UNIVERSITA' DEGLI STUDI DI PADOVA
FACOLTA' DI SCIENZE STATISTICHE

CORSO DI LAUREA TRIENNALE
IN
STATISTICA E TECNOLOGIE INFORMATICHE

BUSINESS INTELLIGENCE E ANALISI DELLA DURATA DELLE CHIAMATE DI ASSISTENZA IN WINTECH

RELATORE: Dott. Bruno Scarpa
LAUREANDO: Roberta Mengato

SOMMARIO

0 Introduzione

La tesi da me sviluppata tratta dello stage svolto presso l'azienda di servizi informatici Wintech S.P.A.

Il problema principale indicato da Wintech consisteva nel dover fare una fotografia della sua situazione aziendale sfruttando i dati relativi alle attività che ogni giorno i dipendenti registrano su un software di nome GECO. Ogni attività è collegata ad una commessa e riporta il tempo che il dipendente ha impiegato per svolgerla.

Inoltre l'azienda aveva un secondo problema, ovvero la comprensione di quali fossero i fattori incidenti sulla durata delle assistenze che effettua per i clienti.
Per fare ciò, si è cercato di adattare ai dati un modello di Cox che sfruttasse anche l'informazione del risolutore del problema tramite l'introduzione di un *effetto frailty*, ovvero un modello che tenesse conto della relazione che lega le assistenze fatte da uno stesso impiegato.
Tutti i dati relativi alle assistenze sono presenti all'interno del database del software KAYAKO.

Lo strumento fornito da Wintech per soddisfare le sue richieste è il software di business intelligence QlikView, che permette l'estrazione e la manipolazione di dati provenienti da qualsiasi fonte (database relazionali, fogli Excel ecc).

Nelle fasi di stima del modello vengono riportati gli output del software R utilizzato per l'analisi, tali risultati vengono evidenziati da `questo tipo di font`.

In ultima istanza l'azienda ha chiesto di fornirgli delle linee guida per la combinazione del software statistico R e QlikView, in modo da poter sfruttare le capacità di analisi del primo e quelle grafiche del secondo.

Prima della trattazione degli argomenti di stage, vi è una piccola introduzione ai concetti base delle basi di dati e di QlikView (capitoli 1 e 2).

1 Introduzione ai concetti informatici e di marketing usati

La tecnologia che permette ad una base di dati di essere strutturata in modo tale da gestire efficientemente e con una discreta affidabilità i dati "in linea" prende il nome di OLTP (On Line Transaction Processing). Essa permette di acquisire volumi elevati di operazioni di modifica e aggiunta dati, necessarie per la gestione degli stessi.
Con l'evoluzione dei sistemi informatici, via via sempre con maggiore capacità di calcolo, la tecnologia OLTP è stata affiancata da quella OLAP (On Line Analytical Processing) che permette un'analisi dei dati con strumenti interattivi (appunto, "in linea"), ad esempio, se usato in azienda fornisce gli strumenti per l'analisi delle vendite.

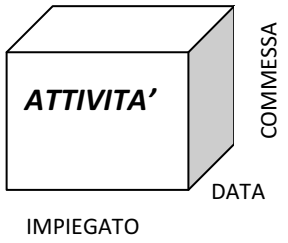
Nelle aziende, la tecnologia OLAP fornisce un supporto alle decisioni e i sistemi OLTP svolgono la funzione di sorgenti di dati per essa.
Mentre i sistemi OLTP sono impiegati in singoli database, le informazioni utilizzate per il supporto alle decisioni in genere vengono raccolte da un data warehouse ovvero un "magazzino di dati" dove convergono dati da più database.
Inoltre, un'altra differenza saliente tra le due tecnologie è la quantità e la tipologia dell'utenza, i sistemi OLTP hanno molti utenti finali, detti "teminalisti", che svolgono semplici operazioni di modifica e aggiunta delle informazioni, invece i sistemi OLAP sono destinati a poche persone che ricoprono la funzione di analisti e preparano le informazioni per la dirigenza.

L'analisi descrittiva che può essere fatta attraverso un data warehouse è l'analisi multidimensionale e consiste in operazioni interattive di aggregazione/disaggregazione lungo opportune "*dimensioni*".
In questo ambito importante è il concetto di **dimensione**, che in informatica si tratta di un dato che categorizza ogni elemento in un set di

dati. Volendo paragonare la dimensione ad un oggetto statistico si può prendere ad esempio una variabile categoriale.
Se le analisi dei dati di un'azienda vengono fatte per aree di competenza (Business Unit) allora la dimensione è proprio quest'ultima.
La funzione di una dimensione è triplice:

