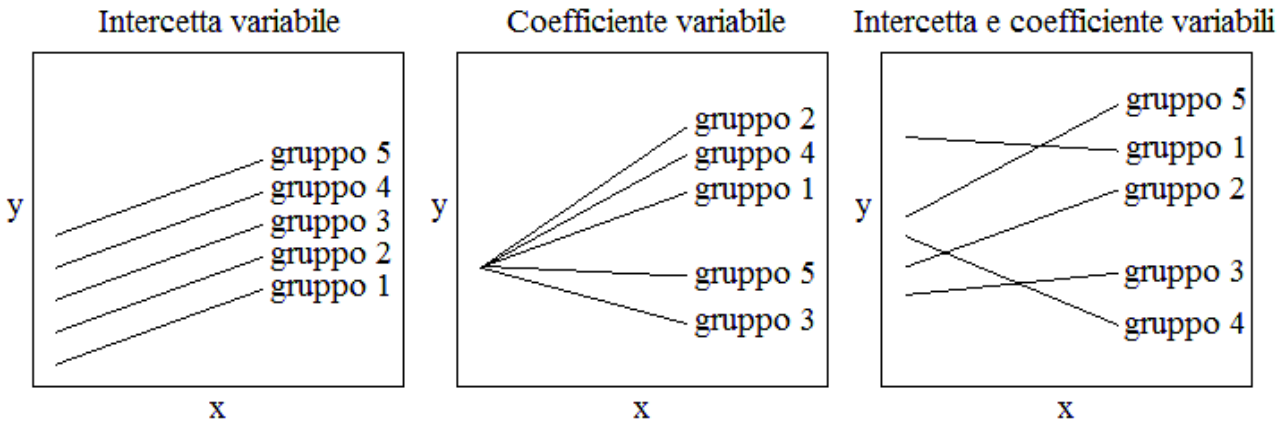

Figura 3.1 (a) Modello di regressione lineare con intercetta variabile ($\boldsymbol{y = \alpha_j + \beta x}$), (b) modello con coefficiente del predittore x variabile ($\boldsymbol{y = \alpha + \beta_j x}$) ed (c) entrambi ($\boldsymbol{y = \alpha_j + \beta_j x}$).

Si può pensare al multilivello come una regressione che include delle variabili categoriali che descrivono l'appartenenza o meno di ogni unità ad un gruppo. In questa prospettiva si utilizza un fattore con J livelli (J è il numero dei gruppi) corrispondenti a J predittori nel modello di regressione.

In questo modo vengono aggiunti al modello J-1 predittori lineari. Il risultato è equivalente a replicare il termine costante della regressione ottenendo J intercette distinte, una per ogni gruppo.

Il modello gerarchico è caratterizzato dal fatto che i J coefficienti sono anch'essi stimati da un modello, o più semplicemente, da una comune distribuzione per i J parametri α_j. In pratica si esegue un secondo modello di regressione per gli α_j usando come predittori le variabili riferite al livello di gruppo. I coefficienti di questo modello vengono chiamati *iperparametri*.

Questo secondo modello è stimato simultaneamente con la regressione al primo livello per y.

I modelli gerarchici sono un compromesso tra due estremi: ***complete pooling***, nel quale gli indicatori di gruppo non sono inclusi nel modello, e ***no pooling***, nel quale viene stimato un modello per ogni gruppo.

Si voglia ora stimare una quantità y con un modello lineare gerarchico. I dati saranno raggruppati in J gruppi di numerosità n_j. In questo semplice scenario senza predittori la stima del multilivello per un dato gruppo j può essere approssimata con una media pesata delle osservazioni in quel gruppo (la stima del no pooling, $\bar{y}_j$) e la media di tutte le osservazioni di tutti i gruppi (la stima del complete pooling, $\bar{y}_{all}$):

$$\hat{\alpha}_j^{multilevel} \approx \frac{\frac{n_j}{\sigma_y^2}\bar{y}_j + \frac{1}{\sigma_\alpha^2}\bar{y}_{all}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}}$$

dove σ_y^2 è la varianza entro i gruppi e σ_α^2 è la varianza fra i gruppi. Le medie dei gruppi con scarsa numerosità portano meno informazione, quindi la stima del multilivello si avvicina a quella totale. Nel caso limite che n_j sia uguale a zero la stima del multilivello coinciderà con quella totale, $\bar{y}_{all}$.

Al contrario, per i gruppi numerosi, l'informazione disponibile sarà maggiore, quindi la stima del modello gerarchico sarà prossima a quella del singolo gruppo. Nel caso limite che n_j sia infinito essa coincide con la media del gruppo j, $\bar{y}_j$.

Il più semplice modello gerarchico lineare è il seguente:

$$y_i \sim N(\alpha_{j[i]} + \beta x_i, \sigma_y^2) \qquad \text{per i = 1,…,n,}$$

Nel modello no pooling gli α_j sono stimati con il metodo dei minimi quadrati, significa che le intercette sono stimate in un modello separato per ogni gruppo (con con il vincolo che i β siano uguali in tutti i modelli). In un multilivello è invece assegnata una distribuzione di probabilità agli α_j:

$$\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2) \qquad \text{per j = 1,…,J}$$

con μ_α e σ_α^2 entrambi stimati dai dati. Quando $\sigma_\alpha \to \infty$ si ritorna al caso del no pooling, mentre se $\sigma_\alpha \to 0$ si ripropone il caso del complete pooling.

La stima multilivello degli α_j può essere espressa con la media pesata della stima no pooling per quel gruppo ($\bar{y}_j - \beta\bar{x}_j$) e la media, μ_α:

stima degli $\alpha_j \approx \frac{\frac{n_j}{\sigma_y^2}}{\frac{n_j}{\sigma_y^2}+\frac{1}{\sigma_\alpha^2}}(\bar{y}_j - \beta\bar{x}_j) + \frac{\frac{1}{\sigma_\alpha^2}}{\frac{n_j}{\sigma_y^2}+\frac{1}{\sigma_\alpha^2}}\mu_\alpha$.

Si continua aggiungendo un nuovo predittore u al secondo livello, il quale descriverà qualche caratteristica comune a tutte le unità del medesimo gruppo. Introducendo un predittore a livello di gruppo si avrà che $\alpha_j \sim N(\gamma_0 + \gamma_1 u_j, \sigma_\alpha^2) = N(U_j\gamma + \eta_j)$ con $\eta_j \sim N(0, \sigma_\alpha^2)$ dove γ_0 rappresenta l'intercetta e γ_1 e il coefficiente per la variabile u. La stima multilivello degli α_j è qui una media pesata della stima no pooling di gruppo $(\bar{y}_j - \beta\bar{x}_j)$ e della previsione della regressione $\hat{\alpha}_j$:

stima degli $\alpha_j \approx \frac{\frac{n_j}{\sigma_y^2}}{\frac{n_j}{\sigma_y^2}+\frac{1}{\sigma_\alpha^2}}(stima\ del\ gruppo\ j) + \frac{\frac{1}{\sigma_\alpha^2}}{\frac{n_j}{\sigma_y^2}+\frac{1}{\sigma_\alpha^2}}(stima\ della\ regressione)$

analogamente gli errori a livello di gruppo η_j sono parzialmente raggruppati intorno allo zero:

stima degli $\eta_j \approx \frac{\frac{n_j}{\sigma_y^2}}{\frac{n_j}{\sigma_y^2}+\frac{1}{\sigma_\alpha^2}}(\bar{y}_j - \bar{X}_j\beta - U_j\gamma) + \frac{\frac{1}{\sigma_\alpha^2}}{\frac{n_j}{\sigma_y^2}+\frac{1}{\sigma_\alpha^2}}0.$

Benefici del multilivello

L'analisi multilivello si basa sull'idea che le osservazioni appartenenti al medesimo gruppo siano più vicine o abbiano dei comportamenti tra loro più simili di quanto non accada con osservazioni appartenenti a reti di relazioni diverse. Come è noto (Aitkin e Longford, 1986; Burstein *et al*., 1978), l'analisi ad un solo livello di dati gerarchici porta ad una distorsione nella stima dei parametri e/o dei loro errori standard. Appare quindi necessario il ricorso a tecniche di analisi che considerano esplicitamente entrambi i livelli, come i modelli gerarchici (Snijders e Bosker, 1999).

L'uso dei multilivello ci permette di:

1) Eliminare la distorsione nella stima degli errori standard dei parametri;
2) Stimare l'effetto del gruppo scomponendo la variabilità in due componenti: quota interna al gruppo (varianza entro i gruppi) e tra gruppi (varianza tra gruppi);
3) Introdurre variabili esplicative a livello di gruppo cercando così di dare una descrizione della variabilità tra gruppi;
4) Modellare gli effetti di interazione;
5) Fare previsioni di nuove osservazioni appartenenti a gruppi sia nuovi che già esistenti.

Considerato che si dispone di dati sul traffico di schede telefoniche registrati in diciotto mesi consecutivi, si cerca di prevedere il traffico sulla base di tutte le variabili disponibili.

I dati sono strutturati su due livelli distinti:

- primo livello: le singole registrazioni del traffico mese per mese (chiamate, sms e mms)
- secondo livello: le variabili relative alla singola Sim (età, sesso, piano tariffario, zona)

Per arrivare a comprendere al meglio un multilivello vengono implementate all'inizio due semplici regressioni: il **complete pooling** e il **no pooling**. La prima sarà elaborata ignorando il fatto che le chiamate sono state effettuate da Sim differenti. La seconda invece sarà implementata introducendo la variabile **id** come un fattore a J livelli. Questa operazione consiste, in pratica, nell'effettuare J differenti regressioni, una per ogni Sim, con il vincolo che i coefficienti delle variabili utilizzate siano gli stessi per tutti i J modelli.
I modelli verranno stimati con le sole informazioni dei primi diciassette mesi, l'ultimo verrà utilizzato per confrontare i valori predetti dai modelli con i relativi valori reali.

3.2 Complete pooling

Il modello che si va a costruire è praticamente uguale alla regressione lineare descritta nel capitolo precedente, con la sola differenza che ora si utilizzano solo le variabili relative al traffico (al primo livello gerarchico).

	stima	**st. error**
intercetta	37.397	0.233
mese	-0.831	0.022
SMS	0.202	0.002
MMS	0.409	0.031
$\hat{\sigma}_y$	49.870	

Tabella 3.1 Stime e st. error per i coefficienti del modello Complete-pooling con le sole variabili al primo livello

3.3 No pooling

Viene aggiunto il fattore **id** con 12243 livelli.

	stime	**st.error**
mese	-0.753	0.015
SMS	0.176	0.002
MMS	0.671	0.029
$\hat{\sigma}_y$	31.160	

Tabella 3.2 Stime e st. error per i coefficienti del modello No-pooling con le sole variabili al primo livello

	minimo	**1 quartile**	**mediana**	**media**	**3 quartile**	**massimo**
intercette	-189.820	13.934	23.637	36.857	44.674	829.490

Tabella 3.3 Sintesi delle intercette per il modello no-pooling

A differenza del complete-pooling si ha qui un'intercetta diversa per ogni scheda telefonica, stimata sulle sole osservazioni riguardanti la Sim stessa. Grandi valori per l'intercetta corrispondono alle stime di un elevato numero di telefonate. Le stime dei coefficienti per le altre variabili sono cambiate leggermente.

3.4 Multilivello ad intercetta variabile con predittori al primo livello

Si consideri ora il modello gerarchico con intercetta variabile. I mesi individuali saranno indicati con t e *j[t]* la scheda j contenente il mese *t*:

$$y_t \sim N\left(\alpha_{j[t]} + \beta_1 mese_t + \beta_2 SMS_t + \beta_3 MMS_t, \sigma_y^2\right) \qquad \text{per } t = 1, \dots, n$$

$$\alpha_j \sim N\left(\mu_\alpha, \sigma_\alpha^2\right) \qquad \text{per } j = 1, \dots, J$$

dove μ_α e σ_α^2 sono stimati dai dati. Con questo modello si ottiene un pooling parziale in ogni gruppo, cioè un compromesso tra le stime ottenute nei due modelli precedenti. Nei casi limiti in cui $\sigma_\alpha \to \infty$ le stime sono uguali a quelle del no pooling, mentre se $\sigma_\alpha \to 0$ esse sono uguali a quelle del complete pooling.
Si usa la teoria bayesiana per calcolare le stime dei coefficienti del modello.
Ogni parametro necessita di una distribuzione a priori. Per β_1, β_2 e β_3 si è scelta una normale con media zero e deviazione standard pari a 100. Per σ_y si utilizza invece una a priori uniforme in 0 - 100. Anche gli iperparametri devono avere una propria distribuzione che sarà normale con media zero e deviazione standard 100 per μ_α e una uniforme in 0 - 100 per σ_α. Queste a priori sono proprie ma contengono pochissima informazione.
Le stime a posteriori ottenute tramite algoritmi iterativi sono riportate in tabella 3.4:

	stima	**st. error**		**stima**	**st. error**
β_1, mese	-0.841	0.014	$\hat{\mu}_\alpha$	37.871	0.400
β_2, SMS	0.174	0.002	$\hat{\sigma}_y$	31.106	0.049
β_3, MMS	0.639	0.026	$\hat{\sigma}_\alpha$	39.006	0.257

Tabella 3.4 stime e st. error per le posteriori del modello gerarchico con le sole variabili al primo livello

La tabella 3.5 riporta alcuni indici di sintesi per gli effetti casuali $\boldsymbol{\alpha_j}$**:**

	minimo	**1 quartile**	**mediana**	**media**	**3 quartile**	**massimo**
intercette	-216.613	-22.315	-12.879	0	7.486	767.732

Tabella 3.5 Sintesi degli effetti casuali $\boldsymbol{\alpha_j}$ per il multilivello con intercetta variabile e predittori al primo strato

Le stime e le deviazioni standard sono molto simili al modello no pooling.

La figura 3.2 confronta il valore delle intercette ± il proprio st. error per 100 Sim selezionate casualmente tra le 12243 che compongono il campione di stima, appartenenti al modello no-pooling e al modello gerarchico. I due grafici sono praticamente identici. Le stime prodotte dal multilivello corrispondono a quelle che si ottengono usando una regressione lineare semplice su ogni singola Sim, con il vincolo che i coefficienti per le variabili **mese**, **SMS** e **MMS** risultino costanti per tutte le Sim.

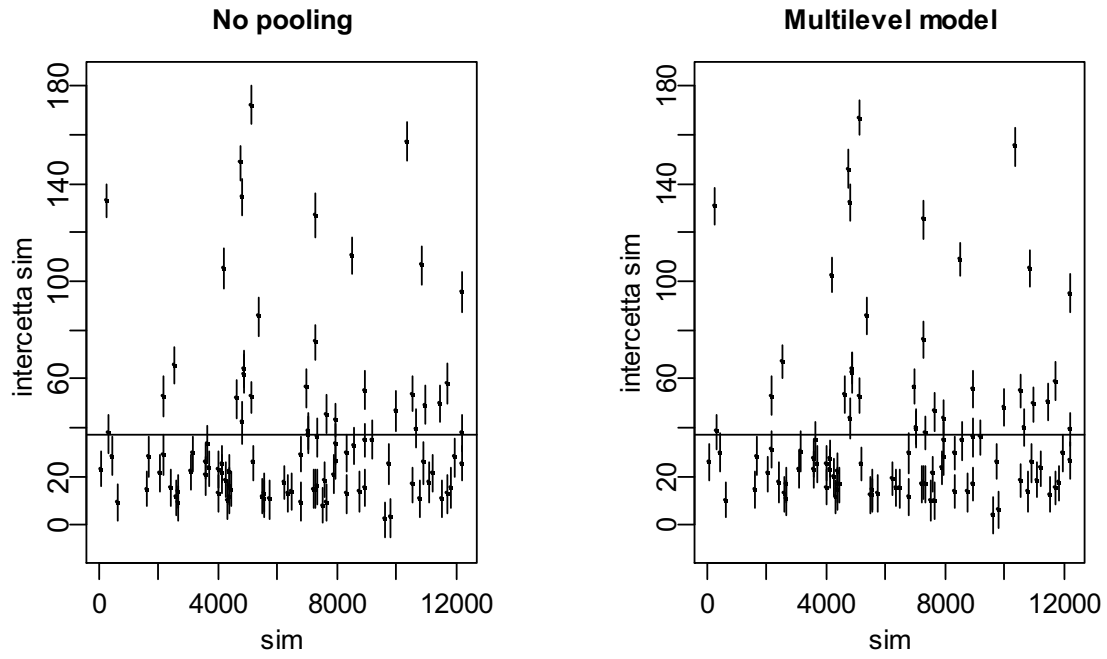

Figura 3.2 (a) Stime delle intercette per il modello no pooling, $\alpha_t \pm$ stand. error per 100 Sim selezionate casualmente dalle 12243 disponibili. (b)stime delle intercette per le medesime Sim calcolate dal multilivello. La linea orizzontale rappresenta la stima invece del modello complete pooling.

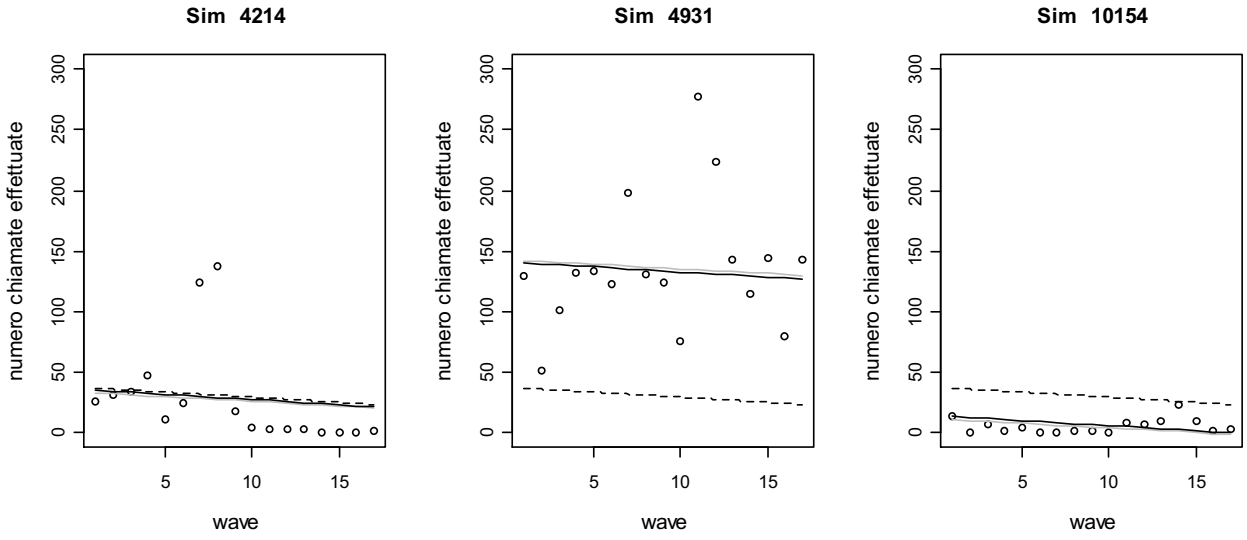

Figura 3.3 Sono qui rappresentate, per 3 diverse Sim, le chiamate effettuate nei 17 mesi disponibili. La linea tratteggiata rappresenta la stima del modello complete pooling, uguale per tutte le Sim. La linea grigia rappresenta la stima del modello no pooling mentre quella nera la stima del multilivello.

La figura 3.3 mostra le chiamate effettuate nei primi diciassette mesi da tre differenti Sim che non hanno inviato né sms, né mms. La linea tratteggiata rappresenta la stima del modello complete-pooling, quella grigia la stima del modello no-pooling e la nera quella del modello gerarchico. Quest'ultima è per costruzione sempre compresa tra le altre due. In tutti e tre i grafici le stime del modello no-pooling e del multilivello sono quasi sovrapposte a dimostrare che i due modelli sono praticamente identici. Solo nel primo grafico esse sono adiacenti alla linea tratteggiata. Per la Sim 4214 è infatti equivalente usare uno dei tre modelli. Il secondo grafico mostra le chiamate della Sim 4931: le stime sono questa volta maggiori rispetto a quella del complete-pooling. Significa che la Sim ha effettuato nei diciassette mesi più chiamate rispetto alla media.

Il terzo grafico illustra un caso contrario, in cui la scheda 10154 ha effettuato meno chiamate della media, infatti le linee grigia e nera risultano essere sotto la linea tratteggiata.

In pratica le stime del multilivello sono molto simili a quelle del modello no pooling. Questo perche la varianza $\hat{\sigma}_\alpha^2$ è molto grande (circa 1521, si ricorda che se $\sigma_\alpha^2 \to \infty$ il modello gerarchico corrisponde al no pooling). Per una generica Sim la stima del numero di chiamate è data da:

$$y_t = intercetta_t - 0.755 mese_t + 0.209 SMS_t + 0.605 MMS_t$$

con deviazione standard 31.106 al primo livello e 39.006 al secondo. Si ha una grande variabilità sia entro i gruppi sia tra i gruppi. I valori delle due varianze sono talvolta espressi dalla correlazione interclasse definita come: $\hat{\rho} = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_y^2)$ che vale in questo caso 0.61. Questa correlazione misura la proporzione di varianza totale spiegata dall'effetto delle schede telefoniche. $\hat{\rho}$ può assumere valori tra 0 e 1.

0 quando non esistono differenze tra i gruppi, quindi conviene usare una normale regressione (complete pooling). Maggiore è il valore di $\hat{\rho}$ maggiore è il guadagno dell'utilizzo del multilivello.

1 nel caso in cui tutte le osservazioni dei medesimi gruppi siano identiche.

La correlazione interclasse rappresenta inoltre la correlazione tra i valori delle chiamate effettuate in due mesi scelti a caso appartenenti a una stessa Sim, anch'essa scelta a caso.

La varianza totale è ottenuta tramite la formula: $var(Y_{tj}) = \sqrt{\sigma_\alpha^2 + \sigma_y^2} = \sqrt{1521.5 + 967.6} = 49.9.$

La figura 3.4 riporta il grafico quantile-quantile e il confronto tra valori predetti e valori reali.

