UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA SPECIALISTICA IN SCIENZE STATISTICHE,
ECONOMICHE, FINANZIARIE E AZIENDALI

PREVEDERE IL CHURN:
UN APPROCCIO LONGITUDINALE

RELATORE: Prof. SILVANO BORDIGNON

CORRELATORE: Prof. BRUNO SCARPA

LAUREANDA: MAELA BONETTO

A Leo,
e alla mia famiglia

Indice dei capitoli

1. Introduzione

Secondo le ultime rilevazioni Istat (2006), il telefono cellulare è uno dei beni tecnologici più diffusi in Italia, dopo il televisore: ormai è presente nel 82.3% delle famiglie, in aumento rispetto al 2005. Il settore degli operatori di telefonia mobile si rivolge ad un mercato sempre più saturo, nel quale è difficile trovare nuovi clienti. L'attenzione delle aziende si è spostata sul tentativo di attirare i clienti dei concorrenti, ma contemporaneamente è diventato fondamentale trattenere i propri clienti. In un campo in cui i costi di acquisizione di nuovi clienti sono alti, è necessario molto tempo prima che un cliente diventi profittevole, ma se egli cambia spesso fornitore, per le aziende diventa impossibile recuperare i costi sostenuti.

Il fenomeno si chiama *churn*, dal termine inglese che significa agitare, mescolare. Il *churn rate* misura il tasso di abbandono di un servizio da parte di un cliente, e si riferisce principalmente ai servizi di telefonia, sia fissa che mobile, ma anche ai servizi internet, bancari, o di assicurazione. Le aziende possono cercare di ridurre il *churn* creando barriere che scoraggiano il cliente dal passare ad un altro fornitore; ne sono esempio, nel campo bancario, il costo di chiusura del conto, e, fino a pochi anni fa, proprio nel mondo della telefonia mobile, l'impossibilità di tenere il proprio numero di telefono nel contratto con un altro fornitore. Altrimenti le aziende possono avviare programmi di fidelizzazione, ed incentivare i clienti a rimanere fedeli, ad esempio con una raccolta punti.

La prevenzione del churn può essere molto meno costosa del suo rimedio, vale a dire dell'acquisizione di nuovi clienti, perciò è ormai necessario che le aziende capiscano quali clienti sono a rischio di abbandono, per indirizzare loro azioni di trattenimento e fidelizzazione. Sono molti gli strumenti

statistici che permettono alle aziende di elaborare i dati che già conoscono riguardo ai propri clienti, allo scopo di capire il loro comportamento futuro.

Nelle aziende si possono perciò costruire modelli statistici, od utilizzare strumenti di data mining, che spiegano la relazione tra i dati osservati sui clienti, e l'eventuale disattivazione (Azzalini, Scarpa, 2004). Alcuni dati osservati sono di tipo sociodemografico, altri riguardano il contratto che il cliente ha sottoscritto con l'azienda, ed altri ancora registrano il traffico telefonico. Di questi, i primi due sono tipi di dati statici, nel senso che sono solitamente rilevati una sola volta, all'attivazione del contratto, mentre il terzo è dinamico, perché è misurato a cadenza periodica, spesso mensile, e può variare ad ogni misurazione.

Generalmente, gli strumenti utilizzati per capire il *churn* trattano tutti i tipi di dati come se fossero statici, ed inseriscono anche i dati relativi al traffico di mesi successivi tra le variabili che spiegano la disattivazione dei clienti (Nath, Behara, 2003). Questi dati, però, non sono indipendenti tra loro, e spesso seguono una traiettoria immaginaria, tipica per ogni cliente, che potrebbe caratterizzare il suo comportamento molto meglio dei dati singoli. Se immaginiamo di possedere il traffico di un anno per ogni cliente, abbiamo dodici variabili da utilizzare per la previsione, ognuna delle quali, probabilmente, piuttosto simile alla precedente e alla successiva. L'informazione che è portata da una variabile è quindi in parte espressa anche dalla successiva, quindi non sappiamo più identificare a quale variabile è dovuto il comportamento del cliente. Una traiettoria invece può essere sintetizzata in pochi parametri, spesso solo due, che potrebbero contenere buona parte dell'informazione utile all'analisi.

In letteratura si trovano facilmente strumenti di tipo statico, proposti allo scopo di capire come gli individui si dividano in due gruppi (attivi o disattivi, malati o guariti), ma gli strumenti che utilizzano le sintesi dei dati dinamici sono ancora poco trattati, e sono soprattutto applicati in campo biomedico, per spiegare la sopravvivenza o la guarigione dei pazienti. Si veda ad esempio Wang, Wang, Wang (2000), e Li, Zhang, Davidian (2004). Noi appli-

cheremo questo tipo di analisi alla previsione del *churn*, per verificare se consente di ottenere risultati migliori rispetto ad un'analisi statica.

In questa tesi eseguiremo l'analisi su un campione di circa 32 mila clienti di un'azienda di telefonia mobile, per i quali conosciamo alcune caratteristiche (dati statici), ed il traffico di diciotto mesi consecutivi; inoltre sappiamo chi di loro ha lasciato l'azienda, per passare ad un concorrente, nel corso dell'ultimo mese.

Il Capitolo 2 è dedicato alle analisi grafiche ed esplorative, che abbiamo eseguito sui dati per intuire una qualche relazione tra le variabili ed il comportamento finale. Nel Capitolo 3 costruiamo due strumenti di previsione che trattano i dati in modo statico, un modello di regressione logistica ed una classificazione ad albero, strumenti classici che potrebbero essere già comunemente usati nelle aziende. Nel Capitolo 4, invece, ipotizziamo due tipi di traiettorie per i dati dinamici, il modello ad effetti casuali ed il lisciamento esponenziale, e per entrambi otteniamo alcuni parametri che sintetizzano la traiettoria. Questi parametri sono poi usati nel Capitolo 5 come variabili esplicative, negli stessi modelli costruiti nel Capitolo 3, per integrare l'informazione statica con l'informazione longitudinale. I modelli ottenuti nei Capitoli 3 e 5 sono confrontati rispetto alla loro capacità di prevedere i clienti che si disattiveranno, per capire se il nuovo metodo è davvero utile per migliorare la comprensione del comportamento dei clienti.

L'attività di preparazione del dataset e di stesura della tesi è stata eseguita con le applicazioni di Microsoft Office 2003, mentre tutte le analisi sono state eseguite con il software statistico R. Maggiori informazioni riguardo questo software ed i pacchetti aggiuntivi utilizzati si possono trovare in bibliografia.

2. Caratteristiche del campione e prime analisi esplorative

Nell'introduzione abbiamo esposto il problema che vogliamo risolvere con questa tesi, ora possiamo presentare i dati su cui sarà effettuata l'analisi. Il campione è stato estratto dal database clienti di un'azienda di telefonia mobile, e contiene 32 524 clienti, scelti casualmente tra coloro che hanno sottoscritto il contratto con l'azienda nell'anno 2003, e che alla fine di marzo 2006 erano ancora attivi. Di questi è stato registrato il traffico telefonico per diciotto mesi, da novembre 2004 ad aprile 2006; inoltre si conoscono alcuni dati sociodemografici e le caratteristiche del contratto firmato, ed è noto se ad aprile 2006 hanno lasciato l'azienda.

In questo capitolo spiegheremo i tipi di variabili che abbiamo a disposizione, ed eseguiremo alcune analisi esplorative, soprattutto grafiche, sui dati osservati, che ci permetteranno di capire alcuni comportamenti e in particolare la relazione con l'eventuale disattivazione.

2.1. Variabili non dipendenti dal tempo

Nella Tabella 2.1 elenchiamo le variabili, con i valori che esse assumono, e di seguito descriviamo le loro caratteristiche, anche commentando i grafici della Figura 2.1 e seguenti.

a) La variabile Stato è la variabile risposta, perché indica se l'utente, al termine del mese di aprile 2006, è ancora cliente dell'azienda (attivo), oppure si è disattivato proprio in quel mese. Una persona non può essere contemporaneamente attiva e disattiva, quindi i valori si escludono a vicenda

e dividono il campione in due gruppi distinti, il gruppo degli Attivi, che contiene 30454 utenti, e il gruppo dei Disattivi, del quale fanno parte 2070 utenti. Ricordiamo però che i due gruppi sono noti solo a posteriori, perché fino a marzo 2006 tutti gli utenti sono ancora attivi; quello che dovremo capire, in seguito, sarà la regola che divide gli utenti nei due gruppi.

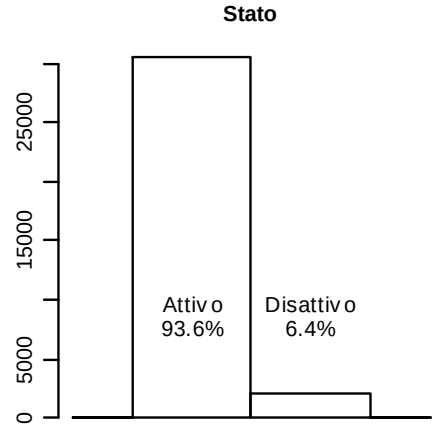

Figura 2.1. Frequenze dei valori assunti dalla variabile risposta Stato. Le percentuali sono calcolate sul totale del campione.

	Nome	Tipo di variabile	Valori assunti	Figura
a)	Stato	Fattore a 2 livelli	Disattivi, Attivi	2.1
b)	MNP	Fattore a 2 livelli	No, Sì	2.2, a sinistra
c)	PianoTariffario	Fattore a 8 livelli	A, B, C, D, E, F, G, H	2.2, a destra
d)	CanaleVendita	Fattore a 6 livelli	A, B, C, D, E, F	2.3, in alto
e)	Marca	Fattore a 5 livelli	LG, Motorola, Nec, Nokia, Sony Ericsson	2.3, al centro
f)	Età	Quantitativa	{17, 18, …, 91}	2.3, in basso a sinistra
g)	Sesso	Fattore a 2 livelli	Maschio, Femmina	2.3, in basso a destra
h)	Provincia	Fattore a 104 livelli	AG, AL, AN, …, VV	2.4
i)	Zona	Fattore a 5 livelli	Nordest, Nordovest, Centro, Sud, Isole	2.5

Tabella 2.1. Variabili osservate sugli individui e loro caratteristiche.

b) La seconda variabile della Tabella 2.1 si chiama MNP, che significa *Mobile Number Portability*. Questo è un servizio che consente di mantenere il proprio numero di telefono cellulare nel passaggio da un operatore di telefonia mobile ad un altro. La variabile indica se l'utente, all'attivazione del contratto, ha usufruito o meno di tale servizio. Nella Figura 2.2, nel pannello a destra, sono rappresentate le frequenze con cui la variabile assume i due valori nei gruppi. Notiamo che le percentuali per entrambi i valori sono simili tra i gruppi, come se non ci fosse alcuna relazione tra la variabile e lo Stato.

c) La variabile Piano Tariffario indica il nome del contratto firmato dall'utente, al quale corrispondono le tariffe telefoniche ed il tipo di pagamento. Ogni azienda telefonica ha diversi tipi di piani tariffari, ma di questi non sappiamo nulla, perché la variabile è schermata, cioè codificata con le lettere dell'alfabeto, dalle quali non possiamo risalire al nome del piano tariffario. Se l'analisi fosse eseguita in azienda, invece, sarebbe molto semplice collegare ogni piano tariffario alle sue caratteristiche. Ad esempio, il piano A è il più diffuso, ma con una percentuale molto maggiore tra i Disattivi che tra gli Attivi: potrebbe essere utile capire se qualche sua caratteristica spinge i clienti a disattivarsi.

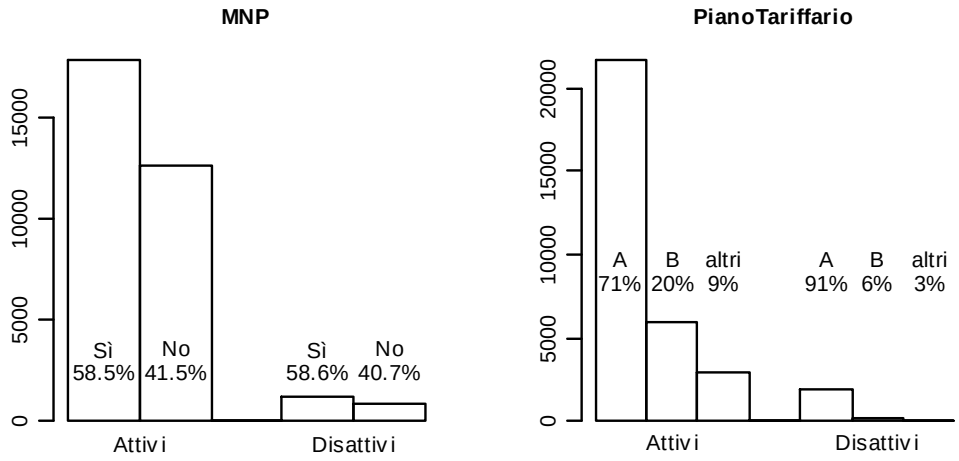

Figura 2.2. Frequenze dei valori assunti dalle variabili MNP e PianoTariffario, divise per i due gruppi, Attivi e Disattivi. Le percentuali riportate si intendono sul totale del gruppo; non sempre sommano a 100, e ciò è dovuto alla mancanza dei dati per alcune persone.

d) Il Canale di Vendita è la variabile che registra il tipo di negozio in cui l'utente ha attivato il contratto, e può riferirsi ad esempio ai negozi in franchising, o agli ipermercati. Anche per questa variabile non conosciamo i valori, perché sono schermati.

e) La marca e il modello del telefonino sono dati presenti solo per gli utenti che hanno attivato il contratto all'acquisto di un nuovo telefono, e non per quelli che hanno cambiato gestore mantenendo il proprio cellulare, perciò non è disponibile per la maggior parte delle persone. Tuttavia possiamo riportare in Figura 2.3 la distribuzione delle marche per i pochi valori a disposizione. Se fossimo in azienda potremmo anche raggruppare i modelli per similitudine, in collaborazione con l'ufficio marketing, per ottenere ulteriori informazioni riguardo alle caratteristiche dei telefoni che non soddisfano i clienti.

f) Nel dataset è contenuta anche la data di nascita di quasi tutti gli utenti, dalla quale abbiamo potuto calcolare l'età al 31 marzo 2006. L'età media è di 39 anni per il gruppo Attivi, e di 37 per i Disattivi. Nel pannello in basso a sinistra di Figura 2.3 abbiamo rappresentato la distribuzione dell'Età con due boxplot, per gli Attivi e i Disattivi. In entrambi i grafici un rettangolo (la "scatola") racchiude i valori compresi tra il primo e il terzo quartile, mentre la linea orizzontale più spessa è la mediana. Le linee tratteggiate che escono dal riquadro si chiamano "baffi", e rappresentano le code della distribuzione. Molti punti sono più in alto del baffo superiore, perché l'età non ha una distribuzione perfettamente simmetrica, ed assume anche valori molto alti. Notiamo che i Disattivi sono leggermente più giovani degli Attivi, perché la scatola corrispondente è situata più in basso. Quindi è possibile che la variabile abbia una certa relazione con lo Stato, tale che i giovani tendono a disattivarsi di più.

g) Al momento della firma del contratto è stato chiesto agli utenti di indicare il loro sesso, e le frequenze dei due valori che esso assume sono nel pannello in basso a destra di Figura 2.3. La maggior parte del campione è formata da uomini, ed entrambi i sessi sembrano dividersi casualmente tra Attivi e Disattivi.

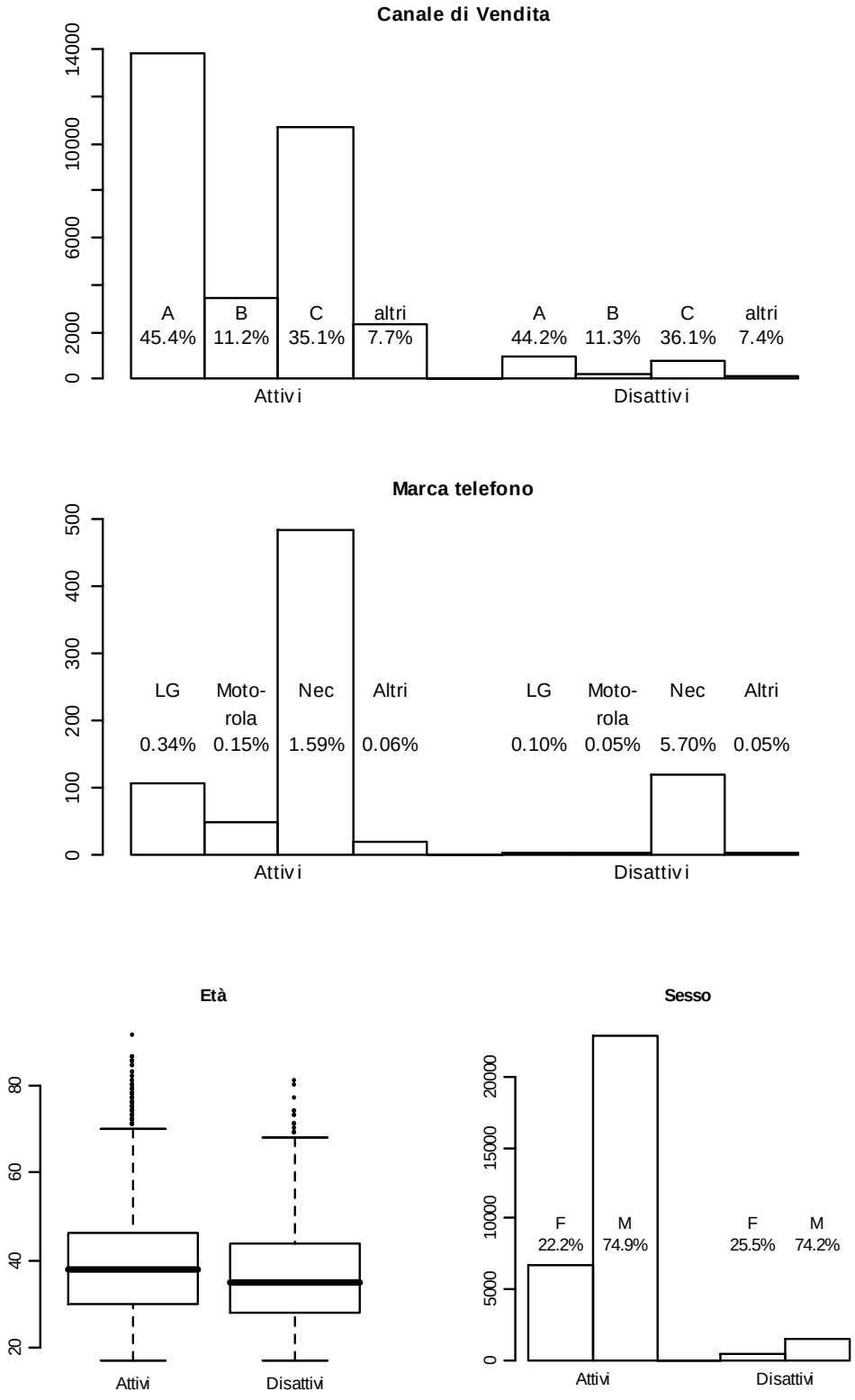

Figura 2.3. Valori assunti da altre variabili del dataset. Le percentuali sono da intendere sul totale del gruppo (Attivi o Disattivi), e la somma è inferiore al 100% per mancanza di dati per alcuni individui.

h) Le province di provenienza degli utenti sono indicate in Figura 2.4, dove a colori più scuri corrisponde una percentuale più alta di Disattivi rispetto al totale di clienti della provincia. In legenda abbiamo riportato alcuni valori soglia; colori intermedi corrispondono a valori intermedi della percentuale. La maggior parte delle province è molto vicina alla media nazionale, che è circa 6%, e nessuna di esse è inferiore al 3%, con l'eccezione di Enna ed Oristano, nelle quali nessun cliente si è disattivato. Per comodità di analisi, comunque, la variabile è raggruppata in zone ed assume solo cinque valori, che rappresentiamo in Figura 2.5.

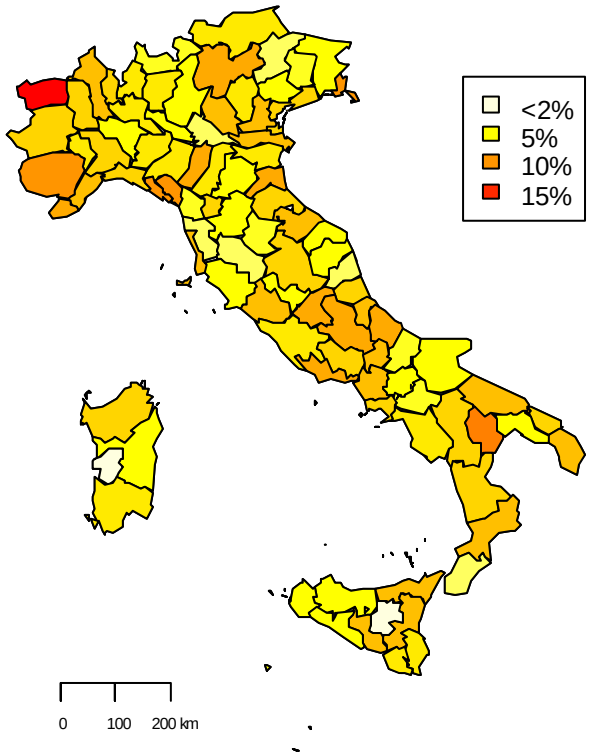

Figura 2.4. Percentuale di disattivi sul numero di clienti, divisa per province. Bisogna però notare che la mappa è aggiornata al 1998, perciò nella cartina sono assenti le province create in seguito.

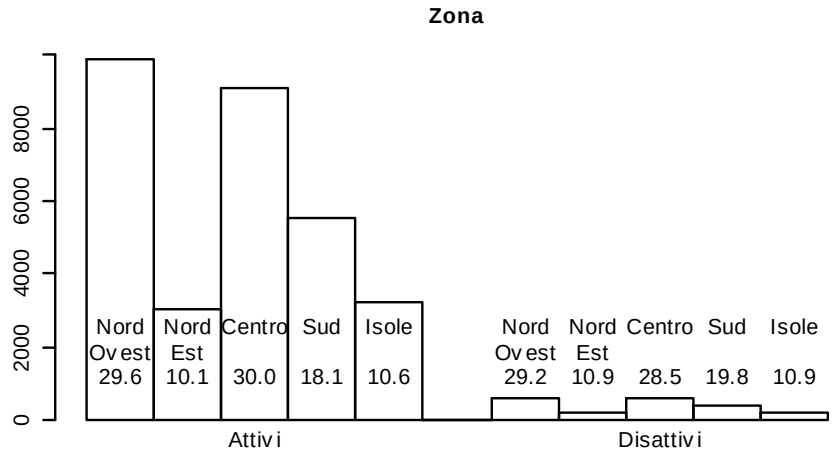

Figura 2.5. Distribuzione dei valori della variabile Zona, divisa per gruppi.

Molte delle variabili presentate in questo paragrafo saranno utilizzate nei modelli dei prossimi capitoli come variabili esplicative per la risposta Stato, perciò è utile eseguire ora un test per verificare l'eventuale indipendenza marginale tra le prime e la seconda. Il test χ^2 è utilizzato a questo scopo, perché confronta la distribuzione di una variabile qualitativa rispetto alla risposta, anch'essa qualitativa, con la distribuzione stimata sotto l'ipotesi di indipendenza tra le due. I risultati del test sono riportati in Tabella 2.2. Notiamo che tre variabili hanno valori del test significativamente grandi, quindi ci aspettiamo una certa correlazione con la variabile risposta. Le altre due variabili invece hanno un valore del test piccolo, perciò affermiamo che sono marginalmente indipendenti dalla risposta.

	χ^2	Gradi di libertà	p-value	Indipendenza
MNP	0.228	1	0.632	Sì
CanaleVendita	24.347	5	0	No
PianoTariffario	403.329	7	0	No
Zona	5.543	4	0.235	Sì
Sesso	7.94	1	0.004	No

Tabella 2.2. Valori del test χ^2 per le variabili qualitative.