- Filtrare
- Raggruppare
- Etichettare

Concettualmente, ad "affiancare" la dimensione vi è il "**fatto**", ovvero un valore o una misurazione che riguarda un'entità o un sistema. Le tabelle dei fatti ("fact tables") contegono tipicamente argomenti numerici, al contrario delle tabelle delle dimensioni che hanno molti attributi descrittivi, tipicamente campi testuali o variabili categoriali.
Una tipica rappresentazione grafica dei fatti è il "*cubo*" dove gli spigoli sono le dimensioni.
Un esempio di cubo proiettato nell'ambito aziendale di stage ha come fatto l'attività svolta dal dipendente e come dimensioni la commessa a cui è collegata, la data in cui è stata svolta e il nome dell'impiegato:



Per la realizzazione di un data warehouse vi sono due vie (diverse dal punto di vista tecnologico):

- OLAP relazionali (ROLAP): i dati vengono memorizzati tramite tabelle e le operazioni di analisi vengono tradotte in istruzioni SQL. I dati di cui si ha bisogno vengono estratti tramite *query* (richieste di informazioni alla base di dati) istantaneamente.

- OLAP multidimensionali (MOLAP): i dati vengono “preaggregati” secondo alcune dimensioni importanti per l’utente e memorizzati in speciali strutture multidimensionali

Quando una base di dati gestisce informazioni commerciali può essere introdotto il termine **CRM** (Customer Relationship Management, tradotto “gestione delle relazioni con i clienti”), che è legato al concetto di fidelizzazione del cliente in un mondo in cui il mercato non è più formato solo dal cliente ma anche dall’ambiente circostante. Il CRM ricopre l'insieme delle funzioni dell'impresa che mirano a conquistare e conservare la propria clientela effettuando operazioni commerciali ad hoc diverse da cliente a cliente.

2 QlikView

Qlikview (della QlikTech) è un programma di *business intelligence* che usa un sistema di analisi innovativo rispetto ai tradizionali sistemi OLAP.
La business intelligence è l'insieme dei processi che servono per raccogliere e analizzare le informazioni sul business aziendale

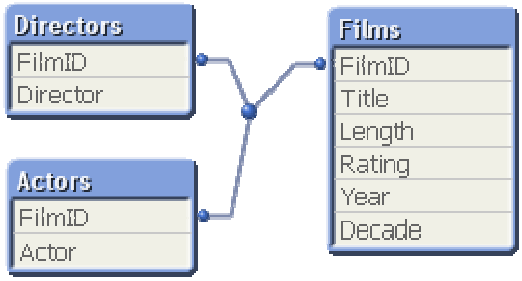
A differenza delle tradizionali analisi OLAP, con QlikView non sono necessarie definizioni a priori di dimensioni di aggregazioni e di selezioni di dati o strutture gerarchiche, inoltre grazie all'approccio visuale, anche gli utenti meno esperti possono facilmente utilizzare l'applicazione, con una notevole riduzione di tempi e costi di implementazione ed apprendimento.
Il software offre la flessibilità di un ROLAP (dati aggregati al momento secondo i bisogni) e la velocità di un MOLAP (accesso veloce ai dataset di aggregazione), ciò è permesso grazie al fatto che il programma lavora unicamente usando la RAM (tecnologia "in-memory") creandosi una sorta di "database virtuale".
Gli strumenti offerti da un sistema MOLAP si basano su un motore multidimensionale con query relazionali a richiesta, mentre Qlik View offre un motore relazionale fornendo la possibilità di creare cubi "on-demand".

I cubi dei sistemi OLAP hanno alcuni difetti a cui QlikView permette di ovviare:

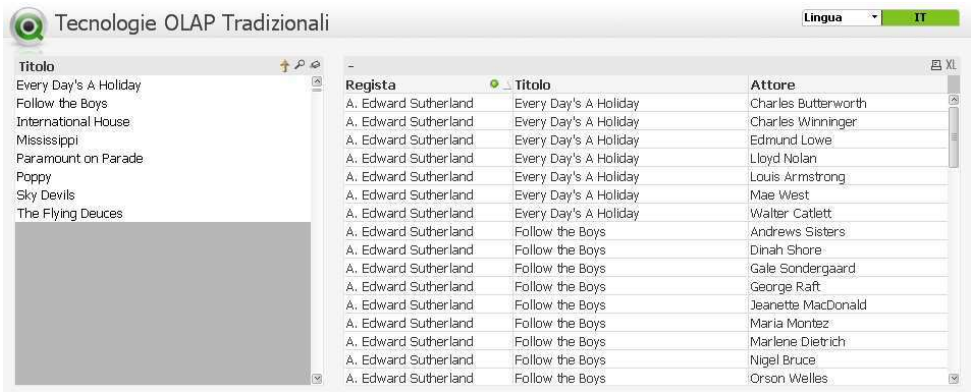
- La loro creazione è un processo complicato che richiede molto tempo e alte competenze
- Non è possibile avere una fotografia dell'azienda al momento, in quanto i cubi in genere vengono aggiornati durante il fine settimana o la sera.