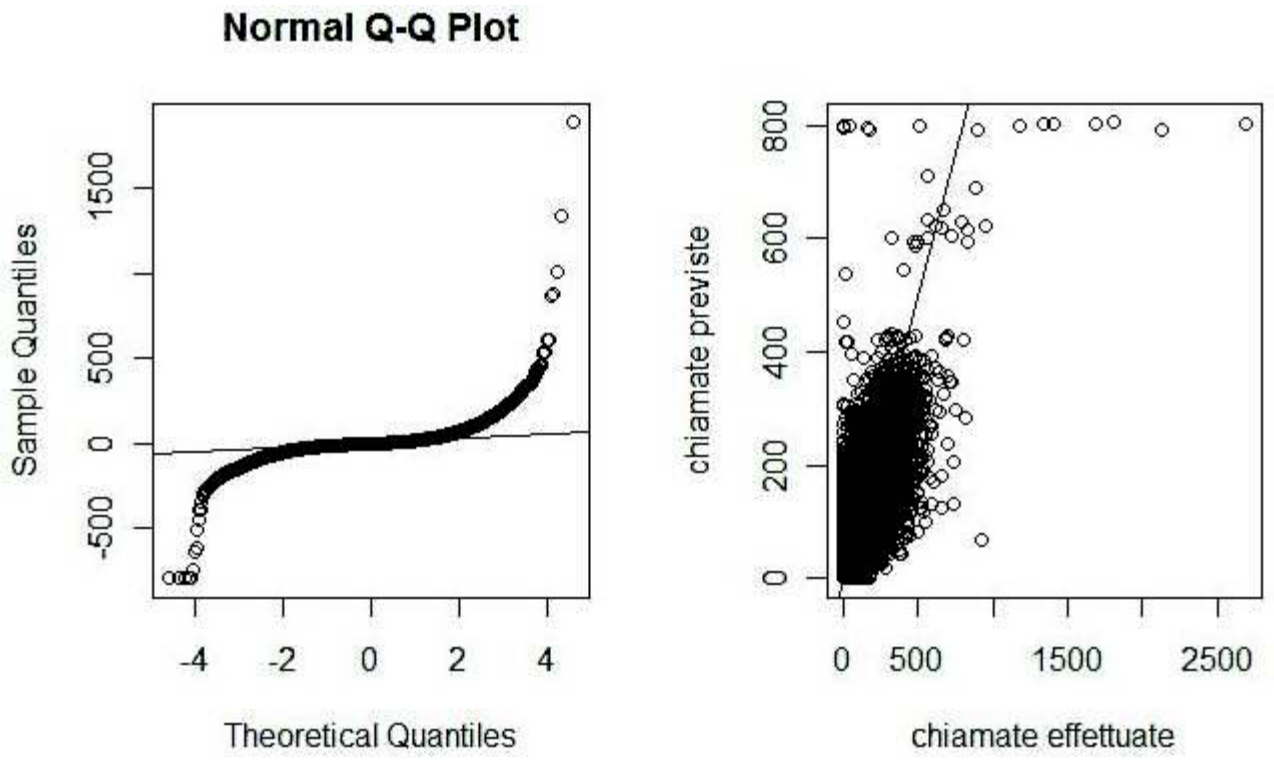

Figura 3.4 (a) Residui del modello (b) valori osservati verso i valori predetti.

I residui del multilivello non risultano essere normali. La deviazione standard residua, calcolata confrontando i valori predetti con quelli reali, relativi ad aprile 2006, vale 31.872. Si sono ottenuti sostanziali miglioramenti usando il modello gerarchico rispetto ai risultati a cui si era giunti con la regressione semplice.

3.5 Multilivello ad intercetta variabile con predittori al primo e al secondo livello

Vengono ora aggiunte al modello gerarchico le variabili relative alla scheda telefonica e al suo utente. Il modello utilizzato è il seguente:

$$y_t \sim N\left(\alpha_{j[t]} + \beta_1 mese_t + \beta_2 SMS_t + \beta_3 MMS_t\,, \sigma_y^2\right) \qquad \text{per mesi } t = 1, \dots, n$$

$$\alpha_t \sim N\left(\gamma_0 + \gamma_1 età + \gamma_2 piano + \gamma_3 zona + \gamma_4 sesso\,, \sigma_\alpha^2\right) \qquad \text{per Sim } j = 1, \dots, J$$

Le variabili introdotte dovrebbero diminuire la varianza tra i gruppi σ_α^2.

Per σ_y e σ_α si utilizzano delle a priori uniformi in 0 – 100, mentre per β_1, β_2, β_3, γ_0, γ_1, γ_2, γ_3, …, γ_{16} una normale con media zero e deviazione standard 100.

La tabella 3.6 riporta le stime a posteriori dei parametri calcolate mediante algoritmi iterativi:

	stime	**st.error**		**stime**	**st.error**
γ_0, intercetta	40.197	0.997	γ_{11}, piano G	-25.588	20.214
γ_1, età 30-40	5.161	0.963	γ_{12}, piano H	27.815	4.679
γ_2, età 40-50	0.948	1.040	γ_{13}, sud	-4.139	1.268
γ_3, età 50-60	0.0533	1.319	γ_{14}, nord	-5.091	0.868
γ_4, età 60-70	0.928	2.178	γ_{15}, isole	-0.740	1.237
γ_5, età +70	-3.068	4.341	γ_{16}, femmine	-1.349	0.825
$\gamma_{6,}$ piano B	-0.275	1.353	β_1, mese	-0.838	0.016
$\gamma_{7,}$ piano C	-1.612	2.649	β_2, MSM	0.174	0.002
$\gamma_{8,}$ piano D	0.417	2.748	β_3, MMS	0.633	0.024
γ_9, piano E	-5.013	0.818	σ_y	31.112	0.048
γ_{10}, piano F	-5.393	9.637	σ_α	38.822	0.274

Tabella 3.6 Stime e st. error per i coefficienti del modello gerarchico con le sole variabili al primo livello

	minimo	**1 quartile**	**mediana**	**media**	**3 quartile**	**massimo**
α_j	-219.715	-21.688	-12.246	0	7.718	770.752

Tabella 3.7 Sintesi delle intercette per il multilivello con intercetta variabile e predittori al primo e secondo strato

Anche se alcune modalità delle nuove variabili risultano significative il modello non ha subito sostanziali miglioramenti ($\hat{\sigma}_\alpha$ vale ora 38.8 mentre prima era 39).

Questo può essere spiegato dalla formula di inizio capitolo:

stima degli $\alpha_j \approx \frac{\frac{n_j}{\sigma_y^2}}{\frac{n_j}{\sigma_y^2}+\frac{1}{\sigma_\alpha^2}}(stima\ del\ gruppo\ j) + \frac{\frac{1}{\sigma_\alpha^2}}{\frac{n_j}{\sigma_y^2}+\frac{1}{\sigma_\alpha^2}}(stima\ della\ regressione)$

Dove $\frac{\frac{n_j}{\sigma_y^2}}{\frac{n_j}{\sigma_y^2}+\frac{1}{\sigma_\alpha^2}}$ è pari a 0.96, mentre $\frac{\frac{1}{\sigma_\alpha^2}}{\frac{n_j}{\sigma_y^2}+\frac{1}{\sigma_\alpha^2}}$ vale 0.04. Questo spiega perché le stime siano simili a quelle del modello gerarchico senza variabili al secondo livello.

La figura 3.5 mostra il grafico quantile-quantile dei residui e i valori predetti dal modello verso quelli osservati. Si può facilmente notare la somiglianza con i due modelli precedenti.

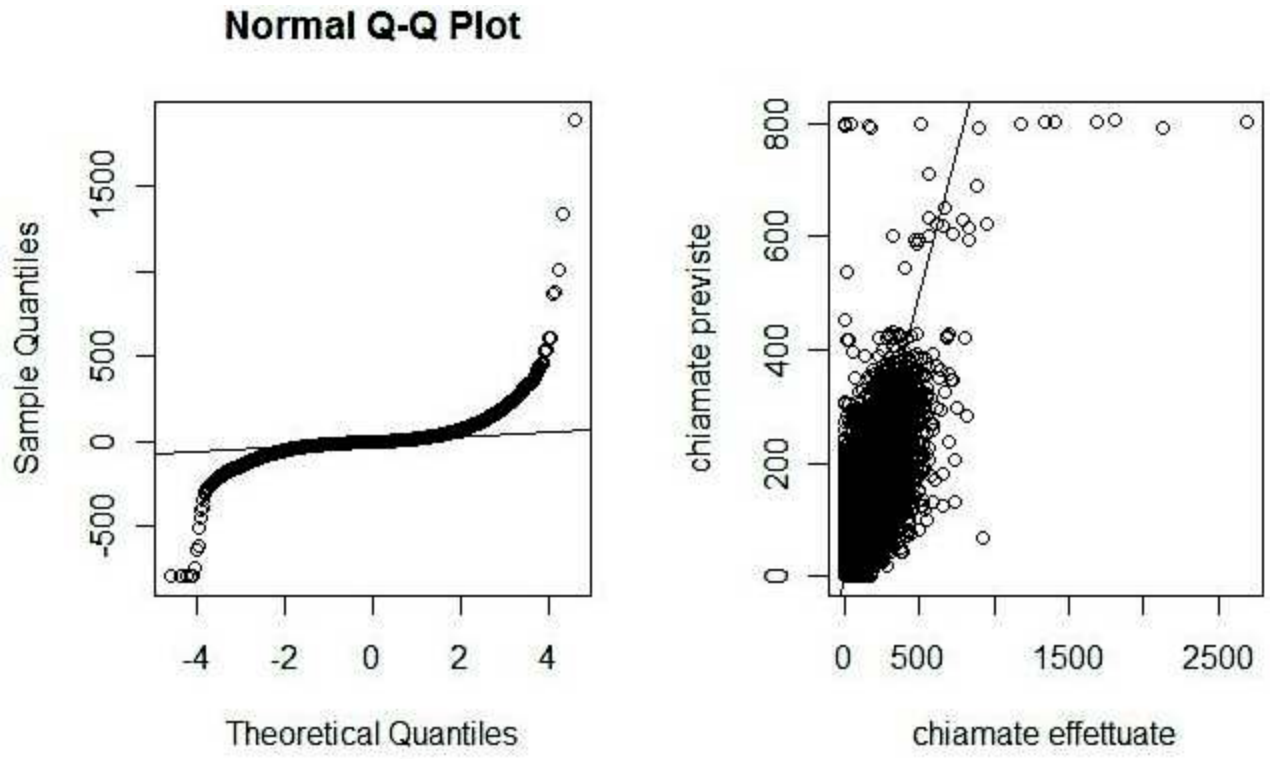

Figura 3.5 (a) Residui del modello (b) valori osservati verso i valori predetti.

La varianza totale $var(Y_{tj}) = \sqrt{\sigma_\alpha^2 + \sigma_y^2} = \sqrt{1507.17 + 967.96} = 49.75$. Anche per questo modello viene stimato il numero di chiamate effettuate nel mese di aprile 2006, i valori predetti vengono confrontati con i valori reali disponibili. La deviazione standard residua $\hat{\sigma}$ vale 31.959. Con l'utilizzo dei predittori al secondo strato non si ha avuto un miglioramento tale da giustificarne il loro utilizzo.

4 MODELLI LINEARI GERARCHICI CON INTERCETTA E COEFFICIENTI VARIABILI

4.1 Introduzione

In questo capitolo si ricorrerà a dei modelli gerarchici dove, non solo l'intercetta, ma anche altri coefficienti possono variare per gruppo. L'idea è rappresentata dalla figura 3.1 di pagina 25.
Nel capitolo precedente si sono sviluppati dei multilivello con solo l'intercetta che varia nei gruppi e si è visto che questi modelli sono molto simili ad una regressione stimata sulle sole osservazioni di ogni singola scheda telefonica, ignorando tutte le altre. Alla luce di questo risultato è ragionevole ridurre il campione di stima con l'obiettivo di alleggerire i calcoli e i tempi di esecuzione dei modelli in **Bugs**.
I multilivello qui di seguito saranno quindi eseguiti su un campione di 1121 schede telefoniche scelte casualmente tra le 12243 Sim che compongono il dataset di stima. I coefficienti sono ottenuti utilizzando delle stime bayesiana, per maggiori dettagli si rimanda al capitolo 8.

4.2 Multilivello con intercetta e coefficiente variabile per mese con predittori al primo strato

Concettualmente con questo modello ogni singola Sim può adesso avere un andamento differente dalle altre schede nei mesi di studio. Nei modelli precedenti il valore per il predittore **mese** era sempre stato negativo. Questo significa che per tutte le schede telefoniche il traffico telefonico stimato doveva per forza diminuire con il passare dei mesi. In questo nuovo modello ciò non è vero, perché per ogni Sim verrà stimato un opportuno valore per la variabile **mese**. Non siamo più obbligati a stimare un singolo valore per tutte le 1121 schede disponibili nel campione di stima.
Escludiamo per il momento le variabile relative al secondo livello gerarchico.
Il modello adottato è il seguente:

$$y_t \sim N\left(\alpha_{j[t]} + \beta_{j[t]} mese_t + \beta_2 SMS_t + \beta_3 MMS_t\,, \sigma_y^2\right) \qquad \text{per mesi } t = 1, \dots, n = 19057$$

$$\begin{pmatrix}\alpha_t \\ \beta_t\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_\alpha \\ \mu_\beta\end{pmatrix}, \begin{pmatrix}\sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2\end{pmatrix}\right) \qquad \text{per Sim } j = 1, \dots, J = 1121$$

Dove σ_α e σ_β sono le deviazioni standard per l'intercetta e per il mese mentre ρ è il parametro di correlazione tra i due.

Per σ_y, σ_α e σ_β si utilizzano delle a priori uniformi in 0 – 100, per β_2, β_3, μ_α, μ_β una normale con media zero e deviazione standard 100, mentre per ρ una uniforme in -1, 1.

Per approfondimenti sull'algoritmo di calcolo e le distribuzioni a priori si rimanda al capitolo 8.

	stima	**st. error**		**Stima**	**st. error**
μ_α, intercetta	37.767	1.496	σ_y	27.142	0.154
μ_β, mese	-0.721	0.091	σ_α	48.197	1.101
β_2, SMS	0.191	0.005	σ_β	2.960	0.076
β_3, MMS	1.111	0.120	ρ	-0.602	0.021

Tabella 4.1 Stime e st. error per i coefficienti per il multilivello con intercetta e coefficiente per mese variabile e predittori al primo strato

	minimo	**1 quartile**	**mediana**	**media**	**3 quartile**	**massimo**
α_j	-140.379	-28.105	-14.425	0	12.222	289.794
β_j	-15.858	-0.769	0.302	0	0.993	16.100

Tabella 4.2 Sintesi degli effetti casuali per il multilivello con intercetta e coefficiente per mese variabile e predittori al primo strato

In questo modello la varianza entro i gruppi ha una deviazione standard stimata pari a 27.1; la stima della deviazione standard per l'intercetta è pari a 48.2 mentre quella del coefficiente variabile per mese vale quasi 3. La stima del parametro di correlazione vale -0.6; μ_α viene stimato pari a 37.8 e μ_β pari a -0.7.

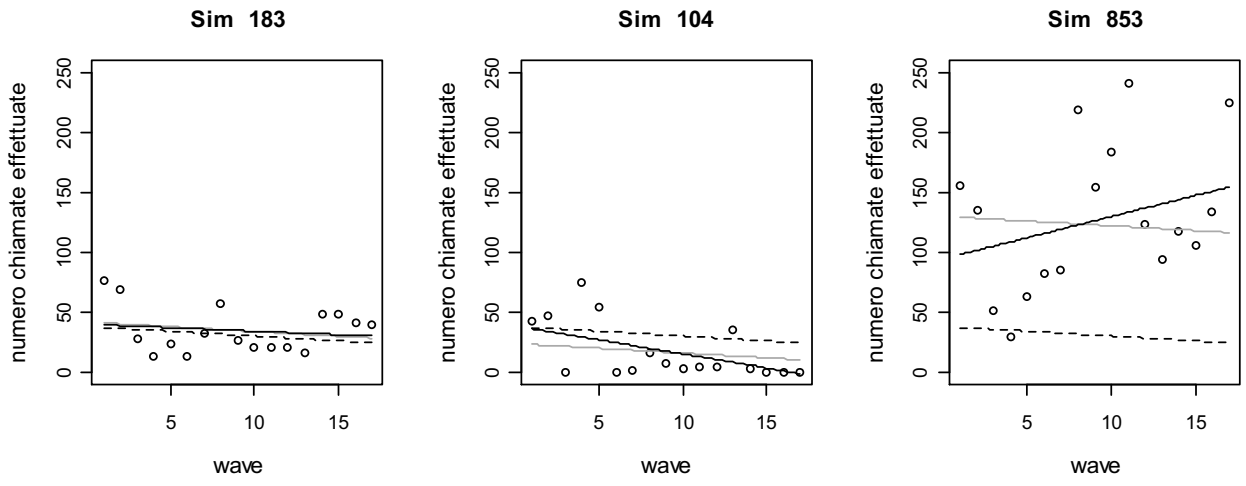

Figura 4.1 Chiamate effettuate nei 17 mesi disponibili per 3 diverse Sim. La linea tratteggiata rappresenta la stima del modello complete pooling, uguale per tutte le Sim. La linea grigia rappresenta la stima del modello no pooling mentre quella nera la stima del multilivello.

La figura 4.1 mostra, per tre differenti Sim che non hanno inviato né sms né mms, il numero di chiamate effettuate nei diciassette mesi, le stime dei modelli complete-pooling (linea tratteggiata, uguale per tutte le

Sim), no-pooling (linea grigia, stessa pendenza per tutte le Sim) e del modello gerarchico appena stimato (linea nera). Nel primo grafico le stime dei modelli sono molto simili. Nel secondo grafico il numero di chiamate diminuisce con il passare dei mesi e la stima del modello gerarchico assume una pendenza negativa. Nel terzo grafico la situazione è opposta, aumentano le chiamate e la retta assume una pendenza positiva.

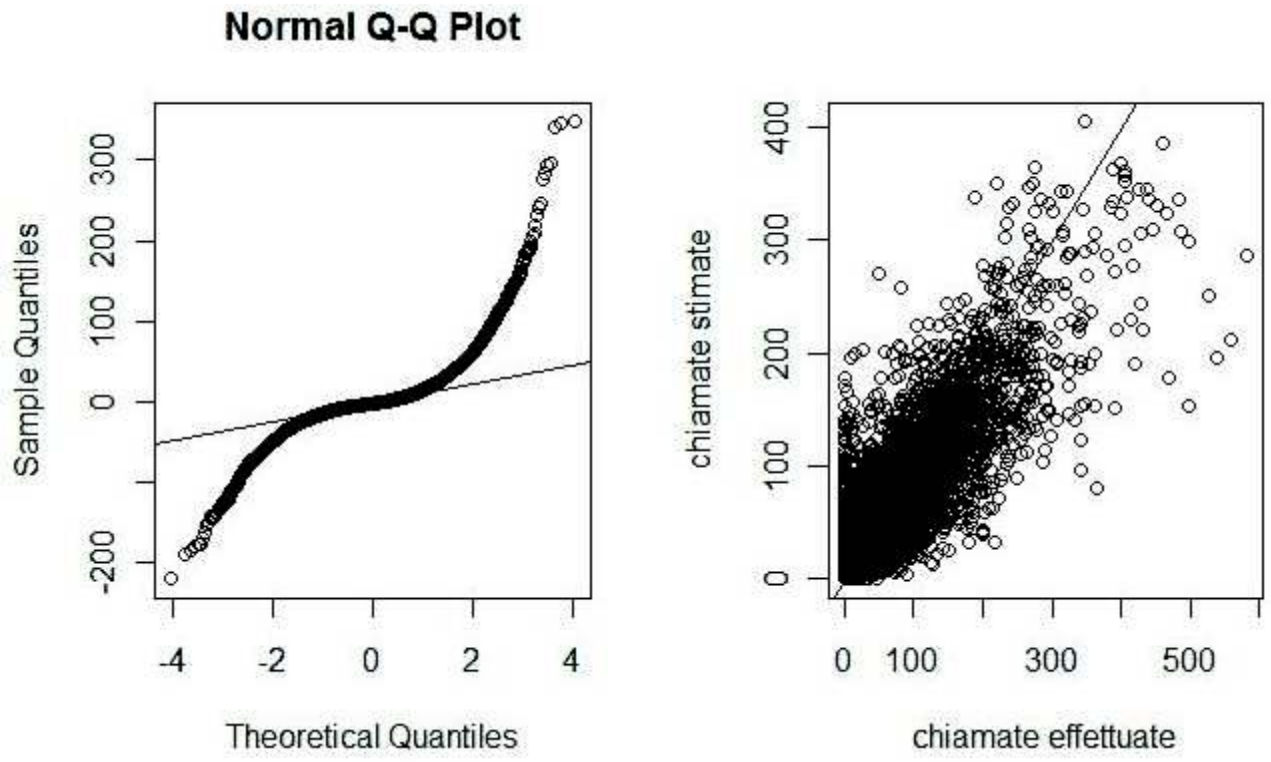

Figura 4.2 (a) qqnorm dei residui, (b) veri valori verso valori predetti

La figura 4.2 mostra il grafico quantile-quantile e il numero di chiamate effettuate dalle Sim verso i valori stimati dal modello. La deviazione standard residua, calcolata confrontando i valori predetti nel mese di aprile 2006, vale 27.680. L'uso del coefficiente variabile per **mese** ha ridotto notevolmente questo valore che risulta essere di gran lunga minore rispetto a quello calcolato nei modelli precedenti. Si vedrà nel capitolo 6 un buon metodo per confrontare tutti i modelli stimati nell'analisi utilizzando il dataset di verifica.

4.3 Multilivello con intercetta e coefficiente variabile per mese con predittori al primo e secondo strato

Vengono ora aggiunte le informazioni riferite all'età, al piano tariffario, alla zona geografica e al sesso dell'utente a cui appartiene la scheda telefonica. Queste variabili saranno collocate nel secondo strato del modello gerarchico. Il modello stimato è il seguente:

$$y_t \sim N\left(\alpha_{j[t]} + \beta_{j[t]} mese_t + \beta_2 SMS_t + \beta_3 MMS_t\,, \sigma_y^2\right) \qquad \text{per mesi } t = 1, \dots, n$$

$$\begin{pmatrix}\alpha_t \\ \beta_t\end{pmatrix} \sim N\left(\begin{pmatrix}\gamma_0^\alpha + \gamma_1^\alpha età + \gamma_2^\alpha piano + \gamma_3^\alpha zona + \gamma_4^\alpha sesso \\ \gamma_0^\beta + \gamma_1^\beta età + \gamma_2^\beta piano + \gamma_3^\beta zona + \gamma_4^\beta sesso\end{pmatrix}, \begin{pmatrix}\sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2\end{pmatrix}\right) \qquad \text{per Sim } j = 1, \dots, J$$

Le variabili introdotte sono tutte fattori, con due livelli per i predittori **sesso**, quattro per **zona,** cinque per **piano** (i livelli di **piano** sarebbero otto ma i piani tariffari F, G e H non sono presenti nel campione selezionato) e sei per **età**. Quindi si ottengono quattordici coefficienti da stimare per α_t e quattordici per β_t.