2.2. Variabili longitudinali

La seconda parte del dataset riguarda il traffico telefonico di ogni utente nei diciotto mesi dell'osservazione. Per ogni utente abbiamo osservato cinque serie storiche mensili, che ora analizzeremo preliminarmente. Per la prima serie esporremo ogni passo dell'analisi, mentre per le altre riporteremo solo i risultati essenziali, derivati dallo stesso processo di analisi, ed eventuali deviazioni dal percorso.

2.2.1. Messaggi di testo (sms)

Cominciamo dunque ad analizzare la prima serie, che conta il numero di sms spediti, ogni mese, dagli utenti.

Il modo più semplice per capire come una serie storica cambia è guardare il grafico della variabile rispetto al tempo. Quando però il numero di serie da analizzare è alto, non è possibile rappresentarle tutte in grafici separati, ed anche disegnarle nello stesso grafico aiuta poco, come si vede nella Figura 2.6.

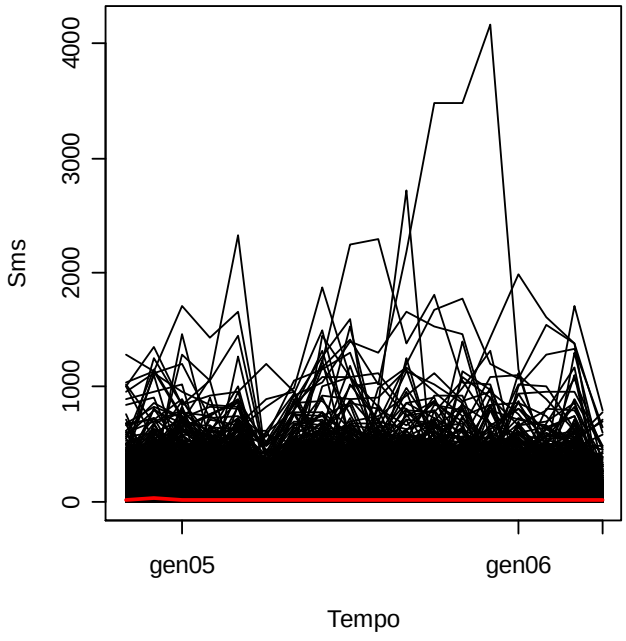

Figura 2.6. Rappresentazione della variabile Sms rispetto al tempo.

Il grafico riporta il numero di messaggi spediti da ogni utente per ogni mese. Le linee sembrano concentrarsi sui valori bassi, perché per valori inferiori a 500 si vede solo una macchia nera compatta; poche invece raggiungono valori alti. La linea rossa in basso è la media, ma ha valori troppo bassi, rispetto alla scala del grafico, perché si capisca la sua forma, perciò la rappresentiamo a parte, in Figura 2.7, insieme alle medie dei due gruppi.

La media complessiva (in rosso), è molto simile alla media degli utenti attivi, ed è quello che ci possiamo aspettare, dato che i Disattivi sono solo una piccola parte del campione. Nella figura si possono notare anche alcuni comportamenti tipici, come il picco in corrispondenza del mese di dicembre, in entrambi gli anni per gli Attivi, e nel primo anno solamente per i Disattivi, spiegabile facilmente con le festività natalizie. In entrambi gli anni si nota anche un calo notevole nel mese di aprile, e questo probabilmente coincide con qualche fenomeno che non conosciamo, o magari con il termine di una promozione.

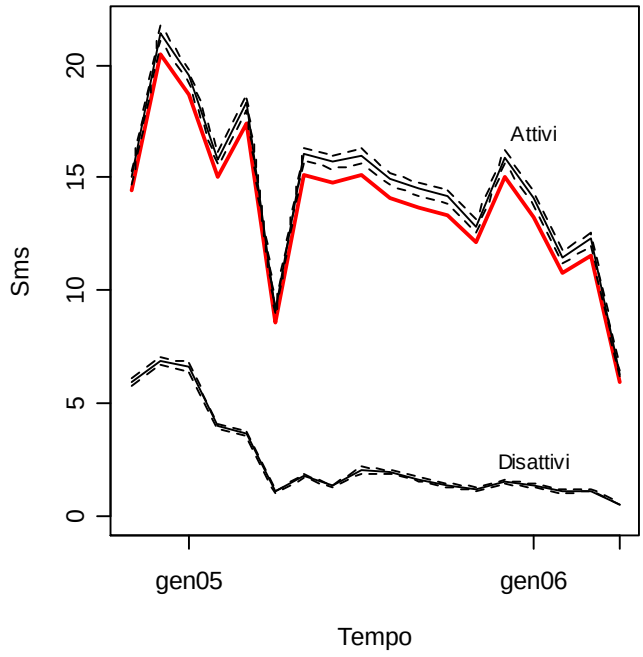

Figura 2.7. Grafico della media della variabile Sms, in rosso, e delle medie dei due gruppi (Attivi e Disattivi). Le linee tratteggiate sono alla distanza di uno standard error dalla media corrispondente.

Infine, si vede alla prima occhiata che le medie dei due gruppi sono piuttosto distanti. Un futuro disattivo scrive in media un quarto degli sms rispetto ad un attivo, ed il divario diventa più evidente a partire dal mese di aprile 2005. Infatti, dopo il "crollo" avvenuto quel mese, gli sms degli attivi tornano a salire, mentre quelli dei Disattivi si appiattiscono ad una quota molto vicina allo zero.

2.2.1.1. Traiettorie empiriche

Un'analisi sulle medie, però, non è sufficiente a cogliere i comportamenti, e spesso è utile analizzare un piccolo campione di utenti. Scegliamo casualmente dodici utenti tra tutti quelli che spediscono meno di 50 sms al mese, e rappresentiamo l'andamento dei loro gli sms in Figura 2.8.

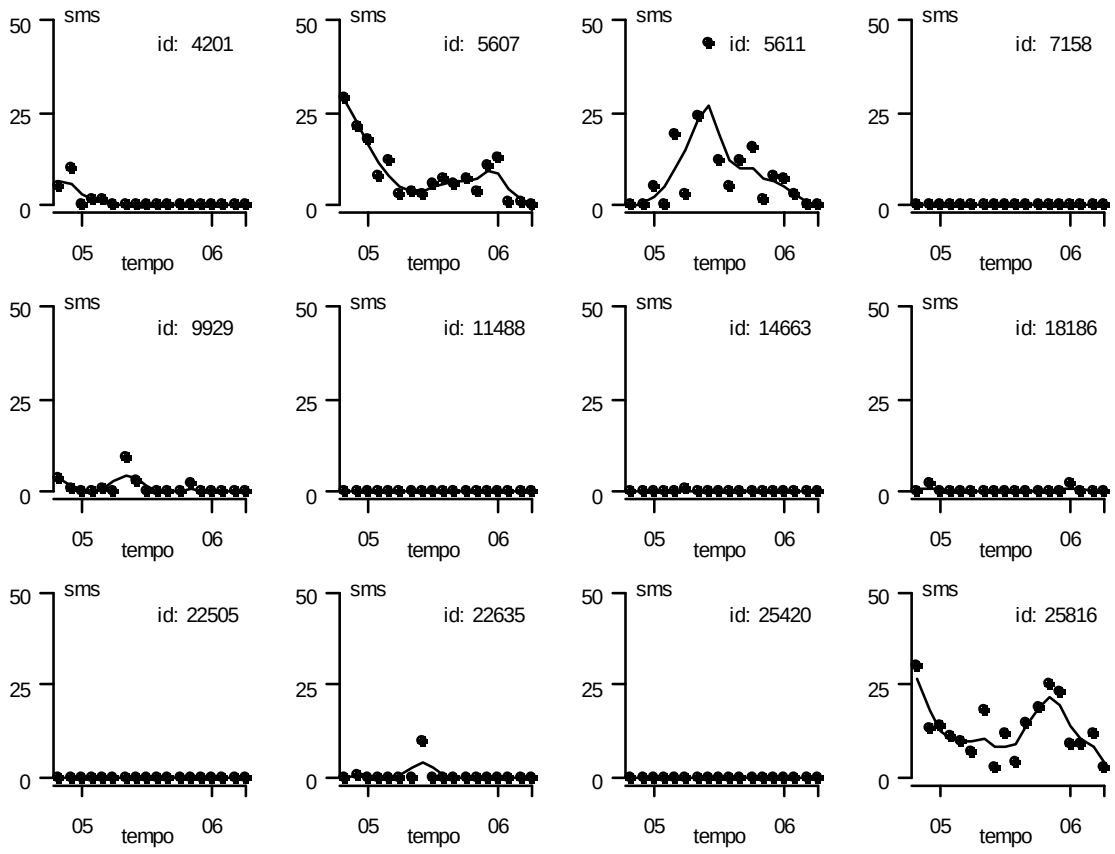

Figura 2.8. Grafico degli sms contro il tempo per dodici utenti scelti casualmente tra coloro che spediscono meno di 50 sms al mese, ed il lisciamento non parametrico dei dati. In ogni pannello, in alto a destra è indicato il numero di id del cliente al quale si riferisce la serie.

Tutti i grafici sono riportati sulla stessa scala, per essere confrontati, e notiamo che prevale un andamento piatto, o quasi, vicino allo zero. Gli utenti che, invece, spediscono sms, tendono a diminuire il numero degli stessi nel corso del tempo, come avveniva pure nelle medie del grafico precedente.

Per comprendere meglio gli andamenti abbiamo deciso di riassumere il grafico di ogni persona in una traiettoria empirica. Qui abbiamo scelto un lisciamento non parametrico, per "far parlare i dati" senza imporre loro alcuna struttura, ma possiamo utilizzare anche una forma funzionale parametrica semplice, come una retta o una curva quadratica. Dalla figura però è evidente che, ad esclusione degli utenti che non scrivono alcun sms, l'andamento è difficilmente parametrizzabile, perché le curve empiriche non sono lineari, ed hanno frequenti cambi di direzione.

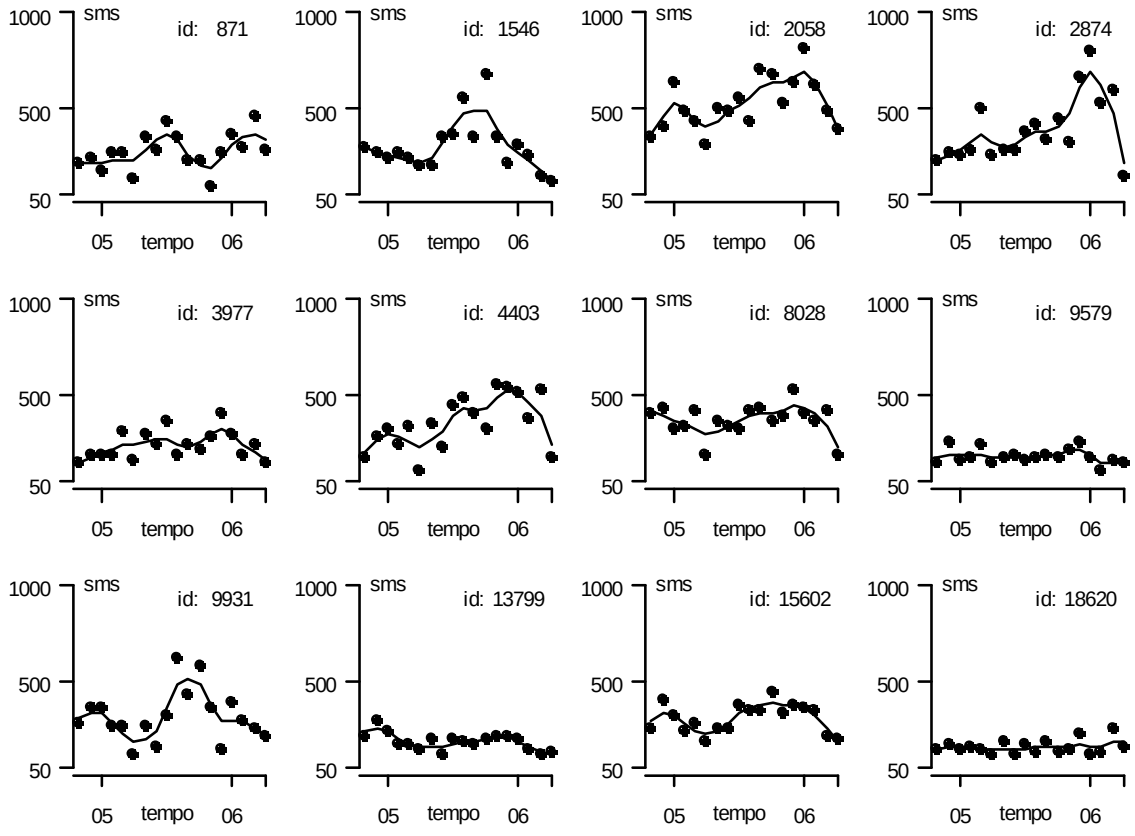

Figura 2.9. Grafico degli sms contro il tempo per altri dodici utenti, scelti casualmente tra coloro che spediscono almeno 100 sms al mese, ed il lisciamento non parametrico dei dati.

Questo è particolarmente vero se scegliamo, invece, un campione casuale tra gli utenti che spediscono più di 100 sms al mese, come abbiamo fatto in Figura 2.9. Più che sulla figura precedente, quindi, è su questa che dovremo stimare un buon modello parametrico.

La scelta di una forma funzionale, però, sembra complicata, perché l'andamento appare lineare per alcuni individui, e curvilineo per altri, ma non è sensato imporre forme diverse a serie diverse, perché i risultati non sarebbero più confrontabili. Spesso la scelta migliore è una forma semplice, che può essere complicata poi secondo la necessità, quindi scegliamo una traiettoria lineare, che potrebbe descrivere abbastanza bene gli andamenti dei dodici utenti, e semplifica sia il confronto, sia l'interpretazione: l'intercetta rappresenta il livello iniziale di consumo, e la pendenza è il tasso mensile di variazione della quantità di sms.

La prima forma funzionale scelta è la seguente:

$$\text{Sms}_{it} = \beta_{0i} + \beta_{1i}\,\text{tempo}_{it} + \varepsilon_{it}\,,$$

dove i indica l'utente e t il tempo. In Figura 2.10 sovrapponiamo la retta di regressione al grafico di alcuni utenti.

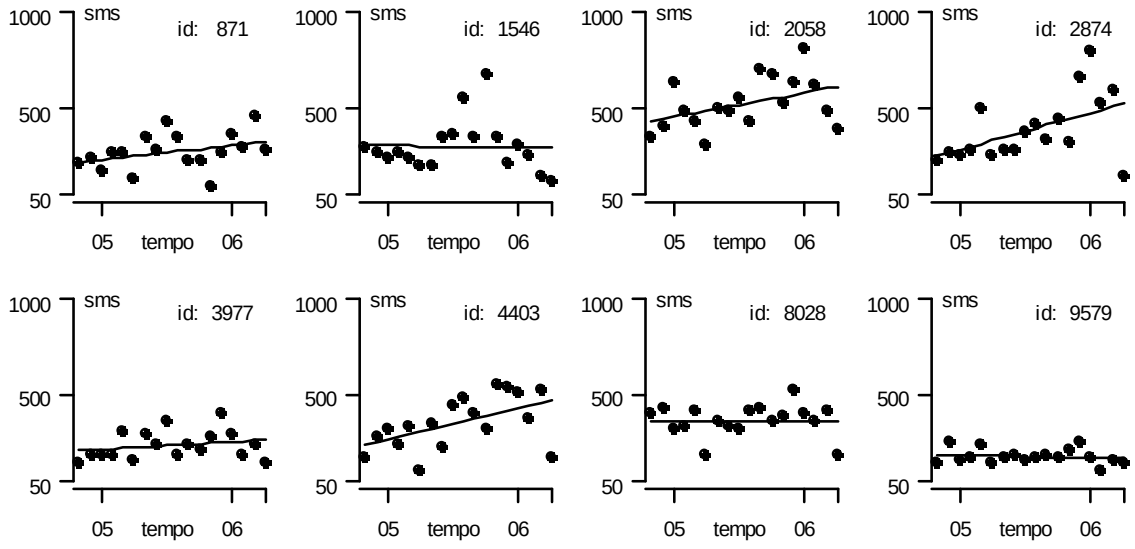

Figura 2.10. Gli sms di otto dei dodici utenti, con la rispettiva retta di regressione stimata.

Nella Tabella 2.3, invece, riportiamo i coefficienti della regressione stimati per i dodici utenti della Figura 2.9. I parametri stimati variano di molto tra gli individui, come ci aspettavamo dopo aver visto le traiettorie molto varie della Figura 2.10. L'R^2 delle regressioni è decisamente basso, perché i punti sono piuttosto dispersi per la maggior parte dei dodici utenti.

id_i	$\hat{\beta}_{0i}$	s.e.	$\hat{\beta}_{1i}$	s.e.	R^2
871	217.693	44.267	6.249	4.089	0.127
1546	311.876	68.144	-0.817	6.295	0.001
2058	428.889	61.519	10.404	5.683	0.173
2874	237.464	75.408	16.191	6.966	0.252
3977	206.719	36.448	3.357	3.367	0.058
4403	230.144	59.944	13.166	5.537	0.261
8028	368.503	41.807	0.099	3.862	0.000
9579	192.229	19.685	-1.240	1.818	0.028
9931	304.869	67.996	1.125	6.281	0.002
13799	221.327	19.994	-3.847	1.847	0.213
15602	295.523	37.876	1.097	3.499	0.006
18620	141.294	18.667	1.671	1.724	0.055

Tabella 2.3. Coefficienti della regressione lineare per i dodici utenti scelti casualmente, con i relativi standard error stimati, e bontà della regressione.

Stimiamo comunque il modello lineare per tutti gli utenti. I grafici in Figura 2.11 rappresentano la distribuzione dei β_0 stimati per i due gruppi. Il picco di massima frequenza si ha intorno allo zero, ma gli Attivi assumono un range di valori molto grande, mentre i Disattivi hanno un range più piccolo.

In Figura 2.12 raffiguriamo la distribuzione dei β_1 stimati. Anche questo parametro ha frequenza massima per valori nulli, ed assume valori più vari per gli Attivi che per i Disattivi, per i quali è molto più concentrato attorno allo zero. Questo significa che il valore stimato per β_1 potrebbe non essere significativo per molti individui, e quindi che le loro traiettorie potrebbero essere semplicemente delle rette orizzontali che non dipendono dal tempo.

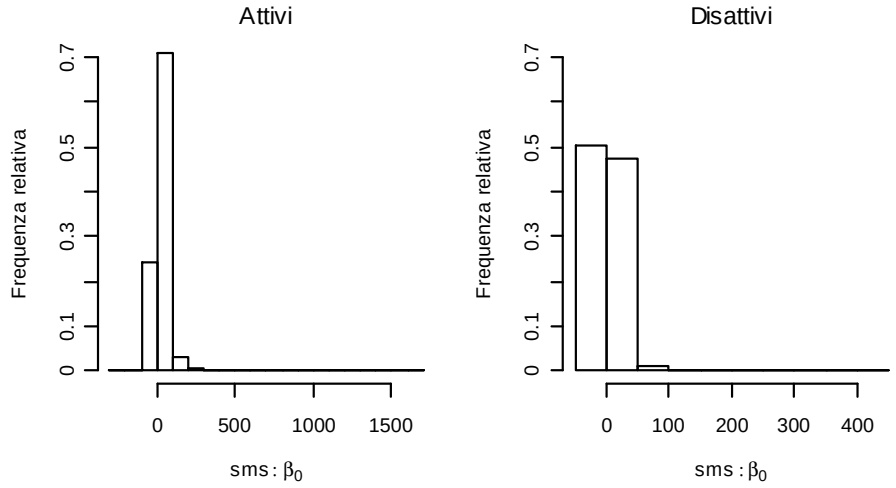

Figura 2.11. La distribuzione dei β_0 stimati per il modello lineare, a sinistra per gli Attivi, a destra per i Disattivi.

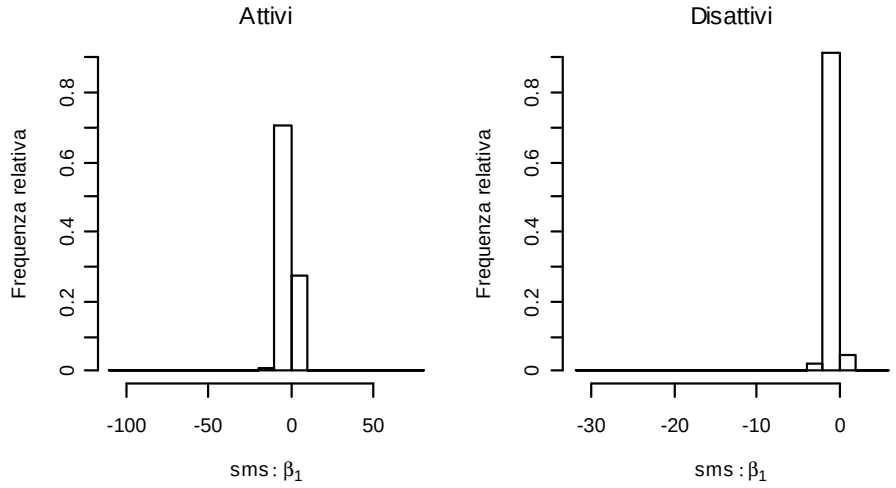

Figura 2.12. Distribuzione empirica dei β_1 stimati per il modello lineare, divisi nei due gruppi.

Notiamo inoltre che, mentre per gli Attivi la pendenza assume quasi simmetricamente sia valori positivi sia negativi, per i Disattivi assume soprattutto valori negativi, che corrispondono ad un andamento in calo.

Resta un'ultima questione, che riguarda la bontà dei modelli così stimati. Nella Tabella 2.3 avevamo notato che l'R^2 era generalmente basso, ma ora vogliamo vedere cosa accade per gli altri utenti. Raffiguriamo quindi la distribuzione del R^2 in un istogramma, non prima di aver apportato una pic-

cola modifica ai dati stimati. Infatti, per gli utenti che non scrivono alcun sms, un modello lineare può stimare perfettamente l'andamento senza errore. In questo caso, l'R^2 non è definito perché ha come denominatore la devianza residua, che è pari a zero. Però questo modello non è stimato male, anzi è perfettamente adattato ai dati, e decidiamo di porre il valore del R^2 uguale ad uno.

Raffiguriamo in Figura 2.13 la distribuzione degli R^2 così modificati. Notiamo che la maggioranza degli utenti Attivi è stimata molto male, poiché il picco più alto dell'istogramma è in corrispondenza dello zero, e che in fondo solo gli utenti che sono fissi a zero sms sono stimati bene. Per i Disattivi, è vera la stessa affermazione, ma qui gli utenti fissi a zero sono più frequenti, quasi metà gruppo, e perciò la barra corrispondente al valore uno è molto più alta delle altre.

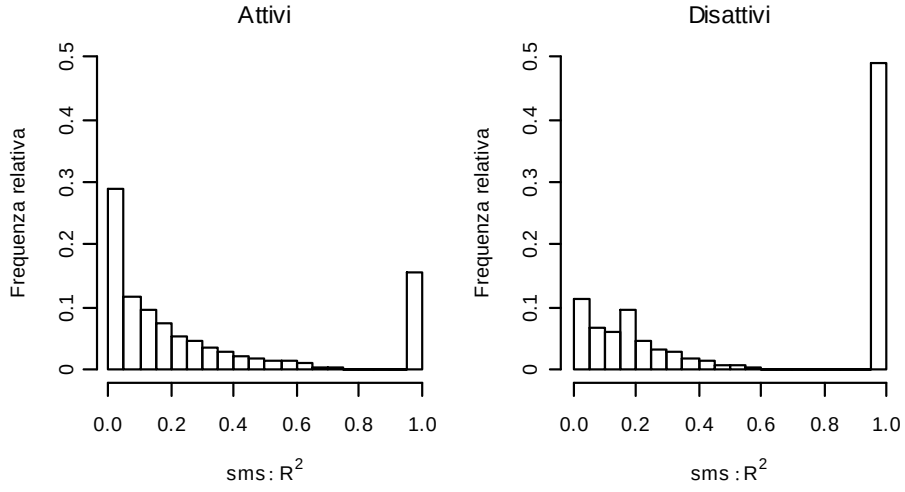

Figura 2.13. Distribuzione empirica del R^2 stimato per il modello lineare, diviso nei due gruppi.