Il programma riesce a lavorare velocemente anche grazie ad un formato particolare, .qvd, che permette di trasformare delle tabelle enormi in file binari discretamente piccoli: una tabella che nel classico formato testo (.csv) ha una dimensione di 110 KB, scende a 58 KB, poco più della metà.

Le tabelle vengono associate tra di loro attraverso i campi chiave che devono semplicemente avere lo stesso nome:

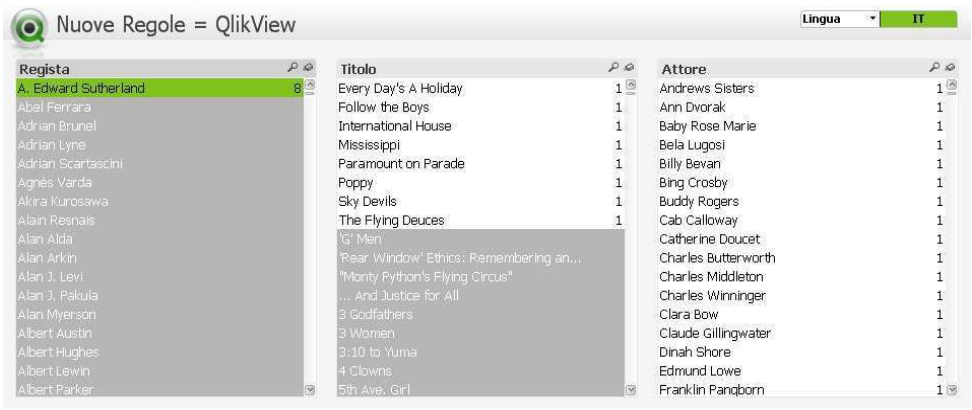


Per capire meglio come si differenzia QlikView dai sistemi tradizionali viene riportato un semplice esempio di una base di dati sui film e loro relative informazioni descrittive; se volessimo avere informazioni sugli attori dei film di un certo regista dopo diverse operazioni di *join* il risultato che si otterrebbe sarebbe:



Come si nota, ci sono numerose ridondanze, ad esempio il nome del regista è ripetuto tante volte quanti sono gli attori dei suoi film (lo stesso vale per gli stessi titoli).

Nel software in studio, selezionando il regista il risultato invece è il seguente:



Qlik View è in grado di estrapolare i dati da diversi formati di lavoro (Excel, Access, XML, qvd ecc) riuscendo a metterli tutti in relazione tra di loro.

3 La Wintech

3.1 Lo schema della gestione dati per la Wintech

Quando un'azienda entra in contatto con Wintech per avere informazioni sui campi di cui essa si occupa diventa automaticamente un "cliente potenziale" e i suoi dati vengono immagazzinati in un database gestito dal software "TDB" (acronimo di "Tabloid De Board"). TDB è un software di CRM operativo documentale ed estende le funzionalità di "Lotus Notes" (di IBM, programma per la gestione della posta elettronica). Quindi, oltre a tenere un'anagrafica di tutti i clienti, TDB registra tutti i contatti avvenuti con essi tramite e-mail (compresi gli allegati).

Se un cliente potenziale, oltre a richiedere informazioni sui prodotti, si dimostra effettivamente interessato alle offerte di Wintech, viene classificato come "*opportunità commerciale*" e su di essa vengono svolte delle operazioni commerciali, registrate come "*attività*", le quali si traducono in investimenti. Le informazioni sulle attività svolte, come i nomi dei commerciali che le seguono e i referenti delle opportunità vengono sempre registrate e consuntivate in TDB.

Consuntivare significa quantificare l'attività che si sta svolgendo, ovvero dargli un valore. In questo caso il valore è espresso in ore.

La chiusura di un'opportunità commerciale può avvenire in due modi:

- Opportunità commerciale persa = il potenziale cliente non si dimostra più interessato all'offerta di Wintech e decide di non stipulare alcun contratto.
- Opportunità commerciale vinta = l'azienda interessata ai prodotti di Wintech decide di aprire un rapporto commerciale con essa.

Nel secondo caso viene aperta una "*commessa*", in cui l'azienda diventa un vero cliente ed effettua un ordine.

Il database gestionale che immagazzina le informazioni relative a tutto ciò che viene fatturato è “eSolver” e viene aggiornato una volta al mese. Quando un’attività è in corso e non è ancora stata fatturata, viene tenuto traccia di tutto ciò che la riguarda (attività, commesse, commerciali che la seguono . . .) nella base di dati gestita dal software “GECO”. L’acronimo GECO sta per “GEstione e COnsuntivazione”, infatti gestisce ciò che è pianificato, ma non ancora fatturato, e lo consuntiva (oltre che tenere memoria del passato che è stato fatturato).