Per σ_y, σ_α e σ_β si utilizzano delle a priori uniformi in 0 – 100, per β_2, β_3, γ_0^α, γ_1^α, …, γ_{14}^α, $\gamma_0^\beta, \gamma_1^\beta, \dots, \gamma_{14}^\beta$ una normale con media zero e deviazione standard 100, mentre per ρ una uniforme in -1, 1.

	stima	**st.error**		**stima**	**st.error**
γ_0^α, intercetta	48.761	3.938	γ_3^β, età 50-60:mese	0.228	0.371
γ_1^α, età 30-40	-2.078	4.234	γ_4^β, età 60-70:mese	0.581	0.607
γ_2^α, età 40-50	-4.347	4.607	γ_5^β, età +70:mese	0.702	1.019
γ_3^α, età 50-60	-5.768	5.443	γ_6^β, piano B:mese	0.836	0.342
γ_4^α, età 60-70	-15.606	8.771	γ_7^β, piano C:mese	2.337	0.816
γ_5^α, età +70	-15.225	15.883	γ_8^β, piano D:mese	0.189	0.857
γ_6^α, piano B	-9.329	5.146	γ_9^β, piano E:mese	0.425	0.264
γ_7^α, piano C	-25.297	13.382	γ_{11}^β, sud:mese	-0.093	0.381
γ_8^α, piano D	-7.356	13.374	γ_{12}^β, nord:mese	-0.447	0.239
γ_9^α, piano E	-8.602	3.874	γ_{13}^β, isole:mese	-0.383	0.289
γ_{11}^α, sud	-10.639	6.188	γ_{14}^β, femmine:mese	0.252	0.222
γ_{12}^α, nord	-6.807	4.117	β_2, SMS	0.192	0.006
γ_{13}^α, isole	-3.712	4.884	β_3, MMS	1.088	0.129
γ_{14}^α, femmine	-0.053	3.554	σ_y	27.138	0.150
γ_0^β, mese	-0.820	0.245	σ_α	48.242	1.092
γ_1^β, età 30-40:mese	0.201	0.271	σ_β	2.961	0.076
γ_2^β, età 40-50:mese	0.050	0.292	ρ	-0.608	0.020

Tabella 4.3 Stime e st.error del multilivello con intercetta e coefficiente per mese variabile e predittori al primo e al secondo strato

	minimo	1 quartile	mediana	media	3 quartile	massimo
α_j	-141.910	-27.808	-13.030	0	13.198	289.385
β_j	-15.547	-0.750	0.318	0	1.078	15.843

Tabella 4.4 Sintesi degli effetti casuali per il multilivello con intercetta e coefficiente per mese variabile e predittori al primo e al secondo strato

I coefficienti γ_j^{β} corrispondono all'interazione tra le variabili al secondo livello con il predittore a coefficiente variabile **mese**.

Le informazioni aggiunte non sembrano aver migliorato il modello, le deviazioni standard sono infatti molto simili al caso precedente senza variabili al secondo livello gerarchico. La figura 4.3 riporta il grafico quantile-quantile e i valori predetti verso i valori reali. Anche questi grafici non si scostano molto da quelli precedenti.

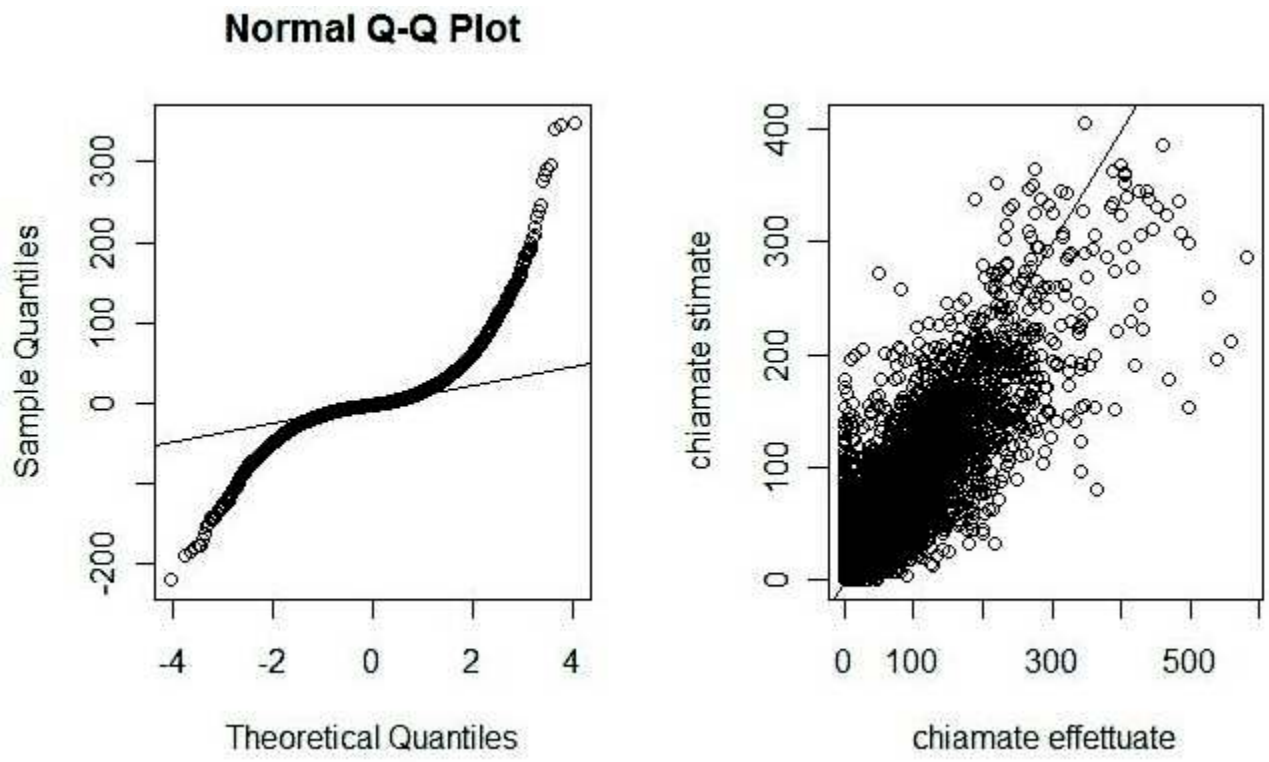

Figura 4.3 (a) qqnorm dei residui, (b) veri valori verso valori predetti

La deviazione standard residua calcolata confrontando i valori predetti nel mese di aprile 2006 vale 27.619. Si può quindi concludere che le variabili **età**, **piano**, **zona** e **sesso** non hanno apportato alcun beneficio al modello. Si noti infatti come le stime per $\alpha_{j[t]}$ e $\beta_{j[t]}$ in figura 4.4 e 4.5 siano molto simili.

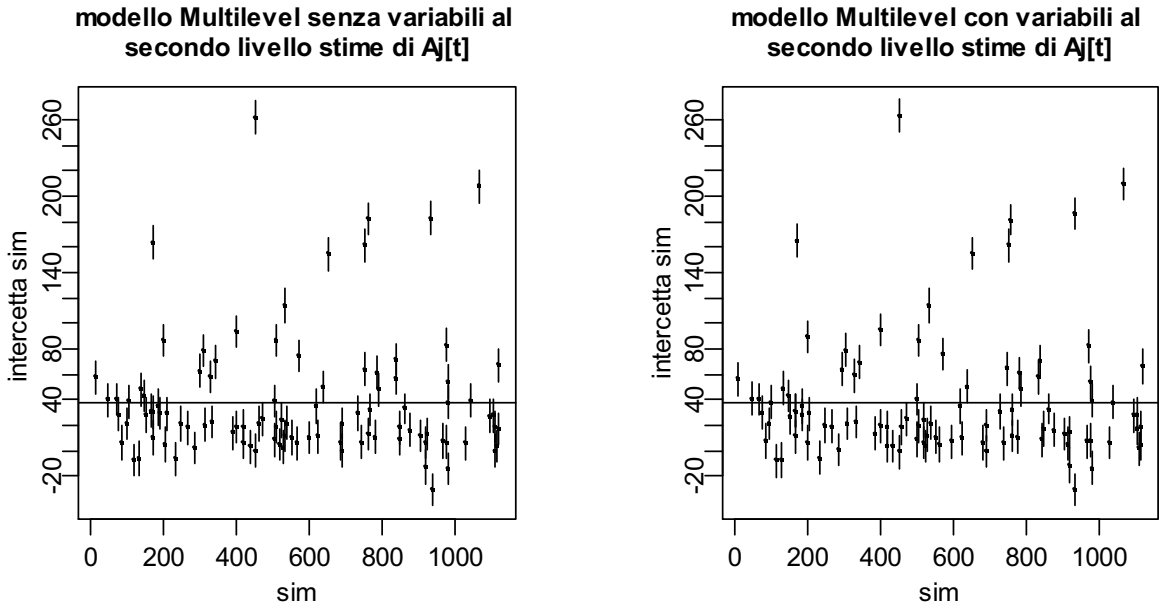

Figura 4.4 (a) stime $\alpha_{j[t]}$ per il modello senza variabili al secondo livello. (b) stime $\alpha_{j[t]}$ per il modello con variabili al secondo livello

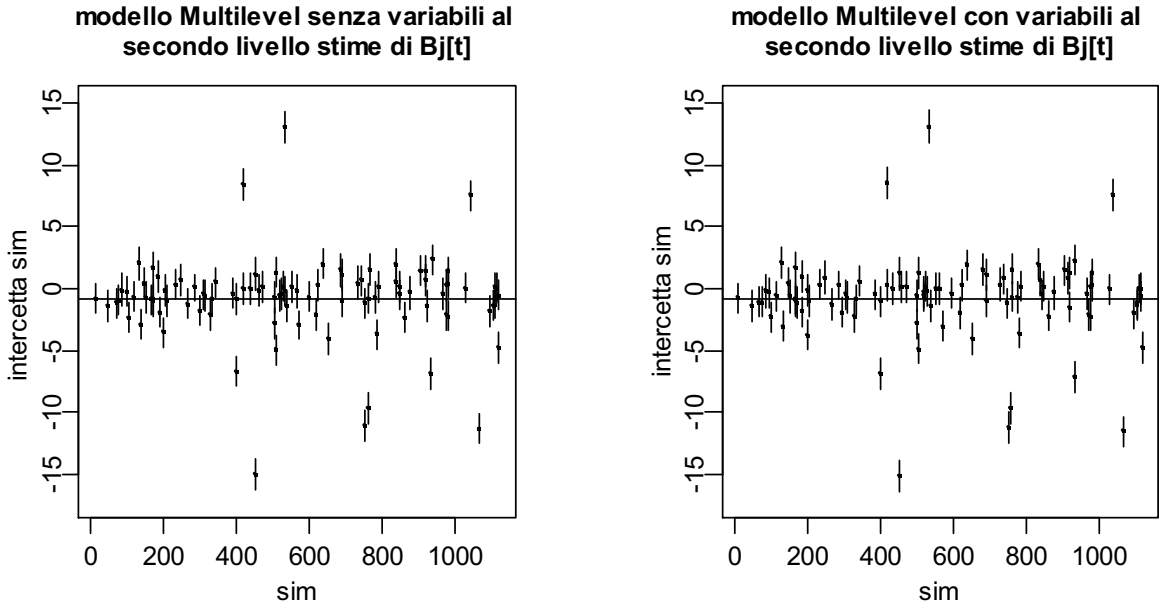

Figura 4.5 (a) stime $\boldsymbol{\beta}_{j[t]}$ per il modello senza variabili al secondo livello. (b) stime $\boldsymbol{\beta}_{j[t]}$ per il modello con variabili al secondo livello

4.4 Multilivello con intercetta e coefficiente variabile per mese sms e mms con predittori al primo strato

Viene ora costruito un nuovo modello in cui possono variare anche i coefficienti per SMS e MMS, cosi da avere quattro termini che variano con i gruppi. Si tralasciano al momento le variabili al secondo livello gerarchico.

$$y_t \sim N\left(\alpha_{j[t]} + \beta_{j[t]} mese_t + \delta_{j[t]} SMS_t + \theta_{j[t]} MMS_t, \sigma_y^2\right) \qquad \text{per mesi } t = 1, \dots, n$$

$$\begin{pmatrix} \alpha_t \\ \beta_t \\ \delta_t \\ \theta_t \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \\ \mu_\delta \\ \mu_\theta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho_{\alpha\beta}\sigma_\alpha\sigma_\beta & \rho_{\alpha\delta}\sigma_\alpha\sigma_\delta & \rho_{\alpha\theta}\sigma_\alpha\sigma_\theta \\ \rho_{\alpha\beta}\sigma_\alpha\sigma_\beta & \sigma_\beta^2 & \rho_{\beta\delta}\sigma_\delta\sigma_\beta & \rho_{\beta\theta}\sigma_\theta\sigma_\beta \\ \rho_{\alpha\delta}\sigma_\alpha\sigma_\delta & \rho_{\beta\delta}\sigma_\delta\sigma_\beta & \sigma_\delta^2 & \rho_{\delta\theta}\sigma_\delta\sigma_\theta \\ \rho_{\alpha\theta}\sigma_\alpha\sigma_\theta & \rho_{\beta\theta}\sigma_\theta\sigma_\beta & \rho_{\delta\theta}\sigma_\delta\sigma_\theta & \sigma_\theta^2 \end{pmatrix} \right) \qquad \text{per Sim } j = 1, \dots, J$$

Per $\sigma_y, \sigma_\alpha, \sigma_\delta, \sigma_\beta$ e σ_θ si utilizzano delle a priori uniformi in 0 – 100, per μ_α, β_3, μ_β, γ_1^α, μ_δ, e μ_θ una normale con media zero e deviazione standard 100, mentre per $\rho_{\alpha\beta}, \rho_{\alpha\delta}, \rho_{\alpha\theta}, \rho_{\beta\delta}, \rho_{\beta\theta}$ e $\rho_{\delta\theta}$ una uniforme in -1, 1.

Quando i coefficienti che variano per gruppo sono più di due diventa difficile stimare i parametri di correlazione ρ_{kl}. Essi infatti devono assumere valori tra -1 e 1 mentre la matrice delle covarianze $\sum$ deve essere definita positiva. Per agevolare le operazioni si ricorre al modello inverse-Wishart.

Modello inverse-Wishart

Questo modello ha il vantaggio di essere computazionalmente conveniente anche se di difficile interpretazione. Nel modello $\sum \sim Inv - Wishart_{K+1}(I)$ che comprende i gradi di libertà $K + 1$(dove K è il numero di coefficienti che variano) e il parametro di scala (nel nostro caso la matrice identità K x K).

$\sum$ è una matrice K x K i cui elementi sulla diagonale principale corrispondono alle varianze σ_{kk}^2 mentre gli altri elementi valgono $\rho_{kl}\sigma_k\sigma_l$. Se si impostano i gradi di libertà pari a $K + 1$ si ottiene una distribuzione uniforme per ogni singolo parametro di correlazione ρ_{kl} (si assume che essi siano ugualmente distribuiti fra -1 e 1).

Nell'analisi verrà utilizzata una variante di questo modello chiamato modello Scaled inverse-Wishart.

Modello Scaled inverse-Wishart

Impostare i gradi di libertà pari a $K + 1$ è sensato per stimare i coefficienti di correlazione, ma così facendo si ottengono delle forti restrizioni sui parametri σ_k. Questo risulta essere un problema perché è necessario che i σ_k vengano stimati dai dati. Cambiando i gradi di libertà sarebbe possibile stimare σ_k più liberamente ma si otterrebbero delle restrizioni su ρ_{kl}.

Si provi ad evitare l'ostacolo introducendo un'estensione del modello *inverse-Wishart* con un nuovo vettore di scala ε_k: $\sum = Diag(\varepsilon)QDiag(\varepsilon)$ dove $Q \sim Inv - Wishart_{K+1}(I)$.

Le varianze, che corrispondono agli elementi della diagonale principale, vengono quindi moltiplicati per un appropriato fattore di scala ε:

$$\sigma_k^2 = \sum_{kk} = \varepsilon_k^2 Q_{kk} \qquad \text{per } k = 1, \ldots, K$$

e le covarianze diventano:

$$\sum_{kl} = \varepsilon_k \varepsilon_l Q_{kk} \qquad \text{per } k, l = 1, \ldots, K$$

Espressi in termini di deviazione standard: $\sigma_k = |\varepsilon_k|\sqrt{Q_{kk}}$ e le correlazioni $\rho_{kl} = \sum_{kl}/(\sigma_k \sigma_l)$.
I parametri ε e Q non possono essere interpretati separatamente, anche se gli elementi di interesse restano σ_k e ρ_{kl}. Anche il modello *Scaled inverse-Wishart* comporta una distribuzione uniforme per i parametri di correlazione.

	stima	**st. error**		**stima**	**st. error**
μ_α, intercetta	33.204	1.317	σ_θ	3.645	0.021
μ_β, mese	-0.597	0.083	$\rho_{\alpha\beta}$	-0.577	0.011
μ_δ, SMS	0.763	0.041	$\rho_{\alpha\delta}$	0.147	0.009
μ_θ, MMS	1.880	0.271	$\rho_{\alpha\theta}$	0.289	0.012
σ_y	24.245	0.172	$\sigma_{\beta\delta}$	-0.146	0.017
σ_α	41.534	1.103	$\sigma_{\beta\theta}$	-0.032	0.015
σ_β	2.463	0.048	$\rho_{\delta\theta}$	0.240	0.010
σ_δ	0.897	0.032			

Tabella 4.5 Stime e st.error del multilivello con 4 coefficienti variabili e predittori al primo livello gerarchico

	minimo	**1 quartile**	**mediana**	**media**	**3 quartile**	**massimo**
α_j	-58.059	-24.135	-12.855	0	9.963	279.100
β_j	-12.440	-0.607	0.241	0	0.843	12.599
δ_j	-1.575	-0.322	-0.059	0	0.126	5.076
θ_j	-5.729	-0.888	-0.473	0	0.415	19.230

Tabella 4.6 Sintesi degli effetti casuali per il multilivello con 4 coefficienti variabili e predittori al primo livello gerarchico

La figura 4.6 riporta il grafico quantile-quantile dei residui e i valori stimati verso i valori reali. Facendo un confronto con la figura 4.2 non sembrano esserci sostanziali differenze per preferire questo modello al precedente. La deviazione standard residua calcolata per i valori relativi ad aprile 2006 vale 27.799 (mentre per il modello con solo due coefficienti variabili era 27.680).

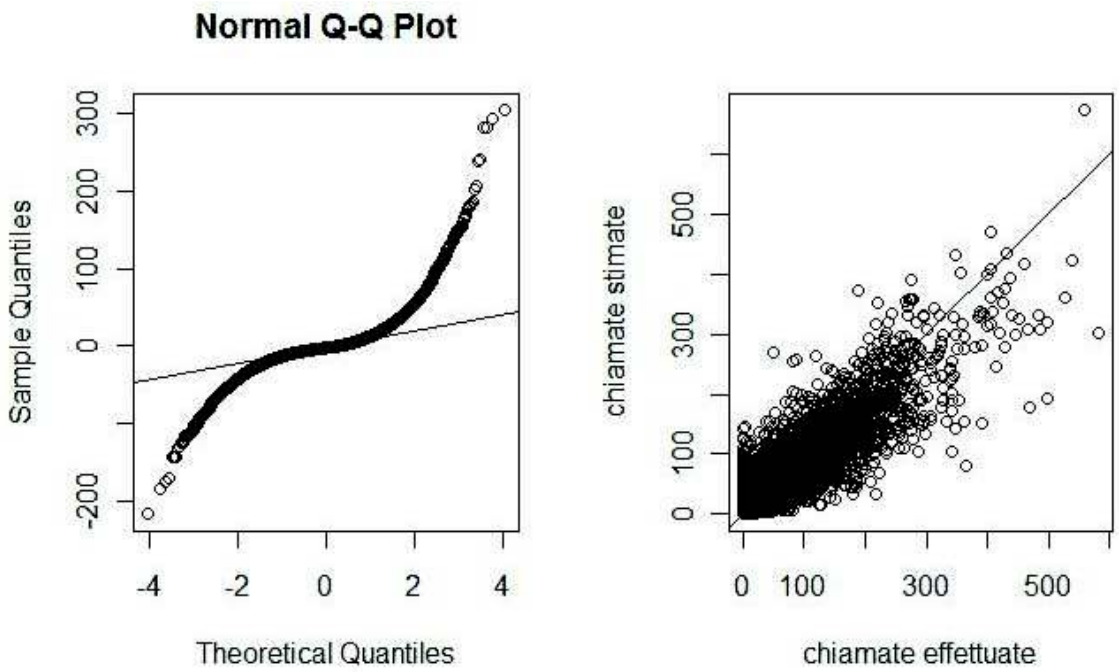

Figura 4.6 (a) residui del modello, (b) veri valori verso valori predetti

4.5 Multilivello con intercetta e coefficiente variabile per mese sms e mms con predittori al primo e al secondo strato

Vengono ora aggiunti i predittori al secondo livello:

$$y_t \sim N\left(\alpha_{j[t]} + \beta_{j[t]} mese_t + \delta_{j[t]} SMS_t + \theta_{j[t]} MMS_t\,, \sigma_y^2\right) \qquad \text{per mesi } t = 1, \dots, n$$

$$\alpha t \beta t \delta t \theta t \sim N$$

$$\left(\begin{pmatrix} \gamma_0^\alpha + \gamma_1^\alpha età + \gamma_2^\alpha piano + \gamma_3^\alpha zona + \gamma_4^\alpha sesso \\ \gamma_0^\beta + \gamma_1^\beta età + \gamma_2^\beta piano + \gamma_3^\beta zona + \gamma_4^\beta sesso \\ \gamma_0^\delta + \gamma_1^\delta età + \gamma_2^\delta piano + \gamma_3^\delta zona + \gamma_4^\delta sesso \\ \gamma_0^\theta + \gamma_1^\theta età + \gamma_2^\theta piano + \gamma_3^\theta zona + \gamma_4^\theta sesso \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho_{\alpha\beta}\sigma_\alpha\sigma_\beta & \rho_{\alpha\delta}\sigma_\alpha\sigma_\delta & \rho_{\alpha\theta}\sigma_\alpha\sigma_\theta \\ \rho_{\alpha\beta}\sigma_\alpha\sigma_\beta & \sigma_\beta^2 & \rho_{\beta\delta}\sigma_\delta\sigma_\beta & \rho_{\beta\theta}\sigma_\theta\sigma_\beta \\ \rho_{\alpha\delta}\sigma_\alpha\sigma_\delta & \rho_{\beta\delta}\sigma_\delta\sigma_\beta & \sigma_\delta^2 & \rho_{\delta\theta}\sigma_\delta\sigma_\theta \\ \rho_{\alpha\theta}\sigma_\alpha\sigma_\theta & \rho_{\beta\theta}\sigma_\theta\sigma_\beta & \rho_{\delta\theta}\sigma_\delta\sigma_\theta & \sigma_\theta^2 \end{pmatrix} \right)$$

per Sim $j = 1, \dots, J$

Si avranno quindi quattordici iperparametri da stimare per la media di ogni coefficiente variabile.