Le traiettorie di molti utenti, però, sono stimate male, quindi possiamo ipotizzare un modello più complesso per spiegarle, ad esempio un modello quadratico nel tempo, che approssimi la traiettoria con una forma parabolica, del tipo:

$$Sms_{it} = \beta_{0i} + \beta_{1i}\, tempo_{it} + \beta_{2i} tempo_{it}^2 + \varepsilon_{it}$$

Stimiamo per prima cosa il nuovo modello per i dodici utenti campione scelti per la Figura 2.9. I coefficienti stimati sono riportati in Tabella 2.4, con i rispettivi standard error e l'R^2 della regressione.

I parametri di questo modello sono meno facili da interpretare rispetto a quelli del modello lineare, perché il termine lineare e quello quadratico possono avere effetti opposti. Si può invece notare come l'R^2 sia aumentato rispetto ai valori della Tabella 2.3, pur rimanendo in genere piuttosto basso.

id_i	$\hat{\beta}_{0i}$	s.e.	$\hat{\beta}_{1i}$	s.e.	$\hat{\beta}_{2i}$	s.e.	R^2
871	218.375	73.785	6.044	17.880	0.011	0.914	0.127
1546	126.843	95.901	54.692	23.240	-2.922	1.189	0.288
2058	331.103	97.367	39.739	23.595	-1.544	1.207	0.255
2874	171.941	123.832	35.848	30.008	-1.035	1.535	0.274
3977	115.265	52.782	30.793	12.791	-1.444	0.654	0.289
4403	126.230	93.889	44.340	22.752	-1.641	1.164	0.348
8028	349.561	69.407	5.782	16.820	-0.299	0.860	0.008
9579	173.115	32.205	4.494	7.804	-0.302	0.399	0.064
9931	218.407	109.712	27.064	26.587	-1.365	1.360	0.065
13799	229.628	33.214	-6.337	8.049	0.131	0.412	0.219
15602	233.677	59.766	19.651	14.483	-0.977	0.741	0.109
18620	161.000	30.432	-4.241	7.375	0.311	0.377	0.096

Tabella 2.4. Coefficienti stimati per il modello quadratico sui dodici utenti campione.

La curva di regressione stimata per otto dei dodici utenti è rappresentata in Figura 2.14. Il nuovo modello sembra cogliere molto bene alcuni comportamenti, ma non riesce a spiegare i tracciati più irregolari.

Stimiamo ora questo modello per tutti gli utenti, e svolgiamo le stesse analisi presentate per il modello lineare. Le distribuzioni empiriche dei parametri stimati sono in Figura 2.15.

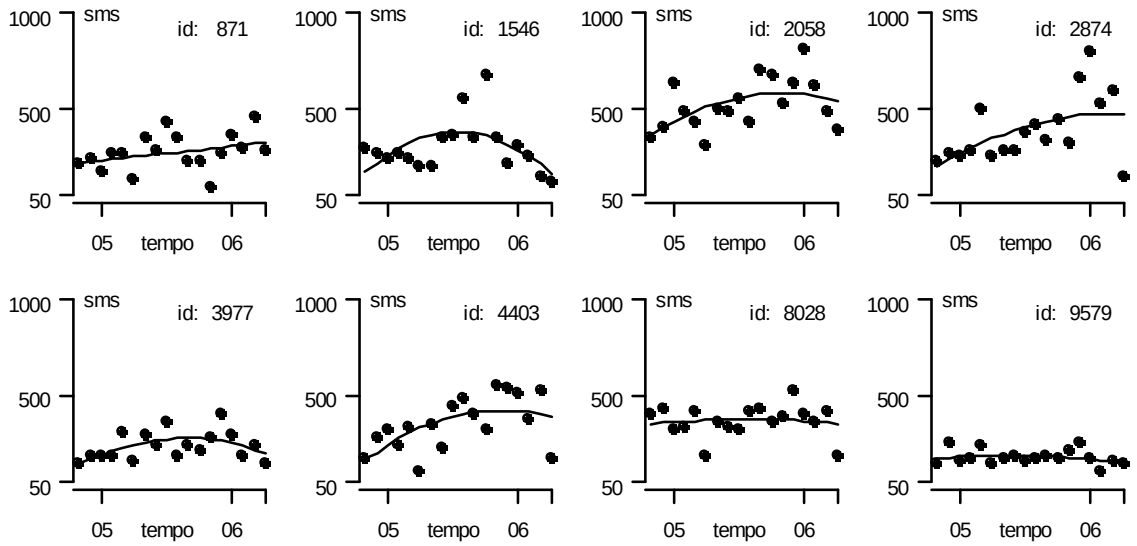

Figura 2.14. La curva quadratica stimata per otto dei dodici utenti campione.

Anche in questo modello i parametri stimati si concentrano fortemente attorno allo zero, e sembra che per buona parte dei Disattivi il miglior modello potrebbe essere ancora un modello costante nel tempo, perché i coefficienti non sono significativi; per gli Attivi invece le stime sono più varie. La distribuzione degli R^2, corretti allo stesso modo del modello lineare, è rappresentata in Figura 2.16. Ora il picco più alto, per Attivi e per Disattivi, si ha in corrispondenza dell'uno, ad indicare perfetto adattamento del modello ai dati, ma sono ancora numerosi i modelli con scarso adattamento.

Sappiamo tuttavia che non è possibile, senza aumentare eccessivamente il numero di parametri, riuscire a spiegare tutti i cambiamenti di direzione delle serie che stiamo analizzando. Nei prossimi capitoli utilizzeremo altri strumenti adatti allo scopo, ma, per l'analisi esplorativa, ci accontentiamo dei risultati ottenuti.

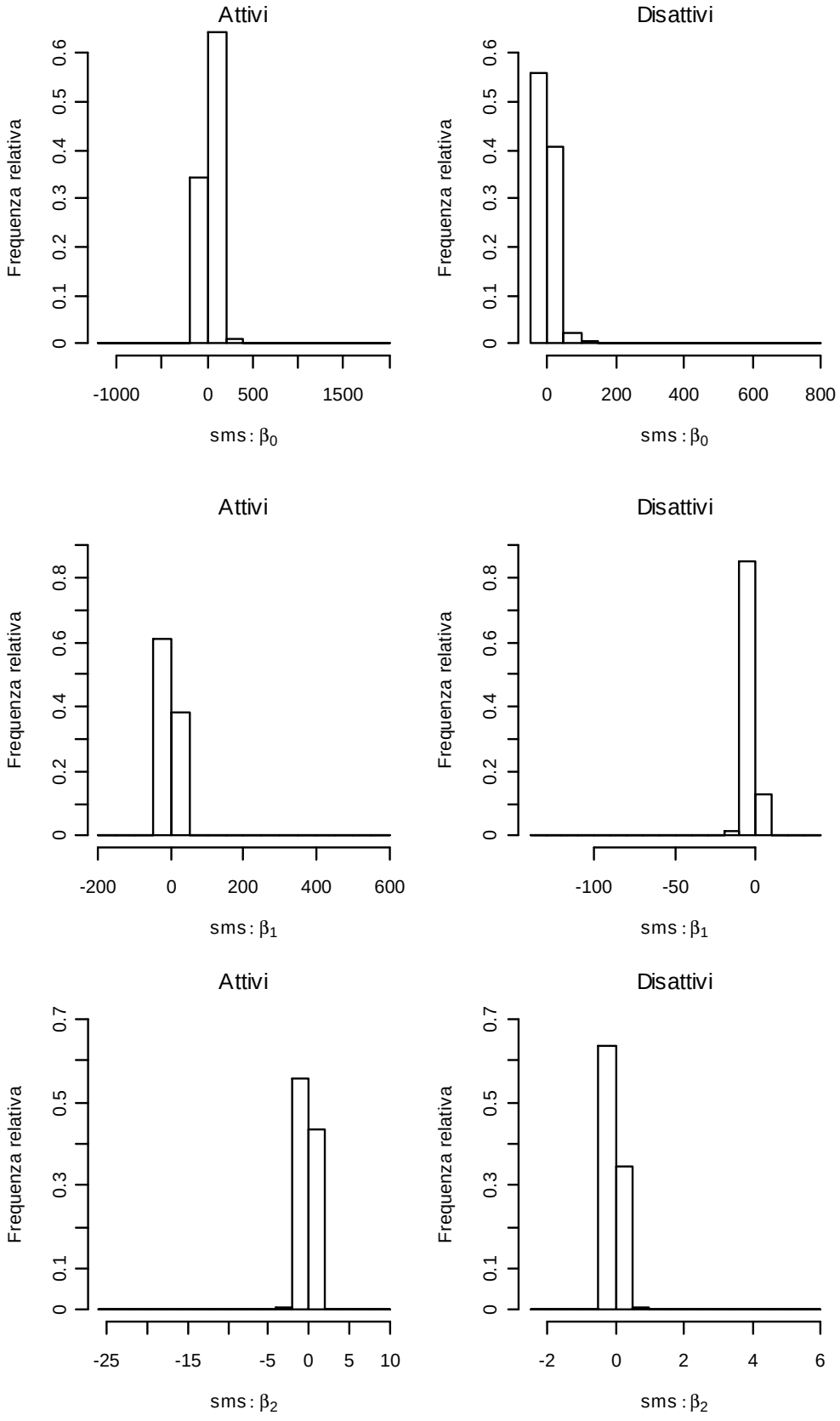

Figura 2.15. Distribuzione dei parametri stimati del modello quadratico, divisi per gruppi.

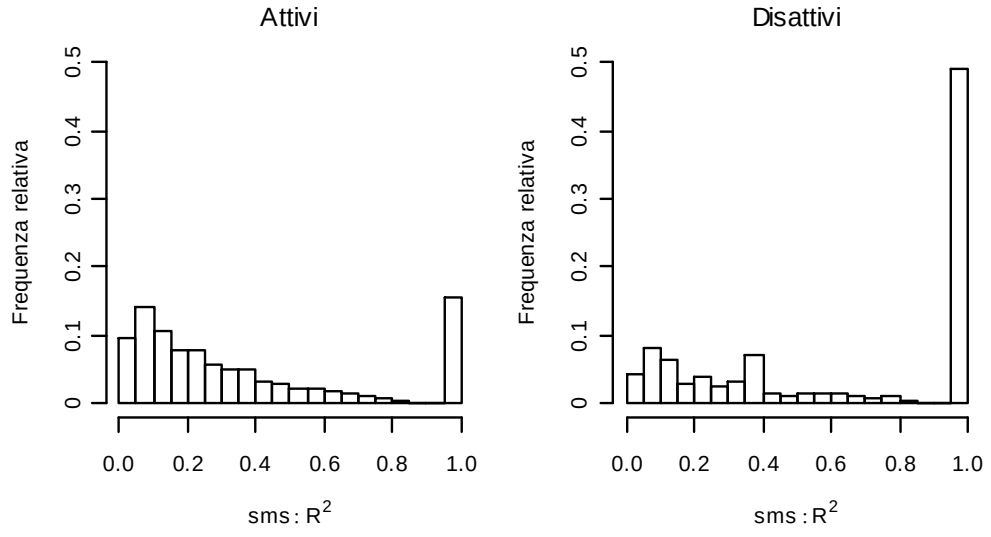

Figura 2.16. Distribuzione empirica del R^2 per il modello quadratico, diviso per i gruppi.

2.2.1.2. Esplorare le differenze tra persone

Possiamo ora cominciare a capire quanto gli individui sono diversi nelle loro traiettorie. Abbiamo infatti alcuni elementi per tratteggiare una prima analisi sui parametri della regressione parametrica.

Osserveremo perciò alcune quantità:

a) Media campionaria dei parametri stimati. I parametri stimati con il metodo dei minimi quadrati sono stime non distorte dei veri parametri per ogni persona, perciò possiamo definire la loro media campionaria come la stima non distorta dei veri parametri della traiettoria media.
b) Varianza campionaria dei parametri. Questi valori quantificano la diversità rispetto alla media tra le persone.
c) Correlazione campionaria tra i parametri stimati, che potrebbe essere sintomo di una multicollinearità tra le variabili esplicative.

I risultati di queste analisi per il modello lineare sono riportate in Tabella 2.5. Dalle medie dei parametri stimati concludiamo che all'inizio gli utenti scrivono in media 18 sms al mese, ed hanno una leggera tendenza a diminuire. La grandezza della deviazione standard ci informa che gli indivi-

dui si discostano fortemente dalle medie stimate, soprattutto per quanto riguarda le intercette.

	$\overline{\beta}_0$	$\overline{\beta}_1$
Media	17.886	-0.424
Deviazione standard	49.979	2.974

Tabella 2.5. Statistiche descrittive per i parametri stimati con il modello lineare.

I valori ottenuti dal modello quadratico sono invece riportati in Tabella 2.6. Mentre la media di β_0 è cambiata di poco rispetto al modello precedente, la media di β_1 si è avvicinata allo zero, ed ha ora una varianza decisamente più grande. La media di β_2 è piccola in valore assoluto, ha meno variabilità, però è fortemente correlata con β_1, e questo spiega la differenza tra le stime del primo e del secondo modello: esiste multicollinearità tra le variabili, e questo peggiora la precisione delle stime, quindi sarebbe meglio inserirne una sola, ed accontentarci del modello lineare.

	$\overline{\beta}_0$	$\overline{\beta}_1$	$\overline{\beta}_2$
Media	16.775	-0.091	-0.017
Deviazione standard	53.073	10.070	0.482
Correlazione tra $\overline{\beta}_1$ e $\overline{\beta}_2$			-0.956

Tabella 2.6. Statistiche descrittive per i parametri stimati con il modello quadratico.

2.2.1.3. Autocorrelazione

L'ipotesi fondamentale su cui si basano i modelli stimati nel paragrafo precedente è l'indipendenza dei residui della regressione. Nel caso che stiamo analizzando l'ipotesi può non essere verificata, perché i dati longitudinali tendono ad essere autocorrelati. Se questo fosse il caso, dovremmo abban-

donare i modelli stimati in precedenza per passare a modelli diversi, ad esempio i processi stocastici.

Guardiamo allora alcuni strumenti grafici utili a capire se esiste autocorrelazione in una serie. Scegliamo casualmente otto utenti tra coloro che spediscono almeno un sms al mese, e visualizziamo in Figura 2.17 il grafico del correlogramma per gli utenti. In ascissa è riportato il *lag*, la distanza tra due osservazioni, e in ordinata c'è la correlazione tra le due. Le linee orizzontali tratteggiate sono le bande di confidenza, all'interno delle quali il valore della correlazione non è significativo. Il primo valore, corrispondente al lag zero, è sempre uguale ad uno, perché è la correlazione di un'osservazione con se stessa. Delle altre linee verticali, invece, poche sono più alte delle bande di confidenza, perciò per la maggior parte degli utenti possiamo supporre un processo di tipo White Noise (Di Fonzo, Lisi, 2005).

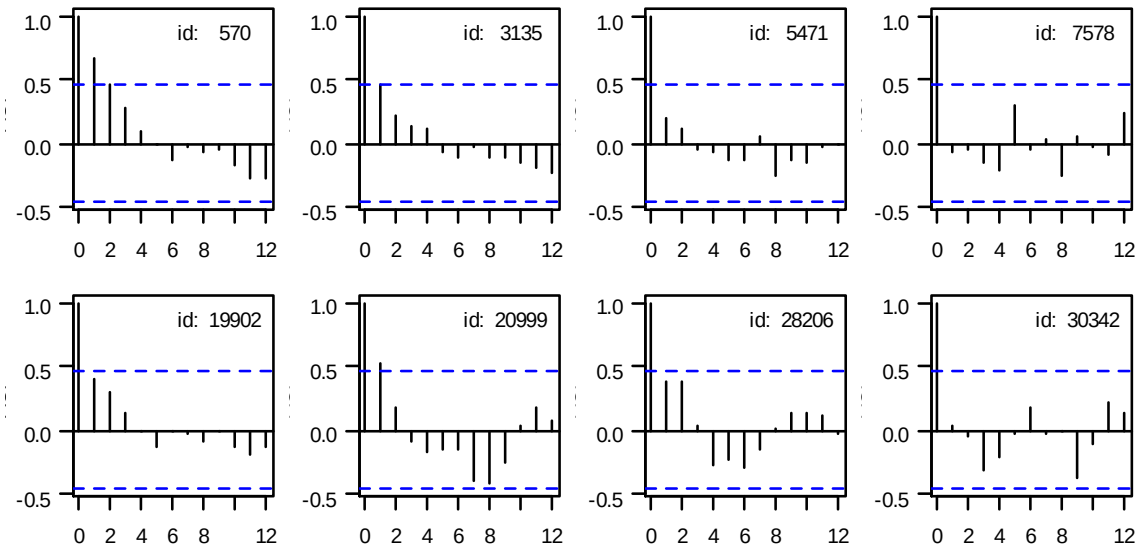

Figura 2.17. Grafico del correlogramma (ACF) per otto utenti scelti casualmente dal campione. In ogni pannello, in alto a destra, si legge l'Id dell'utente.

Un altro strumento grafico è il correlogramma parziale, che misura la correlazione tra due osservazioni al netto di quelle intermedie; i grafici per gli otto utenti sono in Figura 2.18. Qui la prima barra verticale corrisponde al

primo lag, ed è significativamente maggiore di zero solo per tre utenti, gli stessi che avevano valori significativi per il correlogramma.

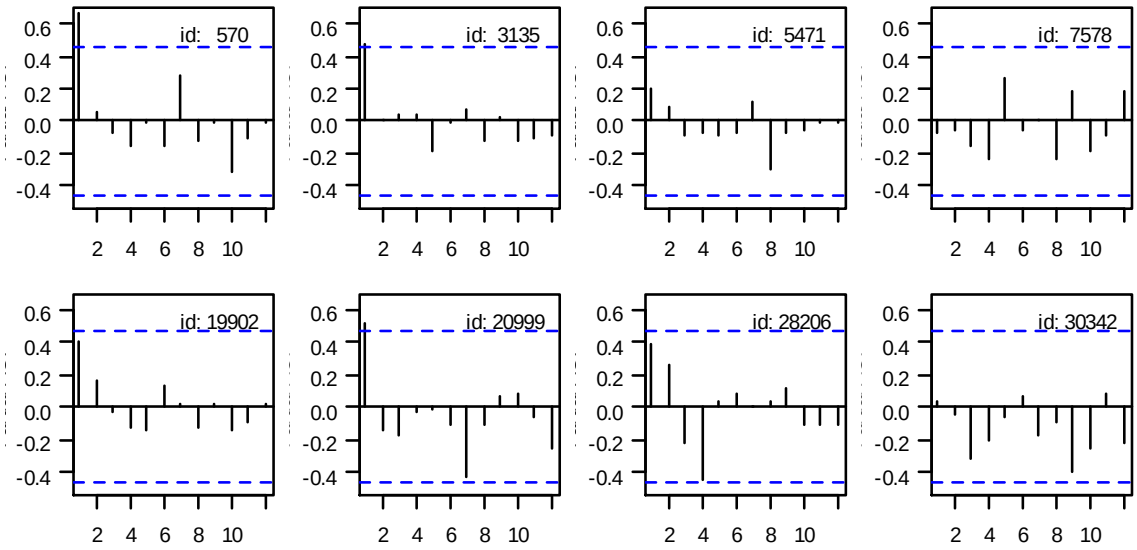

Figura 2.18. Grafico del correlogramma parziale (PACF) per gli otto utenti della figura precedente.

Purtroppo le serie sono troppo corte per essere ben analizzate da questi strumenti, ed infatti le bande di confidenza, che dipendono dalla lunghezza della serie, sono all'altezza di 0.5, troppo alte per essere superate dalla maggior parte delle barre, e per permetterci di capire la forma delle serie. Quando una serie non contiene abbastanza osservazioni, succede spesso che gli strumenti di serie storiche non siano utilizzabili.

2.2.2. Messaggi multimediali (mms)

La seconda variabile longitudinale del dataset riguarda il numero di mms spediti al mese. Come anticipato all'inizio del capitolo, non tratteremo tutta l'analisi anche per questa serie, ma riporteremo i risultati essenziali.

Il grafico in Figura 2.19 rappresenta l'andamento delle medie campionarie della variabile, divisa per i due gruppi. Si nota come i valori siano molto più piccoli di quelli degli sms: in media, una persona spedisce un mms ogni

due mesi. La forma delle serie medie ricorda molto quella degli sms, e forse questa serie obbedisce alle stesse regole della precedente.

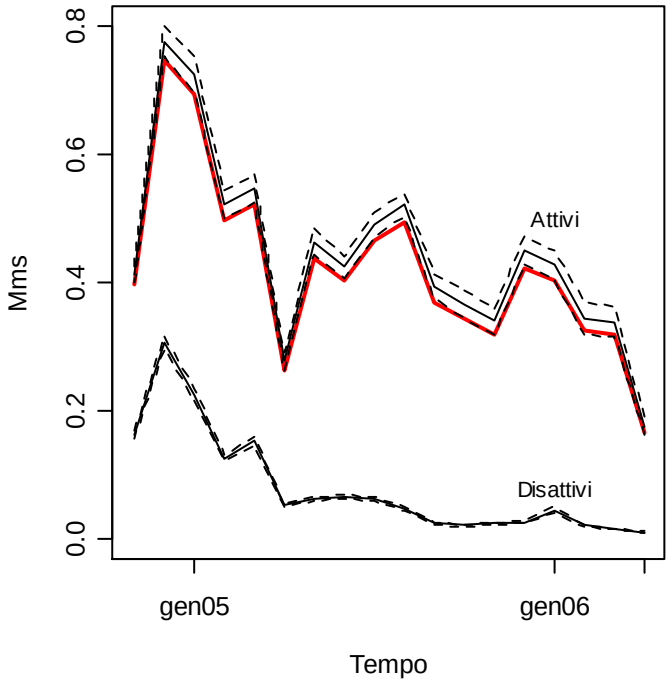

Figura 2.19. Andamento della media della variabile Mms, in rosso, e le medie dei due gruppi (Attivi e Disattivi). Le linee tratteggiate sono alla distanza di uno standard error dalla media corrispondente.

Abbiamo stimato anche per questa variabile un modello lineare, e le distribuzioni dei parametri si concentrano attorno allo zero, per questo motivo non li rappresentiamo. Riportiamo invece la distribuzione del R^2, in Figura 2.20. Per convenzione, abbiamo deciso di porre a 1 gli R^2 delle serie costanti sullo zero, e l'altezza della barra corrispondente a 1 ci informa che sono molti gli utenti di entrambi i gruppi che non spediscono alcun mms, quasi l'80% dei Disattivi.

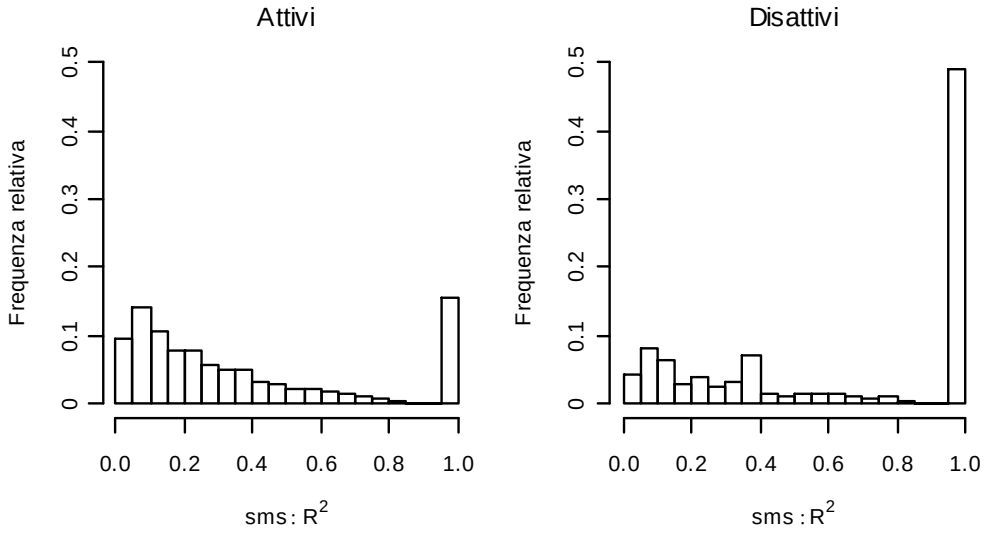

Figura 2.20. Distribuzione del R^2 per il modello lineare stimato sugli Mms.