L’azienda, oltre a vendere servizi, ai suoi clienti offre anche contratti di assistenza la cui gestione ruota attorno al concetto di “*ticket*”. Ogni qualvolta un cliente necessita del supporto degli operatori viene “aperto un ticket”, che visivamente lo si può immaginare come un “cartellino” con scritto un identificativo, chi lo ha richiesto, il problema e chi lo può risolvere

Il ticket può essere aperto tramite contatto telefonico o telematicamente e a seconda del problema è risolto dalla business unit “Sviluppo Software” o “Assistenza tecnica”. La base di dati che tiene traccia di ogni ticket è gestita dal software “KAYAKO”.

GECO, KAYAKO, TDB e eSolver comunicano tra di loro scambianosi informazioni.

3.2 L’obbiettivo di stage e la nostra soluzione

L’obbiettivo dello stage consisteva nella costruzione di una dashboard utile per effettuare una fotografia della situazione di Wintech giornalmente tenendo presente che il fulcro attorno a cui deve ruotare il lavoro è composto dai concetti di attività e ticket descritti nel paragrafo precedente. Tutte le informazioni relative ad essi sono contenute all’interno dei database di GECO e KAYAKO.

Ogni attività è identificata da un ID ed è collegata ad una commessa che a sua volta può avere più attività collegate.

Per identificare univocamente una commessa bisogna utilizzare il suo codice preceduto dalla sigla del gruppo competente (la Wintech è costituita da 3 gruppi: Wintech, Albasoft, Format).

Dato che non esistono tipi di commessa prestabiliti in un elenco (come invece succede per le attività), si può estrapolare una classificazione dalla sigla iniziale del codice commessa:

- C_ = contratto
- S_ = spot (servizio effettuato e saldato)
- T_= prevendita
- O_= ordine
- W_= commessa interna
- G_ = in garanzia
- L_ = consuntivo

La tabella delle attività è collegata ad altre che contengono informazioni su ogni record: il tipo di attività, il cliente per cui è stata svolta (anche i cliente passati), lo stato della fatturazione, da chi è stata svolta e il rispettivo gruppo, se è stata notificata per e-mail o meno.
Lo stato di fatturazione è importante perché ci permette di vedere come va l'azienda per il mese in corso, dato che le attività vengono fatturate a fine mese.
La quantificazione di quanto un'attività ha richiesto è effettuata in ore.

Collegate alle tabelle di GECO ce ne sono anche alcune di KAYAKO, che riguardano le assistenze fatte ai clienti dalla business unit "Assistenza Tecnica". I dati disponibili per i ticket riguardano solo il gruppo Wintech e non Format o Albasoft.

Durante l'attività di esplorazione dei dati è stato riscontrato che la business unit "Sviluppo Software" usa KAYAKO per l'operazione di chiusura dei ticket, mentre l'assistenza tecnica registra tutte le attività svolte su un ticket e la sua conclusione su GECO.

La nostra proposta per soddisfare la richiesta di Wintech consiste in 3 dashboard sviluppate con QlikView:

- Una per monitorare i ticket dell'assistenza tecnica
- Una per monitorare i ticket dello sviluppo software
- Una per la gestione delle attività e delle commessa

3.2.1 Applicazione per monitoraggio ticket

Il file consiste in 4 fogli, dove è possibile ricercare i ticket per cliente, impiegato e id. E' stato fatto anche un collegamento alla pagina web di KAYAKO che permette di vedere tutta la storia del ticket direttamente dall'applicazione QlikView.

E' stata aggiunta un'informazione riassuntiva rispetto a quella fornita da KAYAKO, dato che lo stato del ticket può presentare diversi stadi (aperto, chiuso, intervento . . .) , circa una ventina, per comodità si sono create due semplici categorie: "ticket_aperto" e "ticket_chiuso".

Foglio n° 1. Visione d'insieme dell'andamento della durata di un ticket

a. Funzione di ripartizione delle durate
b. Funzione di densità delle durate
c. Specchietto con statistiche riassuntive su n°ticket aperti/chiusi e loro durate relativo al mese in corso

Foglio n° 2. Contiene le informazioni sui ticket aperti/chiusi e loro durate secondo la dimensione "cliente"

Foglio n° 3. Informazioni dettagliate del cliente selezionato

Foglio n° 4. Andamento storico per mese e anno dei ticket aperti/chiusi e le durate

Foglio n° 5. Informazioni sulle attività svolte per risolvere i ticket e gli impiegati che se ne sono occupati (questa pagina è presente solo per il file dell'assistenza tecnica perché per quello che riguarda le altre B.U. non vi è la possibilità di distinguere gli impiegati che lavorano ad un ticket).

Foglio n° 6. Trasferta

In coda ai fogli descrittivi ne viene aggiunto un altro in cui vengono segnalati i casi in cui la data di apertura è più piccola della data di chiusura, la motivazione può essere che è stata svolta qualche attività prima dell'apertura del ticket.