I coefficienti γ sono ottenuti con l'interazione tra le variabili al primo livello e il rispettivo coefficiente variabile. Per $\sigma_y, \sigma_\alpha, \sigma_\delta, \sigma_\beta$ e σ_θ si utilizzano delle a priori uniformi in 0 – 100, per γ_0^α, …, $\gamma_{14}^\alpha, \gamma_0^\beta$, …, γ_{14}^β, …, $\gamma_0^\delta, \dots, \gamma_{14}^\delta$, γ_0^θ, …, γ_4^θ una normale con media zero e deviazione standard 100, mentre per $\rho_{\alpha\beta}, \rho_{\alpha\delta}, \rho_{\alpha\theta}$, $\rho_{\beta\delta}, \rho_{\beta\theta}$ e $\rho_{\delta\theta}$ una uniforme in -1, 1.

	stima	**st. error**		**stima**	**st. error**
γ_0^α, intercetta	40.191	3.759	γ_6^δ, piano B: SMS	-0.306	0.134
γ_1^α, età 30-40	-0.036	3.393	γ_7^δ, piano C: SMS	-0.369	0.310
γ_2^α, età 40-50	-0.551	3.796	γ_8^δ, piano D: SMS	0.563	0.366

γ_3^{α}, età 50-60	0.484	4.842	γ_9^{δ}, piano E: SMS	-0.153	0.105
γ_4^{α}, età 60-70	-7.050	8.419	γ_{11}^{δ}, sud: SMS	0.210	0.173
γ_5^{α}, età +70	-6.916	13.440	γ_{12}^{δ}, nord: SMS	-0.288	0.100
γ_6^{α}, piano B	-7.041	4.632	γ_{13}^{δ}, isole: SMS	0.092	0.131
γ_7^{α}, piano C	-17.080	11.544	γ_{14}^{δ}, femmine: SMS	-0.039	0.093
γ_8^{α}, piano D	-15.442	11.699	γ_0^{θ}, MMS	3.293	0.817
γ_9^{α}, piano E	-6.124	3.385	γ_1^{θ}, età 30-40: MMS	-0.709	0.689
γ_{11}^{α}, sud	-12.057	5.082	γ_2^{θ}, età 40-50: MMS	-0.687	0.825
γ_{12}^{α}, nord	-5.051	3.294	γ_3^{θ}, età 50-60: MMS	-2.439	1.080
γ_{13}^{α}, isole	-3.818	4.187	γ_4^{θ}, età 60-70: MMS	-2.743	2.486
γ_{14}^{α}, femmine	-0.223	3.125	γ_5^{θ}, età +70: MMS	-1.184	3.608
γ_0^{β}, mese	-0.576	0.236	γ_6^{θ}, piano B: MMS	-1.392	0.942
γ_1^{β}, età 30-40:mese	0.141	0.213	γ_7^{θ}, piano C: MMS	0.029	1.813
γ_2^{β}, età 40-50:mese	-0.035	0.239	γ_8^{θ}, piano D: MMS	-0.470	2.016
γ_3^{β}, età 50-60:mese	0.043	0.304	γ_9^{θ}, piano E: MMS	-0.645	0.766
γ_4^{β}, età 60-70:mese	0.320	0.528	γ_{11}^{θ}, sud: MMS	0.094	1.183
γ_5^{β}, età +70:mese	0.496	0.849	γ_{12}^{θ}, nord: MMS	-0.287	0.731
γ_6^{β}, piano B:mese	0.786	0.292	γ_{13}^{θ}, isole: MMS	-0.304	0.908
γ_7^{β}, piano C:mese	2.027	0.735	γ_{14}^{θ}, femmine: MMS	-0.882	0.630
γ_8^{β}, piano D:mese	0.744	0.733	σ_y	24.233	0.161
γ_9^{β}, piano E:mese	0.344	0.213	σ_{α}	41.462	1.080
γ_{11}^{β}, sud:mese	-0.104	0.320	σ_{β}	2.447	0.075
γ_{12}^{β}, nord:mese	-0.426	0.207	σ_{δ}	0.881	0.015
γ_{13}^{β}, isole:mese	-0.454	0.263	σ_{θ}	3.765	0.154
γ_{14}^{β}, femmine:mese	0.177	0.196	$\rho_{\alpha\beta}$	-0.578	0.011
γ_0^{δ}, SMS	1.018	0.110	$\rho_{\alpha\delta}$	0.134	0.009
γ_1^{δ}, età 30-40: SMS	0.036	0.100	$\rho_{\alpha\theta}$	0.258	0.020
γ_2^{δ}, età 40-50: SMS	-0.210	0.114	$\rho_{\beta\delta}$	-0.148	0.019
γ_3^{δ}, età 50-60: SMS	-0.241	0.146	$\rho_{\beta\theta}$	-0.001	0.008
γ_4^{δ}, età 60-70: SMS	-0.335	0.391	$\rho_{\delta\theta}$	0.198	0.013
γ_5^{δ}, età +70: SMS	-0.695	0.468			

Tabella 4.7 Stime e st.error del multilivello con 4 coefficienti variabili con predittori al primo e al secondo livello gerarchico

	minimo	1 quartile	mediana	media	3 quartile	massimo
α_j	-55.057	-23.567	-11.311	0	11.427	277.388
β_j	-12.540	-0.684	0.248	0	0.905	12.285
δ_j	-1.776	-0.268	-0.050	0	0.135	4.845
θ_j	-6.746	-0.866	-0.440	0	0.460	19.056

Tabella 4.8 Sintesi degli effetti casuali per il multilivello con 4 coefficienti variabili con predittori al primo e al secondo livello gerarchico

La figura 4.7 riporta i residui del modello e il confronto tra valori stimati e valori reali.

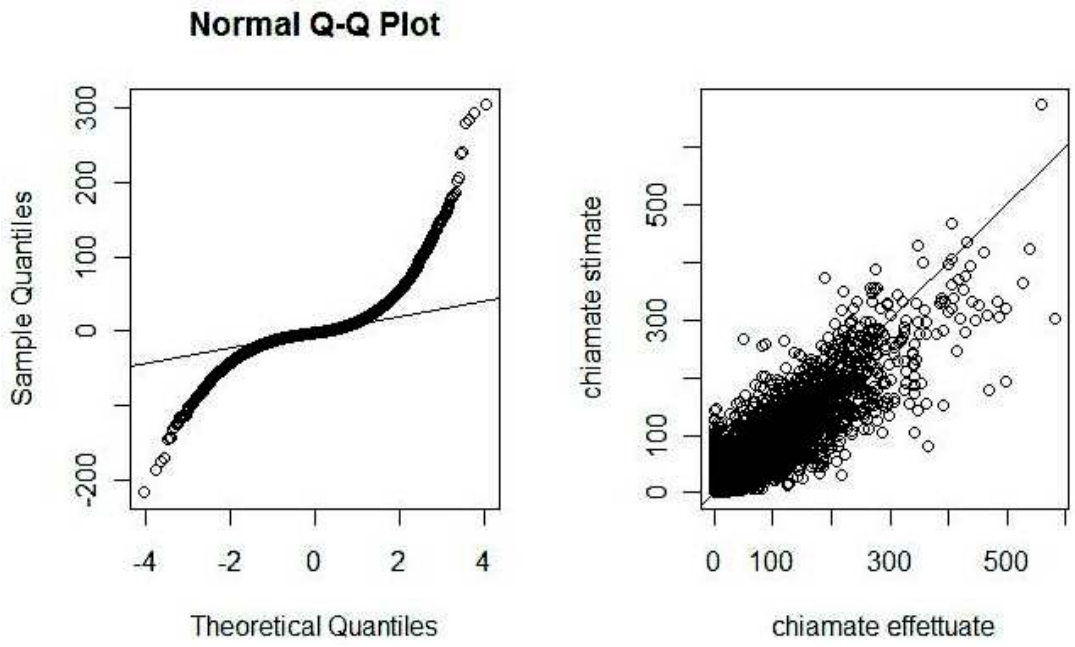

Figura 4.7 (a) residui del modello, (b) veri valori verso valori predetti

Il modello costituisce il caso più complesso a due livelli gerarchici, con intercetta e tre coefficienti variabili.
La deviazione standard residua calcolata per le chiamate del mese di aprile 2006 vale 28.876.
Le figure 4.8 e 4.9 confrontano le stime per $\alpha_{j[t]}$, $\beta_{j[t]}$, $\delta_{j[t]}$, e $\theta_{j[t]}$ per gli ultimi due modelli stimati.
Dalle figure si nota la somiglianza tra le stime per tutti e quattro i coefficienti variabili. In tutti i grafici infatti i punti sono molto bene allineati sulla retta bisettrice.

Non sembrano quindi esserci vantaggi nell'utilizzare le variabili a livello di gruppo. Esse avrebbero dovuto una significativa diminuzione delle varianze, e ciò non si è verificato né per i modelli con due predittori variabili né per quelli con quattro. Si può perciò concludere che le variabili **età**, **piano**, **zona** e **sesso** non ci sono di alcun aiuto per la stima del numero di telefonate effettuate dalle schede telefoniche. Anche l'utilizzo

di quattro coefficienti variabili non ha portato il miglioramento sperato. A questo punto è preferibile adottare il modello stimato ad inizio capitolo con la sola intercetta e il predittore per mese variabili, visto che gli altri modelli forniscono gli stessi risultati ma sono molto più complessi.

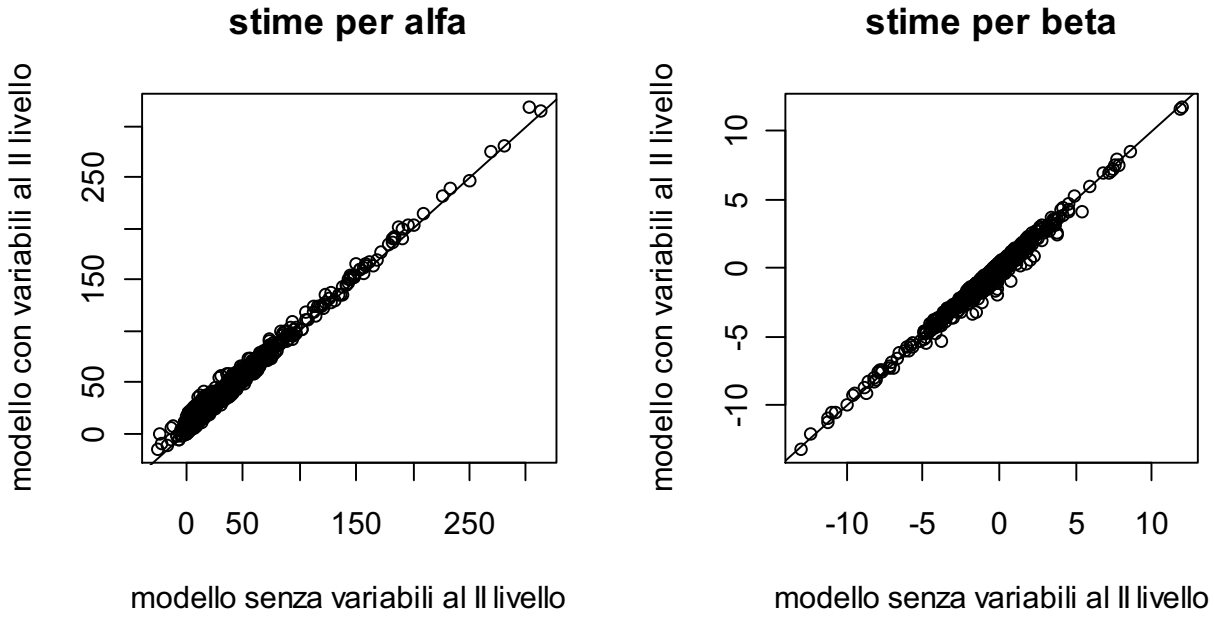

Figura 4.8 (a) stime per alfa per il modello con 4 coefficienti variabili senza e con predittori al secondo livello. (b) stime per beta per il modello con 4 coefficienti variabili senza e con predittori al secondo livello.

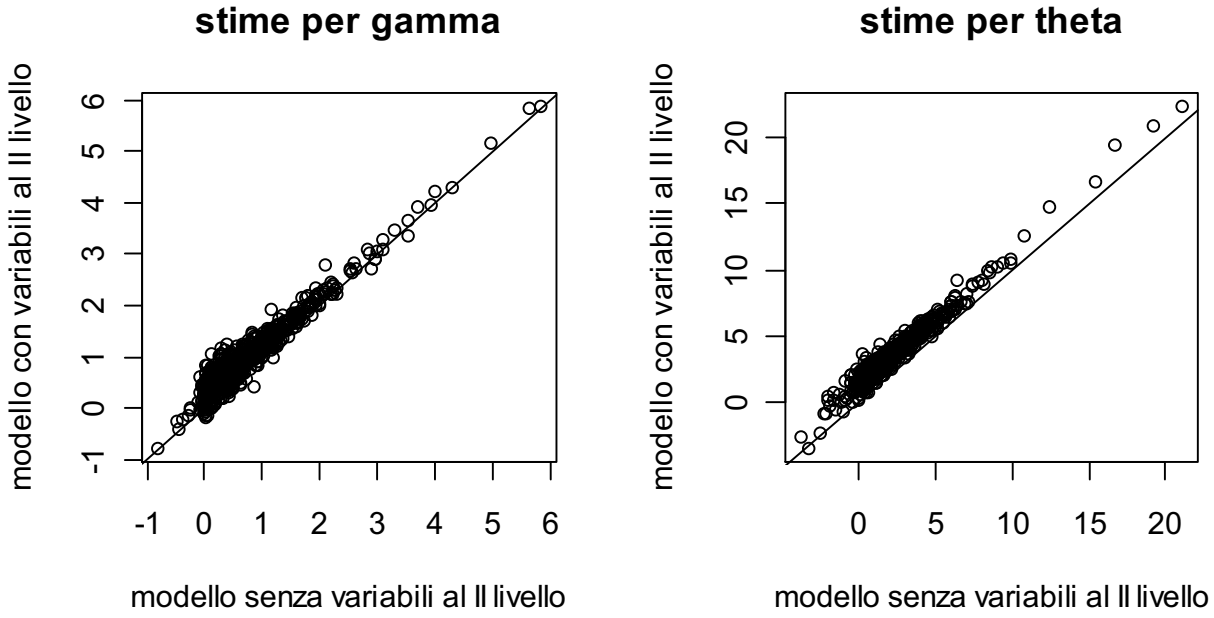

Figura 4.9 (a) stime per gamma per il modello con 4 coefficienti variabili senza e con predittori al secondo livello. (b) stime per theta per il modello con 4 coefficienti variabili senza e con predittori al secondo livello.

5 MODELLI LINEARI GERARCHICI A TRE LIVELLI

5.1 Multilivello a tre livelli con intercetta variabile senza predittori

In questo capitolo si utilizzeranno dei modelli gerarchici a tre livelli. I primi due sono già stati utilizzati in precedenza e rappresentano le singole registrazioni del traffico telefonico e le schede con cui sono state effettuate le chiamate. Il terzo strato consiste nel considerare il possessore delle Sim. Nel campione (formato nel capitolo precedente) ci sono 1121 schede telefoniche che appartengono a 993 clienti diversi. Di cui 872 ne posseggono una, 117 due, 3 utenti ne hanno tre e uno ne ha sei.
Per il primo modello non si utilizzerà alcuna variabile esplicativa ma solo quelle di raggruppamento:

$$y_{tjl} \sim N\left(\alpha_{j[t]}, \sigma_y^2\right) \qquad \text{per } t = 1, \dots, n = 19057$$
$$\alpha_{jl} \sim N\left(\gamma_{l[j]}, \sigma_\alpha^2\right) \qquad \text{per } j = 1, \dots, J = 1121$$
$$\gamma_l \sim N\left(\varphi_\gamma, \sigma_\gamma^2\right) \qquad \text{per } l = 1, \dots, L = 993$$

α_{jl} rappresenta l'intercetta della *j*-esima Sim, γ_l invece l'intercetta dell' l -esimo cliente.

Per σ_y, σ_α e σ_γ si utilizzano delle priori uniformi in 0 – 100, mentre per φ_γ una normale con media zero e deviazione standard 100. La tabella 5.1 riporta le stime a posteriori dei parametri calcolate mediante algoritmi iterativi:

	stima	**st. error**
φ_γ	35.689	1.291
σ_y	32.335	0.176
σ_α	39.686	0.904
σ_γ	2.546	2.068

Tabella 5.2 stime e st. error per le distribuzioni a posteriori dei parametri per il modello gerarchico con intercetta variabile senza predittori

Per questo particolare modello senza variabili esplicative $\alpha_{j[t]}$ è la stima del numero di chiamate effettuate dalla j-esima Sim. Ogni scheda avrà la stessa stima per tutti e diciassette i mesi. γ_l corrisponde alla stima del numero medio di telefonate effettuate da ogni cliente, mentre φ_γ è la stima del numero medio di chiamate per tutti i possessori di Sim.

La stima della deviazione standard per $\alpha_{j[t]}$ vale 39.686: ci sono grandi differenze tra il numero di chiamate medio effettuate tra una scheda e l'altra. La stima della deviazione standard per γ_l vale invece 2.546, significa che tra clienti non ci sono grosse differenze in termini di numero di telefonate.

Sembrerebbe quindi che gli utenti siano omogenei tra loro mentre le Sim no.

La spiegazione sta nel fatto che le stime γ_l sono una media pesata tra le chiamate effettuate da ogni singolo utente e la media delle telefonate di tutti gli utenti con pesi rispettivamente pari a $\frac{1}{\sigma_\gamma^2}/\left(\frac{n_l}{\sigma_\alpha^2}+\frac{1}{\sigma_\gamma^2}\right)$ e $\frac{n_l}{\sigma_\gamma^2}/\left(\frac{n_j}{\sigma_\alpha^2}+\frac{1}{\sigma_\gamma^2}\right)$ dove n_lè il numero di Sim per ogni cliente. Dato che la maggior parte degli utenti possiede una o al massimo due schede (n_l vale 1 o 2) i pesi valgono 0.09 e 0.91, di conseguenza la stima degli γ_l è prossima alla media generale.

Così si spiega perché le stime del numero medio di chiamate per gli utenti siano molto simili tra loro.

	minimo	**I quartile**	**mediana**	**media**	**III quartile**	**massimo**
α_j	2.923	10.610	22.180	35.560	47.160	335.900
γ_l	30.888	33.669	33.762	33.762	33.855	35.932

Tabella 5.2 Sintesi delle distribuzioni di $\boldsymbol{\alpha_j}$ e $\boldsymbol{\gamma_l}$

La tabella 5.2 riporta alcuni indici di sintesi per $\alpha_{j[t]}$ e $\gamma_{l[j]}$; si può notare come i valori di $\gamma_{l[j]}$ siano molto più concentrati attorno alla loro media rispetto a quelli di $\alpha_{j[t]}$ a dimostrazione del fatto che il numero medio di chiamate effettuate dai clienti è simile, mentre ciò non accade per le Sim.

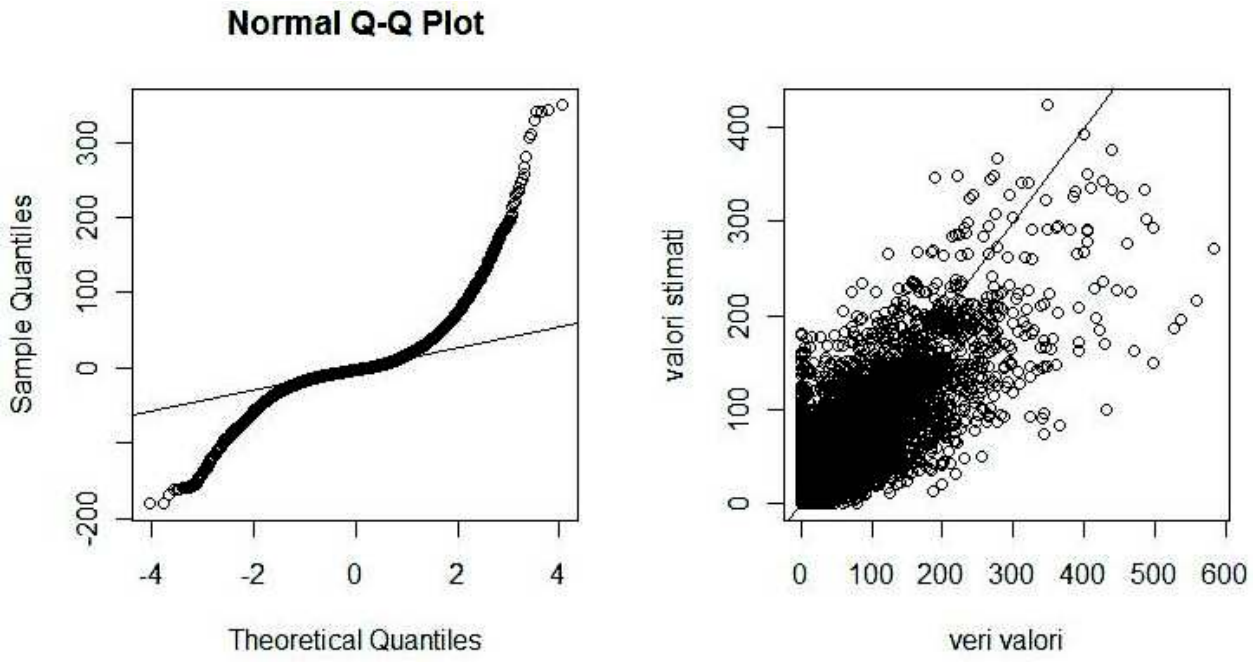

Figura 5.1 (a) residui del modello (b) valori stimati verso valori reali

La figura 5.1 mostra il grafico quantile-quantile dei residui e il confronto tra valori stimati e valori reali. I residui non hanno un andamento normale. La deviazione standard residua calcolata stimando i valori per il mese di aprile 2006 vale 33.296. Per questo modello le stime del traffico sono le stesse per tutti i mesi e coincidono con $\alpha_{j[t]}$.

La variabilità totale è data dalla somma di σ_y^2, σ_α^2 e σ_γ^2. Della varianza totale,($\sigma_\gamma^2/\left(\sigma_y^2+\sigma_\alpha^2+\sigma_\gamma^2\right)$ = 0.003) lo 0.3% è situato al livello utente, ($\sigma_\alpha^2/\left(\sigma_y^2+\sigma_\alpha^2+\sigma_\gamma^2\right)$ = 0.599) il 59.9% è collocato nel livello Sim e il restante 39,8% nel primo livello. Si può inoltre calcolare la correlazione tra due Sim dello stesso utente con la formula $\sigma_\gamma^2/\left(\sigma_\alpha^2+\sigma_\gamma^2\right)$ = 0.004 e tra due osservazioni della stessa Sim tramite $\sigma_\alpha^2/\left(\sigma_y^2+\sigma_\alpha^2\right)$ = 0.601.

Si ha un grande vantaggio nell'utilizzare un modello dove il secondo livello è composto dalle schede telefoniche, mentre non risulta proficuo l'impiego del terzo livello gerarchico formato dagli utenti delle Sim.

5.2 Multilivello a tre livelli con intercetta variabile e predittori in tutti i livelli

Si passa ora a stimare il modello precedente con l'aggiunta delle variabili esplicative. L'informazione contenuta in esse dovrebbe diminuire le deviazioni standard per tutti e tre i livelli.

$$y_{tjl} \sim N\left(\alpha_{j[t]} + \beta_1 mese + \beta_2 SMS + \beta_3 MMS,\ \sigma_y^2\right) \qquad \text{per } t = 1, \dots, n = 19057$$

$$\alpha_{jl} \sim N\left(\gamma_{l[j]} + \theta piano,\ \sigma_\alpha^2\right) \qquad \text{per } j = 1, \dots, J = 1121$$

$$\gamma_l \sim N\left(\varphi_\gamma^0 + \varphi_\gamma^1 zona + \varphi_\gamma^2 sesso,\ \sigma_\gamma^2\right) \qquad \text{per } l = 1, \dots, L = 993$$

Le distribuzioni a priori per σ_y, σ_α e σ_γ sono uniformi in 0 – 100, mentre quelle per $\beta_1,\ \beta_2,\ \beta_3, \vartheta_1,\ \vartheta_2,\ \vartheta_3,\ \vartheta_4, \varphi_\gamma^0,\ \varphi_\gamma^1,\ \varphi_\gamma^2,\ \varphi_\gamma^3,\ \varphi_\gamma^4$ sono normali con media 0 e deviazione standard 100. Le stime a posteriori calcolate mediante algoritmi iterativi sono riportate in tabella 5.3:

	stima	st. error		stima	st. error
β_1, mese	-0.717	0.044	φ_γ^1, sud	-1.833	5.173
β_2, SMS	0.196	0.006	φ_γ^2, nord	7.299	2.963
β_3, MMS	1.350	0.137	φ_γ^3, isole	2.124	3.436
ϑ_1, piano B	-2.221	3.958	φ_γ^4, femmine	-0.947	3.043
ϑ_2, piano C	-1.267	10.121	σ_y	30.977	0.167
ϑ_3, piano D	-4.114	9.917	σ_α	38.299	0.888
ϑ_4, piano E	-3.500	2.903	σ_γ	2.779	1.717
φ_γ^0, intercetta	35.061	2.559			

Tabella 5.3 stime e st. error per le distribuzioni a posteriori dei parametri del modello gerarchico ad intercetta variabile e predittori in tutti i livelli

La variabile **età** non è stata utilizzata perché le stime presentavano un errore standard troppo elevato e non è stato perciò possibile assegnare loro un preciso valore.