Non stupisce quindi che l'intercetta media sia molto vicina allo zero, come si nota in Tabella 2.7. L'intercetta ha una varianza piuttosto grande, ma non è così per la pendenza, e decidiamo che è inutile procedere con il modello quadratico. Anzi, il modello migliore potrebbe essere costante nel tempo per molti utenti.

	$\overline{\beta}_0$	$\overline{\beta}_1$
Media	0.589	-0.018
Deviazione standard	2.842	0.216

Tabella 2.7. Valori medi stimati per i parametri del modello lineare.

2.2.3. Telefonate verso telefoni fissi

La terza variabile longitudinale riguarda le telefonate effettuate dagli utenti e la cui destinazione era un telefono fisso. Rappresentiamo in Figura 2.21 l'andamento delle medie di questa variabile.

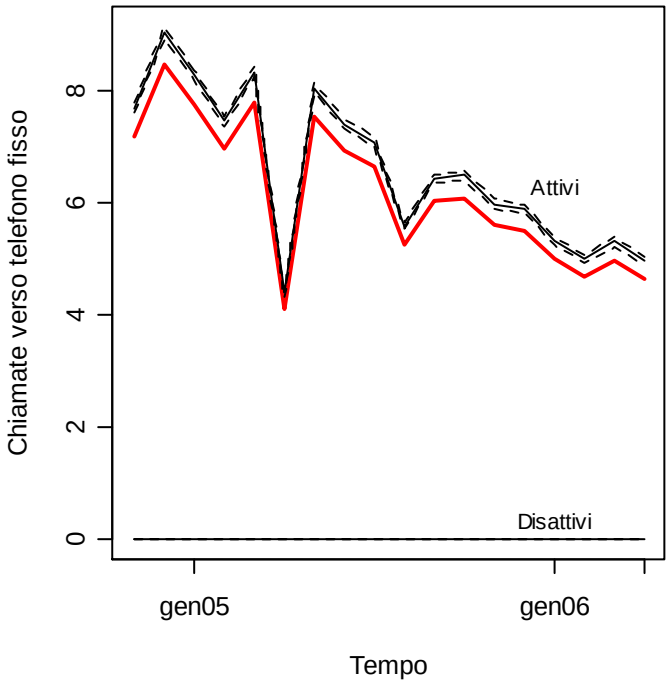

Figura 2.21. Media della variabile Chiamate verso fisso, divisa per i due gruppi.

Le medie dei gruppi sono decisamente diverse rispetto a quelle delle due variabili precedenti: tutti gli utenti del gruppo Disattivi, cioè coloro che alla fine della finestra di osservazione lasceranno l'azienda, non hanno effettuato alcuna chiamata. Questo semplifica di molto la nostra analisi, perché possiamo riassumere l'informazione di queste diciotto variabili longitudinali in una sola variabile fattore, che indica se l'utente ha effettuato almeno una telefonata verso telefoni fissi:

$$\text{Fisso}_i = \begin{cases} \text{N} & \text{se}\,\max(\text{ch.verso.fisso}_{it}) = 0 \\ \text{S} & \text{altrimenti} \end{cases}$$

Tuttavia, prima di utilizzare il modello in chiave previsiva, su altri dati, dovrebbe essere eseguita una nuova analisi, per verificare se i futuri disattivi non effettuano alcuna chiamata verso telefoni fissi; il modello, infatti, si fonda su un'ipotesi ricavata dal campione, e per il momento non possiamo verificarla.

Ora la variabile è diventata qualitativa, quindi possiamo guardarne la distribuzione nei due gruppi (Figura 2.22). Già ad occhio si nota che questa variabile discrimina molto bene tra i due gruppi, e ci aspettiamo che abbia grande influenza nei modelli dei prossimi capitoli. Abbiamo calcolato anche per questa variabile il test χ^2, per l'indipendenza marginale dalla variabile Stato, ed il test rifiuta l'ipotesi nulla con qualsiasi livello di significatività.

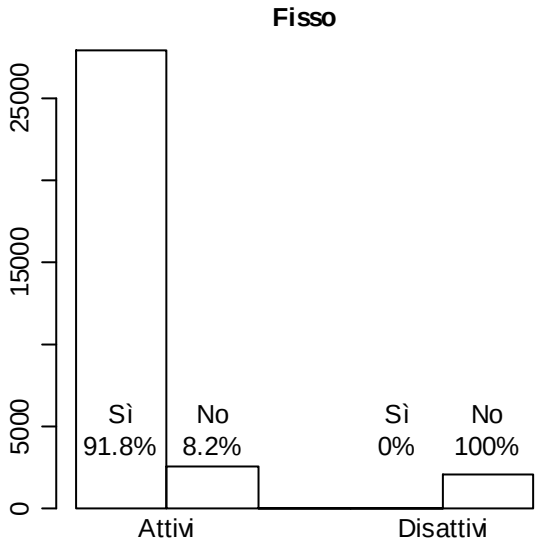

Figura 2.22. Distribuzione empirica della variabile Fisso per i due gruppi.

2.2.4. Telefonate verso altri operatori

Le ultime variabili che abbiamo osservato riguardano le telefonate verso i cellulari. Spesso le tariffe telefoniche variano a seconda che il cellulare destinatario della telefonata sia cliente dello stesso operatore telefonico o di un operatore concorrente. Per questo manterremo distinte le due tipologie di chiamate. Analizzeremo per prime le serie che riguardano le telefonate verso altri operatori.

Raffiguriamo in Figura 2.23 la media della variabile e le medie dei due gruppi. Anche questa variabile, come le chiamate verso telefono fisso, prende solo valore zero ad ogni tempo per il gruppo dei Disattivi. Possiamo quindi

seguire lo stesso ragionamento, e trasformare anche questa variabile in un fattore, naturalmente con le precauzioni già citate:

$$\text{Altri}_i = \begin{cases} \text{N} & \text{se } \max(\text{ch.verso.altri}_{it}) = 0 \\ \text{S} & \text{altrimenti} \end{cases}$$

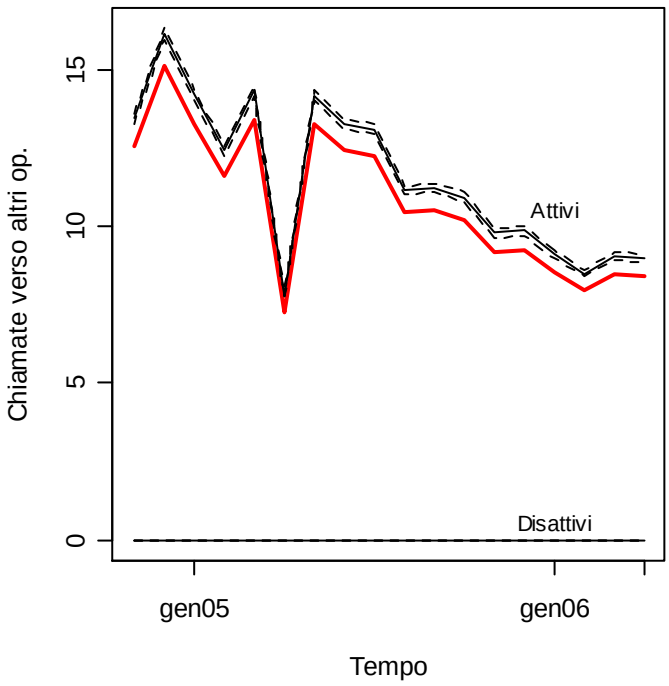

Figura 2.23. Grafico della media della variabile Chiamate verso altri operatori, anche divisa per gruppi.

L'istogramma con la distribuzione empirica della variabile è in Figura 2.24. Anche questa variabile discrimina molto bene tra i gruppi, e il test χ^2 rifiuta l'ipotesi di indipendenza marginale tra la variabile e lo Stato.

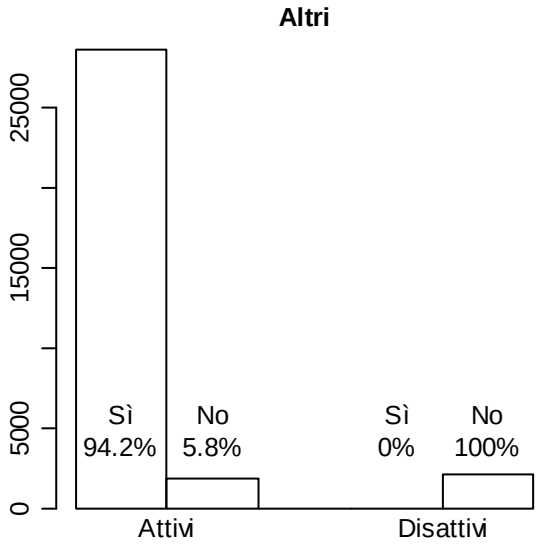

Figura 2.24. Distribuzione empirica della variabile fattore Altri divisa per gruppi.

2.2.5. Telefonate verso lo stesso operatore

Concludiamo questa analisi preliminare con le telefonate verso cellulari dello stesso operatore. In Figura 2.25 raffiguriamo l'andamento delle medie.

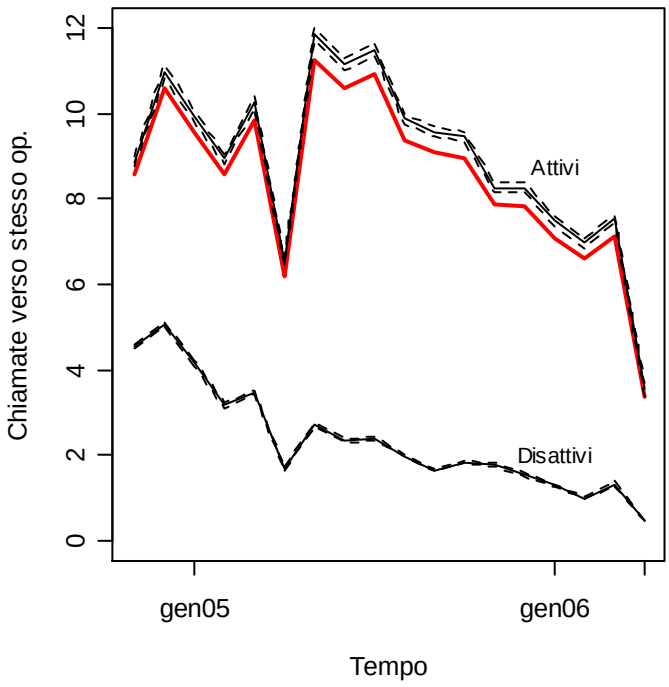

Figura 2.25. Andamento della media della variabile Chiamate verso stesso operatore, e delle medie dei due gruppi.

Notiamo che anche il gruppo degli utenti Disattivi effettua telefonate di questo tipo, perciò l'analisi della variabile è più simile all'analisi degli Sms, rispetto alle altre variabili Chiamate viste in precedenza.

Proviamo a stimare un modello lineare per la variabile e, nonostante intercetta e pendenza si concentrino attorno allo zero, sono ben pochi gli utenti che non effettuano alcuna telefonata nei diciotto mesi di osservazione. Questo si nota soprattutto nel grafico del R^2, in Figura 2.26, dove, in corrispondenza del valore uno, sono conteggiati tutti gli utenti per i quali il modello è perfettamente stimato, cioè chi non effettua alcuna chiamata. Per le variabili Sms e Mms questa barra è molto alta, ma per le Chiamate verso stesso operatore è bassa.

Questo fenomeno può essere causato dal fatto che spesso le persone tendono a scegliere lo stesso operatore telefonico di amici o parenti, perché le tariffe tra cellulari possono essere molto più alte se il destinatario è utente di altri operatori. Inoltre si sta diffondendo l'abitudine di possedere più schede telefoniche, se non addirittura più telefoni cellulari, e di usarli a seconda del destinatario della telefonata.

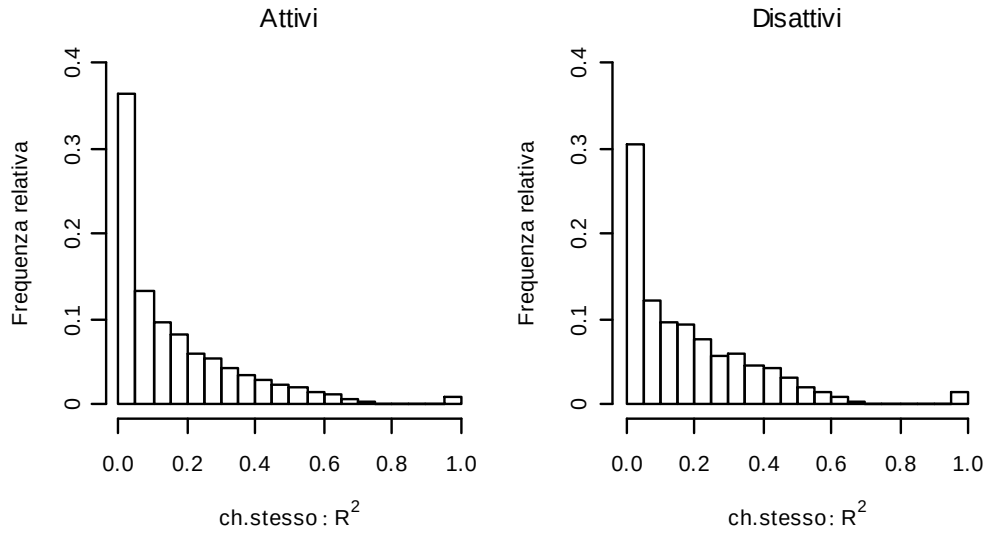

Figura 2.26. Distribuzione del R^2 per il modello lineare stimato sulla variabile Chiamate verso stesso operatore.

Riportiamo in Tabella 2.8 e in Tabella 2.9 i valori medi dei parametri della regressione lineare e di quella quadratica. Come per le variabili precedenti, l'aggiunta di un parametro aumenta la variabilità degli altri, e la correlazione tra $\overline{\beta}_1$ e $\overline{\beta}_2$ è molto alta, segno di multicollinearità.

	$\overline{\beta}_0$	$\overline{\beta}_1$
Media	10.697	-0.227
Deviazione standard	21.453	1.340

Tabella 2.8. Valori riassuntivi per i parametri della regressione lineare.

	$\overline{\beta}_0$	$\overline{\beta}_1$	$\overline{\beta}_2$
Media	8.207	0.519	-0.039
Deviazione standard	25.644	5.037	0.249
Correlazione tra $\overline{\beta}_1$ e $\overline{\beta}_2$			-0.964

Tabella 2.9. Valori riassuntivi per i parametri della regressione quadratica.

2.3. Ultime modifiche

Prima di procedere al prossimo capitolo e alla costruzione dei primi modelli, è necessario prendere alcune decisioni sui dati: come dividere il campione per la stima e per la verifica, e quali variabili longitudinali utilizzare nella stima.

2.3.1. Campione di stima, campione di verifica

Quando si vuole verificare la bontà di previsione di un modello sullo stesso campione su cui tale modello è stato stimato, si tende a sottostimare l'errore di previsione, perciò è sempre consigliabile effettuare la verifica su

una parte del campione non utilizzata per la stima. Se la dimensione campionaria è grande, questa operazione può essere eseguita con una perdita di informazione minima rispetto al campione completo, semplicemente dividendo casualmente il campione a metà: il campione di stima contiene quindi 16 262 utenti, ed altrettanti ne contiene il campione di verifica.

Non si può però dimenticare che la variabile risposta Stato non assume le due modalità con la stessa probabilità. Come abbiamo visto nei grafici iniziali, infatti, solo il 6.36% degli utenti del campione sono Disattivi, valore molto vicino al percentile 5% con il quale si giudica la bontà dei modelli. Ne segue che un "buon" modello potrebbe essere quello che assegna probabilità 1 a tutti gli individui di rimanere attivi. L'errore di questo modello sarebbe quindi del 6.36%, vale a dire tutti i Disattivi, e sarebbe un errore certamente basso.

Il problema qui esposto emerge spesso, quando il fenomeno che si vuole modellare è raro nella popolazione. L'obiettivo dell'analisi, infatti, è prevedere quali utenti hanno più probabilità di disattivarsi, anche se questi sono una percentuale molto piccola rispetto al totale di utenti, e siamo disposti ad accettare un errore di previsione totale più grande, pur di raggiungere l'obiettivo.

La scelta dei campioni, quindi, non può essere completamente casuale. Il campione di stima che abbiamo appena creato contiene 1030 utenti Disattivi, ma 15 232 utenti Attivi. Per ottenere un campione bilanciato, dove Attivi e Disattivi sono presenti con la stessa probabilità, dobbiamo prelevare casualmente 1030 utenti dagli Attivi del campione, insieme a tutti gli utenti Disattivi dello stesso campione. Il nuovo campione di stima bilanciato contiene 2060 utenti, e sarà utilizzato per stimare i modelli dei prossimi capitoli. Il campione di verifica, invece, deve rimanere inalterato, perché la bontà di previsione deve essere misurata sulle percentuali di disattivazione osservate.

2.3.2. Variabili longitudinali

Un'altra questione importante riguarda gli ultimi mesi delle variabili longitudinali. In particolare, l'ultimo mese di cui abbiamo i dati è aprile 2006, e questo è il mese in cui avviene (o non avviene) la disattivazione. Il numero di telefonate o di messaggi è calcolato alla fine del mese, perciò è contemporaneo alla variabile risposta, e non possiamo utilizzare uno per prevedere l'altro; chiaramente, se uno è disponibile, sarà disponibile anche l'altro, rendendo inutile la previsione. L'ultimo mese, quindi, non può far parte delle variabili esplicative.

Il penultimo mese, marzo, è osservato appena un mese prima della disattivazione, e i dati sul traffico del mese potrebbero arrivare troppo tardi per essere utilizzati a scopo previsivo, e soprattutto tardi per effettuare eventuali azioni di trattenimento del cliente. Tuttavia, in molte grandi aziende, sono automatizzate sia la raccolta dei dati, sia l'invio di materiale promozionale. Possiamo perciò decidere di conservare tra le esplicative le variabili longitudinali di tutti i mesi compresi tra novembre 2004 (il primo mese), e marzo 2006 (il diciassettesimo).

3. Modelli con variabili statiche

Nel capitolo precedente abbiamo compiuto alcune analisi preliminari sulle variabili che sono presenti nel dataset, ed abbiamo notato che alcune si comportano in maniera differente a seconda che l'individuo a cui si riferiscono appartenga al gruppo degli Attivi o a quello dei Disattivi, cioè per i diversi valori assunti dalla variabile Stato. Ora vogliamo scoprire se questa correlazione ci permette, in qualche modo, di usare queste variabili per spiegare lo Stato. Costruiremo quindi alcuni modelli in cui Stato è la variabile risposta ed altre variabili sono le esplicative, e li stimeremo sulla parte del campione che abbiamo scelto come campione di stima. Poi verificheremo poi la capacità dei modelli di prevedere la risposta per gli individui sui quali non è stato costruito il modello, per scegliere il miglior modello da applicare nella pratica aziendale.

In questo capitolo costruiremo due strumenti statistici di tipo statico, che fanno parte dell'insieme di strumenti già ampiamente utilizzati da molte aziende per condurre analisi sui propri clienti. Infatti, spesso è sufficiente un'analisi ben condotta, anche se semplice, per aumentare la probabilità di successo rispetto alla scelta casuale, ad esempio, di una parte della clientela cui offrire una promozione.

In Appendice A si possono trovare i comandi per R che abbiamo usato nella costruzione degli strumenti statistici, e le stime dei coefficienti di alcuni modelli, che non riportiamo nel testo per non appesantirlo.

3.1. Modello Logistico

La variabile risposta, Stato, è una variabile dicotomica, che può assumere solo due valori, Attivo e Disattivo, che codifichiamo rispettivamente come uno e zero:

$$\text{Stato}_i = \begin{cases} 0 & \text{se ad Aprile 2006 l'utente i si è disattivato} \\ 1 & \text{altrimenti} \end{cases}$$

Il modello che costruiremo per questa variabile parte dall'ipotesi che lo Stato di ogni cliente prenda valori da una variabile aleatoria binomiale, la cui probabilità di successo, cioè la probabilità di essere Attivo, può essere diversa per ogni individuo, ed è indipendente tra gli utenti.

In particolare, la probabilità di successo è espressa in funzione di un predittore lineare. Il predittore lineare è espresso da una regressione lineare delle variabili esplicative, e può assumere qualsiasi valore reale; la relazione con la probabilità di successo è espressa dalla funzione logit, che riporta il predittore nella stessa scala della probabilità, cioè nell'intervallo compreso tra zero e uno.

Si possono usare altre trasformazioni con lo stesso scopo, come il probit, ma il logit è da preferire perché è interpretabile come log-rapporto tra la probabilità di successo e la probabilità di insuccesso (log odds), e semplifica la funzione di verosimiglianza associata al modello, perciò è detto legame canonico. In realtà c'è generalmente poca differenza tra la forma del logit e del probit, tranne che nelle code della distribuzione (Zelterman, 1999).

Il modello così descritto è un modello lineare generalizzato di tipo logistico (Pace, Salvan, 2001), e si può scrivere:

$$\begin{aligned} \text{Stato}_i &\sim \text{Bi}(1, \pi_i) \\ \pi_i &= \frac{e^{\eta_i}}{1 + e^{\eta_i}} \\ \eta_i &= \mathbf{x}_i^T \boldsymbol{\beta} \end{aligned} \qquad (3.1)$$

dove π_i è la probabilità di successo, η_i il predittore lineare, $\mathbf{x_i}$ il vettore delle variabili esplicative e $\boldsymbol{\beta}$ il vettore dei coefficienti.

La prima riga della (3.1) dichiara che la variabile Stato segue la distribuzione binomiale; la seconda definisce il legame tra la probabilità di successo e il predittore lineare, che è la funzione inversa del logit, mentre la ter-

za riga riporta la regressione lineare nei parametri. Qui è scritta in forma matriciale, ma potrà di volta in volta essere specificata con le variabili esplicative che decideremo di utilizzare, e sarà l'unica parte che cambierà per specificare i diversi modelli. I parametri della regressione lineare hanno asintoticamente una distribuzione normale centrata nel loro vero valore, perciò con il metodo della massima verosimiglianza otterremo per essi stime non distorte.

3.1.1. Stima dei modelli logistici

Nel corso del paragrafo stimeremo alcuni modelli sul campione di stima che abbiamo definito nel Capitolo 2, e verificheremo la loro capacità previsiva: li applicheremo sul campione di verifica per stimare la probabilità di successo, assegneremo ad ogni utente lo Stato più probabile, e confronteremo questo risultato con la vera risposta di ogni individuo. Inoltre i modelli saranno confrontati in base ai criteri di informazione, scritti seguendo la notazione di Azzalini, Scarpa (2004):

$$\begin{aligned} \text{IC} &= -2\log L + \text{penalità}(p) \\ \text{AIC} &\quad \text{penalità}(p) = 2p \\ \text{BIC} &\quad \text{penalità}(p) = p\log n \end{aligned}$$

dove L è la verosimiglianza del modello, e p è il numero di parametri stimati dal modello. Gli indici sono misure della bontà di adattamento del modello, ma penalizzano l'utilizzo di troppe variabili esplicative. Infatti, se continuiamo ad aggiungere variabili possiamo migliorare sempre di più il modello, ma lo adattiamo troppo ai dati sui quali abbiamo stimato, con il rischio che il modello non sia più valido per i dati relativi ad altre persone. I criteri di informazione crescono al peggiorare del modello, perciò cercheremo di scegliere il modello che ottiene i valori minimi per gli indici, o almeno per uno dei due.

Cominciamo ora a selezionare le variabili esplicative da inserire nel predittore lineare. Per il momento consideriamo solo le variabili che non dipen-

dono dal tempo, presentate nel Capitolo 2. Abbiamo notato allora che alcune di queste variabili hanno distribuzioni marginalmente indipendenti dalla risposta, con il test χ^2 in Tabella 2.2, quindi possiamo presumere che non avranno grande significato nel modello. Altre invece hanno un valore del test molto grande, e saranno queste le variabili più importanti per l'analisi: il canale di vendita, il piano tariffario, e il sesso. Ci sono inoltre l'età, una variabile quantitativa, e i due fattori Fisso ed Altri, definiti nello stesso capitolo, che a loro volta hanno valori del test χ^2 molto alti.