Altre anomalie possono esserci nei conteggi delle ore in quanto ci sono dei casi in cui le ore di lavoro delle attività sono a zero, le motivazioni sono 2:

1. Quando si decide di non far pagare il cliente per il servizio svolto, quindi si hanno attività non fatturabili
2. Attività relative a reperibilità: a turni della durata di una settimana, ogni tecnico fa una settimana di reperibilità 24 ore su 24 e il cliente viene valorizzato a Wintech (se il tecnico viene effettivamente chiamato e svolge l'intervento registra una nuova attività il cui il cliente è chi lo ha chiamato).

3.2.2 Applicazione per gestione commesse

La struttura del file che monitorizza le attività e le commesse è la seguente:

Foglio n° 1. Abbiamo la percentuale di commesse per

a. Business Unit

b. Tipo commessa

   (Si possono avere le due informazioni congiunte, percentuale di tipi commessa per B.U., semplicemente cliccando un pulsante)

Foglio n° 2. Tabella informativa con riferimenti temporali e delle B.U. del n° di ore che devono ancora essere fatturate. Cliccando su un pulsante è possibile andare in una pagina che visualizza anche le ore già fatturate.

Foglio n° 3. Pagina dedicata ai contratti a forfait, in quanto per questi si hanno a disposizione :

a. Informazione iniziale sul n° di ore previste per il loro svolgimento
b. Informazione aggiornata sul n° di ore che sono volute/ci stanno volendo per svolgere il contratto

Foglio n° 4. Pagina dedicata alle commesse in garanzia

a. Grafico contenente il n° di attività svolte per commessa
b. Tabella contenente i dettagli temporali e dello stato della commesa

Foglio n° 5. Informazioni clienti attivi e n° di clienti per:

a. Provincia
b. Regione
c. Paese

   E' stato inserito anche un collegamento a *google maps* per poter visualizzare l'indirizzo del cliente selezionato.

Foglio n° 6. Informazioni clienti non più attivi

Foglio n° 7. In questo foglio si ha una visione d'insieme del numero di clienti per provincia attraverso una cartina con degli spot che hanno colore diverso a seconda della quantità di clienti.

Foglio n° 8. Foglio dedicato alle informazioni dei clienti non attivi, viene riportato il loro stato precedente (cliente, cliente potenziale, fornitore ecc.)

Foglio n° 9. Foglio con tutte le informazioni dei dipendenti

4 Combinazione di R e QlikView

Un'ulteriore richiesta di Wintech è stata quella di combinare le conoscenze universitarie di R con QlikView in modo da poter sfruttare la capacità del primo programma di effettuare previsioni e del secondo di rappresentazione grafica dinamica ed esteticamente gradevole.
Purtroppo non ci sono stati forniti i dati del fatturato aziendale così abbiamo concordato di fare un'applicazione dimostrativa attraverso dei dati didattici utilizzati durante il corso di "Analisi delle Serie Temporali" (anno accademico 2009/2010).

QlikView offre la possibilità di essere interfacciato con altri programmi eseguibili da riga di comando. Ai nostri fini è possibile creare un legame con R attraverso la funzionalità *Rscript* che permette di lanciare il programma direttamente dal prompt dei comandi.

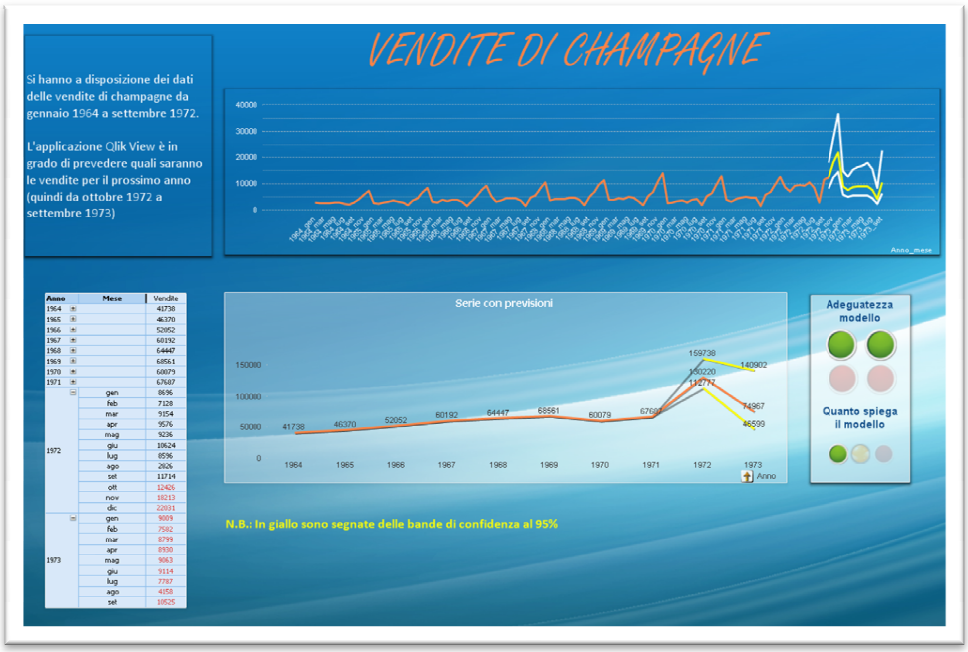
Ciò avviene inserendo nello script QlikView il seguente codice:

```
EXECUTE cmd.exe /C "rscript previsioniR.R";
```

dove `previsioniR.R` è il classico script di R già preconfezionato.