Con l'introduzione delle variabili nei tre livelli gerarchici le stime per σ_y, σ_α, e σ_γ sono diminuite leggermente rispetto al caso precedente.

La figura 5.2 confronta i valori di α_j e γ_l per i due modelli. Nel primo riquadro la maggior parte dei punti è allineata sulla bisettrice. Per il multilivello senza variabili le stime sono tutte positive, mentre per il modello con i predittori le stime di α_j assumono talvolta valori negativi e risultano perciò minori dei rispettivi α_j per il multilivello senza variabili. Nel secondo riquadro sull'asse delle ascisse i punti sono ben concentrati attorno al valore 36, mentre sull'asse delle ordinate le stime di $\gamma_{l[j]}$ per il modello con variabili esplicative si raggruppano in otto diversi valori distinti. Partendo dall'alto la prima coppia di gruppi appartiene al nord, la seconda alle isole, la terza al centro e l'ultima al sud. Per ogni regione il gruppo più in alto corrisponde ai maschi e l'altro alle femmine. Le differenze sono date dai segni dei coefficienti del modello. Le stime delle

femmine, che hanno un segno negativo nel modello, risultano sempre minori dei maschi indipendentemente dalle regioni. Nord e isole che hanno segni positivi risultano maggiori del centro e il sud invece minore.

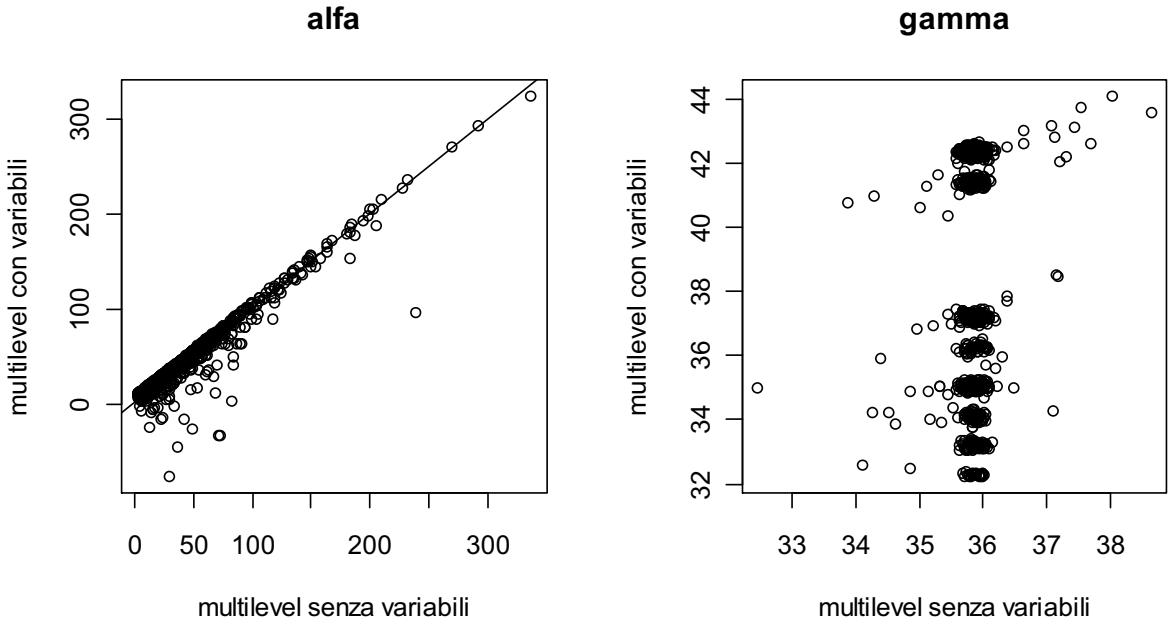

Figura 5.2 Confronto stime alfa (e) e gamma (f) per i multilivello con e senza variabili

La figura 5.3 mostra il grafico quantile-quantile dei residui e i valori stimati confrontati con i valori reali.
La deviazione standard residua calcolata per le chiamate effettuate nel mese di aprile 2006 vale 31.277. Nel modello senza predittori la stessa quantità valeva 33.296. Le variabili hanno quindi portato un miglioramento nella stima del numero di telefonate effettuate.

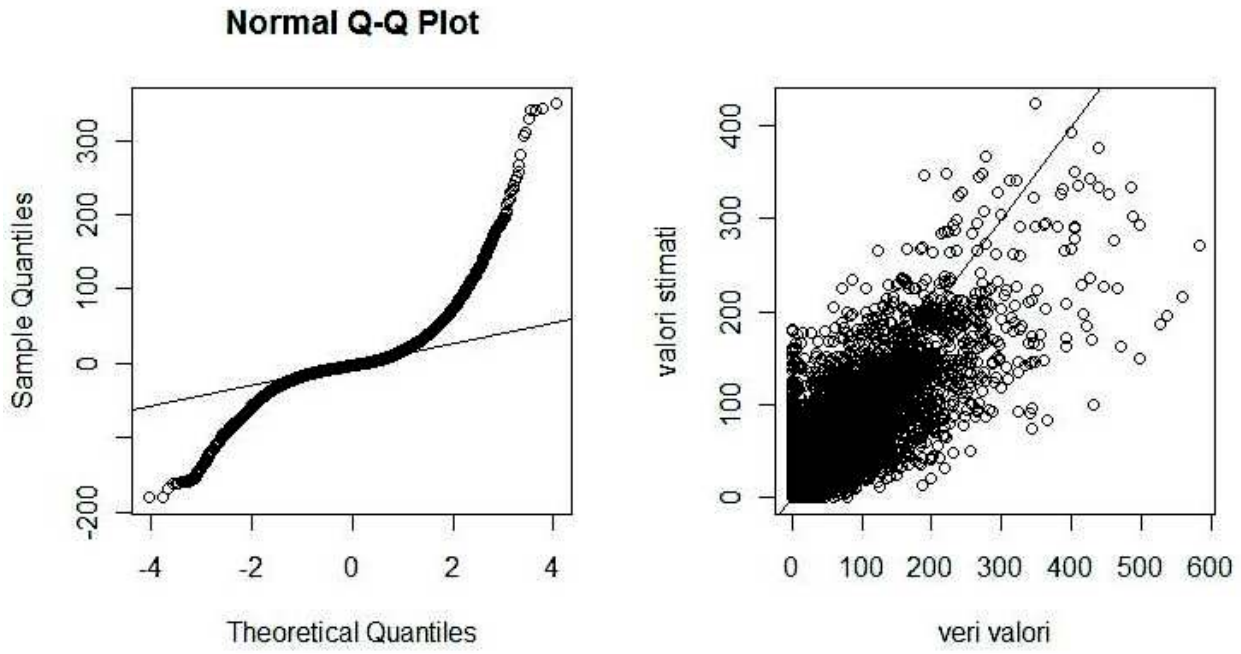

Figura 5.3 (a) residui del modello (b) valori stimati verso valori reali

5.3 Multilivello a tre livelli con intercette e coefficiente variabile per mese con predittori al primo livello

Si utilizzano le sole variabili **mese**, **SMS** e **MMS** al primo livello per costruire un modello a tre strati dove variano:

- intercetta al primo livello
- intercetta al secondo livello
- coefficiente per la variabile **mese** al primo livello

$$y_{tjl} \sim N\left(\alpha_{j[t]} + \beta_{j[t]} mese + \beta_2 SMS + \beta_3 MMS,\ \sigma_y^2\right) \qquad \text{per } t = 1, \dots, n = 19057$$

$$\begin{pmatrix}\alpha_{jl}\\ \beta_{jl}\end{pmatrix} \sim N\left(\begin{pmatrix}\gamma^{\alpha}_{l[j]}\\ \gamma^{\beta}_{l[j]}\end{pmatrix}, \begin{pmatrix}\sigma_\alpha^2 & \rho_1\sigma_\alpha\sigma_\beta\\ \rho_1\sigma_\alpha\sigma_\beta & \sigma_\beta^2\end{pmatrix}\right) \qquad \text{per } j = 1, \dots, J = 1121$$

$$\begin{pmatrix}\gamma^{\alpha}_{l}\\ \gamma^{\beta}_{l}\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_{\gamma\alpha}\\ \mu_{\gamma\beta}\end{pmatrix}, \begin{pmatrix}\sigma^2_{\gamma\alpha} & \rho_2\sigma_{\gamma\alpha}\sigma_{\gamma\beta}\\ \rho_2\sigma_{\gamma\alpha}\sigma_{\gamma\beta} & \sigma^2_{\gamma\beta}\end{pmatrix}\right) \qquad \text{per } l = 1, \dots, L = 993$$

Le distribuzioni a priori per $\sigma_y, \sigma_\alpha, \sigma_\beta, \sigma_{\gamma\alpha}$ e $\sigma_{\gamma\beta}$ sono uniformi in 0 – 100, quelle per β_2, β_3, $\mu_{\gamma\alpha}$ e $\mu_{\gamma\beta}$ sono normali con media 0 e deviazione standard 100 e quelle per ρ_1 e ρ_2 sono uniformi in -1 e 1.
Le stime a posteriori calcolate mediante algoritmi iterativi sono riportate in tabella 5.4:

	stima	**st. error**		**stima**	**st. error**
β_2, SMS	0.192	0.006	σ_β	2.964	0.079
β_3, MMS	1.118	0.126	$\sigma_{\gamma\alpha}$	3.691	1.716
$\mu_{\gamma\alpha}$, intercetta	37.676	1.589	$\sigma_{\gamma\beta}$	0.144	0.106
$\mu_{\gamma\beta}$, mese	-0.736	0.096	ρ_1	-0.604	0.021
σ_y	27.139	0.150	ρ_2	-0.314	0.571
σ_α	48.121	1.117			

Tabella 5.4 Stime e st. error per il multilivello con intercette e coefficienti variabili e predittori al primo livello

In questo modello $\mu_{\gamma\alpha}$ rappresenta l'intercetta e $\mu_{\gamma\beta}$ il coefficiente per **mese**. La tabella 5.5 riassume le distribuzioni per gli effetti casuali $\alpha_{j[t]}$, $\beta_{j[t]}$, $\gamma^{\alpha}_{l[j]}$ e $\gamma^{\beta}_{l[j]}$:

	minimo	**I quartile**	**mediana**	**media**	**III quartile**	**massimo**
α_j	-103.000	9.612	23.320	37.770	49.730	327.500
β_j	-16.500	-1.487	-0.435	-0.725	0.260	15.350
γ^{α}_{l}	34.197	37.579	37.679	37.675	37.761	41.030
γ^{β}_{l}	-0.848	-0.740	-0.736	-0.737	-0.733	-0.667

Tabella 5.5 Sintesi delle distribuzioni di $\boldsymbol{\alpha_j}$, $\boldsymbol{\beta_j}$, $\boldsymbol{\gamma^{\alpha}_{l}}$ e $\boldsymbol{\gamma^{\beta}_{l}}$.

$\gamma^{\alpha}_{l[j]}$ è il numero di chiamate medie effettuate da ogni singolo utente. I valori per $\gamma^{\beta}_{l[j]}$ rappresentano il valore **mese** per ogni possessore di Sim, essendo tutti negativi significa che tutti gli utenti con il passare dei mesi diminuiscono il numero di telefonate.

$\alpha_{j[t]}$ è il numero di chiamate medie effettuate da ogni singola Sim, il suo campo di variazione è maggiore rispetto a quello di $\gamma^{\alpha}_{l[j]}$. La situazione si ripete anche per $\beta_{j[t]}$ che rappresenta il valore **mese** per ogni singola scheda telefonica. $\beta_{j[t]}$ assume, a differenza di $\gamma^{\beta}_{l[j]}$ sia valori positivi che negativi, significa che alcune schede hanno un trend negativo mentre altre positivo.

La maggior parte degli utenti ha al massimo due schede e il modello gerarchico non riesce a cogliere eventuali differenze tra le Sim appartenenti alla stessa persona. Il risultato è che i clienti risultano essere tutti omogenei tra loro.

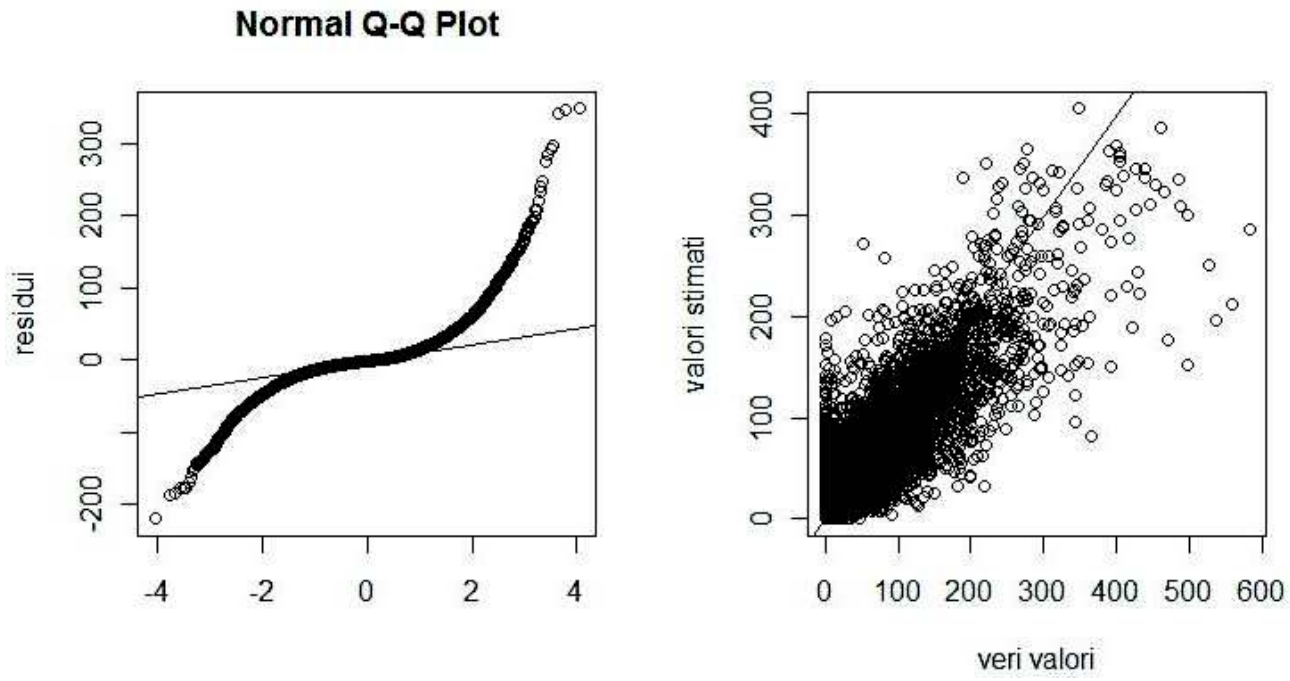

Figura 5.4 (a) qqnorm dei residui (b) valori stimati verso valori reali

La figura 5.4 mostra il grafico quantile-quantile dei residui e i valori predetti verso i valori reali. Come nei modelli precedenti i residui non hanno un andamento normale. La deviazione standard residua calcolata per le chiamate effettuate nel mese di aprile 2006, vale 27.726. Con l'uso del coefficiente variabile per il predittore **mese** si ha avuto un notevole miglioramento rispetto ai casi precedenti.

5.4 Multilivello a tre livelli con intercette e coefficiente variabile e predittori in tutti i livelli

Si aggiungano ora al modello costruito nella sezione 5.3 le variabili relative al secondo e terzo strato gerarchico:

$$y_{tjl} \sim N\left(\alpha_{j[t]} + \beta_{j[t]} mese + \beta_2 SMS + \beta_3 MMS,\ \sigma_y^2\right) \qquad \text{per } t = 1, \dots, n = 19057$$

$$\begin{pmatrix}\alpha_{jl}\\ \beta_{jl}\end{pmatrix} \sim N\left(\begin{pmatrix}\gamma^{\alpha}_{l[j]} + \theta^{\alpha}_1 piano\\ \gamma^{\beta}_{l[j]} + \theta^{\beta}_1 piano\end{pmatrix}, \begin{pmatrix}\sigma^2_{\alpha} & \rho_1\sigma_{\alpha}\sigma_{\beta}\\ \rho_1\sigma_{\alpha}\sigma_{\beta} & \sigma^2_{\beta}\end{pmatrix}\right) \qquad \text{per } j = 1, \dots, J = 1121$$

$$\begin{pmatrix}\gamma^{\alpha}_{l}\\ \gamma^{\beta}_{l}\end{pmatrix} \sim N\left(\begin{pmatrix}\varphi^{\alpha}_0 + \varphi^{\alpha}_1 piano + \varphi^{\alpha}_2\\ \varphi^{\beta}_0 + \varphi^{\beta}_1 piano + \varphi^{\beta}_2\end{pmatrix}, \begin{pmatrix}\sigma^2_{\gamma\alpha} & \rho_2\sigma_{\gamma\alpha}\sigma_{\gamma\beta}\\ \rho_2\sigma_{\gamma\alpha}\sigma_{\gamma\beta} & \sigma^2_{\gamma\beta}\end{pmatrix}\right) \qquad \text{per } l = 1, \dots, L = 993$$

Le distribuzioni a priori per $\sigma_y, \sigma_\alpha, \sigma_\beta, \sigma_{\gamma\alpha}, \sigma_{\gamma\beta}$ sono uniformi in 0 – 100, quelle per $\beta_2, \beta_3, \theta_1^{\alpha}, \theta_2^{\alpha}, \theta_3^{\alpha}, \theta_4^{\alpha}, \theta_1^{\beta}, \theta_2^{\beta}, \theta_3^{\beta}, \theta_4^{\beta}, \varphi_0^{\alpha}, \varphi_1^{\alpha}, \varphi_2^{\alpha}, \varphi_3^{\alpha}, \varphi_4^{\alpha}, \varphi_0^{\beta}, \varphi_1^{\beta}, \varphi_2^{\beta}, \varphi_3^{\beta}, \varphi_4^{\beta}$ sono normali con media 0 e deviazione standard 100, mentre quelle per ρ_1 e ρ_2 sono uniformi in -1,1. Le stime a posteriori calcolate mediante algoritmi iterativi sono riportate in tabella 5.6:

	stima	**st. error**		**stima**	**st. error**
φ_0^{α}, intercetta	33.233	2.721	θ_1^{β}, piano B:mese	0.749	0.327
φ_1^{α}, sud	-2.407	6.441	θ_2^{β}, piano C:mese	2.161	0.836
φ_2^{α}, nord	11.009	3.129	θ_3^{β}, piano D:mese	0.134	0.854
φ_3^{α}, isole	5.736	3.804	θ_4^{β}, piano E:mese	0.383	0.229
φ_4^{α}, femmine	0.712	4.851	β_2, SMS	0.192	0.006
φ_0^{β}, mese	-0.627	0.131	β_3, MMS	1.111	0.124
φ_1^{β}, sud:mese	0.135	0.358	σ_y	27.146	0.150
φ_2^{β}, nord:mese	-0.317	0.192	σ_α	47.929	1.090
φ_3^{β}, isole:mese	-0.312	0.221	σ_β	2.949	0.074
φ_4^{β}, femmine:mese	-0.256	0.271	$\sigma_{\gamma\alpha}$	3.016	1.823
θ_1^{α}, piano B	-8.054	5.242	$\sigma_{\gamma\beta}$	0.194	0.078
θ_2^{α}, piano C	-19.170	13.267	ρ_1	-0.600	0.021
θ_3^{α}, piano D	-4.484	12.748	ρ_2	-0.174	0.622
θ_4^{α}, piano E	-6.168	3.793			

Tabella 5.6 Stime e st. error per il multilivello con intercette e coefficienti variabili e predittori in tutti i livelli

I coefficienti θ_i^{β} e φ_i^{β} corrispondono all'interazione tra le variabili del secondo e del terzo livello con il coefficiente del predittore **mese**. I predittori aggiunti non hanno portato alcuna diminuzione delle varianze nel modello.

	minimo	**I quartile**	**mediana**	**media**	**III quartile**	**massimo**
α_j	-102.300	9.722	23.510	37.740	49.790	327.400
β_j	-16.590	-1.477	-0.397	-0.719	0.254	15.460
γ_l^{α}	30.112	33.368	39.790	39.593	44.278	46.678
γ_l^{β}	-1.241	-0.948	-0.940	-0.887	-0.740	-0.457

Tabella 5.7 Sintesi delle distribuzioni di $\boldsymbol{\alpha_j}$, $\boldsymbol{\beta_j}$, $\boldsymbol{\gamma_l^{\alpha}}$ e $\boldsymbol{\gamma_l^{\beta}}$.

La tabella 5.7 riassume brevemente le distribuzioni per i quattro coefficienti variabili del modello. L'uso delle variabili **zona**, **piano** e **sesso** non ha modificato la distribuzione degli effetti casuali rispetto al caso precedente.

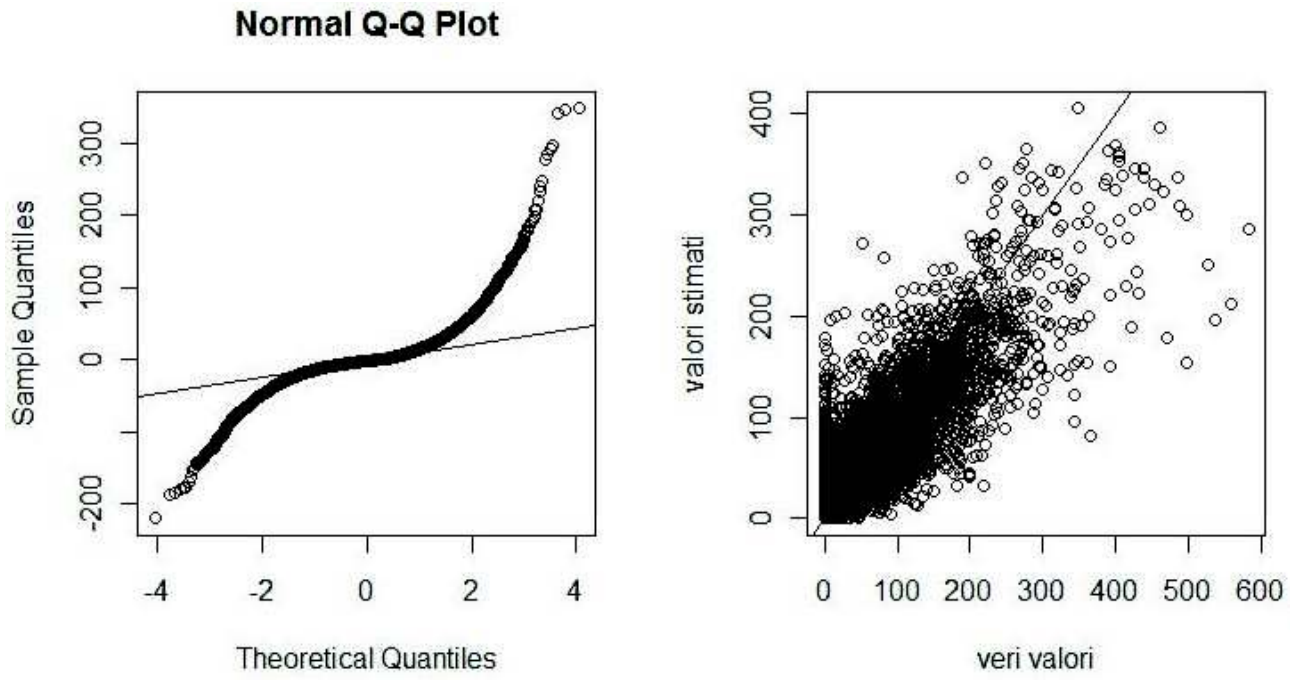

Figura 5.5 (a) qqnorm dei residui (b) valori stimati verso valori reali

La figura 5.5 mostra il grafico quantile-quantile dei residui e i valori predetti verso i valori reali. Entrambi sono simili al modello precedente della sezione 5.3. La deviazione standard calcolata per le chiamate nel mese di aprile 2006 vale 27.683. Si può quindi concludere che le variabili al secondo e terzo livello non consentono di ottenere stime migliori. Pertanto il modello migliore a tre strati risulta essere quello con intercette e coefficiente per **mese** variabile con predittori solo al primo strato.