Costruiamo un modello con una sola variabile esplicativa, nel quale inseriamo le variabili appena citate, una alla volta. Riportiamo i criteri di informazione per ognuno dei sei modelli in Tabella 3.1. Il modello con la variabile Altri ottiene i criteri di informazione più bassi, mentre il modello peggiore contiene il canale di vendita.

η_i	AIC	BIC
$\beta_0 + \beta_1 \text{Altri}_i$	557.675	568.935
$\beta_0 + \beta_1 \text{Fisso}_i$	674.979	686.239
$\beta_0 + \beta_1 \text{PianoTariffario}_i$	2718.035	2734.927
$\beta_0 + \beta_1 \text{Età}_i$	2799.473	2810.734
$\beta_0 + \beta_1 \text{Sesso}_i$	2816.511	2827.772
$\beta_0 + \beta_1 \text{CanaleVendita}_i$	2840.535	2874.318

Tabella 3.1. Valore assunto dai criteri di informazione per i modelli il cui predittore lineare è nelle righe.

Naturalmente qui non sarebbe necessario presentare entrambi gli indici di informazione, perché la penalità cambia solo con il numero di parametri del modello, e tutti i modelli della tabella hanno un solo parametro. Tuttavia, per consentire il confronto anche con i modelli successivi, li abbiamo presentati entrambi.

Scegliamo quindi il primo modello della tabella, e proviamo a migliorarlo aggiungendo un'altra variabile esplicativa. Procediamo allo stesso modo, aggiungendo solo una variabile alla volta e confrontando i criteri di informazione, finché non siamo soddisfatti di quello che possiamo definire modello finale.

In Tabella 3.2 si può seguire il criterio logico con il quale abbiamo costruito i modelli. Abbiamo scelto innanzitutto di utilizzare solo due variabili esplicative, e dalla tabella precedente ci aspettiamo che la variabile Fisso sia la più informativa, dopo la variabile Altri. Infatti i criteri della prima riga della Tabella 3.2 sono più bassi di quelli della tabella precedente, ed anche di quelli che non abbiamo riportato, nei quali compaiono altre variabili. Poi abbiamo voluto aggiungere una terza variabile, ed abbiamo scelto il piano tariffario, perché era la prima delle variabili rimaste della Tabella 3.1. I criteri di informazione, però, sono risultati più bassi per il modello la cui terza variabile esplicativa è l'età.

η_i	AIC	BIC
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i$	447.130	464.022
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{PianoTariffario}_i$	445.034	473.186
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i$	430.318	452.840
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i + \beta_4 \text{PianoTariffario}_i$	431.333	465.116
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i + \beta_4 \text{Sesso}_i$	430.620	458.773
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i + \beta_4 \text{CanaleVendita}_i$	432.885	483.560

Tabella 3.2. Valore dei criteri di informazione per i modelli il cui predittore lineare è nelle righe della tabella.

Il modello nella terza riga della tabella è il migliore di quelli che abbiamo stimato, ed aggiungendo altre variabili non riusciamo ad ottenere criteri di informazione più bassi.

In realtà le variabili a nostra disposizione non sono ancora finite. Abbiamo lasciato da parte tutte le variabili longitudinali, gli sms, gli mms e le chiamate verso lo stesso operatore. Il motivo è dovuto al numero di tali variabili, ben diciassette per ogni tipo, che renderebbe piuttosto lento il processo di scelta se volessimo provare ad aggiungere al modello una variabile per volta.

Possiamo piuttosto ragionare sul fatto che l'eventuale disattivazione avviene, per il nostro campione, nel mese di Aprile 2006, perciò siamo propensi ad immaginare che il numero di sms spediti diciotto mesi prima sia meno informativo di quelli spediti il mese precedente, ad esempio. Scegliamo quindi alcuni degli ultimi mesi per ognuna delle variabili, e li aggiungiamo al terzo modello della Tabella 3.2, procedendo, come prima, una variabile per volta.

Non riportiamo qui i risultati ottenuti perché, pur proseguendo a ritroso fino a giugno dell'anno precedente, non abbiamo trovato modelli migliori per entrambi i criteri. Solo il criterio AIC ha ottenuto a volte valori più bassi, per alcuni mesi della variabile Chiamate verso stesso operatore, ma è noto che tale criterio tende a sovraparametrizzare, cioè ad accettare modelli con troppe variabili esplicative, perciò quando i due criteri erano in disaccordo abbiamo preferito seguire il criterio BIC, ed accettare il modello più ridotto.

D'altra parte le variabili longitudinali non possono essere del tutto inutili per l'analisi. Esse infatti descrivono il comportamento di ogni cliente, ma non sono statiche come la variabile Fisso o la variabile Altri: esprimono un cambiamento continuo, nel tempo, che potrebbe essere molto utile per il modello. Abbiamo appena scoperto che inserire un mese singolo nel modello non lo migliora, ma, se avessimo a disposizione poche variabili che riassumono il cambiamento, forse potremmo avere davvero un modello migliore. Nel capitolo precedente abbiamo calcolato un modello lineare per ogni traiettoria individuale, ed abbiamo per ogni persona i parametri del modello. Questi ed altri indici di cambiamento che stimeremo nel prossimo capitolo saranno utilizzati nel capitolo 5 per costruire nuovi modelli, che saranno confrontati con i migliori modelli stimati in questo capitolo.

3.1.2. Bontà del modello

Abbiamo scelto come miglior modello logistico per i dati il terzo modello della Tabella 3.2, che è il seguente:

$$\begin{aligned} \text{Stato}_i &\sim \text{Bi}(1, \pi_i) \\ \pi_i &= \frac{e^{\eta_i}}{1 + e^{\eta_i}} \\ \eta_i &= \beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i \end{aligned} \qquad (3.2)$$

Da questo modello è possibile calcolare la probabilità di successo per ogni utente, anche per coloro che non fanno parte del campione di stima, vale a dire gli utenti del campione di verifica, ma anche i clienti delle analisi future.

Assegniamo ad ogni cliente lo Stato stimato di Attivo se la sua probabilità di successo stimata $\hat{\pi}_i$ è superiore a 0.5, e Disattivo se è inferiore; in questo modo possiamo confrontare lo Stato stimato con quello osservato, e calcolare l'errore di previsione. La Tabella 3.3 riporta la matrice di confusione, dove nelle righe abbiamo lo Stato stimato, e nelle colonne lo Stato osservato. Le caselle della diagonale principale indicano che sono 1037 i Disattivi, e 14179 gli Attivi, che sono stati correttamente previsti. Nel campione di verifica, però, i Disattivi sono 1040, e gli Attivi 15222, perciò non abbiamo previsto per tutti la giusta risposta.

	Risposta osservata	
Risposta prevista:	disattivo	attivo
disattivo	1037	599
attivo	0	14179

Tabella 3.3. Matrice di confusione per il modello logistico in (3.2).

Gli elementi fuori della diagonale principale contengono le errate classificazioni. Non abbiamo previsto come Attivo nessun utente che in realtà fos-

se Disattivo, ma 599 utenti Attivi sono stati previsti come Disattivi. Dalla tabella mancano inoltre 447 utenti, per i quali non conosciamo l'età, e perciò non possiamo calcolare il valore del predittore lineare e stimare lo Stato.

Se questa mancata classificazione fosse considerata importante, si potrebbe decidere di calcolare, solo per gli utenti dei quali non conosciamo l'età, un modello meno buono ma che non contiene questa variabile.

Torniamo piuttosto agli errori. La frequenza d'errore si può calcolare come la somma delle errate classificazioni, divisa per il numero totale degli utenti classificati, ed ha valore 0.038. Un modello può essere considerato buono in senso previsivo se ha un errore di classificazione inferiore al 5%, e questo modello ha previsto un risultato sbagliato solo per il 3.8% degli utenti del campione, perciò il modello, oltre ad essere il migliore finora calcolato, ha buone proprietà di previsione.

Dalla Tabella 3.3 possiamo calcolare altri due tipi di errore, la frequenza di falsi positivi e la frequenza di falsi negativi. I falsi positivi sono coloro che, pur essendo Disattivi, sono previsti come Attivi, e per questo motivo non saranno raggiunti dalle attività di trattenimento dell'azienda. Questo è un errore molto grave, perché l'obiettivo dell'analisi è proprio raggiungere gli utenti Disattivi. I falsi negativi, invece, sono gli utenti Attivi che noi abbiamo previsto come Disattivi, perciò saranno oggetto delle attività di trattenimento inutilmente. Pur essendo questo uno spreco di risorse, l'errore è meno grave del precedente. Fortunatamente, il modello non stima alcun falso positivo, e l'errore più grave non avviene mai, almeno per il campione di verifica. I falsi negativi invece sono il 4.1% degli Attivi, che è comunque un valore basso.

La matrice di confusione e gli indici di errore ad essa collegati sono indicatori molto sintetici della capacità previsiva del modello, ma possiamo introdurre anche uno strumento più analitico, la funzione *lift*, che misura quanto il modello migliora la classificazione rispetto ad una scelta casuale (Azzalini, Scarpa, 2004). La Figura 3.1 mostra il grafico della funzione *lift* calcolata per il modello stimato. In ascissa abbiamo la percentuale di clienti ai quali potrebbe essere indirizzata l'azione di trattenimento; in ordinata po-

niamo una misura di quanto è più probabile trovare utenti disattivi, nel gruppo scelto, rispetto alla scelta casuale.

I valori del miglioramento sono maggiori per piccole frazioni del campione, ed è ciò che vogliamo, perché al termine dell'analisi dovremo scegliere una piccola parte della clientela alla quale rivolgere azioni di trattenimento, e spereremo di aver selezionato soprattutto futuri clienti disattivi. Se eseguiremo la selezione con il modello logistico, sarà molto più facile scegliere gli utenti giusti.

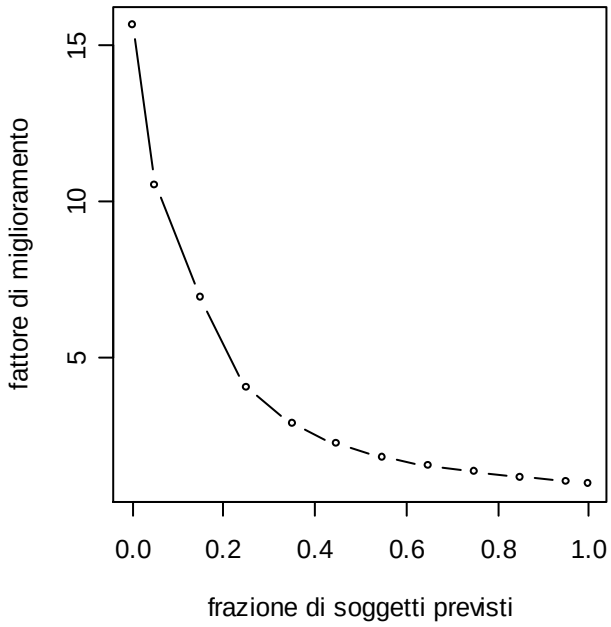

Figura 3.1. Curva della funzione lift per il modello logistico.

3.2. Classificazione ad albero

Finora abbiamo supposto di conoscere la funzione che regola la divisione degli utenti tra i due gruppi, e di non conoscere solo alcuni parametri di tale funzione. In realtà non conosciamo nemmeno la forma della funzione. Possiamo imporre ai dati una forma nota, come abbiamo fatto nel paragrafo

precedente, scegliendo la funzione logistica; oppure possiamo stimare una forma adatta in maniera non parametrica.

Gli alberi di classificazione approssimano la forma della funzione con una funzione a gradini, cioè costante su intervalli, e discontinua. Più che con una formula, il modello si può rappresentare graficamente in un albero binario (Figura 3.2). L'albero è strutturato a nodi, che sono i punti da cui partono due linee, dette rami, in direzioni opposte. Ad ogni nodo corrisponde un'affermazione di tipo logico, che confronta una sola variabile esplicativa con uno o più valori che essa può assumere. Le osservazioni per le quali l'affermazione è vera percorrono il ramo alla sinistra del nodo, le altre passano alla destra. I punti in cui l'albero finisce si chiamano foglie, e ad esse è associata una probabilità $\hat{\pi}$, alla quale corrisponde la classe 1 se la probabilità è superiore a 0.5, a cui corrisponde il gruppo Attivi, e 0 altrimenti, per il gruppo Disattivi.

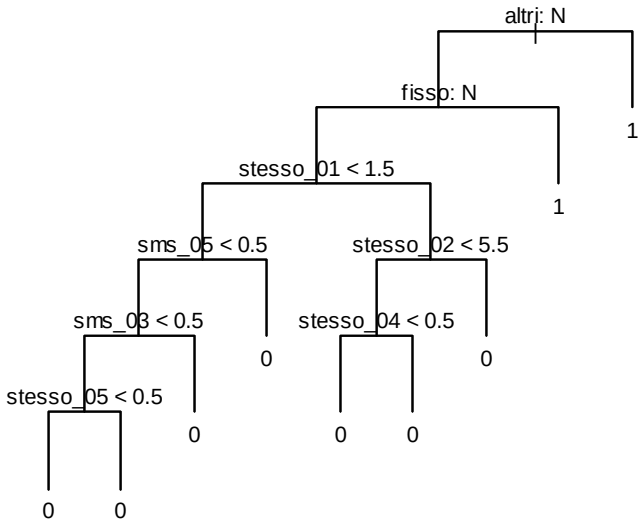

Figura 3.2. Classificazione ad albero per la variabile Stato.

I rami situati più in basso nell'albero di Figura 3.2 terminano tutti in foglie con la stessa classe, Disattivi, perciò portano informazione inutile a discriminare tra i due gruppi, e vanno eliminati con la "potatura". La potatura consiste nell'eliminare un nodo, e quindi unire le due foglie in una sola. Se

eseguiamo la potatura guardando all'insieme di stima, si potrebbe scegliere un numero troppo alto di nodi, perché così il modello si adatta meglio ai dati, ma è meno utile per essere applicato su individui diversi; la situazione è parallela ad un modello lineare con troppe variabili.

Invece, eseguendo la potatura su una piccola parte del campione, che non è utilizzata per la stima, si può scegliere il numero di nodi che minimizza una funzione di devianza, e di solito si ottengono pochi nodi, che contengono solo le variabili più discriminanti. Qui scegliamo di minimizzare la funzione di devianza della distribuzione binomiale:

$$D = -2\sum_{i=1}^{n} \{Stato_i \log \hat{\pi}_i + (1 - Stato_i) \log(1 - \hat{\pi}_i)\} \qquad (3.3)$$

ma potremmo anche usare l'indice di Gini, o una misura della frequenza di errori di classificazione (Azzalini, Scarpa, 2004). Decidiamo inoltre di usare il campione di stima per costruire l'albero, e un piccolo campione di 1000 utenti, provenienti dal campione di verifica, per potarlo. Questo perché vogliamo eseguire la potatura con le frequenze di Attivi e Disattivi osservate, e non quelle modificate da noi per la stima.

In Figura 3.3 rappresentiamo il grafico della funzione di devianza. La funzione ha minimo per il secondo nodo, poi aumenta sempre di più, perciò l'albero migliore ha due nodi, ed è rappresentato in Figura 3.4. In questa figura, diversamente dalla Figura 3.2, la lunghezza dei rami è proporzionale alla riduzione di devianza portata dal nodo a cui si riferiscono. Nella figura precedente, invece, i rami avevano lunghezza uniforme, per motivi di leggibilità: i rami inferiori erano molto corti, perché l'introduzione degli ultimi nodi non portava sostanziali miglioramenti nel modello.

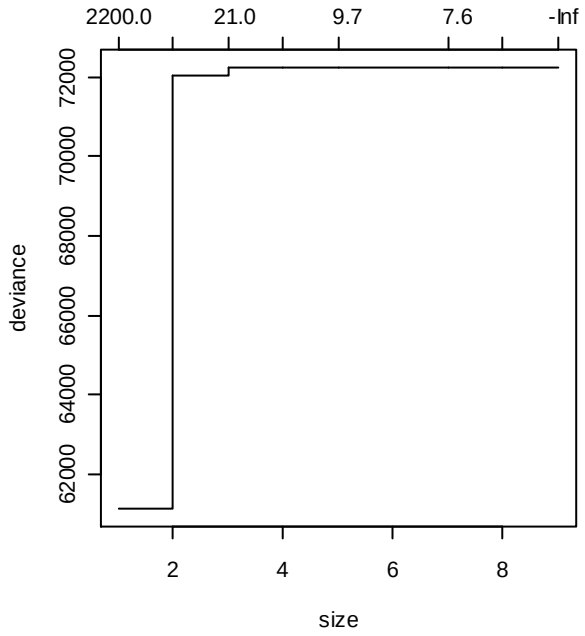

Figura 3.3. Misura della devianza per il modello ad albero, calcolata per diversi numeri di nodi (size). In alto sono rappresentati i valori della penalizzazione per il costo-complessità che si deve adottare per scegliere un albero della dimensione corrispondente.

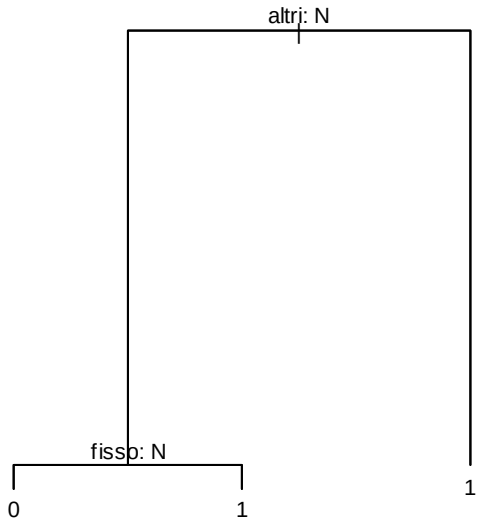

Figura 3.4. Miglior modello per la classificazione ad albero.

In quest'albero, invece, l'introduzione del primo nodo effettua già una buona partizione tra i due gruppi, poiché sappiamo dal Capitolo 2 che nessun Disattivo effettua chiamate verso altri operatori. La seconda variabile di-

vide di nuovo nettamente i clienti restanti, e nelle due foglie di destra sono presenti solo utenti Attivi. Nella foglia di sinistra, invece, sono presenti sia Disattivi che Attivi, ma questi ultimi sono solo una piccola parte, e alla foglia è assegnata la classe Disattivi.

Come per il modello logistico, possiamo usare l'albero di classificazione per prevedere lo Stato per gli utenti che appartengono al campione di verifica. La matrice di confusione è rappresentata in Tabella 3.4.

	Risposta osservata	
Risposta prevista:	disattivo	attivo
disattivo	1040	622
attivo	0	14600

Tabella 3.4. Matrice di confusione per l'albero di classificazione di Figura 3.4.

La differenza sostanziale rispetto alla corrispondente matrice del modello logistico è che non manca nessun utente, tutti sono stati classificati. Un motivo è il fatto che le due variabili, Altri e Fisso, non hanno alcun valore mancante, mentre nel modello logistico compariva anche l'età, che non era disponibile per tutti. Inoltre, la classificazione ad albero è molto flessibile nel trattare dati mancanti, perché li considera come un altro valore della variabile, e riesce a far scendere sempre tutte le osservazioni in una foglia. Quindi, se l'albero contenesse, ad esempio, la variabile età, potrebbe classificare lo stesso anche gli utenti la cui età non è nota.

Invece, non cambia rispetto al modello logistico la frequenza di errori di classificazione, che rimane al 3.8%. Anche questo modello non classifica alcun falso positivo, ed i falsi negativi sono il 4.1% degli Attivi.

Se dal punto di vista della previsione i due modelli sono equivalenti, possiamo confrontarli secondo i criteri di informazione, poiché il modello logistico ha quattro parametri, e l'albero di classificazione solo due. Costruiamo quindi per i due criteri di informazione un adattamento al caso non parametrico (Venables, Ripley, 1999):

$$IC = -2\log L + \text{penalità}(p)$$
$$\cong D + \text{penalità}(p)$$

Approssimiamo cioè la prima parte della formula, che conterrebbe la log-verosimiglianza, con la devianza della classificazione, calcolata dalla (3.3).

Otteniamo quindi i criteri di informazione che riportiamo in Tabella 3.5, dove sono confrontati con i corrispondenti indici del miglior modello logistico che abbiamo stimato. Ricordiamo che il modello logistico aveva tre parametri, contro i due nodi dell'albero di classificazione, ed anche per questo motivo otteniamo indici più bassi.

	AIC	BIC
Modello logistico, in (3.2)	430.318	452.840
Albero di classificazione, in Figura 3.4	418.641	429.902

Tabella 3.5. Confronto tra i criteri di informazione stimati.

Inoltre, in Figura 3.5 sovrapponiamo la curva *lift* calcolata per l'albero, in blu, alla curva *lift* del modello logistico, in rosso. Tale curva è l'ingrandimento della Figura 3.1, della quale guardiamo solo le frazioni di soggetti più piccole, che corrispondono alla percentuale della clientela alla quale sarà rivolta l'attività di trattenimento e fidelizzazione. Notiamo che la curva *lift* per l'albero di classificazione raggiunge valori più alti di quella per il modello logistico, quando le frazioni sono inferiori a 0.25, mentre le due curve coincidono per frazioni superiori.

Il miglior modello stimato in questo capitolo, per i dati in esame, è quindi l'albero di classificazione.

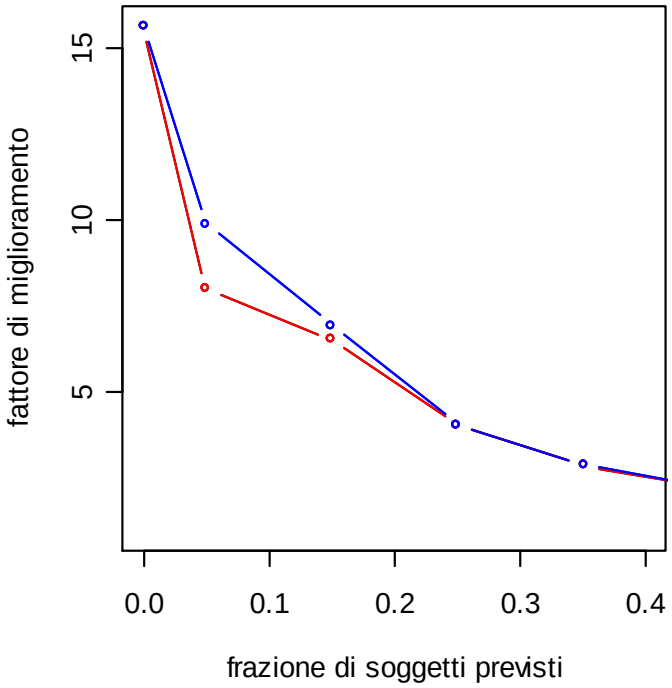

Figura 3.5. Curva della funzione lift, in rosso per il modello logistico ed in blu per l'albero di classificazione.

4. Modelli per dati longitudinali

Nel Capitolo 3 abbiamo stimato alcuni modelli di tipo statico, il modello logistico e l'albero di classificazione, ed abbiamo verificato che è buona la loro capacità di prevedere gli utenti a rischio di abbandono. Tuttavia, è possibile che esista una dipendenza temporale della quale non abbiamo tenuto conto nei modelli, ma che potrebbe spiegare meglio il comportamento degli utenti.

L'idea di fondo è rendere meno statico il modello per la variabile risposta, introducendo nuove variabili che descrivono le traiettorie del traffico telefonico. Tali variabili sintetizzano l'informazione temporale contenuta nei dati osservati ogni mese, e possono essere utilizzate come esplicative nel modello per la variabile Stato. Se l'informazione temporale è importante per spiegare lo Stato, otterremo modelli migliori, e quindi potremo prevedere con maggior precisione quali clienti sceglieranno di lasciare l'azienda.

In questo capitolo, quindi, lasciamo da parte i modelli per la variabile Stato, e passiamo a costruire modelli di diverso tipo per le variabili longitudinali. Questi modelli non sono il vero obiettivo dell'analisi, ma possono essere considerati modelli accessori. Nel prossimo capitolo torneremo a stimare i modelli per la variabile Stato, ma potremo arricchirli aggiungendo le informazioni che derivano dalle analisi accessorie di questo capitolo.