La combinazione con R può essere utile per la rappresentazione di previsioni sul fatturato che utilizzano tutta la teoria sulle serie storiche. Si sottolinea il fatto che l'analisi della serie deve essere fatta manualmente attraverso R e poi ciò che ne risulta può essere rappresentato tramite QlikView.

Ad esempio, applicando un modello ARIMA (Di Fonzo e Lisi, 2001) ad una serie trimestrale di vendite, che va dal 1976 al 2005, si può avere una visione d'insieme tramite QlikView:



L'applicazione è stata creata con l'idea che i dati della serie vengano presi da QlikView, dal suo script è possibile proiettare i dati in un file di testo, che sarà letto da R per poter effettuare le analisi.
Una volta svolta l'analisi dei dati, R scrive un file di testo (csv) che contiene le previsioni. QlikView legge la tabella con le previsioni e le incorpora all'interno dell'applicazione.

Per creare un programma dimostrativo per Wintech è stata implementata una procedura semiautomatica per la modellistica di serie temporali dove viene effettuata una scelta tra un modello ARIMA ed uno a lisciamento esponenziale sulla base dei test di Ljung-Box e della variabilità spiegata.
Oltre alla previsione, all'utilizzatore finale viene fornito un intervallo di confidenza al 95% calcolato sulla base di residui che possono essere normali e non. Per poter fornire un intervallo di confidenza anche quando i residui non si distribuiscono normalmente si è utilizzata una procedura di ricampionamento "*bootstrap*" (Efron e Tibshirani 1994).

Tale procedura consiste in:

1. Estrarre un campione con reinserimento tra i residui di numerosità pari alla numerosità del campione originale (nell'ambito delle serie storiche può essere la frequenza della serie).
2. Effettuare delle previsioni attraverso il modello ARIMA stimato sommando i "nuovi" residui appena estratti.
3. Ripetere i punti 1. e 2. 10000 volte tenendo in memoria (in una matrice) tutti i valori ottenuti.
   Si otterrà una matrice che ha numero di righe pari al numero di osservazioni di cui fornire la previsione (ad esempio la frequenza della serie) e 10000 colonne.
4. Calcolare i quantili desiderati (ad esempio al 2.5% e al 97.5%) da ogni riga per ciascuna osservazione.

La procedura è *semi*automatica perché alcune operazioni si devono svolgere manualmente:

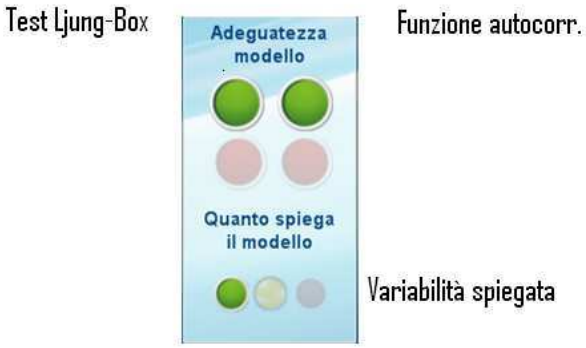
- Scelta di una trasformata, come il logaritmo, per ridurre la variabilità della serie
- Scelta del numero di differenziazioni stagionali e non

E' stata creata una piccola guida che dovrebbe servire ad insegnare ad usare l'applicazione a chi conosce poco di statistica.

Per l'utilizzatore finale della dashboard contenente le previsioni, sono stati creati 3 segnali visivi, sottoforma di semafori, che hanno il compito di allarmarlo nel caso venga a mancare una buona capacità predittiva del modello.
Tali allarmi si attivano quando:

- vi è un test di Ljung-Box (e. g. Di Fonzo e Lisi, 2001) per l'incorrelazione dei residui che risulta significativo al 5% contro l'ipotesi nulla.

- la funzione di autocorrelazione a diversi *lag* esce da intervalli di confidenza costruiti sulla base della funzione di autocorrelazione di una “random walk”
- la variabilità spiegata dal modello è troppo bassa (il semaforo diventa arancione quando è compresa tra l’80% e il 50%, e rosso quando è minore del 50%)



5 Analisi della durata di un ticket

5.1 Introduzione

Per effettuare valutazioni sul trattamento delle chiamate di assistenza dei clienti, la Wintech è interessata a sapere quali sono i fattori che influenzano la durata di un ticket, utilizzando i dati a disposizione su KAYAKO. Si entra quindi nel campo dell'analisi dei dati di durata, in quanto sono presenti alcuni dati censurati a destra (ticket non ancora chiusi).