6 SCELTA DEL MODELLO

6.1 Introduzione

In questo capitolo si confronteranno i modelli sviluppati nell'indagine per selezionarne il migliore sulla base dei due seguenti criteri:

- previsioni del numero di chiamate per il mese di aprile 2006 per le Sim appartenenti al gruppo di stima
- previsioni del numero di chiamate per i mesi da novembre 2004 ad aprile 2006 per le 9515 Sim del gruppo di verifica.

Tutti i modelli verranno stimati utilizzando i dati appartenenti al campione di stima costituita da 1121 schede telefoniche. I valori predetti dai modelli vengono confrontati con quelli reali utilizzando la deviazione standard residua definita come $\hat{\sigma} = \sum_{t=1}^{n} (\hat{y}_t - y_t)^2 / n$.
Le previsioni vengono ottenute facendo delle simulazioni dal modello di riferimento.
Per la regressione lineare si crea una matrice $\tilde{X}$, $\tilde{n}$ per k (dove $\tilde{n}$ è il numero di osservazioni da stimare e k il numero di predittori del modello) contenete i valori per le variabili delle osservazioni da prevedere.
Per ogni valore $\hat{y}_t$ si effettuano 1000 simulazioni da una normale avente per media il prodotto scalare della t-esima riga di $\tilde{X}_{\tilde{n},k}$ per il vettore $\hat{\lambda}$ dei coefficienti stimati dal modello e varianza pari a $\hat{\sigma}_y^2$: $\hat{y}_t \sim N(\hat{\lambda}_k \tilde{X}_t, \hat{\sigma}_y)$. La media dei 1000 numeri generati è la previsione per $\hat{y}_t$. Vengono simulati 1000 dati per cercare di catturare al meglio l'incertezza nella previsione dei nuovi valori.
Il medesimo procedimento verrà adottato anche per la regressione con trasformata radice quadrata.
Per i modelli gerarchici si deve distinguere il caso in cui si prevede il numero di chiamate per le Sim già utilizzate nel dataset di stima e quelle invece "nuove". Nel primo caso si utilizzano gli effetti casuali già stimati dal modello, mentre per le schede appartenenti al dataset di verifica è necessario simulare prima i valori per i coefficienti variabili e solo poi il traffico in uscita.
Per i multilivello a due strati quando si prevede il numero di telefonate effettuate nel mese di aprile 2006 le stime degli effetti casuali (α_j e se previsti dal modello anche β_j, δ_j e θ_j) sono già state calcolate dal modello e sarà quindi necessario simulare solo i valori della variabile risposta da una normale con media ottenuta dal prodotto scalare della t-esima riga di $\tilde{X}_{\tilde{n},k}$ per il vettore $\hat{\lambda}$ dei coefficienti e varianza $\hat{\sigma}_y^2$: $\hat{y}_t \sim N(\hat{\lambda}_k \tilde{X}_t, \hat{\sigma}_y)$. ($\tilde{X}_{\tilde{n},k}$ contiene sole le informazioni relative alle variabili del primo strato: **mese**, **SMS** e **MMS**; qualora fossero incluse nel modello).

Per le Sim del dataset di verifica è invece necessario stimare prima gli effetti casuali tramite simulazione da una normale: $\hat{\alpha}_j \sim N(\hat{\gamma}_l^{\alpha}\ \widetilde{U}_j, \hat{\sigma}_{\alpha})$ ed eventualmente $\hat{\beta}_j \sim N\left(\hat{\gamma}_l^{\beta}\ \widetilde{U}_j, \hat{\sigma}_{\beta}\right)$, $\hat{\delta}_j \sim N(\hat{\gamma}_l^{\delta} \widetilde{U}_j, \hat{\sigma}_{\delta})$ e $\hat{\theta}_j \sim N(\hat{\gamma}_l^{\theta} \widetilde{U}_j, \hat{\sigma}_{\theta})$. ($\widetilde{U}_{m,l}$ è una matrice che contiene le informazioni per le l variabili del secondo livello gerarchico per le m Sim da simulare e $\hat{\gamma}$ è la stima dei coefficienti per le variabili al secondo strato).

Si effettuano poi delle simulazioni per il traffico telefonico da una normale con media ottenuta dal prodotto scalare della t-esima riga di $\tilde{X}_{\tilde{n},k}$ per il vettore $\hat{\lambda}$ dei coefficienti e varianza $\hat{\sigma}_y^2$: $\hat{y}_t \sim N(\hat{\lambda}_k \tilde{X}_t, \hat{\sigma}_y)$. ($\tilde{X}_{\tilde{n},k}$ contiene sole le informazioni relative alle variabili del primo strato).

Per ogni nuova osservazione da prevedere vengono generati 1000 valori di $\hat{\alpha}_j$ e se previsti anche 1000 di $\hat{\beta}_j$, di $\hat{\delta}_j$ e di $\hat{\theta}_j$. Vengono poi simulati 1000 valori per ogni $\hat{y}_t$. Gli elementi del vettore $\hat{\lambda}$ contengono le stime dei coefficienti variabili ottenuti dalle precedenti simulazioni. La media di quest'ultimi numeri generati è la stima del numero di telefonate effettuate dalla nuova Sim. Così facendo si riesce a catturare la propagazione dell'incertezza nelle varie fasi della previsione.

Lo stesso procedimento si ripete per i modelli gerarchici a tre strati: per le previsioni del mese di aprile 2006 simuliamo solo il traffico in uscita utilizzando le stime già disponibili dal modello: $\hat{y}_t \sim N(\hat{\lambda}_k \tilde{X}_t, \hat{\sigma}_y)$, mentre per le nuove schede sono necessarie tre simulazioni. Prima si stimano gli effetti casuali nel terzo livello: $\hat{\gamma}_l^{\alpha} \sim N(\hat{\varphi}_k^{\alpha} \tilde{P}_l, \hat{\sigma}_{\gamma}^{\alpha})$ e se previsto $\hat{\gamma}_l^{\beta} \sim N\left(\hat{\varphi}_k^{\beta} \tilde{P}_l, \hat{\sigma}_{\gamma}^{\beta}\right)$, poi quelli del secondo $\hat{\alpha}_j \sim N(\hat{\gamma}_l^{\alpha}\ \widetilde{U}_j, \hat{\sigma}_{\alpha})$ ed eventualmente $\hat{\beta}_j \sim N\left(\hat{\gamma}_l^{\beta}\ \widetilde{U}_j, \hat{\sigma}_{\beta}\right)$ e infine i valori per la variabile risposta: $\hat{y}_t \sim N(\hat{\lambda}_k \tilde{X}_t, \hat{\sigma}_y)$.

($\tilde{P}$ contiene le variabili del terzo strato, $\widetilde{U}$ quelle del secondo e $\tilde{X}$ quelle del primo se previste dai modelli).

Per tutti i modelli contenenti le variabili **SMS** e **MMS** è necessario stimare il numero di sms e mms inviati da ogni Sim per ogni mese. Per le previsioni del traffico del mese di aprile 2006 per le schede telefoniche del dataset di stima i valori di **SMS** e **MMS** vengono posti uguali alla media degli sms e mms inviati dalla medesima Sim nei diciassette mesi precedenti.

Per le previsioni delle Sim "nuove" i valori di **SMS** e **MMS** vengono posti uguale alla media del numero di sms e mms inviati da tutte le schede del gruppo di stima in quel particolare mese. Questo viene fatto per i mesi da novembre 2004 ad a marzo 2006. Per il mese di aprile 2006, **SMS** e **MMS,** vengono uguagliati alla media del numero di sms e mms inviati da tutte le Sim del dataset di stima in tutti i diciassette mesi precedenti.

6.2 Risultati delle simulazioni

La tabella sottostante riporta per i modelli dei capitoli precedenti il numero di parametri e le deviazioni standard residue per il traffico delle Sim. Per le schede appartenenti al gruppo di stima si prevedono le chiamate del solo mese di aprile 2006 mentre per quelle del gruppo di verifica si stimano le chiamate per tutto il periodo in esame (da novembre 2004 ad aprile 2006).

	modello	n° parametri	$\hat{\sigma}$ per il mese di aprile 2006	$\hat{\sigma}$ per le Sim di verifica
1.1	**Regressione lineare normale**	17	39.028	50.076
1.2	**Regressione lineare normale con trasformata radice quadrata**	17	39.362	50.903
2.1	**Multilivello con intercetta variabile e predittori solo al primo livello**	1127	32.231	49.899
2.2	**Multilivello con intercetta variabile e predittori al primo e al secondo livello**	1138	32.241	59.766
2.3	**Multilivello con intercetta e coefficiente variabile per mese con predittori solo al primo livello**	2250	27.580	49.917
2.4	**Multilivello con intercetta e coefficiente variabile per mese con predittori al primo e al secondo livello**	2263	27.987	75.666
2.5	**Multilivello con intercetta e coefficiente variabile per mese, SMS e MMS con predittori solo al primo livello**	4499	27.949	51.237
2.6	**Multilivello con intercetta e coefficiente variabile per mese, SMS e MMS con predittori al primo e al secondo livello**	4512	28.960	85.845
3.1	**Multilivello a tre strati con intercetta variabile senza predittori**	2118 (1087)	33.306	50.208
3.2	**Multilivello a tre strati con intercetta variabile e tutti i predittori**	2129 (1084)	31.410	50.109

3.3	**Multilivello a tre strati con intercetta e coefficiente per mese variabile e predittori solo al primo livello**	2251 (2022)	27.689	50.098
3.4	**Multilivello a tre strati con intercetta e coefficiente per mese variabile e tutti i predittori**	4255 (2023)	27.691	49.907

Tabella 6.1 numero di parametri, deviazione standard residua per il mese di aprile 2006 per le Sim del dataset di stima, e di tutti i mesi per quelle del dataset di verifica per i modelli proposti nei capitoli precedenti.

Il numero di parametri nei modelli gerarchici dipende dall'ammontare di pooling: i J parametri dei coefficienti variabili equivalgono a uno solo se c'è complete pooling (se il multilivello è uguale alla regressione lineare) mentre sono effettivi se non c'è pooling (se il modello multilivello è uguale al no-pooling). Generalmente il numero di parametri effettivi è un compromesso tra questi due estremi.
Nella tabella è riportato il numero di parametri massimo per i modelli e tra parentesi quello effettivo.

6.3 Conclusioni

Per le previsioni del mese di aprile 2006 per le Sim del dataset di stima i modelli di regressione lineare risultano essere i peggiori tra quelli provati. Questo significa che i multilivello riescono ad approssimare meglio il fenomeno oggetto di studio. Il vantaggio consiste nel considerare il "fattore Sim", ovvero il fatto che l'ammontare di traffico in uscita dipende, oltre che dalle variabili esplicative disponibili, dalla scheda telefonica che le effettua.
Per i modelli gerarchici a due livelli un significativo miglioramento si ha avuto con l'utilizzo del coefficiente variabile per **mese**, mentre ciò non si è verificato per le variabili **SMS** e **MMS**. Inoltre, confrontando a due a due i modelli, si nota come le variabili al secondo strato non portino ad una diminuzione della deviazione standard residua.
Anche per i multilivello a tre strati risulta essere decisivo il coefficiente variabile per la variabile **mese** che permette al terzo e al quarto modello di ottenere previsioni migliori rispetto ai primi due.
Il modello 2.3 e 3.3 hanno errori di stima molto simili. Questo significa che l'introduzione del terzo livello non ha portato alcun miglioramento. Si conclude quindi che, con i modelli utilizzati, non si è registrato nessun "fattore utente". Conoscere il cliente che utilizza la Scheda telefonica per effettuare chiamate non ci è di alcun aiuto per la stima della variabile risposta. Anche le variabili **età**, **zona**, **piano** e **sesso** non apportano alcuna informazione utile alla stima del numero di chiamate effettuate dalle Sim.
Il modello a due livelli con intercetta e coefficiente variabile per **mese** con predittori solo al primo livello risulta essere il migliore.

Per le previsioni del dataset di verifica la situazione è diversa. I modelli hanno infatti errori di stima molto simili (fatta eccezione per il 2.2, 2.4, 2.6: i multilivello a due strati con tutti i predittori). Non si registra alcun vantaggio nell'utilizzare i modelli gerarchici rispetto alla regressione lineare. Quest'ultima è pertanto la scelta migliore.

7 ULTERIORI MODELLI

Modelli gerarchici non annidati

I modelli non annidati permettono di considerare le osservazioni raggruppate in strutture complesse.
I dati sono stati finora esaminati secondo la scheda telefonica che ha effettuato le chiamate, ma è possibile valutare anche altri fattori di selezione come ad esempio la provincia di appartenenza o il mese a cui corrispondono le telefonate.
In questo capitolo verranno utilizzati alcuni tipi di modelli non annidati per ricercare ulteriori elementi che permettano di ridurre l'errore di stima.

7.1 Multilivello non annidato per Sim e provincia

Nel primo modello le singole osservazioni sono considerate raggruppate in base alla Sim e alla provincia di appartenenza. Si utilizza un multilivello a due strati con due effetti casuali e il coefficiente per **mese** variabile:

$$y_t \sim N\left(\mu + \alpha_{j[t]} + \gamma_{k[t]} + \beta_{j[t]} mese_t + \beta_2 SMS_t + \beta_3 MMS_t\,,\, \sigma_y^2\right) \qquad \text{per } t = 1, \dots, n = 19057$$

$$\begin{pmatrix}\alpha_j \\ \beta_j\end{pmatrix} \sim N\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2\end{pmatrix}\right) \qquad \text{per Sim } j = 1, \dots, J = 1121$$

$$\gamma_k \sim N\left(0\,,\, \sigma_\gamma^2\right) \qquad \text{per città } k = 1, \dots, K = 103$$

α_j e γ_k rappresentano l'effetto della Sim e della provincia. La loro distribuzione è centrata in zero perché il modello ha già un'intercetta μ, e ogni altra media per α e γ viene inglobata in μ.
Per σ_y, σ_α e σ_β si utilizzano delle a priori uniformi in 0 – 100, per μ una normale con media zero e deviazione standard 100, mentre per ρ una uniforme in -1, 1.

La tabella 7.1 riporta le stime a posteriori dei parametri calcolate mediante algoritmi iterativi:

	stima	**st. error**		**stima**	**st. error**
μ, intercetta	38.117	1.856	σ_α	47.539	1.072
β, mese	-0.724	0.097	σ_β	2.962	0.031
β_2, SMS	0.192	0.006	σ_γ	8.360	0.026
β_3, MMS	1.116	0.123	σ_y	27.136	1.178
			ρ	-0.611	0.011

Tabella 7.3 Stime e st. error per il modello non annidato con intercette per Sim e provincia e coef. variabile per mese.

	minimo	**1 quartile**	**mediana**	**media**	**3 quartile**	**Massimo**
α_j	-131.225	-26.570	-13.815	0	12.848	276.394
β_j	-15.859	-0.771	0.301	0	0.990	16.107
γ_k	-9.690	-2.412	-0.736	0	1.769	15.501

Tabella 7.2 Sintesi degli effetti casuali per il modello non annidato con intercette per Sim e provincia e coef. variabile per mese.

La tabella 7.2 riporta alcuni indici di sintesi per le distribuzioni degli effetti casuali del modello. Al netto dell'effetto delle altre variabili la provincia che in media effettua più telefonate è Livorno seguita da Prato e Grosseto, mentre ai livelli più bassi ci sono Pavia, Bari e per ultima Padova.

La figura 7.1 mostra il grafico quantile-quantile dei residui e il confronto tra valori predetti dal modello e valori reali. I residui continuano a non essere normali.

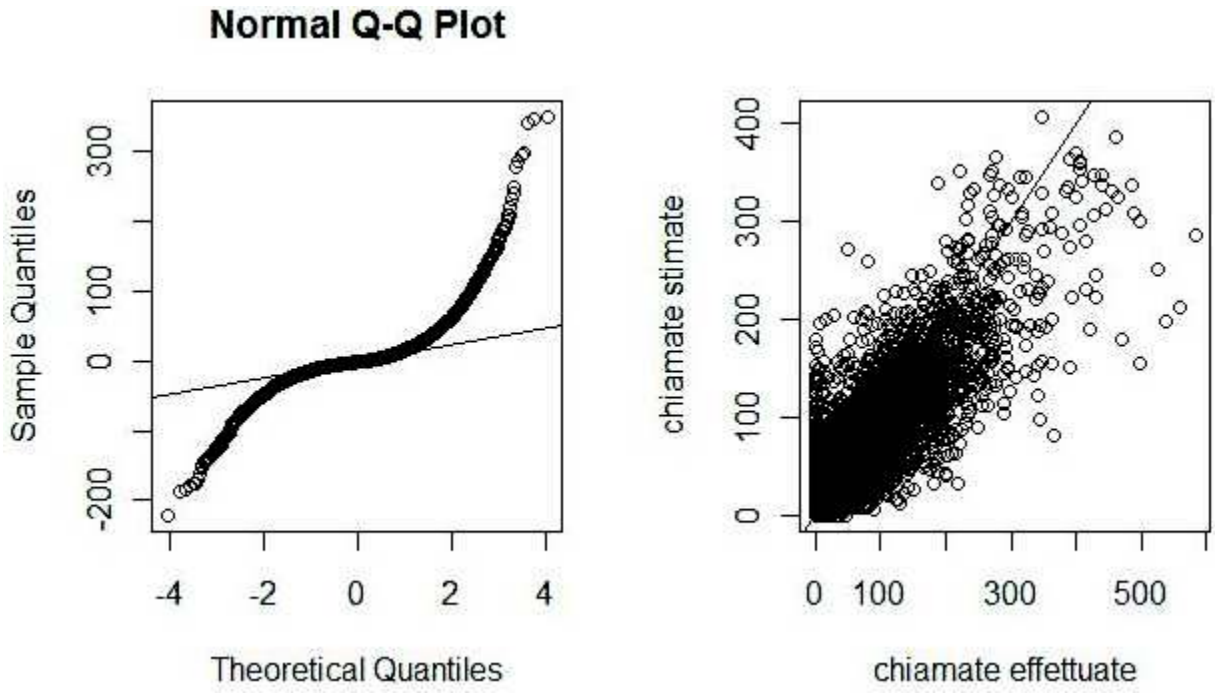

Figura 7.2 (a) qqnorm dei residui (b) valori predetti verso valori reali

La deviazione standard residua calcolata per il mese di aprile 2006 per le Sim del dataset vale 28.234; mentre quella per le Sim di verifica calcolata per tutti i diciotto mesi di studio vale 49.928.

7.2 Multilivello non annidato per Sim e mese

Di seguito si costruisce un modello con le osservazioni raggruppate in base alla Sim e al mese in cui le telefonate sono state effettuate. Si utilizzano solo i predittori al primo livello con due effetti casuali e i coefficienti variabili per le esplicative:

$$y_t \sim N\left(\mu + \alpha_{j[t]} + \beta_{j[t]} SMS_t + \theta_{j[t]} MMS_t + \gamma^0_{k[t]} + \gamma^1_{k[t]} SMS_t, \sigma_y^2\right) \quad \text{per } t = 1, \dots, n = 19057$$

$$\begin{pmatrix}\alpha_j \\ \beta_j \\ \theta_j\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}\sigma_\alpha^2 & \rho_{\alpha\beta}\sigma_\alpha\sigma_\beta & \rho_{\alpha\theta}\sigma_\alpha\sigma_\theta \\ \rho_{\alpha\beta}\sigma_\alpha\sigma_\beta & \sigma_\beta^2 & \rho_{\delta\theta}\sigma_\delta\sigma_\theta \\ \rho_{\alpha\theta}\sigma_\alpha\sigma_\theta & \rho_{\delta\theta}\sigma_\delta\sigma_\theta & \sigma_\theta^2\end{pmatrix}\right) \quad \text{per } j = 1, \dots, J = 1121$$

$$\begin{pmatrix}\gamma^0_k \\ \gamma^1_k\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma^2_{\gamma 0} & \rho_\gamma\sigma_{\gamma 0}\sigma_{\gamma 1} \\ \rho_\gamma\sigma_{\gamma 0}\sigma_{\gamma 1} & \sigma^2_{\gamma 1}\end{pmatrix}\right) \quad \text{per } k = 1, \dots, K = 17$$

Per $\sigma_y, \sigma_\alpha, \sigma_\beta, \sigma_\delta$ e σ_θ si utilizzano delle a priori uniformi in 0 – 100, per μ una normale con media zero e deviazione standard 100, mentre per $\rho_{\alpha\beta}$, $\rho_{\alpha\theta}, \rho_{\delta\theta}, \rho_\gamma$ delle uniformi in -1, 1.

La tabella 7.3 riporta le stime a posteriori dei parametri calcolate mediante algoritmi iterativi:

	stime	st. error		stime	st. error
μ	27.281	1.519	$\rho_{\alpha\beta}$	0.083	0.011
σ_α	32.978	1.054	$\rho_{\alpha\theta}$	0.324	0.014
σ_β	1.221	0.020	$\rho_{\delta\theta}$	0.356	0.020
σ_θ	4.745	0.019	$\sigma_{\gamma 0}$	4.639	0.921
σ_y	26.309	1.129	$\sigma_{\gamma 1}$	0.029	0.005
			ρ_γ	0.195	0.012

Tabella 7.3 Stime e st. error per il modello non annidato per Sim e mese

	minimo	1 quartile	mediana	media	3 quartile	Massimo
α_j	-32.294	-19.735	-10.834	0	8.424	277.370
β_j	-1.595	-0.491	0.072	0	0.689	8.213
θ_j	-9.485	-1.288	-0.659	0	0.619	25.569
γ^0_k	-8.213	-4.823	1.455	0	3.784	5.807
γ^1_k	-0.057	-0.011	0	0	0.009	0.058

Tabella 7.4 Sintesi degli effetti casuali per il modello non annidato per Sim e mese.