Abbiamo già effettuato le prime stime per la traiettoria delle variabili longitudinali nel Capitolo 2, dove abbiamo effettuato una regressione nel tempo, che coincide con la stima di un trend temporale lineare. In Figura 4.1 rappresentiamo le traiettorie stimate per gli Sms di un campione di 50 utenti Attivi, nel pannello di sinistra, e di 50 utenti Disattivi, nel pannello di destra. I due gruppi hanno traiettorie piuttosto diverse, e la traiettoria media degli Attivi è situata più in alto di quella dei Disattivi. Esiste però una forte etero-

geneità anche all'interno di ogni gruppo, che si nota soprattutto nel gruppo degli Attivi, dove esistono traiettorie in calo e traiettorie in aumento.

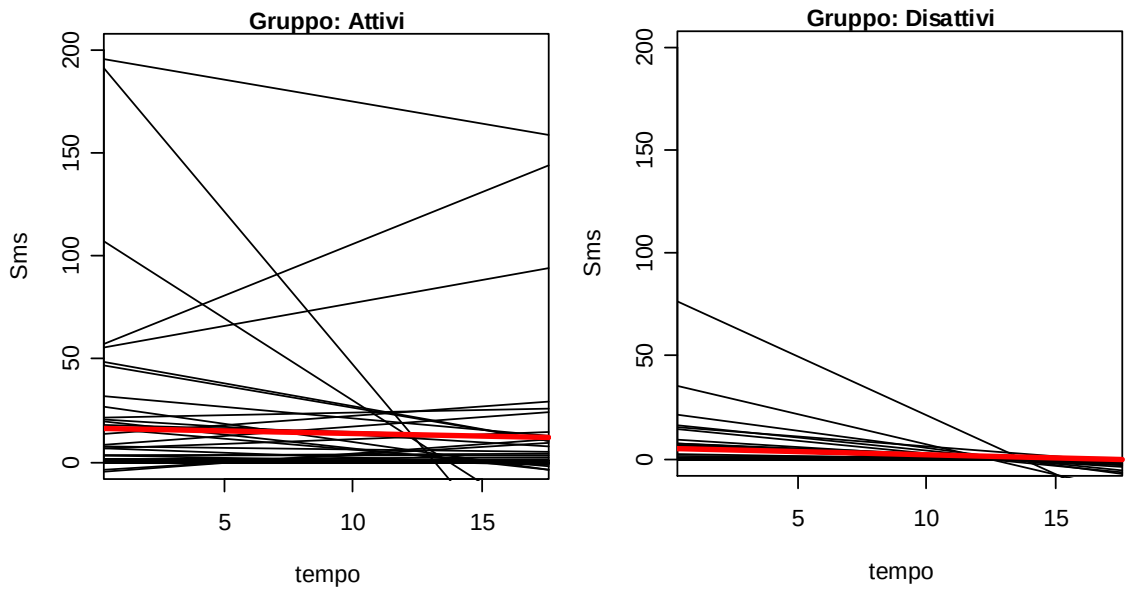

Figura 4.1. Traiettorie stimate per due campioni di 50 individui provenienti dai due gruppi, Attivi e Disattivi. In rosso la traiettoria media del gruppo.

In questo capitolo, introdurremo i modelli ad effetti casuali, nel paragrafo 4.1, e stimeremo nuove traiettorie per i dati longitudinali. Nel paragrafo 4.2, invece, utilizzeremo il lisciamento esponenziale per produrre una previsione del traffico telefonico per il mese di aprile. Come già nel Capitolo 2, per evitare ridondanza, riporteremo i risultati dell'analisi completa solo per la variabile longitudinale Sms, ed eseguiremo gli stessi procedimenti per Mms e per Chiamate verso stesso operatore, ma non li commenteremo.

4.1. Modello ad effetti misti

Volendo spiegare la traiettoria di una variabile longitudinale, si può pensare di costruire un modello di regressione, nel quale la variabile è spiegata da un certo numero di variabili esplicative, tra cui il tempo. Il vettore di parametri stimato da questo modello rappresenta una traiettoria media

dell'intera popolazione. Il modello ad effetti misti, invece, permette ad una parte dei parametri di variare casualmente, per tenere conto di una naturale eterogeneità nella popolazione. La variabile risposta è quindi modellata da una combinazione di effetti fissi, comuni a tutta la popolazione, ed effetti casuali, che variano tra gli individui (Fitzmaurice, Laird, Ware, 2004).

Se chiamiamo $\mathbf{Y_i}$ la variabile longitudinale (il vettore degli Sms per un generico individuo i), possiamo scrivere il modello ad effetti misti come:

$$\mathbf{Y_i} = X_i\boldsymbol{\beta} + Z_i\mathbf{b_i} + \boldsymbol{\varepsilon_i} \qquad i = 1,\ldots,N$$

Qui $\boldsymbol{\beta}$ è il vettore (px1) degli effetti fissi, e $\mathbf{b_i}$ è il vettore (qx1) degli effetti casuali. Le matrici X_i e Z_i contengono le variabili esplicative, e sono rispettivamente di dimensioni (Txp) e (Txq), dove T è il numero di misurazioni ripetute per ogni individuo. Il numero di effetti casuali è inferiore o uguale al numero degli effetti fissi, e le variabili contenute in Z_i sono una parte delle variabili contenute in X_i, in modo che gli effetti casuali sono le variazioni individuali rispetto agli effetti fissi.

La prima colonna di X_i è posta uguale ad uno, perché il primo elemento di $\boldsymbol{\beta}$ è l'intercetta. La seconda colonna è un vettore che misura il tempo, e quindi contiene i valori da 1 a T, mentre le altre colonne possono contenere altre variabili esplicative, sia varianti nel tempo sia statiche. In questa tesi considereremo solo il caso in cui p e q possono essere uguali ad 1 o 2.

Inoltre, gli effetti casuali sono incorrelati con i residui, ed entrambi hanno una distribuzione normale multivariata:

$$\mathbf{b_i} \sim N_q(0, \Gamma)$$
$$\boldsymbol{\varepsilon}_i \sim N_T(0, \Omega)$$

Ricordiamo ora che l'obiettivo dell'analisi è produrre una stima della traiettoria che possa, nel prossimo capitolo, spiegare la futura disattivazione per ogni utente, perciò i parametri di interesse non sono gli effetti fissi, ma

gli effetti casuali. Essendo le $\mathbf{b_i}$ variabili casuali, non ha molto senso pensare di stimare il loro "vero" valore, come faremmo nel caso degli effetti fissi. Possiamo invece considerare una previsione del loro valore per ogni individuo tramite la media condizionale della variabile (Fitzmaurice, Laird, Ware, 2004), dati i valori osservati per la variabile $\mathbf{Y_i}$:

$$\begin{aligned} &E(\mathbf{b_i} \mid \mathbf{Y_i}) = GZ_i' \Sigma_i^{-1}(\mathbf{Y_i} - X_i\hat{\boldsymbol{\beta}}) \\ &\text{dove} \quad \Sigma_i = \text{Cov}(\mathbf{Y_i}) \\ &\qquad\quad G = \text{Cov}(\boldsymbol{\beta}) \end{aligned} \tag{4.1}$$

Questo è reso possibile dal fatto che, dati gli assunti del modello ad effetti misti, $\mathbf{Y_i}$ e $\mathbf{b_i}$ hanno una distribuzione congiunta normale multivariata.

Di seguito cerchiamo di formulare il miglior modello, tale che minimizzi l'indice BIC, provando alcuni valori diversi per p e q, e diverse forme per la matrice di covarianza dei residui Ω.

4.1.1. Modello costante ad intercetta casuale, con residui omoschedastici nel tempo (A)

Il primo modello che stimiamo, seguendo Singer, Willett (2003), non contiene alcuna variabile esplicativa, perciò p e q sono uguali ad 1. Inoltre la matrice di covarianza dei residui è omoschedastica: non esiste correlazione tra i residui a tempi diversi, e la varianza dei residui ad ogni tempo ha lo stesso valore. Il modello è quindi:

$$\begin{aligned} \mathbf{Y_i} &= \beta_0 + b_{i0} + \boldsymbol{\varepsilon_i} \\ \boldsymbol{\varepsilon_i} &\sim N(\mathbf{0}, \sigma_\varepsilon^2 I) \\ b_{i0} &\sim N(0, \sigma_0^2) \end{aligned}$$

Il modello, che chiamiamo A, stima una traiettoria piatta, costante nel tempo, il cui livello può variare casualmente tra gli individui. Nella Figura 4.2 riportiamo gli effetti fissi stimati per questo modello, disegnando la retta

A; il modello suppone che ogni individuo abbia una traiettoria parallela alla retta, più in alto o più in basso a seconda del segno della sua intercetta casuale b_{i0}.

4.1.2. Modello lineare nel tempo ad intercette casuali (B)

Rispetto al modello A, nel modello B è aggiunta la variabile esplicativa tempo solo per la parte ad effetti fissi, perciò p è uguale a 2, mentre q rimane uguale a 1:

$$\mathbf{_i} = \beta_0 + \beta_1 \mathbf{tempo}_i + b_{i0} + \boldsymbol{\varepsilon}_\mathbf{i}$$
$$\boldsymbol{\varepsilon}_\mathbf{i} \sim N(\mathbf{0}, \sigma_\varepsilon^2 I)$$
$$b_{i0} \sim N(0, \sigma_0^2)$$

La traiettoria stimata dagli effetti fissi ha un andamento lineare nel tempo, ed ogni individuo ha una traiettoria parallela alla retta B tracciata in Figura 4.2.

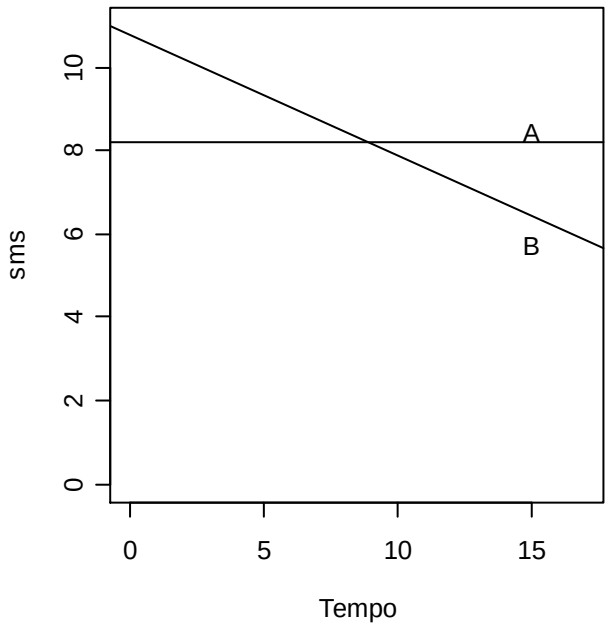

Figura 4.2. Effetti fissi stimati per i modelli A e B.

In Tabella 4.1 riportiamo alcuni valori dei primi due modelli stimati: le varianze dell'intercetta casuale e degli errori, e gli indici di informazione. I due modelli hanno effetti fissi molto diversi, come si nota nella figura, ma la varianza dei residui del modello B non è molto più piccola di quella del modello A, quindi supponiamo che il modello B non riesca a spiegare tutta la variabilità della variabile risposta.

Modello	Varianze stimate	AIC	BIC
A	$\sigma_0^2 = 599.391$ $\sigma_\varepsilon^2 = 487.033$	322 460.9	322 486.3
B	$\sigma_0^2 = 599.529$ $\sigma_\varepsilon^2 = 484.883$	322 323.2	322 357.1

Tabella 4.1. Alcuni valori stimati per i modelli A e B.

4.1.3. Modello lineare nel tempo ad effetti casuali (C)

Il modello C si ottiene complicando ulteriormente il modello B, aggiungendo anche una pendenza casuale, tale che ogni individuo può avere una traiettoria diversa. Abbiamo quindi lo stesso valore per p e q, ed entrambi valgono 2:

$$\mathbf{Y_i} = \beta_0 + \beta_1 \mathbf{tempo}_i + b_{i0} + b_{i1} \mathbf{tempo}_i + \boldsymbol{\varepsilon_i}$$
$$\boldsymbol{\varepsilon_i} \sim N(\mathbf{0}, \sigma_\varepsilon^2 I)$$
$$\mathbf{b_i} \sim N(\mathbf{0}, \Gamma)$$

Questo modello migliora di molto la stima delle traiettorie individuali, perché vediamo in Tabella 4.2 che è diminuita di molto la varianza dei residui. È aumentata, invece, la varianza dell'intercetta, e probabilmente si potrebbe spiegare meglio aggiungendo variabili esplicative invarianti nel tempo, come il sesso dell'utente o la zona di provenienza. Tuttavia, ricordiamo che le analisi di questo capitolo vogliono sintetizzare l'informazione delle variabili longitudinali in pochi parametri da usare successivamente per prevedere lo

Stato, ed è sconsigliabile usare a questo livello variabili che potrebbero entrare nel modello di vero interesse più tardi, perciò decidiamo di non proseguire in questo senso.

Modello	Varianze stimate	AIC	BIC
A	$\sigma_0^2 = 599.391$ $\sigma_\varepsilon^2 = 487.033$	322 460.9	322 486.3
B	$\sigma_0^2 = 599.529$ $\sigma_\varepsilon^2 = 484.883$	322 323.2	322 357.1
C	$\Gamma = \begin{bmatrix} 1100.543 & -45.332 \\ -45.332 & 3.967 \end{bmatrix}$ $\sigma_\varepsilon^2 = 383.747$	317 956.1	318 006.9

Tabella 4.2. Confronto tra i primi modelli del capitolo.

È invece molto importante, per l'analisi successiva, introdurre una struttura di covarianza per i residui diversa da quella omoschedastica. Abbiamo già commentato più volte, in precedenza, che i dati longitudinali possono essere correlati nel tempo, ma finora abbiamo ipotizzato che tutta la correlazione fosse spiegata dalle esplicative, e che i residui fossero incorrelati tra loro. I residui del modello C sono rappresentati in Figura 4.3, e si nota come siano più concentrati per valori positivi che per valori negativi. Non è questo il comportamento che ci aspettiamo da residui omoschedastici.

La struttura della covarianza dei residui può essere analizzata con i metodi proposti da Verbeke e Molenberghs (2000), tuttavia gli autori stessi avvertono che è sufficiente stimare e confrontare i risultati di modelli diversi, quando la forma della covarianza non è di primario interesse, come in questo caso. Per costruire i prossimi modelli, quindi, partiremo dal modello C e specificheremo una struttura per la matrice di covarianza dei residui Ω, che non sia omoschedastica come nei modelli precedenti.

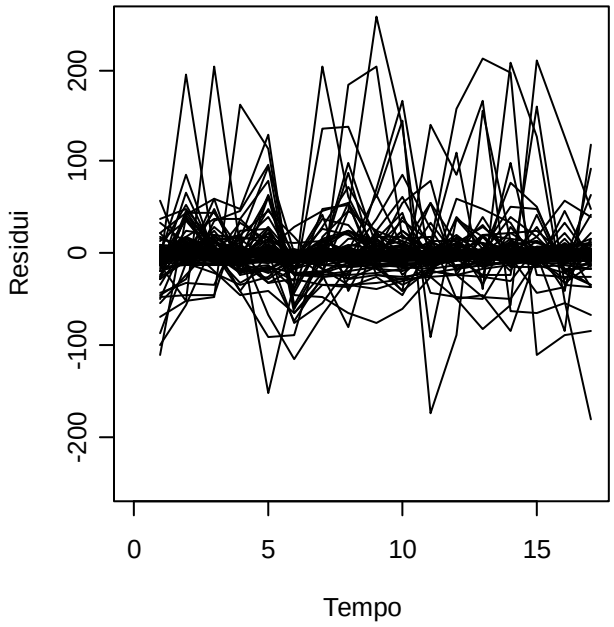

Figura 4.3. Residui per un sottocampione di 200 individui, stimati dal modello C.

4.1.4. Modello lineare ad effetti casuali, con struttura di covarianza autoregressiva di primo ordine (D)

Il modello D si ottiene dal modello C ipotizzando che i residui abbiano una struttura di covarianza autoregressiva del primo ordine:

$$\mathbf{Y_i} = \beta_0 + \beta_1 \mathbf{tempo}_i + b_{i0} + b_{i1} \mathbf{tempo}_i + \boldsymbol{\varepsilon_i}$$

$$\mathbf{b_i} \sim N(\mathbf{0}, \Gamma)$$

$$\boldsymbol{\varepsilon_i} \sim N(\mathbf{0}, \Omega)$$

$$\Omega = \begin{bmatrix} \sigma_\varepsilon^2 & \sigma_\varepsilon^2\phi & \sigma_\varepsilon^2\phi^2 & \dots & \sigma_\varepsilon^2\phi^{16} \\ \sigma_\varepsilon^2\phi & \sigma_\varepsilon^2 & \sigma_\varepsilon^2\phi & & \sigma_\varepsilon^2\phi^{15} \\ \sigma_\varepsilon^2\phi^2 & \sigma_\varepsilon^2\phi & \sigma_\varepsilon^2 & & \sigma_\varepsilon^2\phi^{14} \\ \vdots & & & \ddots & \vdots \\ \sigma_\varepsilon^2\phi^{16} & \sigma_\varepsilon^2\phi^{15} & \sigma_\varepsilon^2\phi^{14} & \dots & \sigma_\varepsilon^2 \end{bmatrix}$$

Prima la matrice di covarianza era diagonale, ora è simmetrica, ma è specificata completamente da due soli parametri: la varianza σ_ε^2, e ϕ, che è la correlazione tra due residui successivi. Nella Tabella 4.3 riportiamo le sti-

me per questi due parametri, oltre alla matrice di covarianza degli effetti casuali. Dato che ϕ è stimato minore di 1, la correlazione decresce esponenzialmente all'aumentare della distanza temporale.

Inoltre, il modello D ha i criteri di informazione più bassi del modello C, e questo significa che la struttura di covarianza ipotizzata migliora la bontà del modello.

Modello	Varianze stimate	AIC	BIC
C	$\Gamma = \begin{bmatrix} 1100.543 & -45.332 \\ -45.332 & 3.967 \end{bmatrix}$ $\sigma^2_\varepsilon = 383.747$	317 956.1	318 006.9
D	$\Gamma = \begin{bmatrix} 858.067 & -27.288 \\ -27.288 & 2.041 \end{bmatrix}$ $\sigma^2_\varepsilon = 480.244$ $\phi = 0.498$	311 675.5	311 734.8

Tabella 4.3. Confronto tra i modelli C e D.

4.1.5. Modello con struttura di covarianza di tipo compound symmetry (E)

Il modello E è formulato come il modello C, ma ha una matrice di covarianza per i residui di tipo *compound symmetry* (Singer, Willett, 2003), una matrice piena e di nuovo specificata da soli due parametri, la varianza σ^2_ε e la covarianza ρ, che è costante ad ogni tempo:

$$\mathbf{Y_i} = \beta_0 + \beta_1 \mathbf{tempo}_i + b_{i0} + b_{i1} \mathbf{tempo}_i + \boldsymbol{\varepsilon_i}$$

$$\mathbf{b_i} \sim N(\mathbf{0}, \Gamma)$$

$$\boldsymbol{\varepsilon_i} \sim N(\mathbf{0}, \Omega)$$

$$\Omega = \begin{bmatrix} \sigma^2_\varepsilon & \rho & \rho & \dots & \rho \\ \rho & \sigma^2_\varepsilon & \rho & & \rho \\ \rho & \rho & \sigma^2_\varepsilon & & \rho \\ \vdots & & & \ddots & \vdots \\ \rho & \rho & \rho & \dots & \sigma^2_\varepsilon \end{bmatrix}$$

Questa struttura di covarianza può essere troppo restrittiva, ed infatti la stima per parametro ρ è zero (terza riga della Tabella 4.5): il miglior modello di tipo E che possiamo stimare sui dati non ha la covarianza costante, ed in pratica coincide con il modello C.

Modello	Varianze stimate	AIC	BIC
C	$\Gamma = \begin{bmatrix} 1100.543 & -45.332 \\ -45.332 & 3.967 \end{bmatrix}$ $\sigma_\varepsilon^2 = 383.747$	317 956.1	318 006.9
D	$\Gamma = \begin{bmatrix} 858.067 & -27.288 \\ -27.288 & 2.041 \end{bmatrix}$ $\sigma_\varepsilon^2 = 480.244$ $\phi = 0.498$	311 675.5	311 734.8
E	$\Gamma = \begin{bmatrix} 1100.542 & -45.332 \\ -45.332 & 3.967 \end{bmatrix}$ $\sigma_\varepsilon^2 = 383.747$ $\rho = 0$	317 958.1	318 017.4

Tabella 4.4. Confronto tra i modelli C, D ed E

4.1.6. Modello con struttura di covarianza gaussiana (F)

In Verbeke, Molenberghs (2000) troviamo altre due forme per la matrice di covarianza, la correlazione esponenziale e la correlazione gaussiana. Queste strutture sono normalmente usate per esprimere una correlazione spaziale, o temporale quando il tempo è misurato in modo continuo, invece che ad intervalli regolari. Nel caso discreto, come quello che stiamo analizzando, la correlazione esponenziale coincide con la correlazione autoregressiva (modello D), ma possiamo applicare una struttura gaussiana alla matrice di covarianza dei residui. Il modello è il seguente:

$$\mathbf{Y_i} = \beta_0 + \beta_1 \mathbf{tempo}_i + b_{i0} + b_{i1} \mathbf{tempo}_i + \boldsymbol{\varepsilon_i}$$
$$\mathbf{b_i} \sim N(\mathbf{0}, \Gamma)$$
$$\boldsymbol{\varepsilon_i} = \boldsymbol{\varepsilon}_{(1)\mathbf{i}} + \boldsymbol{\varepsilon}_{(2)\mathbf{i}}$$
$$\boldsymbol{\varepsilon}_{(1)\mathbf{i}} \sim N(\mathbf{0}, \sigma_\varepsilon^2 I)$$
$$\boldsymbol{\varepsilon}_{(2)\mathbf{i}} \sim N(\mathbf{0}, H)$$

La formulazione è valida sia per il modello con correlazione esponenziale, sia con correlazione gaussiana, per diverse forme della matrice H. Per il modello F, che ha correlazione gaussiana, la matrice si scrive:

$$H = [h_{jk}] = \left[\exp\left(- \frac{| \text{tempo}_{ij} - \text{tempo}_{ik} |}{r} \right)^2 \right]$$

dove il *range* r è il parametro da stimare, ed è il valore con cui vengono pesate le distanze, in modo che la covarianza tra i residui diminuisca all'aumentare della distanza. Confrontiamo i valori stimati per questo modello con quelli dei modelli C e D, in Tabella 4.5, e scopriamo che anche questo modello è migliore rispetto a C, ma non è buono come il modello D, che resta il miglior modello ad effetti casuali che abbiamo stimato.

Infine nella Figura 4.4 rappresentiamo gli effetti fissi stimati per tutti i modelli. Si nota che, eccetto il modello A, tutti i modelli hanno stimato la stessa traiettoria.

Modello	Varianze stimate	AIC	BIC
C	$\Gamma = \begin{bmatrix} 1100.543 & -45.332 \\ -45.332 & 3.967 \end{bmatrix}$ $\sigma^2_\varepsilon = 383.747$	317 956.1	318 006.9
D	$\Gamma = \begin{bmatrix} 858.067 & -27.288 \\ -27.288 & 2.041 \end{bmatrix}$ $\sigma^2_\varepsilon = 480.244$ $\phi = 0.498$	311 675.5	311 734.8
F	$\Gamma = \begin{bmatrix} 991.729 & -37.340 \\ -37.340 & 3.188 \end{bmatrix}$ $\sigma^2_\varepsilon = 406.083$ $r = 0.974$	312 847.8	312 907.1

Tabella 4.5. Confronto tra i modelli C, D, F.

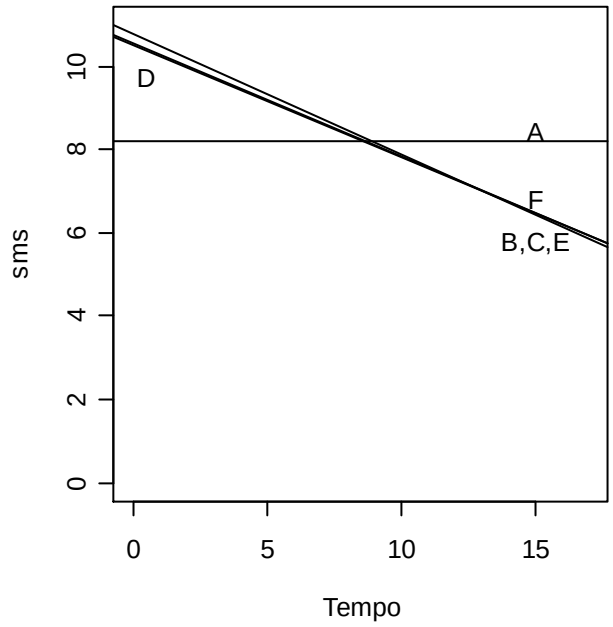

Figura 4.4. Effetti fissi stimati per i modelli del capitolo, relativamente alla variabile Sms.