Rilevante in questo ambito è la conoscenza della funzione di sopravvivenza ad un istante t, *S(t)*[1], che fornisce la probabilità che un ticket rimanga aperto oltre il tempo t, e la funzione di rischio h(t)[2] che dà delle indicazioni sulla probabilità che un ticket non venga chiuso da t a t+Δt sapendo che è stato aperto fino a t. Si sottolinea il fatto che la funzione di rischio non è una probabilità.

Per l'analisi viene qui proposto il modello a rischi proporzionali di Cox, che non necessita di assunti parametrici (se non per la forma del regressore), e la sua estensione con effetti casuali per verificare se vi è differenza di durata del ticket tra quelli svolti da un determinato dipendente.
Sono stati anche provati dei modelli parametrici, ma per essi non viene soddisfatto l'assunto distributivo.

Il modello di Cox modella la funzione di rischio al tempo t, infatti considerando il tempo all'evento τ una variabile casuale e t una sua realizzazione, la sua formulazione è:

$$h(t|Z) = h_0(t)e^{(\beta^t Z)} = h_0(t)e^{(\sum_{k=1}^{p}\beta_k Z_k)} \quad (0)$$

con:

$Z = vettore\ delle\ covariate$

$\beta = (\beta_1, \dots, \beta_p)^t = vettore\ dei\ parametri$

[1] Per la definizione di funzione di sopravvivenza si veda paragrafo 5.1.1
[2] Per la definizione di funzione di rischio si veda paragrafo 5.1.2

$h(t|Z)$ = *funzione di rischio al tempo t per un'osservazione con vettore delle covariate Z*

$h_0(t)$ = *funzione di rischio di base = funzione di rischio per un individuo che ha tutte le covariate a 0*

La caratteristica e il vantaggio del modello di Cox è che non si deve fare alcuna assunzione parametrica sul rischio di base in quanto nella formulazione di Cox la funzione di verosimiglianza per il modello descritto è vista come il prodotto di 2 verosimiglianze:

$$L(t) = L_1(t) * L_2(t)$$

ed è stato dimostrato che per fare inferenza su β si può utilizzare solo $L_2(t)$ (D. R. Cox, 1972), detta "verosimiglianza parziale", la quale non dipende dal rischio di base ma solo da β.

L_2 gode delle stesse proprietà della verosimiglianza completa e le stime di massima verosimiglianza che si ottengono sono consistenti e asintoticamente normali.

Siano $t_1 < \cdots < t_m$ gli istanti di chiusura del ticket ordinati, allora:

$$L_2(t) = \prod_{i=1}^{m} \frac{e^{\sum_{k=1}^{p} \beta_k z_{(i)k}}}{\sum_{v \in R(t_i)} e^{\sum_{k=1}^{p} \beta_k z_{jk}}}$$

con

$m = n°$ *totale di ticket chiusi*

$z_{(i)k} =$ *k-esima covariata associata all'osservazione con tempo di chiusura del ticket* t_i

$R(t_i) =$ *classe di individui a rischio nell'istante* t_i

$p = n°$ *di parametri*

5.1.1 La funzione di sopravvivenza

Sia τ una variabile casuale che descrive le durate relative ai ticket e t una sua realizzazione, la funzione di sopravvivenza S(t) non è altro che la probabilità di sopravvivere (= che il ticket rimanga aperto) oltre il tempo t, in formule:

$$S(t) = \int_t^{+\infty} f(u)du = \Pr(T > t)$$

$S(0) = 1$ $\lim_{t\to+\infty} S(t) = 0$

Quando non si hanno dati censurati, uno stimatore corretto, consistente e asintoticamente normale di S(t) è $\hat{S}(t) = 1 - \hat{F}(t)$ dove $\hat{F}(t)$ è la funzione di ripartizione empirica di τ e $\hat{S}(t)$ è detta *funzione di sopravvivenza empirica*.

Nel caso in cui si abbiano dati censurati, siano *m* gli istanti distinti in cui vengono chiusi i ticket, indicati con $\tau_1 < \cdots < \tau_m$.
Siano m_i il numero di ticket chiusi in τ_i e r_i i ticket "a rischio" nello stesso istante. Per ticket a rischio si intende quei ticket che non sono né chiusi né censurati prima di τ_i.
Considerando τ come una variabile casuale discreta con supporto l'insieme $\{\tau_1, \dots, \tau_m\}$ si ha che uno stimatore fortemente consistente e asintoticamente normale è rappresentato dallo *stimatore di Kaplan-Meier* (o del "prodotto limite") (Klein e Moeschberger, 1997),

$$\hat{S}_{KM}(t) = \prod_{T_i \le t} \left(1 - \frac{m_i}{r_i}\right)$$

$\hat{S}_{KM}(t) \xrightarrow{q.c.} S(t)$
$\sqrt{n}\left(\hat{S}_{KM}(t) - S(t)\right) \xrightarrow{d} N(0, \sigma^2)$

Durante l'analisi verrà utilizzato questo stimatore.