Il modello è composto da un intercetta μ e da cinque effetti casuali. Con l'utilizzo di **mese** come fattore di classificazione il risultato è, rispetto ai casi precedenti, una diminuzione di σ_α che ora vale 27.28. Il mese che ha il valore più basso per γ_k^0 è aprile 2005 come si poteva presumere osservando il grafico 1.2.

La figura 7.2 riporta il grafico quantile-quantile dei residui e il confronto tra valori predetti dal modello e valori reali. Anche per questo modello i residui seguitano a non essere normali.

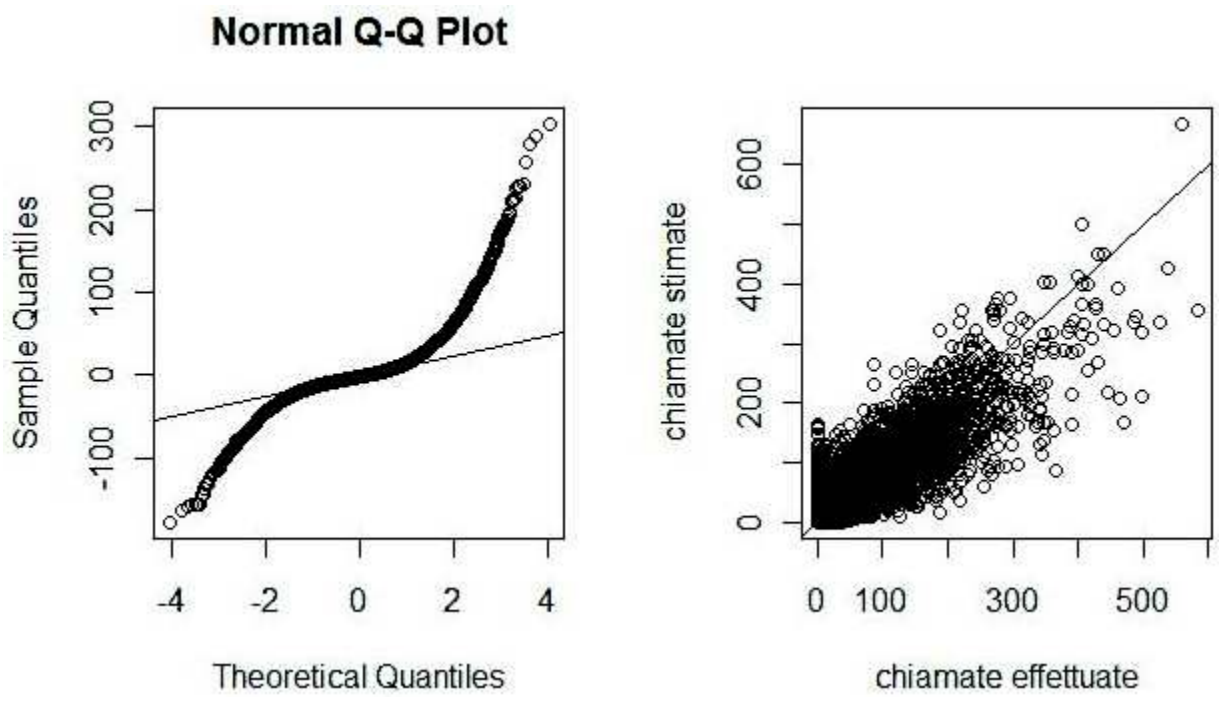

Figura 7.2 (a) qqnorm dei residui (b) valori predetti verso valori reali

La deviazione standard residua calcolata per il mese di aprile 2006 per le Sim del dataset vale 32.011; mentre quella per le Sim di verifica calcolata per tutti i diciotto mesi di studio vale 50.131.
Questo modello ottiene il minor errore di previsione per il traffico delle nuove Sim.

Modelli gerarchici multivariati

I modelli multivariati consentono di stimare più variabili risposta. Nell'indagine si è esaminato il numero totale di chiamate effettuate mensilmente dalle Sim.

Con questo tipo di modelli si prevede il numero di telefonate verso i numeri fissi, verso i cellulari dello stesso e di altri operatori.

Per costruire dei modelli multivariati è necessario aggiungere un ulteriore livello per rappresentare le tre variabili risposta, ottenendo così tre strati gerarchici che identificano le variabili risposta, le osservazioni e le Sim.

7.3 Multilivello trivariato con intercetta variabile

Viene costruito ora un multilivello con intercetta variabile e con predittori **mese**, **SMS** e **MMS**. Per comodità i primi due strati sono rappresentati come un unico livello:

$$\begin{pmatrix} y_{1tj} \\ y_{2tj} \\ y_{3tj} \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha^1_{j[t]} + \beta^1_1 mese_t + \beta^1_2 SMS_t + \beta^1_3 MMS_t \\ \alpha^2_{j[t]} + \beta^2_1 mese_t + \beta^2_2 SMS_t + \beta^2_3 MMS_t \\ \alpha^3_{j[t]} + \beta^3_1 mese_t + \beta^3_2 SMS_t + \beta^3_3 MMS_t \end{pmatrix}, \begin{pmatrix} \sigma^2_1 & \rho_{12}\sigma_{y1}\sigma_{y2} & \rho_{13}\sigma_{y1}\sigma_{y3} \\ \rho_{12}\sigma_{y1}\sigma_{y2} & \sigma^2_2 & \rho_{23}\sigma_{y2}\sigma_{y3} \\ \rho_{13}\sigma_{y1}\sigma_{y3} & \rho_{23}\sigma_{y2}\sigma_{y3} & \sigma^2_{y3} \end{pmatrix} \right)$$

$\alpha^1_j \sim N(\mu_{1\alpha}, \sigma^2_{1\alpha})$ per t = 1, …, n = 1700

$\alpha^2_j \sim N(\mu_{2\alpha}, \sigma^2_{2\alpha})$ per j = 1, …, J = 100

$\alpha^3_j \sim N(\mu_{3\alpha}, \sigma^2_{3\alpha})$

Si sono utilizzate 100 Sim selezionate casualmente dalle 1121 che componevano il campione formato nel capitolo 4. Per $\sigma_{y1}, \sigma_{y2}, \sigma_{y3}, \sigma_{1\alpha}, \sigma_{2\alpha}$ e $\sigma_{3\alpha}$ si utilizzano delle a priori uniformi in 0 – 100, per μ una normale con media zero e deviazione standard 100, mentre per ρ_{12}, ρ_{13} e ρ_{23} delle uniformi in -1, 1.

La tabella 7.5 riporta le stime a posteriori dei parametri calcolate mediante algoritmi iterativi:

	stima	**st. error**		**stima**	**st. error**		**stima**	**st. error**
$\mu_{1\alpha}$	10.790	1.106	$\mu_{2\alpha}$	14.825	1.469	$\mu_{3\alpha}$	16.715	2.763
β^1_1	-0.199	0.107	β^1_2	-0.140	0.138	β^1_3	-0.320	0.255
β^1_2	0.045	0.012	β^1_2	0.001	0.017	β^1_2	0.188	0.025
β^1_3	-0.031	0.214	β^1_3	0.270	0.279	β^1_3	-1.875	0.414
$\sigma_{1\alpha}$	0.459	0.432	$\sigma_{2\alpha}$	1.266	0.931	$\sigma_{3\alpha}$	1.715	1.282
σ_{y1}	13.090	0.408	σ_{y2}	17.400	0.544	σ_{y3}	25.380	0.814
ρ_{12}	0.420	0.037	ρ_{13}	0.451	0.035	ρ_{23}	0.404	0.038

Tabella 7.5 Stime e st. error per i coefficienti del multilivello trivariato con intercetta variabile

La tabella 7.6 riporta alcuni indici di sintesi per gli effetti casuali del modello:

	minimo	1 quartile	mediana	media	3 quartile	massimo
α_j^1	-0.192	-0.060	0.009	0	0.054	0.163
α_j^2	-0.719	-0.254	-0.037	0	0.216	1.056
α_j^3	-1.056	-0.406	-0.001	0	0.358	1.034

Tabella 7.6 distribuzione degli effetti casuali del modello

La tabella 7.7 riporta le deviazioni standard residue per il numero di chiamate effettuate dalle Sim del gruppo di stima e da quelle del gruppo di verifica verso i numeri fissi, i cellulari dello stesso operatore e quelli di altri operatori.

	chiamate ai fissi	cellulari stesso operatore	cellulari altri operatori	chiamate totali
gruppo stima	8.038	10.669	19.185	30.081
gruppo verifica	16.216	18.962	27.046	50.445

Tabella 7.7 deviazioni standard residue

L'ultima colonna permette di confrontare questo multilivello con quelli dei capitoli precedenti.
Con il modello trivariato non si ottengono risultati migliori rispetto ai casi precedenti sia per quanto riguarda le Sim del gruppo di stima, sia per le schede del gruppo di verifica.

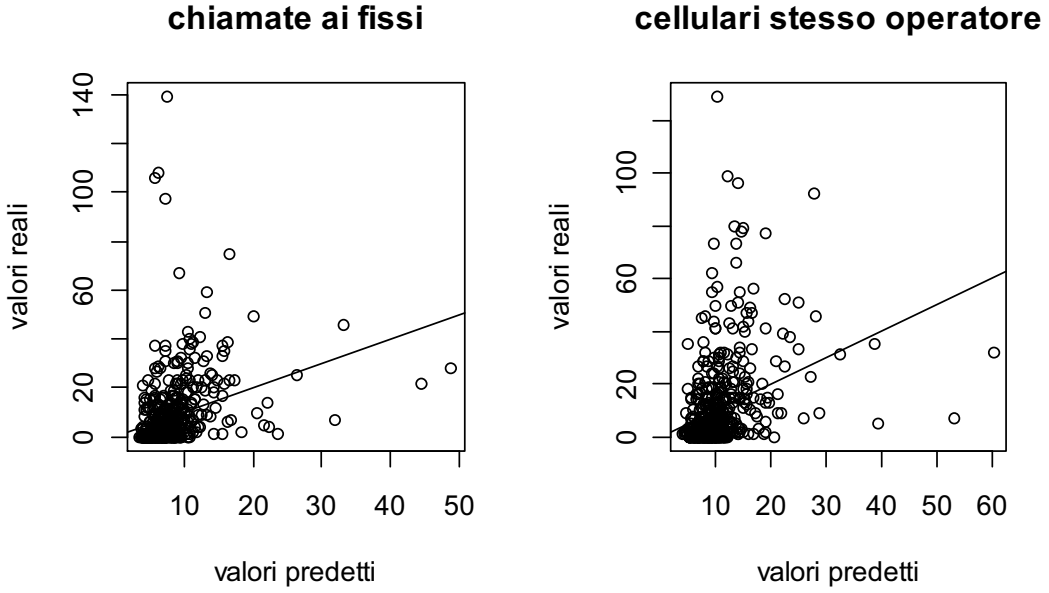

Figura 7.3 (a) chiamate predette verso i numeri fissi verso i valori reali (b) chiamate predette ai cellulari dello stesso operatore verso i valori reali

I grafici 7.3 e 7.4 riportano i valori predetti dal modello trivariato verso i valori reali per le telefonate verso i fissi, verso i cellulari dello stesso e di altri operatori, e le chiamate totali.

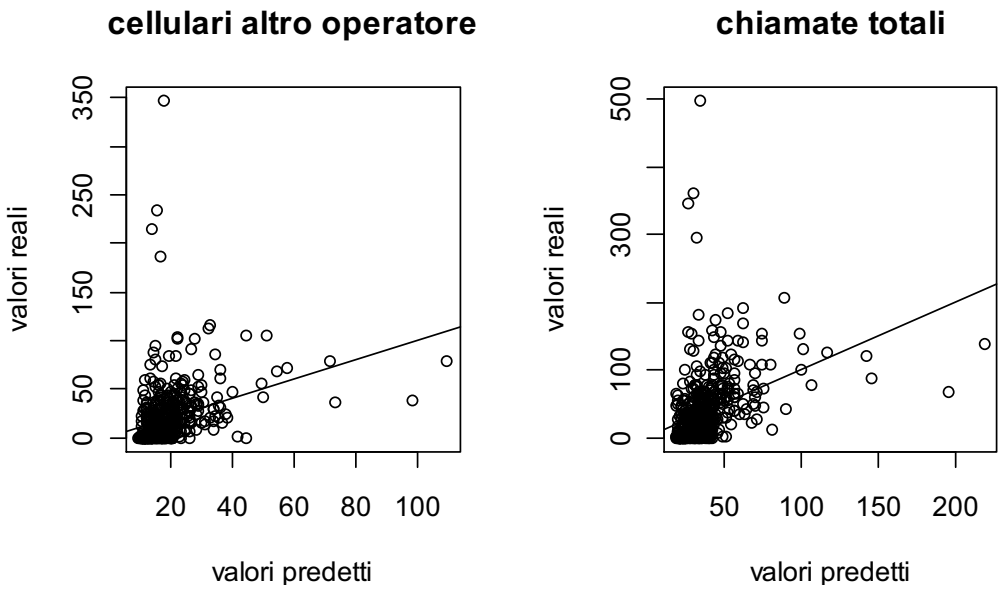

Figura 7.4 (a) chiamate predette ai cellulari degli altri operatori verso i valori reali (b) chiamate predette totali verso i valori reali

8 TEORIA DEI MULTILIVELLO

8.1 Stime dei modelli gerarchici

Per la stima dei modelli si è ricorsi all'uso della teoria Bayesiana. L'inferenza bayesiana si basa sulla distribuzione a posteriori ottenuta dal prodotto della verosimiglianza per le distribuzioni a priori.
Nella regressione gerarchica i modelli dei livelli sottostanti corrispondono alle distribuzioni a priori per i coefficienti variabili del primo livello.

Multilivello senza predittori

Il modello più semplice è quello con intercetta variabile senza predittori:

$$y_t \sim N\left(\alpha_{j[t]}, \sigma_y^2\right) \qquad \text{per t = 1,... , n}$$
$$\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2) \qquad \text{per j = 1,... , J}$$

dove α_j e σ_y^2 sono i parametri del modello e μ_α e σ_α^2 gli iperparametri.
Se fossero noti i valori di σ_y, μ_α e σ_α, gli α_j avrebbero distribuzioni normali indipendenti:
$\alpha_j | y, \mu_\alpha, \sigma_y, \sigma_\alpha \sim N\left(\hat{\alpha}_j, V_j\right)$ dove:

$$\hat{\alpha}_j = \frac{\frac{n_j}{\sigma_y^2}\bar{y}_j + \frac{1}{\sigma_\alpha^2}\mu_\alpha}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}}, \qquad V_j = \frac{1}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}}, \qquad 7.1$$

Le stime $\hat{\alpha}_j$ sarebbero quindi un compromesso tra $\bar{y}_j$ e μ_α, la media del j-esimo gruppo e la media totale delle osservazioni, con pesi che dipendono dalla numerosità del gruppo e dalle varianze di entrambi i livelli.

numerosità del gruppo, n_j	**stime $\hat{\alpha}_j$**
$n_j = 0$	$\hat{\alpha}_j = \mu_\alpha\ (complete\ pooling)$
$n_j < \sigma_y^2/\sigma_\alpha^2$	$\hat{\alpha}_j$ prossima a μ_α
$n_j = \sigma_y^2/\sigma_\alpha^2$	$\hat{\alpha}_j = 1/2\bar{y}_j + 1/2\mu_\alpha$
$n_j > \sigma_y^2/\sigma_\alpha^2$	$\hat{\alpha}_j$ prossima a $\bar{y}_j$
$n_j = \infty$	$\hat{\alpha}_j = \bar{y}_j\ (no\ pooling)$

Tabella 8.4 relazione tra stime con numerosità del gruppo e rapporto varianze

Per tutti i modelli utilizzati nello studio non è noto alcun parametro. E' necessario quindi specificare la distribuzione a priori per tutti quelli impiegati. Non disponendo di alcun tipo di informazione preliminare utile sono state usate in tutti i modelli dell'analisi distribuzioni a priori proprie ma con poca informazione (gli al-

goritmi utilizzati non permettono l'uso di a priori non informative). In particolare per le varianze si è scelta una uniforme tra 0 e 100 mentre per tutti gli altri coefficienti sono state utilizzate delle normali di media 0 e deviazione standard 100.
Con il teorema di Bayes è possibile ottenere le distribuzioni a posteriori in maniera esplicita ma il calcolo risulta molto complesso. Si ricorre allora all'uso di algoritmi iterativi di **gibbs sampling**.

Multilivello con predittori al primo strato

Con l'introduzione dei predittori il modello diventa:

$$y_t \sim N\left(\alpha_{j[t]} + \beta X_t, \sigma_y^2\right) \qquad \text{per t = 1,… , n}$$
$$\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2) \qquad \text{per j = 1,… , J}$$

La distribuzione per gli α_j corrisponde ad una a priori normale uguale per tutti i gruppi J. Per ogni gruppo J la verosimiglianza $p(y|\beta, \sigma, X)$ indica i valori per gli α_j che sono più verosimili con i dati. La distribuzione a posteriori per ogni gruppo è centrata tra il valore di massima verosimiglianza e la media della distribuzione a priori. A seconda della numerosità n_j la stima degli α_j sarà più prossima alla verosimiglianza (n_j elevato) o alla a priori (n_j piccolo).

Multilivello con predittori al primo e al secondo strato

Con l'aggiunta dei predittori al secondo livello il modello diventa:

$$y_t \sim N\left(\alpha_{j[t]} + \beta X_t, \sigma_y^2\right) \qquad \text{per t = 1,… , n}$$
$$\alpha_j \sim N\left(\gamma_k U_j, \sigma_\alpha^2\right) \qquad \text{per j = 1,… , J}$$

La distribuzione per gli α_j è ancora normale ma dipende da U, al variare dei predittori variano anche le a priori per i gruppi j.

8.2 Gibbs sampler per modelli multilivello

Gibbs sampling è il nome di una famiglia di algoritmi iterativi impiegati per stimare modelli usando la teoria Bayesiana. L'algoritmo permette di modellare i parametri, uno alla volta, condizionatamente ai dati e al valore corrente di tutti gli altri. La sequenza di valori generati costituisce una catena di Markov. La prima parte della successione ottenuta corrisponde al *burn-in* e viene perciò scartata. La procedura necessita inoltre di valori iniziali per le quantità di interesse.

Per un multilivello senza predittori l'algoritmo è il seguente:

1. Aggiorna α: per j=1,…, J calcola $\hat{\alpha}_j$ e V_j dalla 7.1 e poi simula α_j da una normale $N(\hat{\alpha}_j, V_j)$.
2. Aggiorna μ_α: calcola $\hat{\mu}_\alpha$ dalla 7.3 e simula μ_α da una normale $N(\hat{\mu}_\alpha, \sigma_\alpha^2/J)$.
3. Aggiorna σ_y: calcola $\hat{\sigma}_y^2$ dalla 7.2 e stima $\sigma_y^2 = \hat{\sigma}_y^2/X_{n-1}^2$, dove X_{n-1}^2 è un Chi-quadrato con $n-1$ gradi di libertà.
4. Aggiorna σ_α: calcola $\hat{\sigma}_\alpha^2$ dalla 7.4 e stima $\sigma_\alpha^2 = \hat{\sigma}_\alpha^2/X_{j-1}^2$, dove X_{n-1}^2 è un Chi-quadrato con $k-1$ gradi di libertà.

Per un multilivello con predittori (X contiene quelli al primo livello e U quelli al secondo) si utilizza la seguente variante:

1. Aggiorna α: si usa l'espressione $\alpha_j = U_j + \eta_j$, dove gli η_j corrispondono agli errori al secondo livello. Si applica la 7.1 opportunamente modificata con l'uso dei predittori.
   Per ogni osservazione si calcola $y_t^{temp} = y_t - X_t\beta - U_{j[t]}\gamma$. Poi per ogni = 1,.. ,J si calcola $\hat{\eta}_j$ e V_j dalla 7.1 usando y^{temp} al posto di y e si ottengono η_j dalla normale $N(\hat{\eta}_j, V_j)$.
   Si completa l'aggiornamento settando $\alpha_j = U_j\gamma + \eta_j$.
2. Aggiorna β: per ogni osservazione si calcola $y_t^{temp} = y_t - \alpha_{j[t]}$. Si calcola una stima $\hat{\beta}$ con una regressione di y^{temp} su X, e la matrice di covarianze $V_\gamma = (X^tX)^{-1}\hat{\sigma}^2$. Si simula poi β da $N(\hat{\beta}, V_\beta)$.
3. Aggiorna γ: si stima $\hat{\gamma}$ con una regressione di α su U, e la matrice di covarianza $V_\alpha = (U^tU)^{-1}\sigma_\alpha^2$.
   Si simula poi γ da $N(\hat{\gamma}, V_\gamma)$.
4. Aggiorna σ_y: si calcola $\hat{\sigma}_y^2 = 1/n\sum_{t=1}^{n}(y_t - \alpha_{j[t]} - X_t\beta)^2$ e poi $\sigma_y^2 = \hat{\sigma}_y^2/X_{n-1}^2$ dove X_{n-1}^2 è un Chi-quadrato con $n-1$ gradi di libertà.
5. Aggiorna σ_α: si calcola $\hat{\sigma}_\alpha^2 = 1/J\sum_{j=1}^{J}(\alpha_j - U_j\gamma)^2$ e poi $\sigma_\alpha^2 = \hat{\sigma}_\alpha^2/X_{J-1}^2$ dove X_{n-1}^2 è un Chi-quadrato con $n-1$ gradi di libertà.

8.3 Confronto metodi di stima

Quando si ricorre a priori poco informative la a posteriori dovrebbe coincidere con la funzione di massima verosimiglianza. I risultati raggiunti tramite l'inferenza Bayesiana con l'impiego di Bugs dovrebbero quindi essere simili a quelli ricavati dal metodo di stima di massima verosimiglianza (REML: residual maximum likelihood) ottenibili utilizzando la funzione lmer di R.