4.1.7. Stima dei coefficienti casuali

Nel paragrafo precedente abbiamo scelto il modello D come miglior modello per i dati, secondo gli indici di informazione, ed ora procediamo alla stima del modello per il campione *completo*, compresi gli utenti che fanno parte del campione di verifica. Infatti, in questo capitolo non stiamo costruendo una regola generale per la variabile risposta Stato, come abbiamo fatto nel Capitolo 3, ma stiamo costruendo un modello accessorio dal quale vogliamo ottenere la stima di alcuni parametri, che nel prossimo capitolo diventeranno variabili esplicative nei modelli per lo Stato.

Se applicassimo il modello accessorio solo al campione di stima, nel Capitolo 5 potremmo stimare un nuovo modello in cui i coefficienti casuali sono variabili esplicative, ma non potremmo verificarlo, perché per gli utenti del campione di stima non sarebbero disponibili i valori di queste variabili esplicative.

Stimiamo allora il modello per tutto il campione, e per ogni utente eseguiamo la previsione degli effetti casuali come introdotto all'inizio del capitolo, con la (4.1). Sempre all'inizio del capitolo abbiamo notato come gli effetti casuali siano le variazioni individuali rispetto agli effetti fissi, quindi la traiettoria di ogni individuo è espressa dalla somma tra gli effetti fissi e gli effetti casuali. Possiamo quindi definire un nuovo vettore, che chiamiamo $\boldsymbol{\beta}_i$, il cui valore è proprio tale somma:

$$\hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\beta}} + \hat{\mathbf{b}}_i$$

Il vettore $\boldsymbol{\beta}_i$ *contiene i coefficienti casuali che utilizzeremo nel prossimo capitolo, quindi guardiamo la distribuzione di questi coefficienti per i due gruppi. Riportiamo in Figura 4.5 gli istogrammi con la distribuzione dei coefficienti stimati per la variabile Sms, in Figura 4.5. Distribuzione dei coefficienti casuali stimati per la variabile Sms.*

per la variabile Mms, in Figura 4.6. Distribuzione dei coefficienti casuali stimati per la variabile Mms. e Figura 4.8 per la variabile Chiamate verso stesso operatore.

Per tutti i parametri la distribuzione è fortemente concentrata su valori vicini allo zero, e la maggiore differenza tra i due gruppi sembra l'ampiezza del *range* di valori assunti dai parametri.

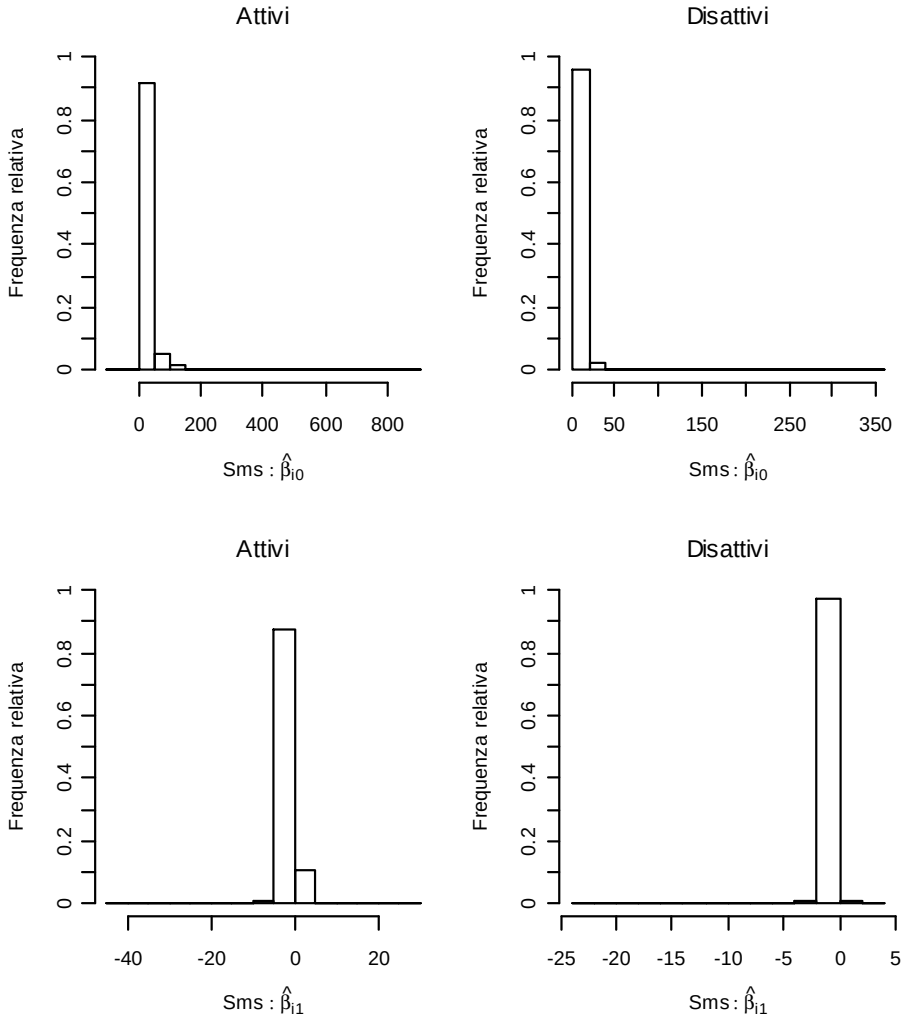

Figura 4.5. Distribuzione dei coefficienti casuali stimati per la variabile Sms.

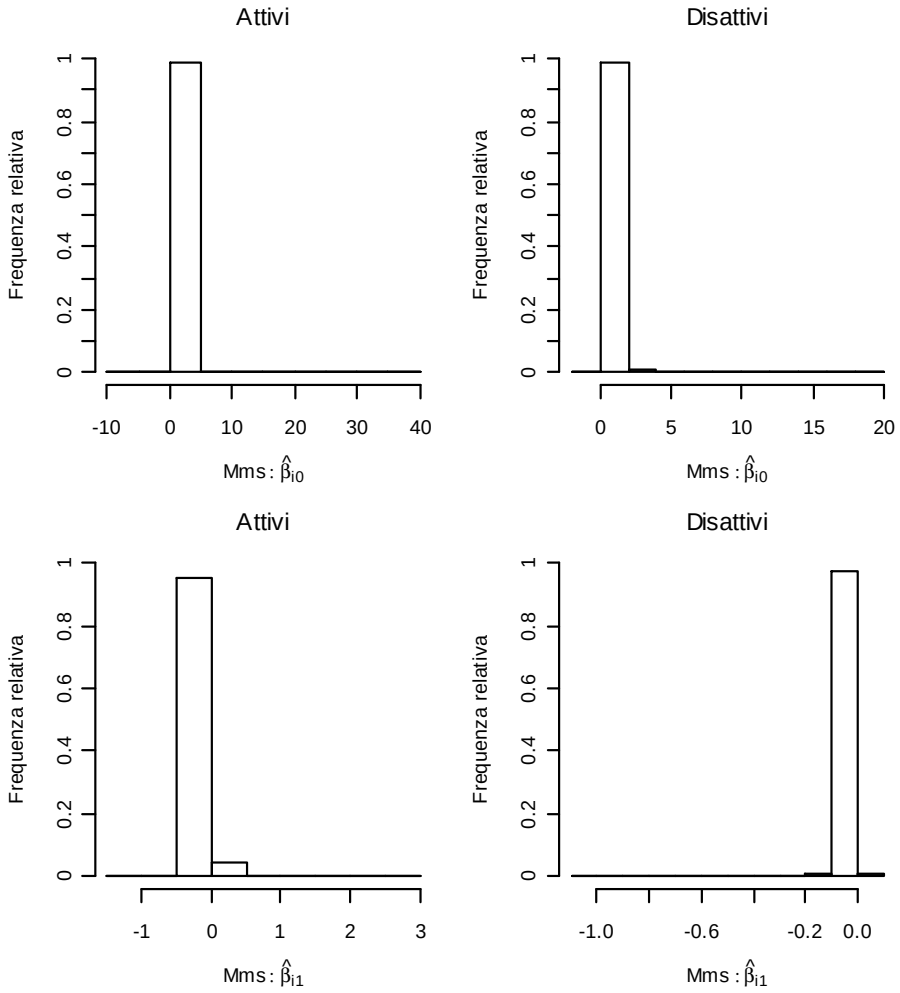

Figura 4.6. Distribuzione dei coefficienti casuali stimati per la variabile Mms.

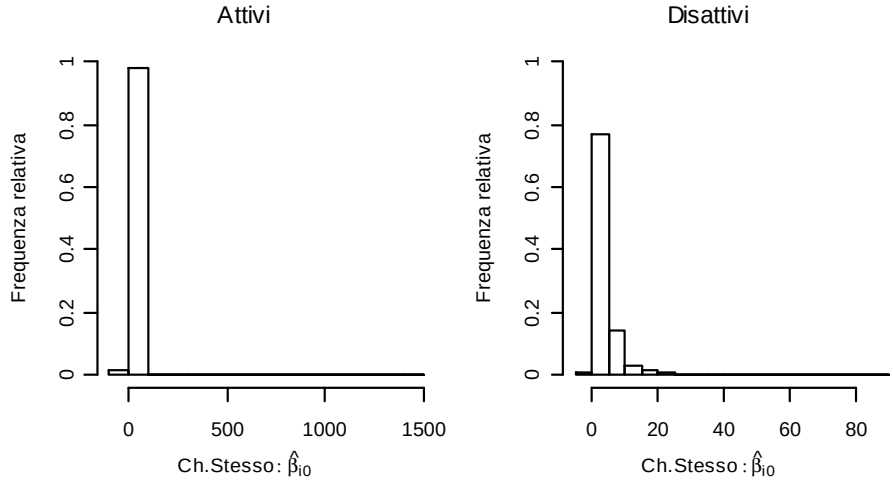

Figura 4.7. Distribuzione dei coefficienti casuali β_{i0} *stimati per la variabile Chiamate verso stesso operatore.*

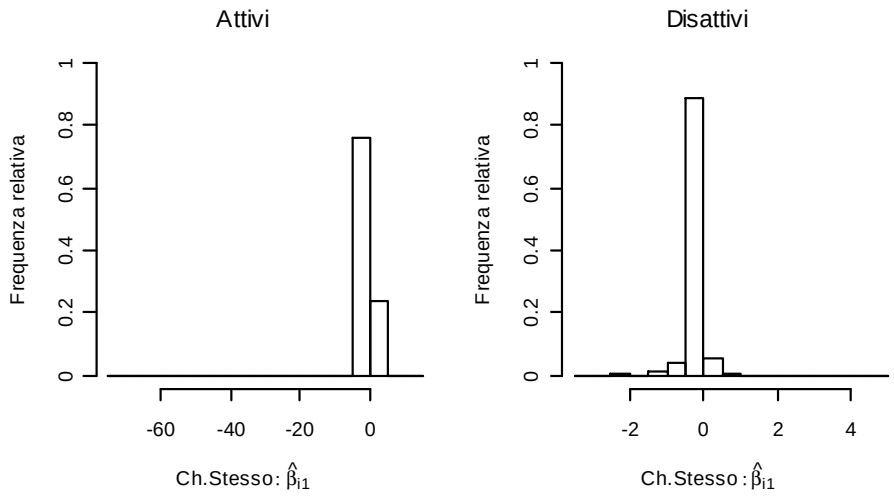

Figura 4.8. Distribuzione dei coefficienti casuali β_{i1} *stimati per la variabile Chiamate verso stesso operatore.*

4.2. Lisciamento esponenziale

Nel Capitolo 2 abbiamo commentato che le serie disponibili per ogni utente sono troppo corte per permetterci di stimare un processo stocastico: sono composte solo da diciassette dati. Questo è un problema che si riscontra molto spesso nell'analisi di dati longitudinali, e per questo motivo solitamente si eseguono analisi di diverso tipo rispetto alle serie storiche, ad esempio modelli ad effetti casuali, come abbiamo fatto nel paragrafo precedente.

Piuttosto di stimare un processo stocastico, decidiamo di applicare un lisciamento alle serie, e di usarlo per prevedere il valore della variabile longitudinale per il mese di aprile 2006, il mese in cui avviene l'eventuale disattivazione. Il metodo che utilizziamo è il lisciamento esponenziale (Di Fonzo, Lisi, 2005).

Il lisciamento esponenziale di una serie storica consiste nella media pesata di tutti i valori passati della serie stessa, nella quale i pesi diminuiscono esponenzialmente all'aumentare della distanza temporale tra le osservazioni.

Se chiamiamo $F_{iT,1}$ il lisciamento della serie $\mathbf{Y}_i$ per l'individuo i al tempo T+1 (non osservato), si può scrivere:

$$F_{iT,1} = (1-\delta)\sum_{j=1}^{T} \delta^j \, Y_{i(T-j)}$$

dove la costante di lisciamento δ assume valori compresi tra 0 ed 1. Uno dei vantaggi di questo metodo di lisciamento è la sua semplicità di aggiornamento, perché ad ogni nuova osservazione non è necessario calcolare nuovamente la media pesata. La formula, infatti, si può riscrivere come:

$$F_{iT,1} = (1-\delta)\, y_T + \delta\, F_{T-1,1}$$

Quindi, la previsione per il tempo T+1 è una media pesata della previsione per il tempo T, fatta con le osservazioni fino a T-1, e l'ultima osservazione.

Possiamo ora stimare il parametro di lisciamento δ per ogni serie osservata con il metodo dei minimi quadrati, e lasciamo che il parametro assuma valori diversi per ogni individuo. Da questa stima prevediamo il valore $F_{iT,1}$, che è il numero di sms che il cliente i spedirà nel mese di aprile 2006. La previsione potrebbe essere importante per spiegare se avviene la disattivazione in quel mese, perciò sarà utilizzata nel prossimo capitolo come variabile esplicativa, che chiamiamo sms_apr. Rappresentiamo la distribuzione dei valori che la variabile assume nei due gruppi in Figura 4.9. Sono stati previsti soprattutto valori bassi, anche se per qualche cliente attivo la previsione supera i 1500 sms.

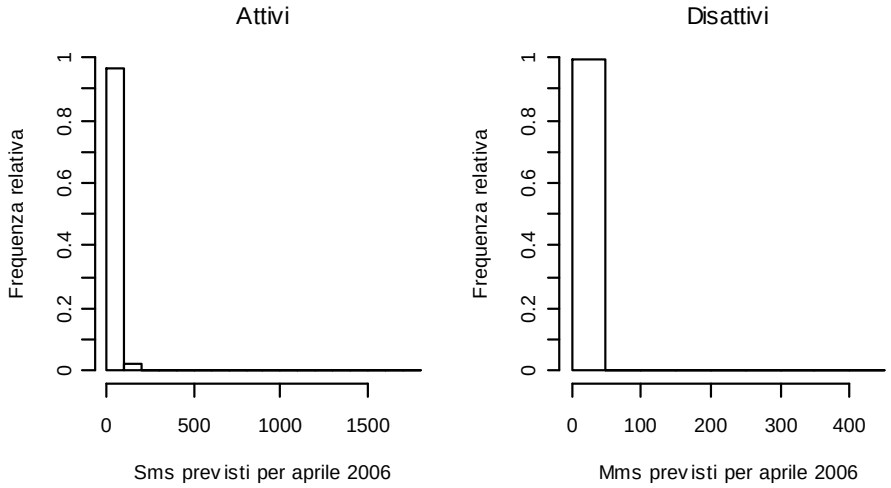

Figura 4.9. Distribuzione delle previsioni per il mese di aprile, nei due gruppi.

In alternativa, possiamo decidere di trasformare la previsione in una variabile fattore, che indica se prevediamo che l'utente spedirà almeno un sms nel mese di aprile 2006. La variabile è definita da:

$$\text{sms\_0}_i = \begin{cases} \text{N} & \text{se sms\_apr}_i = 0 \\ \text{S} & \text{altrimenti} \end{cases}$$

La distribuzione della variabile sms_0 è in Figura 4.10, dove riportiamo anche la distribuzioni delle variabili definite allo stesso modo anche per gli Mms (mms_0) e per le Chiamate verso stesso operatore (ch.stesso_0). Solo la variabile sms_0 ha una distribuzione completamente diversa nei due gruppi, perché per gli Attivi è più frequente il valore S, mentre per i Disattivi è più frequente il valore N.

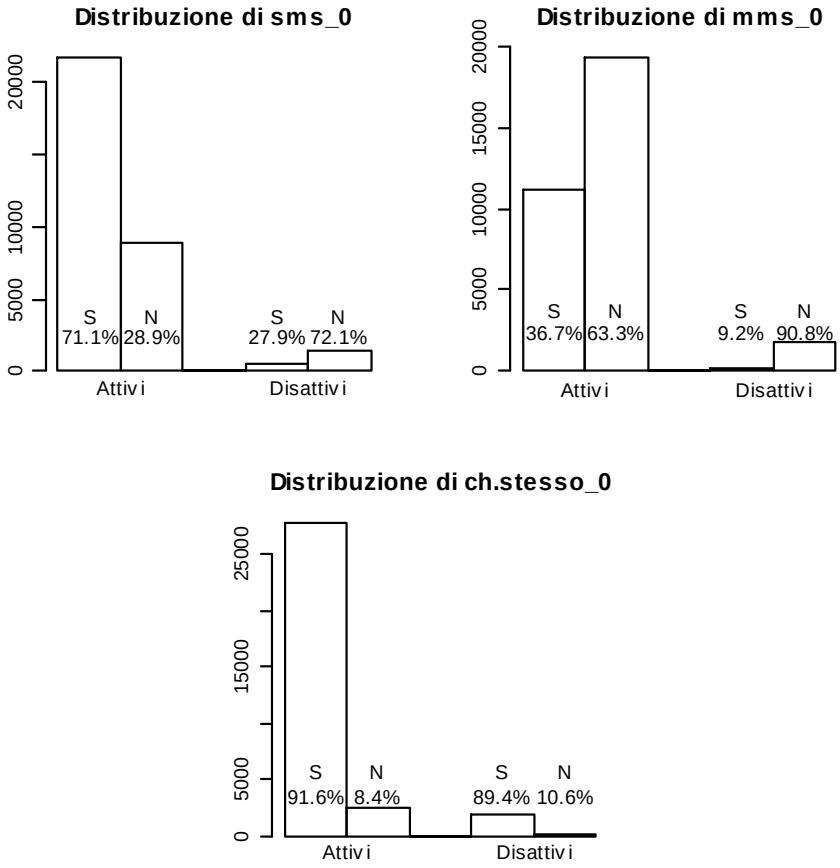

Figura 4.10. Distribuzione per le variabili fattore sms_0, mms_0 e ch.stesso_0.

5. Modelli le cui esplicative sono parametri di modelli longitudinali

In questo capitolo vogliamo tornare ai modelli classici del Capitolo 3, per arricchirli con i risultati ottenuti dalle analisi longitudinali del Capitolo 4. Stimeremo quindi alcuni modelli nei quali le variabili esplicative sono una certa rappresentazione delle traiettorie individuali, e li confronteremo con i modelli precedenti, per decidere se migliorano la previsione del comportamento dei clienti.

Nel primo paragrafo di questo capitolo stimeremo il modello logistico, mentre il paragrafo 5.2 sarà dedicato agli alberi di classificazione. Per entrambi i modelli saranno inseriti, tra le variabili esplicative, i parametri dei modelli lineari nel tempo, stimati nel Capitolo 2, i coefficienti casuali del modello ad effetti misti, stimati nel Capitolo 4, e le previsioni ottenute con il lisciamento esponenziale, dello stesso capitolo.

5.1. Modello logistico

Nel Capitolo 3 abbiamo scelto come miglior modello logistico il seguente:

$$\begin{aligned} \text{Stato}_i &\sim \text{Bi}(1, \pi_i) \\ \pi_i &= \frac{e^{\eta_i}}{1 + e^{\eta_i}} \\ \eta_i &= \beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i \end{aligned} \qquad (5.1)$$

Tra le variabili del predittore lineare η_i aggiungiamo alcune variabili derivate dall'analisi accessoria del Capitolo 4, e delle analisi esplorative del capitolo 2.

5.1.1. Trend lineare e trend quadratico

Il primo modello temporale che abbiamo stimato, nel Capitolo 2, è un modello a trend lineare o quadratico, che allora abbiamo usato solo per sintetizzare le traiettorie e capire qualche andamento generale. Ora invece recuperiamo i parametri stimati per ogni individuo, relativi alle variabili Sms, Mms e Chiamate verso stesso operatore, e li introduciamo nel predittore lineare. Come già nell'analisi del Capitolo 3, aggiungiamo un parametro per volta e confrontiamo la bontà dei modelli con i criteri di informazione, in Tabella 5.1. Il predittore della prima riga è del modello originario, in (5.1). Notiamo subito che alcuni modelli hanno i criteri di informazione più bassi di quello originario, nonostante ognuno di essi contenga una variabile in più.

η_i	AIC	BIC
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i$	430.318	452.840
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i + \beta_4 \text{sms\_beta0}_i$	424.559	452.712
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i + \beta_4 \text{sms\_beta1}_i$	428.782	456.934
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i + \beta_4 \text{mms\_beta0}_i$	431.908	460.061
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i + \beta_4 \text{mms\_beta1}_i$	431.956	460.109
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i + \beta_4 \text{ch.stesso\_beta0}_i$	411.121	439.273
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i + \beta_4 \text{ch.stesso\_beta1}_i$	423.115	451.268

Tabella 5.1. Criteri di informazione e verosimiglianza per i modelli logistici il cui predittore è nelle righe.

Dalla tabella scegliamo il modello con gli indici più bassi, nella sesta riga, il modello in cui è esplicativa anche l'intercetta della variabile Chiamate verso stesso operatore. Ripetiamo il predittore lineare:

$$\eta_i = \beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i + \beta_4 \text{ch.stesso\_beta0}_i \qquad (5.2)$$

Potremmo complicare ancora il modello, aggiungendo anche la pendenza della stessa variabile, o l'intercetta stimata per gli Sms, perché sono i parametri che hanno portato ai modelli migliori della Tabella 5.1. Invece l'aggiunta di una nuova variabile non migliora abbastanza l'adattamento ai dati, probabilmente perché esiste correlazione con la variabile già inserita, e non otteniamo alcun modello migliore secondo il criterio BIC. Non riportiamo quindi i risultati dei nuovi modelli.

Passiamo invece ai modelli che derivano dalla stima di un trend quadratico per le variabili longitudinali. Qui i parametri stimati sono tre per ognuna delle tre variabili. Stimiamo perciò nove diversi modelli logistici, ognuno dei quali aggiunge alle esplicative di quello originario uno solo dei parametri stimati dal trend quadratico. Di questi, solo il modello la cui esplicativa è l'intercetta del trend per le Chiamate verso stesso operatore riesce a migliorare il modello originario, ma è comunque peggiore del modello in (5.2), che rimane il miglior modello di questo paragrafo.

Applichiamo quindi il modello (5.2) agli utenti del campione di verifica, per calcolare gli errori di previsione, che riportiamo in Tabella 5.2, e per tracciare la curva lift del modello, in verde nella Figura 5.1, dove è confrontata con la curva del modello del Capitolo 3. La curva è leggermente più in alto, ma il numero di falsi negativi (nella tabella) è aumentato, perciò il miglioramento non è così evidente come potevamo sperare.

	Risposta	osservata
Risposta prevista:	disattivo	attivo
disattivo	1037	602
attivo	0	14176

Tabella 5.2. Matrice di confusione per il modello logistico il cui predittore lineare è in (5.2).

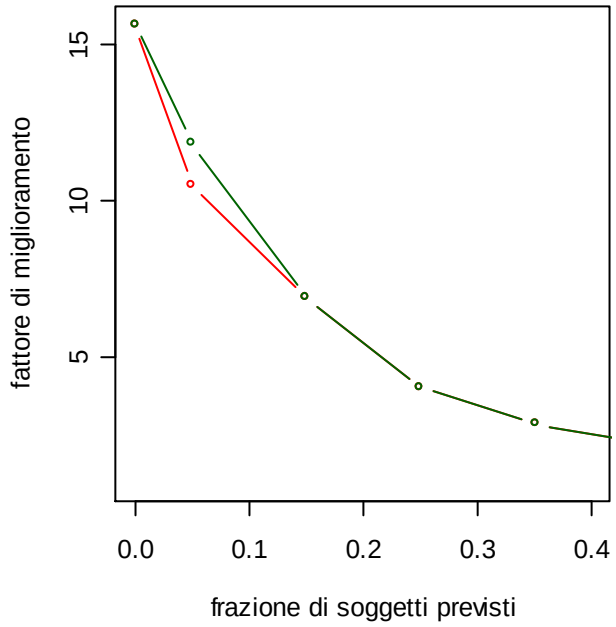

Figura 5.1. Confronto tra le curve lift per i modelli logistici i cui predittori lineare sono in (5.1), in rosso, e in (5.2), in verde.

5.1.2. Effetti casuali

Nella prima parte del Capitolo 4 abbiamo ipotizzato che l'andamento delle variabili longitudinali nel tempo seguisse un modello ad effetti misti, nel quale si potevano riconoscere molte traiettorie individuali, espresse dagli effetti casuali, lineari nel tempo. Il modello quindi potrebbe sembrare simile a quello con trend lineare, che abbiamo usato nel paragrafo precedente; tuttavia la principale differenza riguarda la struttura di correlazione dei residui, e non è affatto di secondaria importanza.

Già nel Capitolo 2 abbiamo notato che poteva aver poco senso costruire un modello lineare per i dati longitudinali, perché tale modello non accetta correlazione nei residui; nel modello ad effetti misti, invece, abbiamo incorporato un'ipotesi aggiuntiva sulla struttura di covarianza, che abbiamo definito di tipo autoregressivo del primo ordine.

Ritorniamo dunque al modello logistico originario (5.1) ed aggiungiamo al predittore lineare l'intercetta casuale e la pendenza casuale delle variabili Sms, Mms, e Chiamate verso stesso operatore. I migliori modelli si ottengono

con i coefficienti delle Chiamate verso stesso operatore, e il modello che contiene l'intercetta casuale di tale variabile (ch.stesso_int) ha i criteri più bassi sia del modello originario sia del modello in (5.2). In Tabella 5.3 confrontiamo infatti gli indici di informazione per i modelli appena citati, e l'ultimo modello ottiene tutti i valori più bassi.

	AIC	BIC
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i$	430.318	452.840
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i + \beta_4 \text{ch.stesso\_beta0}_i$	411.121	439.273
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i + \beta_4 \text{ch.stesso\_int}_i$	408.639	436.791

Tabella 5.3. Confronto tra i modelli in (5.1), in (5.2), ed il modello con l'intercetta casuale per la variabile Chiamate verso stesso operatore.

A questo modello abbiamo provato ad aggiungere altri parametri, ma non otteniamo un indice BIC più basso. Non riportiamo i risultati relativi alle altre variabili perché poco rilevanti. Scegliamo invece come miglior modello quello il cui predittore lineare è:

$$\eta_i = \beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i + \beta_4 \text{ch.stesso\_int}_i \qquad (5.3)$$

Applicando questo modello al campione di verifica, otteniamo una tabella di classificazione identica a quella di Tabella 5.2, e gli errori di classificazione rimangono gli stessi del modello originario. L'errore totale è del 3.8%, non c'è errore di primo tipo ed otteniamo un errore del secondo tipo del 4,1%. Quindi anche questo modello è sostanzialmente equivalente ai precedenti, almeno per quanto riguarda la previsione.

5.1.3. Lisciamento esponenziale

Concludiamo la sezione relativa al modello di regressione logistico considerando le previsioni ottenute con il lisciamento esponenziale. Nell'ultimo

paragrafo del Capitolo 4 abbiamo calcolato la previsione per il valore del traffico telefonico ad aprile 2006, e questa è una variabile che useremo qui come esplicativa. L'altra variabile, invece, indica se il traffico telefonico, ad aprile, è previsto uguale a zero, oppure maggiore.

Abbiamo inserito queste previsioni come variabili esplicative nel predittore lineare, ed abbiamo ottenuto per ogni modello gli indici di informazione che riportiamo in Tabella 5.4. Il modello in cui compare sms_0, la previsione per gli sms, è migliore di quello originale, ma l'indice BIC non è inferiore a quello della (5.3).

η_i	AIC	BIC
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i$	430.318	452.840
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i + \beta_4 \text{sms\_apr}_i$	427.726	455.879
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i + \beta_4 \text{mms\_apr}_i$	432.279	460.431
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i + \beta_4 \text{ch.stesso\_apr}_i$	432.279	460.431
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i + \beta_4 \text{sms\_0}_i$	415.653	443.806
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i + \beta_4 \text{mms\_0}_i$	431.851	460.003
$\beta_0 + \beta_1 \text{Altri}_i + \beta_2 \text{Fisso}_i + \beta_3 \text{Età}_i + \beta_4 \text{ch.stesso\_0}_i$	431.355	459.507

Tabella 5.4. Indici di informazione per i modelli i cui predittori lineari sono nelle righe.

5.2. Classificazione ad albero

Nel Capitolo 3 abbiamo introdotto anche la classificazione ad albero. Vogliamo ora ampliare l'analisi allo stesso modo dei paragrafi precedenti, aggiungendo le variabili che derivano dalle analisi longitudinali. Il processo di stima selezionerà tra le variabili proposte il modello più adatto ai dati, e probabilmente confermerà il modello del Capitolo 3, perché era il miglior modello che abbiamo costruito, ma aggiungerà alcuni nodi nella parte inferiore, tra i quali compariranno forse le nuove variabili. Compiremo poi una potatura

dell'albero per ottenere la minima devianza, e se il modello migliore sarà diverso dal precedente, potremo confrontarne la bontà di previsione.

Il processo di selezione delle variabili, per gli alberi di classificazione, è tale che, ad ogni nodo, sono prese in considerazione solo le variabili che permettono di ottenere una migliore separazione tra i gruppi. I due alberi della Figura 5.2 sono ottenuti dall'aggiunta delle variabili provenienti dall'analisi longitudinale, ed il primo contiene i parametri del trend lineare, il secondo i coefficienti casuali.

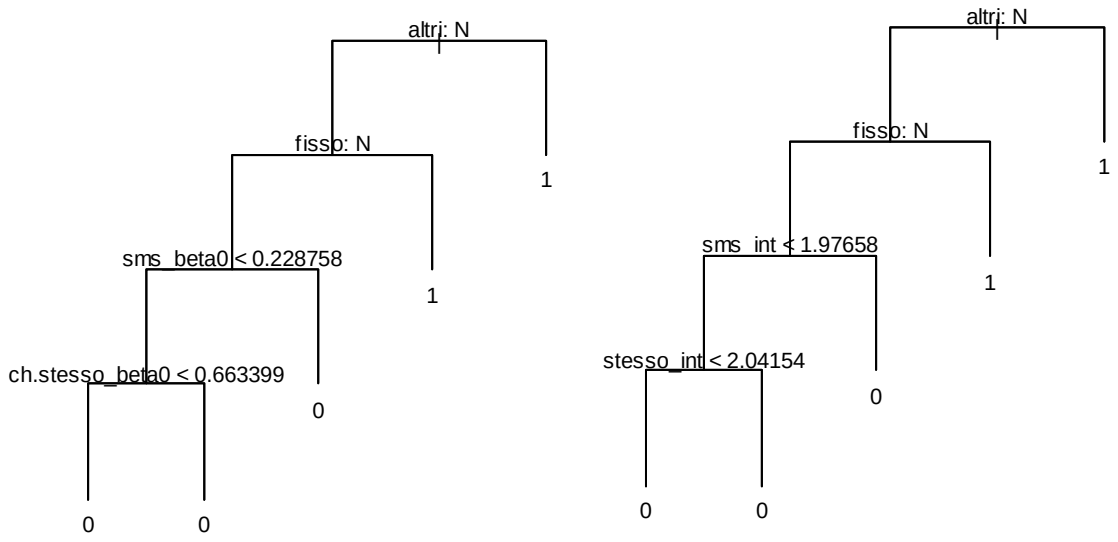

Figura 5.2. Alberi di classificazione con le variabili provenienti dall'analisi longitudinale, a destra per il trend lineare, e a sinistra per gli effetti casuali.

In entrambi gli alberi, le foglie inferiori appartengono alla stessa classe, quindi con la potatura devono essere eliminate, e l'albero finale contiene solo due nodi, cioè è l'albero scelto nel Capitolo 3.

Lo stesso accade per l'albero che contiene le previsioni ottenute dal lisciamento esponenziale, che riportiamo in Figura 5.3.

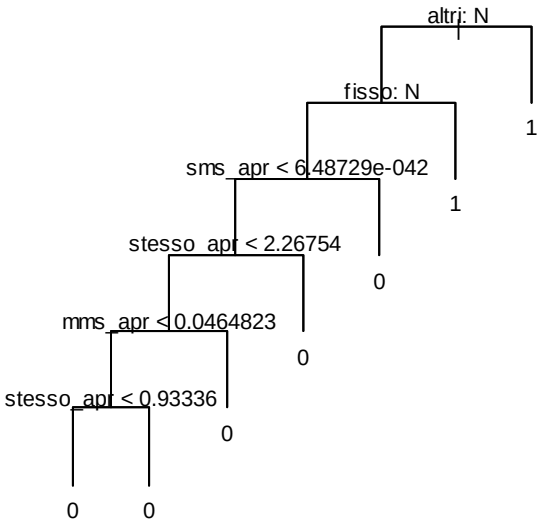

Figura 5.3. Albero di classificazione con le previsioni ottenute con il lisciamento esponenziale.

Nessun modello, né del Capitolo 3 né di questo capitolo, è riuscito ad ottenere un errore di classificazione inferiore al 3.8%, quindi tutti i modelli sono equivalenti per la qualità di previsione. Ne concludiamo che il miglior modello è l'albero di classificazione stimato nel Capitolo 3, perché è il migliore secondo gli indici di classificazione, ed è equivalente agli altri per la bontà di previsione.

6. Conclusioni e ulteriori sviluppi

Nel corso di questa tesi abbiamo costruito alcuni modelli statistici, con lo scopo di prevedere l'eventuale disattivazione dei clienti di un'azienda telefonica. In particolare, avevamo l'obiettivo di verificare se fosse possibile migliorare la previsione, tenendo conto dell'eventuale dipendenza temporale per alcune variabili.

Per i dati che abbiamo a disposizione per l'analisi, abbiamo scoperto che i modelli classici portavano già a risultati soddisfacenti, perché consentivano di prevedere i clienti che si sarebbero disattivati, con un basso errore di classificazione. Si potrebbe tuttavia allargare l'analisi, considerando altri tipi di modelli statistici, sia parametrici, come l'analisi discriminante, sia non parametrici, come ad esempio i *MARS* (*multivariate adaptive regression splines*). Dall'analisi sappiamo, però, che se utilizzeremo la classificazione ad albero per scegliere una parte della clientela, alla quale indirizzare le attività di fidelizzazione e di trattenimento, sarà molto più facile contattare i clienti a rischio di abbandono, piuttosto che affidarsi ad una scelta casuale.

I modelli costruiti tenendo conto della natura longitudinale dei dati, invece, non hanno migliorato significativamente la bontà delle previsioni. Abbiamo ottenuto solo miglioramenti marginali, ma al prezzo di un aumento di complessità. Se è indifferente scegliere tra un modello semplice ed uno più complesso, sia concettualmente che computazionalmente, nella pratica la scelta cadrà sempre sul modello semplice.

Probabilmente, per i dati a disposizione, l'albero di classificazione riesce a spiegare tutta la variabilità che non è casuale, e quindi non è necessario ampliare l'analisi. In situazioni meno ottimali questo potrebbe non succedere, ed allora è più probabile che i modelli di tipo longitudinale riescano a migliorare sostanzialmente la capacità previsiva rispetto ai modelli classici.

D’altra parte si potrebbe sviluppare l’analisi creando variabili fattore anche per gli altri tipi di traffico telefonico, perché è possibile che, proprio come le due variabili Altri e Fisso, che abbiamo trasformato, possano facilitare la previsione.

Appendice A: i comandi in R

Questa appendice è dedicata alla formulazione per R dei modelli stimati nel corso della tesi. Di seguito proporremo i comandi che sono stati utilizzati per stimare i modelli dei capitoli 3 e 4, ed alcuni risultati che non sono stati esposti nel testo.

Per scrivere i comandi abbiamo consultato Iacus, Masarotto (2003), Azzalini, Scarpa (2004), Venables, Ripley (1999), e le risorse on line fornite dal programma. Per la spiegazione dei modelli e per la relativa bibliografia si rimanda invece al testo.

A.1. Modello logistico

Il modello logistico fa parte della famiglia dei modelli lineari generalizzati, ed è stimato dal comando `glm()`. Chiamiamo "stima" il dataset che contiene i dati di tutte le variabili per gli utenti del campione di stima; la formulazione per il modello (3.3) è la seguente:

```
logistico<-glm(Stato~Eta+Fisso+Altri, data=stima, family=binomial())
```

I coefficienti stimati dal modello sono nella Tabella A.1. Notiamo che nessuno dei coefficienti assume valori significativamente diversi da zero. Questo è dovuto alla multicollinearità tra le variabili Fisso e Altri, che però non può essere eliminata togliendo una delle variabili dal modello: il modello, infatti, è stato scelto per minimizzare i criteri di informazione, e non esistono altri modelli equivalenti. Si potrebbe risolvere il problema creando una variabile intersezione tra Fisso ed Altri, che vale N solo quando entrambe valgono N, e vale S altrimenti.

	Stima	Std.Err	z value	Pr(>|z|)
(Intercept)	-3.096	0.265	-11.64	<2e-16
Eta	0.007	0.0126	0.608	0.543
FissoS	20.398	844.689	0.024	0.981
AltriS	21.113	810.952	0.026	0.979

Tabella A.1. Valori stimati per i coefficienti del modello logistico.

Per quanto riguarda gli indici di informazione, solo il criterio AIC è presentato nell'output del modello, mentre se vogliamo visualizzarli entrambi possiamo usare il comando:

```
c(AIC=AIC(logistico,k=2), BIC=AIC(logistico,k=log(nrow(stima))))
```

dove k è la parte della penalità che va moltiplicata al numero di parametri.

Dal modello stimato si ottengono le previsioni, con il comando `predict`, come probabilità di successo. Se chiediamo che tale probabilità sia maggiore del 50%, otteniamo un vettore logico, dove `true` significa attivo e `false` significa disattivo.

```
previsti<-predict(logistico, type="response", newdata=verifica)>=0.5
```

Questo vettore può essere confrontato con il vero Stato degli utenti (matrice di confusione) tramite il comando:

```
table(previsti,verifica$Stato)
```

Infine riportiamo il codice per costruire una curva lift, ottenuto rielaborando la funzione di Azzalini, Scarpa, (2004). Notiamo che, avendo codificato il valore Disattivo con 0, quando normalmente il valore di interesse è codificato con 1, dobbiamo inserire nella funzione il complemento ad uno delle

probabilità di successo previste, per ottenere la probabilità di "insuccesso", cioè di essere un utente Disattivo.

```
library(sm)
g<-as.numeric(verifica$Stato==0)
p<-1-predict(logistico,type="response",newdata=verifica)
ind<-order(p,decreasing=T,na.last=T)
n<-length(g)
x1<-(1:n)/n
x2<-cumsum(g[ind])/(mean(g)*(1:n))
b<-binning(x1,x2,breaks=(-0.001:10/9.999))
x<-c(0,seq(0.05,0.95,by=0.1),1)
plot(x,c(x2[1],b$means,1),type="b",xlim=c(0,1),ylim=c(1,max(x2)),
xlab="frazione di soggetti previsti",
ylab="fattore di miglioramento")
```

A.2. Albero di classificazione

In R esistono due librerie aggiuntive per costruire gli alberi di classificazione, `tree` e `rpart`. Per questa tesi abbiamo utilizzato i comandi contenuti nella prima, con la seguente sintassi:

```
library(tree)
albero<-tree(Stato ~.,data=stima)
```

L'albero così ottenuto è quello che minimizza la devianza per il campione di stima. L'output del modello è in Tabella A.2. Nella prima colonna è riportato l'ordine dei nodi: ai numeri pari corrispondono i rami di sinistra, e ai numeri dispari i rami di destra, e numeri consecutivi come 2 e 3 sono rami provenienti dallo stesso nodo. Nella seconda colonna c'è lo split, l'affermazione logica corrispondente al ramo, dove root significa radice. La terza colonna riporta il numero di individui presenti ad ogni nodo, mentre la quarta contiene la devianza calcolata per ogni nodo. La quinta colonna contiene la probabilità di successo associata, calcolata come frequenza di individui Attivi

rispetto al totale del nodo. L'ultima colonna, infine, contiene un asterisco per i nodi terminali, cioè per le foglie.

nodo)	split	n	devianza	prob.	foglia
1)	root	1988	496.9	0.4945	
2)	altri: N	1077	67.19	0.06685	
4)	fisso: N	1057	49.44	0.0492	*
5)	fisso: S	20	0	1	*
3)	altri: S	911	0	1	*

Tabella A.2. Output dell'albero di classificazione.

Per la potatura è necessario utilizzare una parte del campione di verifica, che chiamiamo "verifica_1", ed il comando è il seguente:

```
potatura<-prune.tree(albero,newdata=verifica_1)
```

La matrice di confusione e la funzione *lift* per l'albero di classificazione si ottengono allo stesso modo di quelli già presentati per il modello logistico.

A.3. Modello ad effetti casuali

Per il programma R esistono due diverse librerie che si possono utilizzare per stimare un modello ad effetti casuali per dati longitudinali, `nlme` e `lme4`. La seconda è un'evoluzione della prima, ma la versione attuale non supporta ancora le strutture di covarianza, perciò nella tesi è stato utilizzato il primo.

Per utilizzare il comando `lme` (*linear mixed effect*) è necessario che la variabile longitudinale sia in una colonna del dataset, del quale le prime T righe appartengono al primo individuo, le righe dalla T+1 alla 2T appartengono al secondo e così via. La colonna Id contiene un codice che identifica il cliente, mentre la colonna tempo ripete i tempi da 1 a T per ogni individuo.

Chiamiamo `long` il dataset così rielaborato, di cui riportiamo un esempio in Tabella A.3.

id	sms	tempo
1	0	1
1	0	2
⋮	⋮	⋮
1	0	17
2	16	1
⋮	⋮	⋮

Il modello si scrive indicando per prima la formula che corrisponde agli effetti fissi, in questo caso gli sms in relazione al tempo. Successivamente, si indicano le variabili per gli effetti casuali, con `random`, e la variabile indicatrice degli individui, che in questo caso è l'Id.

```
library(nlme)
eff<-lme(sms~tempo,random=~tempo|id,data=dati,correlation=corAR1())
```

L'output del modello contiene la stima per gli effetti fissi, che riportiamo in Tabella A.4, e per gli altri parametri, tra cui la matrice di covarianza degli effetti casuali, nella forma che presentiamo in Tabella A.5. In realtà si tratta dello standard error e della correlazione, invece che la varianza e la covarianza, ma la trasformazione è immediata.

La stima, o meglio la previsione, degli effetti casuali per ogni individuo si ottiene con il comando:

```
ranef(eff)
```

Variabile risposta: Sms	Stima	Std.Error	DF	test t	p-value
(Intercept)	10.528	0.743	32959	14.162	0
tempo	-0.269	0.047	32959	-5.727	0

Tabella A.4. Valori stimati per gli effetti fissi del modello ad effetti misti per la variabile Sms, con i rispettivi standard error ed il test di significatività.

Variabile risposta: Sms	StdDev	Corr
(Intercept)	29.292	
Tempo	1.428	-0.652
Residual	21.914	

Tabella A.5. Valori stimati per lo standard error degli effetti fissi e dei residui, nella prima colonna, e correlazione tra gli effetti fissi nella seconda colonna.

A.4. Lisciamento esponenziale

Il lisciamento esponenziale si ottiene con i comandi di serie storiche, che sono compresi nel pacchetto base di R. Per utilizzare tali comandi, però, i dati devono essere trasformati in una serie storica con il comando `ts()`. Il comando agisce sui vettori colonna, per questo abbiamo eseguito la trasposta `t()` dei dati, che invece erano in una riga. Nel comando si devono indicare anche il momento in cui inizia la serie storica, che per noi era novembre 2004, e la frequenza delle osservazioni, che vale 12 quando la serie è mensile:

```
serie_i<-ts(t(sms[i,]),start=c(2004,11),freq=12)
```

Il lisciamento esponenziale si ottiene con il comando `HoltWinters()` ponendo `beta` e `gamma` uguali a zero, perché si riferiscono rispettivamente al trend ed alla stagionalità. L'unico parametro che viene stimato è quindi `alpha`, che corrisponde al δ utilizzato nel testo. La previsione per il tempo suc-

cessivo all'ultimo osservato si ottiene semplicemente con il comando `predict()`.

```
Lisciamento_i<-HoltWinters(serie_i, alpha=NULL, beta=0, gamma=0)
predict(lisciamento_i)
```

Dato che il procedimento deve essere eseguito per ogni utente del campione, lo abbiamo inserito in un ciclo, con il comando:

```
for(i in 1:N)  { ... }
```

Riferimenti bibliografici

Agresti A. (2002), *Categorical data analysis*. Hobocken, NJ, Wiley-interscience.

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

Di Fonzo T., Lisi F. (2005), *Serie storiche economiche*. Roma, Carocci E-ditore.

Diggle P.J. (2002), *Analysis of longitudinal data*. Oxford, Oxford University press.

Fahrmeir L., Tutz G. (2001) *Multivariate Statistical Modelling based on Generalized Linear Models*. New York, Springer.

Fitzmaurice G.M., Laird N.M, Ware J.H. (2004), *Applied Longitudinal Analysis*. Hoboken, New Jersey, John Wiley & Solns.

Iacus S.M, Masarotto G. (2003) *Laboratorio di Statistica con R*. Milano, McGraw-Hill.

Istat, Istituto Nazionale di Statistica (2006), *ICT nelle famiglie e utilizzo degli individui*, Roma. Informazioni reperite sul sito www.istat.it.

Li E., Zhang D., Davidian M. (2004), Conditional estimation for generalized linear models when covariates are subject-specific parameters in a mixed model for longitudinal measurements. *Biometrics* 60, 1-7.

Nath S.V., Behara R.S. (2003), *Customer churn Analysis in the Wireless Industry: A Data Mining Approach*, Proceedings of the 34th meeting of the Decision Sciences Institute.

Pace L., Salvan A. (2001), *Introduzione alla statistica. II Inferenza, verosimiglianza, modelli*. Padova, Cedam.

Singer J.D., Willett J.B. (2003), *Applied Longitudinal Data Analysis*. Oxford, University Press.

Venables W.N., Ripley B.D. (1999), *Modern Applied Statistics with S-PLUS*, Third Edition, New York, Springer-Verlag

Verbeke G., Molenberghs G. (2000), *Linear Mixed Models for Longitudinal Data*. New York, Springer.

Wang C.Y., Wang N., Wang S. (2000), Regression analysis when covariates are regression parameters of a random effects model for observed longitudinal measurements. *Biometrics* 56, 487-495.

Ware J.H. (1985), Linear models for the analysis of longitudinal studies. *The American Statistician* 39, 95-101.

Zelterman D. (1999), *Models for Discrete Data*. Oxford, Oxford University Press.

Software statistico

R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Becker R.A. et al. (2005). *maps: Draw Geographical Maps*. R package version 2.0-27.

Bowman A.W., Azzalini A., Ripley B.D. (2005). *sm: Smoothing methods for nonparametric regression and density estimation*. R package version 2.1-0. http://www.stats.gla.ac.uk/~adrian/sm

Pinheiro J. et al. (2005). *nlme: Linear and nonlinear mixed effects models*. R package version 3.1-65.

Ripley B. (2005). *tree: Classification and regression trees*. R package version 1.0-19.