La tabella 8.2 confronta le stime dei due metodi per il modello con intercetta e coefficiente variabile per mese con predittori solo al primo livello:

	Bugs		**R**	
	stima	**st.error**	**stima**	**st.error**
μ_α, intercetta	37.774	1.504	37.905	1.420
μ_β, mese	-0.724	0.097	-0.728	0.095
SMS	0.192	0.006	0.191	0.005
MMS	1.115	0.123	1.102	0.117
σ_α	48.225		48.235	
σ_β	2.962		2.964	
σ_y	27.136		27.130	
ρ	-0.606		-0.603	

Tabella 8.2 confronto stime modello Bugs con R

Le stime prodotte dai due programmi sono molto simili. Lo stesso vale per gli effetti casuali riportati in tabella 8.3. Il grafico 8.1 mette a confronto le stime α_j e β_j.

		minimo	**I quartile**	**mediana**	**media**	**III quartile**	**massimo**
Bugs	α_j	-101.500	9.878	23.450	37.850	50.180	328.200
	β_j	-16.530	-1.476	-0.434	-0.728	0.280	15.340
R	α_j	-102.600	9.669	23.350	37.770	50.000	327.600
	β_j	-16.580	-1.493	-0.421	-0.724	0.269	15.380

Tabella 8.3 confronto effetti casuali modello Bugs con R

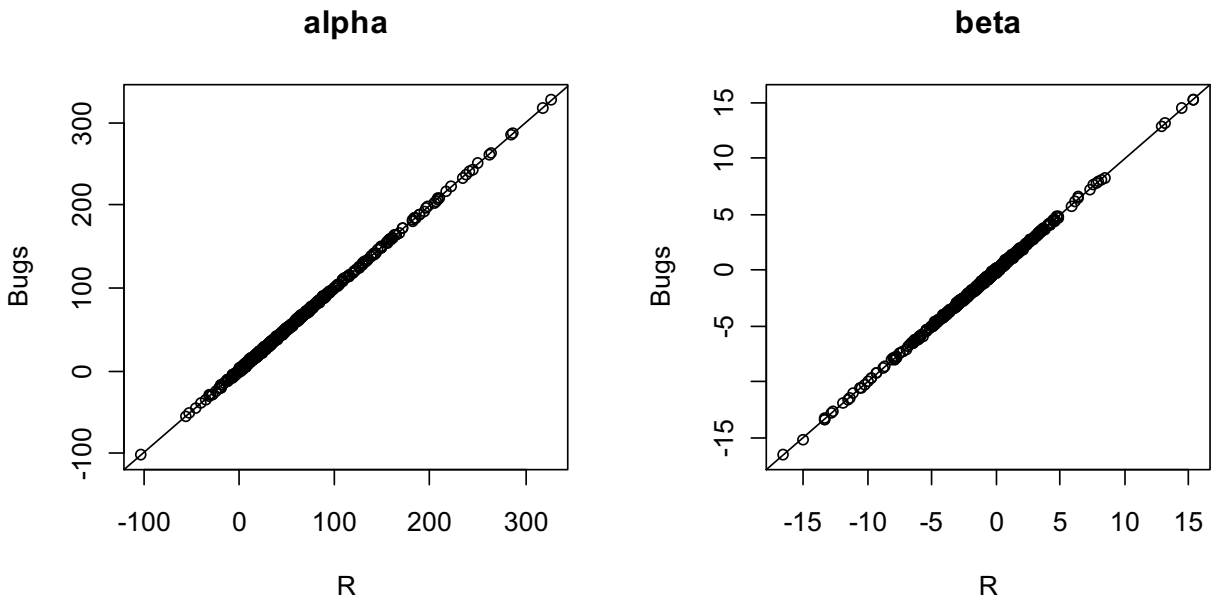

Figura 8.1 confronto effetti casuali modello Bugs con R

Non sono state rilevate differenze significative nell'utilizzare l'inferenza Bayesiana o REML per implementare modelli gerarchici. Bugs offre la possibilità di costruire modelli complessi (come ad esempio quelli a tre livelli) anche se ha tempi di esecuzione molto più lunghi di R.

APPENDICE

Si riportano i comandi utilizzati in **Bugs** per stimare i modelli.

Capitolo 3

Multilivello ad intercetta variabile con predittori al primo livello

```
model {
      for (i in 1:n){
            y[i] ~ dnorm (y.hat[i], tau.y)
            y.hat[i] <- a[id[i]] + inprod(b[],V[i,])
      }
      tau.y <- pow(sigma.y, -2)
      sigma.y ~ dunif (0, 100)

      for(k in 1:K){
            b[k] ~ dnorm(0, 0.0001)
      }

      for (j in 1:J){
            a[j] ~ dnorm (mu.a, tau.a)
      }

      mu.a ~ dnorm (0, .0001)
      tau.a <- pow(sigma.a, -2)
      sigma.a ~ dunif (0, 100)
}
```

Multilivello ad intercetta variabile con predittori al primo e al secondo livello

```
model {
      for (i in 1:n){
            y[i] ~ dnorm (y.hat[i], tau.y)
            y.hat[i] <- a[id[i]] + inprod(b[],V[i,])
      }
      tau.y <- pow(sigma.y, -2)
      sigma.y ~ dunif (0, 100)

      for(k in 1:K){
            b[k] ~ dnorm(0, 0.0001)
      }
      for (j in 1:J){
            a[j] ~ dnorm (a.hat[j], tau.a)
            a.hat[j] <- inprod(g[],GG[j,])
      }
      for (p in 1:G){
            g[p] ~ dnorm (0, .0001)
      }
      tau.a <- pow(sigma.a, -2)
      sigma.a ~ dunif (0, 100)
}
```

Capitolo 4

Multilivello con intercetta e coefficiente variabile per mese con predittori al primo strato

```
model {
  for (i in 1:n){
    y[i] ~ dnorm (y.hat[i], tau.y)
    y.hat[i] <- a[id[i]] + b[id[i]]*wave[i] + c*SMS[i] + d*MMS[i]
  }
  tau.y <- pow(sigma.y, -2)
  sigma.y ~ dunif (0, 100)
  c ~ dnorm(0, .0001)
  d ~ dnorm(0, .0001)

  for (j in 1:J){
    a[j] <- B[j,1]
    b[j] <- B[j,2]
    B[j,1:2] ~ dmnorm (B.hat[j,], Tau.B[,])
    B.hat[j,1] <- mu.a
    B.hat[j,2] <- mu.b
  }
  mu.a ~ dnorm(0, .0001)
  mu.b ~ dnorm(0, .0001)

  Tau.B[1:2,1:2] <- inverse(Sigma.B[,])
  Sigma.B[1,1] <- pow(sigma.a, 2)
  sigma.a ~ dunif (0, 100)
  Sigma.B[2,2] <- pow(sigma.b, 2)
  sigma.b ~ dunif (0, 100)
  Sigma.B[1,2] <- rho*sigma.a*sigma.b
  Sigma.B[2,1] <- Sigma.B[1,2]
  rho ~ dunif (-1, 1)
}
```

Multilivello con intercetta e coefficiente variabile per mese con predittori al primo e secondo strato

```
model {
  for (i in 1:n){
    y[i] ~ dnorm (y.hat[i], tau.y)
    y.hat[i] <- a[id[i]] + b[id[i]]*wave[i] + c*SMS[i] + d*MMS[i]
  }
  tau.y <- pow(sigma.y, -2)
  sigma.y ~ dunif (0, 100)
  c ~ dnorm(0, .0001)
  d ~ dnorm(0, .0001)

  for (j in 1:J){
    a[j] <- B[j,1]
    b[j] <- B[j,2]
    B[j,1:2] ~ dmnorm (B.hat[j,], Tau.B[,])
    B.hat[j,1] <- inprod(g.a[], U[j,])
    B.hat[j,2] <- inprod(g.b[], U[j,])
  }

  for (l in 1:L){
    g.a[l] ~ dnorm (0, .0001)
    g.b[l] ~ dnorm (0, .0001)
  }

  Tau.B[1:2,1:2] <- inverse(Sigma.B[,])
  Sigma.B[1,1] <- pow(sigma.a, 2)
```

```
  sigma.a ~ dunif (0, 100)
  Sigma.B[2,2] <- pow(sigma.b, 2)
  sigma.b ~ dunif (0, 100)
  Sigma.B[1,2] <- rho*sigma.a*sigma.b
  Sigma.B[2,1] <- Sigma.B[1,2]
  rho ~ dunif (-1, 1)
}
```

Multilivello con intercetta e coefficiente variabile per mese sms e mms con predittori al primo strato

```
model {
  for (i in 1:n){
    y[i] ~ dnorm (y.hat[i], tau.y)
    y.hat[i] <- inprod(B[id[i],],V[i,])
  }
  tau.y <- pow(sigma.y, -2)
  sigma.y ~ dunif (0, 100)

  for (j in 1:J){
    for(k in 1:K){
      B[j,k] <- xi[k]*B.raw[j,k]
    }
    B.raw[j,1:K] ~ dmnorm(mu.raw[], Tau.B.raw[,])
  }
  for(k in 1:K){
    mu[k] <- xi[k]*mu.raw[k]
    mu.raw[k] ~ dnorm(0, .0001)
    xi[k] ~ dunif(0, 100)
  }
  Tau.B.raw[1:K,1:K] ~ dwish(W[,],df)
  df <- K + 1
  Sigma.B.raw[1:K,1:K] <- inverse(Tau.B.raw[,])
  for(k in 1:K){
    for(k.prime in 1:K){
      rho.B[k,k.prime] <- Sig-
ma.B.raw[k,k.prime]/sqrt(Sigma.B.raw[k,k]*Sigma.B.raw[k.prime,k.prime])
    }
    sigma.B[k] <- abs(xi[k])*sqrt(Sigma.B.raw[k,k])
  }
}
```

Multilivello con intercetta e coefficiente variabile per mese sms e mms con predittori al primo e al secondo strato

```
model {
  for (i in 1:n){
    y[i] ~ dnorm (y.hat[i], tau.y)
    y.hat[i] <- inprod(B[id[i],],V[i,])
  }
  tau.y <- pow(sigma.y, -2)
  sigma.y ~ dunif (0, 100)

  for (k in 1:K){
    for(j in 1:J){
      B[j,k] <- xi[k]*B.raw[j,k]
    }
    xi[k] ~ dunif(0, 100)
  }
```

```
  for(j in 1:J){
    B.raw[j,1:K] ~ dmnorm(B.raw.hat[j,], Tau.B.raw[,])
    for(k in 1:K){
      B.raw.hat[j,k] <- inprod(G.raw[k,],U[j,])
    }
  }

  for(k in 1:K){
    for(l in 1:L){
      G[k,l] <- xi[k]*G.raw[k,l]
      G.raw[k,l] ~ dnorm(0, .0001)
    }
  }
  Tau.B.raw[1:K,1:K] ~ dwish(W[,],df)
  df <- K + 1
  Sigma.B.raw[1:K,1:K] <- inverse(Tau.B.raw[,])
  for(k in 1:K){
    for(k.prime in 1:K){
      rho.B[k,k.prime] <- Sig-
ma.B.raw[k,k.prime]/sqrt(Sigma.B.raw[k,k]*Sigma.B.raw[k.prime,k.prime])
    }
    sigma.B[k] <- abs(xi[k])*sqrt(Sigma.B.raw[k,k])
  }
}
```

Capitolo 5

Multilivello a tre livelli con intercetta variabile senza predittori

```
model {
      for (i in 1:n){
            y[i] ~ dnorm (y.hat[i], tau.y)
            y.hat[i] <- a[id[i]]
      }
      tau.y <- pow(sigma.y, -2)
      sigma.y ~ dunif (0, 100)

      for (j in 1:J){
            a[j] ~ dnorm (a.hat[j], tau.a)
            a.hat[j] <- ga[id1[j]]
      }
      tau.a <- pow(sigma.a, -2)
      sigma.a ~ dunif (0, 100)

      for (l in 1:L){
            ga[l] ~ dnorm (mu.ga, tau.ga)
      }
      mu.ga ~ dnorm (0, .0001)
      tau.ga <- pow(sigma.ga, -2)
      sigma.ga ~ dunif (0, 100)
}
```

Multilivello a tre livelli con intercetta variabile e predittori in tutti i livelli

```
model {
      for (i in 1:n){
            y[i] ~ dnorm (y.hat[i], tau.y)
            y.hat[i] <- a[id[i]] + b*wave[i] + c*SMS[i] + d*MMS[i]
      }
```

```
      tau.y <- pow(sigma.y, -2)
      sigma.y ~ dunif (0, 100)

      b ~ dnorm(0, 0.0001)
      c ~ dnorm(0, 0.0001)
      d ~ dnorm(0, 0.0001)

      for (j in 1:J){
            a[j] ~ dnorm (a.hat[j], tau.a)
            a.hat[j] <- ga[id1[j]] + inprod(gb[],PP[j,])
      }
      tau.a <- pow(sigma.a, -2)
      sigma.a ~ dunif (0, 100)

      for (p in 1:P){
            gb[p] ~ dnorm (0, .0001)
      }

      for (l in 1:L){
            ga[l] ~ dnorm (ga.hat[l], tau.ga)
            ga.hat[l] <- inprod(k[],KK[l,])
      }
      for (t in 1:T){
            k[t] ~ dnorm (0, .0001)
      }
      tau.ga <- pow(sigma.ga, -2)
      sigma.ga ~ dunif (0, 100)
}
```

Multilivello a tre livelli con intercette e coefficiente variabile per mese con predittori al primo livello

```
model {
  for (i in 1:n){
    y[i] ~ dnorm (y.hat[i], tau.y)
    y.hat[i] <- a[id[i]] + b[id[i]]*wave[i] + c*SMS[i] + d*MMS[i]
  }
  tau.y <- pow(sigma.y, -2)
  sigma.y ~ dunif (0, 100)

  c ~ dnorm(0, .0001)
  d ~ dnorm(0, .0001)

  for (j in 1:J){
    a[j] <- B[j,1]
    b[j] <- B[j,2]
    B[j,1:2] ~ dmnorm (B.hat[j,], Tau.B[,])
    B.hat[j,1] <- ga[id1[j]] + inprod(ta[], PP[j,])
    B.hat[j,2] <- gb[id1[j]] + inprod(tb[], PP[j,])
  }

  for (p in 1:M){
    ta[p] ~ dnorm (0, .0001)
    tb[p] ~ dnorm (0, .0001)
  }

  Tau.B[1:2,1:2] <- inverse(Sigma.B[,])
  Sigma.B[1,1] <- pow(sigma.a, 2)
  sigma.a ~ dunif (0, 100)
  Sigma.B[2,2] <- pow(sigma.b, 2)
  sigma.b ~ dunif (0, 100)
  Sigma.B[1,2] <- rho*sigma.a*sigma.b
```

```
  Sigma.B[2,1] <- Sigma.B[1,2]
  rho ~ dunif (-1, 1)

  for (l in 1:L){
    ga[l] <- B2[l,1]
    gb[l] <- B2[l,2]
    B2[l,1:2] ~ dmnorm (B2.hat[l,], Tau.B2[,])
    B2.hat[l,1] <- inprod(ka[], KK[l,])
    B2.hat[l,2] <- inprod(kb[], KK[l,])

  }
  for (t in 1:T){
    ka[t] ~ dnorm (0, .0001)
    kb[t] ~ dnorm (0, .0001)
  }
  Tau.B2[1:2,1:2] <- inverse(Sigma.B2[,])
  Sigma.B2[1,1] <- pow(sigma.ga, 2)
  sigma.ga ~ dunif (0, 100)
  Sigma.B2[2,2] <- pow(sigma.gb, 2)
  sigma.gb ~ dunif (0, 100)
  Sigma.B2[1,2] <- rho2*sigma.ga*sigma.gb
  Sigma.B2[2,1] <- Sigma.B2[1,2]
  rho2 ~ dunif (-1, 1)
}
```

Multilivello a tre livelli con intercette e coefficiente variabile per mese con predittori in tutti i livelli

```
model {
  for (i in 1:n){
    y[i] ~ dnorm (y.hat[i], tau.y)
    y.hat[i] <- a[id[i]] + b[id[i]]*wave[i] + c*SMS[i] + d*MMS[i]
  }
  tau.y <- pow(sigma.y, -2)
  sigma.y ~ dunif (0, 100)

  c ~ dnorm(0, .0001)
  d ~ dnorm(0, .0001)

  for (j in 1:J){
    a[j] <- B[j,1]
    b[j] <- B[j,2]
    B[j,1:2] ~ dmnorm (B.hat[j,], Tau.B[,])
    B.hat[j,1] <- ga[id1[j]]
    B.hat[j,2] <- gb[id1[j]]
  }

  Tau.B[1:2,1:2] <- inverse(Sigma.B[,])
  Sigma.B[1,1] <- pow(sigma.a, 2)
  sigma.a ~ dunif (0, 100)
  Sigma.B[2,2] <- pow(sigma.b, 2)
  sigma.b ~ dunif (0, 100)
  Sigma.B[1,2] <- rho*sigma.a*sigma.b
  Sigma.B[2,1] <- Sigma.B[1,2]
  rho ~ dunif (-1, 1)

  for (l in 1:L){
    ga[l] <- B2[l,1]
    gb[l] <- B2[l,2]
    B2[l,1:2] ~ dmnorm (B2.hat[l,], Tau.B2[,])
    B2.hat[l,1] <- mu.ga
    B2.hat[l,2] <- mu.gb
```

```
  }
  mu.ga ~ dnorm (0, .0001)
  mu.gb ~ dnorm (0, .0001)

  Tau.B2[1:2,1:2] <- inverse(Sigma.B2[,])
  Sigma.B2[1,1] <- pow(sigma.ga, 2)
  sigma.ga ~ dunif (0, 100)
  Sigma.B2[2,2] <- pow(sigma.gb, 2)
  sigma.gb ~ dunif (0, 100)
  Sigma.B2[1,2] <- rho2*sigma.ga*sigma.gb
  Sigma.B2[2,1] <- Sigma.B2[1,2]
  rho2 ~ dunif (-1, 1)
}
```

Capitolo 7

Multilivello non annidato per Sim e provincia

```
model {
  for (i in 1:n){
    y[i] ~ dnorm (y.hat[i], tau.y)
    y.hat[i] <- mu + ga[prov[i]] + a[id[i]] + b[id[i]]*wave[i] + c*SMS[i] +
d*MMS[i]
  }
  mu.adj <- mu + mean(ga[]) + mean(a[])
  mu ~ dnorm(0, .0001)
  tau.y <- pow(sigma.y, -2)
  sigma.y ~ dunif (0, 100)
  c ~ dnorm(0, .0001)
  d ~ dnorm(0, .0001)

  for(k in 1:K){
    ga[k] ~ dnorm (mu.ga, tau.ga)
    ga.adj[k] <- ga[k] - mean(ga[])
  }
  mu.ga ~ dnorm(0, .0001)
  tau.ga <- pow(sigma.ga, -2)
  sigma.ga ~ dunif (0, 100)

  for (j in 1:J){
    a[j] <- B[j,1]
    b[j] <- B[j,2]
    B[j,1:2] ~ dmnorm (B.hat[j,], Tau.B[,])
    B.hat[j,1] <- mu.a
    B.hat[j,2] <- mu.b
    a.adj[j] <- a[j] - mean(a[])
  }
  mu.a ~ dnorm(0, .0001)
  mu.b ~ dnorm(0, .0001)

  Tau.B[1:2,1:2] <- inverse(Sigma.B[,])
  Sigma.B[1,1] <- pow(sigma.a, 2)
  sigma.a ~ dunif (0, 100)
  Sigma.B[2,2] <- pow(sigma.b, 2)
  sigma.b ~ dunif (0, 100)
  Sigma.B[1,2] <- rho*sigma.a*sigma.b
  Sigma.B[2,1] <- Sigma.B[1,2]
  rho ~ dunif (-1, 1)
}
```

Multilivello non annidato per Sim e mese

```
model {
  for (i in 1:n){
    y[i] ~ dnorm (y.hat[i], tau.y)
    y.hat[i] <- inprod(B[id[i],],V[i,]) + inprod(G[wave[i],],Z[i,])
  }
  tau.y <- pow(sigma.y, -2)
  sigma.y ~ dunif (0, 100)

  for (j in 1:J){
    for(k in 1:K){
      B[j,k] <- xi[k]*B.raw[j,k]
    }
    B.raw[j,1:K] ~ dmnorm(mu.raw[], Tau.B.raw[,])
  }
  for(k in 1:K){
    mu[k] <- xi[k]*mu.raw[k]
    mu.raw[k] ~ dnorm(0, .0001)
    xi[k] ~ dunif(0, 100)
  }
  Tau.B.raw[1:K,1:K] ~ dwish(W[,],df)
  df <- K + 1
  Sigma.B.raw[1:K,1:K] <- inverse(Tau.B.raw[,])
  for(k in 1:K){
    for(k.prime in 1:K){
      rho.B[k,k.prime] <- Sig-
ma.B.raw[k,k.prime]/sqrt(Sigma.B.raw[k,k]*Sigma.B.raw[k.prime,k.prime])
    }
    sigma.B[k] <- abs(xi[k])*sqrt(Sigma.B.raw[k,k])
  }

  for (t in 1:T){
    a[t] <- G[t,1]
    b[t] <- G[t,2]
    G[t,1:2] ~ dmnorm (G.hat[t,], Tau.G[,])
    G.hat[t,1] <- mu.a
    G.hat[t,2] <- mu.b
  }
  mu.a ~ dnorm(0, .0001)
  mu.b ~ dnorm(0, .0001)

  Tau.G[1:2,1:2] <- inverse(Sigma.G[,])
  Sigma.G[1,1] <- pow(sigma.a, 2)
  sigma.a ~ dunif (0, 100)
  Sigma.G[2,2] <- pow(sigma.b, 2)
  sigma.b ~ dunif (0, 100)
  Sigma.G[1,2] <- rho*sigma.a*sigma.b
  Sigma.G[2,1] <- Sigma.G[1,2]
  rho ~ dunif (-1, 1)

}
```

Multilivello trivariato con intercetta variabile

```
model {

      for(i in 1:n){
            y[i,1:3] ~ dmnorm(y.hat[i,], Tau.Y.raw[,])
            y.hat[i,1] <- a1[id[i]] + inprod(b1[],V[i,])
            y.hat[i,2] <- a2[id[i]] + inprod(b2[],V[i,])
            y.hat[i,3] <- a3[id[i]] + inprod(b3[],V[i,])
      }

      for(l in 1:3){
            xi[l] ~ dunif(0, 100)
      }

      Tau.Y.raw[1:3,1:3] ~ dwish(W[,],df)
      df <- 4
      Sigma.Y.raw[1:3,1:3] <- inverse(Tau.Y.raw[,])
      for(l in 1:3){
            for(l.prime in 1:3){
                  rho.Y[l,l.prime] <- Sig-
ma.Y.raw[l,l.prime]/sqrt(Sigma.Y.raw[l,l]*Sigma.Y.raw[l.prime,l.prime])
            }
            sigma.Y[l] <- sqrt(Sigma.Y.raw[l,l])
      }

      for(k in 1:K){
            b1[k] ~ dnorm(0, 0.0001)
            b2[k] ~ dnorm(0, 0.0001)
            b3[k] ~ dnorm(0, 0.0001)
      }

      for (j in 1:J){
            a1[j] ~ dnorm (mu.a1, tau.a1)
            a2[j] ~ dnorm (mu.a2, tau.a2)
            a3[j] ~ dnorm (mu.a3, tau.a3)
      }
      mu.a1 ~ dnorm (0, .0001)
      tau.a1 <- pow(sigma.a1, -2)
      sigma.a1 ~ dunif (0, 100)

      mu.a2 ~ dnorm (0, .0001)
      tau.a2 <- pow(sigma.a2, -2)
      sigma.a2 ~ dunif (0, 100)

      mu.a3 ~ dnorm (0, .0001)
      tau.a3 <- pow(sigma.a3, -2)
      sigma.a3 ~ dunif (0, 100)
}
```

RIFERIMENTI BIBLIOGRAFICI

Gelman A., Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Tom A., Snijders and Roel J. Bosker.(1999), *An introduction to basic and advanced multilevel modeling*. Sage Publications.

Azzalini A., Scarpa B. (2004), *Analisi dei dati e data mining*. Milano, Springer.

Peter Congdon (2001), *Bayesian statistical modeling*. Wiley.

Harvey Goldstein (1999), *Multilevel statistical models*. Londra.

Software utilizzati:

R: Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.r-project.org.

WinBugs: (Bayesian Inference using Gibbs Sampling) software developed jointly with the Imperial College School of Medicine at St Mary's, London. URL http://www.mrc-bsu.cam.ac.uk/bugs/.

C++: linguaggio di programmazione orientato a oggetti. È stato sviluppato (in origine col nome di "C con classi") da Bjarne Stroustrup ai Bell Labs nel 1983 come un miglioramento del linguaggio C.