5.1.2 La funzione di rischio

La funzione di rischio riguarda (non è) la probabilità condizionata che un ticket che è stato aperto fino al tempo t, venga chiuso nel periodo che va da t a t + Δt.

Sia τ una variabile casuale e t una sua realizzazione, allora si definisce funzione di rischio il limite

$$h(t) = \lim_{\Delta t \to 0^+} \frac{\Pr(t < \tau \leq t + \Delta t \mid \tau > t)}{\Delta t}$$

e la probabilità descritta nelle prime righe è $h(t) * \Delta t$.

La densità, la funzione di rischio e la funzione di sopravvivenza sono legate tra di loro perché:

$$h(t) = \lim_{\Delta t \to 0^+} \frac{\Pr(t < T \leq t + \Delta t \mid T > t)}{\Delta t} =$$

$$= \lim_{\Delta t \to 0^+} \frac{\Pr(t < T \leq t + \Delta t)}{\Pr(T > t)\,\Delta t} = \lim_{\Delta t \to 0^+} \frac{\Pr\,(t < T \leq t + \Delta t)}{\mathrm{S(t)}\Delta t}$$

$$= \lim_{\Delta t \to 0^+} \frac{\mathrm{F(t + \Delta t)} - F(t)}{\mathrm{S(t)}\Delta t} = \frac{f(t)}{S(t)}$$

$$==> \quad h(t) = \frac{f(t)}{S(t)} \quad (1)$$

Inoltre

$$S(t) = 1 - F(t)$$

$$S'^{(t)} = -F'(t)$$

$$S'^{(t)} = -f(t) \quad (2)$$

Sostituendo la (2) in (1) si ottiene:

$$h(t) = -\frac{S'^{(t)}}{S(t)} = -\frac{d}{dt}\log\big(S(t)\big) \quad (3)$$

5.1.3 La funzione di rischio cumulato

Sia τ una variabile casuale continua e t una sua realizzazione, allora si definisce funzione di rischio cumulato

$$H(t) = \int_0^t h(u)du$$

Unendo la definizione con il risultato (3) si ottiene:

$$H(t) = \int_0^t h(u)du \; = \int_0^t -\frac{d}{du}\log\big(S(u)\big)\,du \; = \big[\log\big(S(u)\big)\big]_0^t$$

$$-H(t) = \log\big(S(t)\big) - \log\big(S(0)\big)$$

$$S(t) = e^{-H(t)}$$

5.2 Il dataset

Il dataset a disposizione è composto da 895 osservazioni di durate di ticket di cui si sa:

- il cliente che lo ha richiesto
- il mese e il giorno dell'apertura (non consideriamo l'anno perché abbiamo a disposizione solo dati del 2011)
- il n° di attività svolte dai dipendenti per la risoluzione
- tipo di contatto del cliente
    - 0 = richiesta via mail
    - 1 = richiesta via telefonica
- lo stato del problema dove:
    - 0 = problema ancora non risolto = ticket aperto = *dato censurato*
    - 1 = problema risolto = ticket chiuso = *dato completo*
- Se sono state effettuate trasferte

Ticketid	cliente	NOME_IMPIEGATO	MESE_APERTURA	N_ATTIVITA
9564	aaaaaa	-	4	2
9567	ClienteB	Mario	4	1
9568	cccccc	-	4	2
9569	dddddd	Paolo	4	1
9571	ClienteB	Stefano	4	1
9572	ClienteB	Francesco	4	1

STATO	GIORNO_APERTURA	CHIAMATA	DURATA	CLIENTE	TRASFERTA
1	6	0	12.975	altri	no
1	6	0	3.029	ClienteB	no
0	6	1	2.163	altri	si
0	6	0	0.140	altri	si
1	6	1	0.972	ClienteB	si
1	6	1	0.887	ClienteB	no

(I nomi dei tecnici e dei clienti sono stati sostituiti con nomi fittizi)

5.3 Analisi esplorative

Dalle analisi esplorative risulta principalmente che sono due i clienti che richiedono la gran parte de ticket, ClienteT (24%) e ClienteB (50%) (i nomi dei clienti sono fittizi per motivi di privacy), i primi vengono risolti da una ditta esterna, mentre i secondi vengono risolti da Wintech.

Per non dover stimare eccessivi parametri categorizziamo i clienti in 3 classi:

- ClienteB
- ClienteT
- Tutti gli altri

La variabile che conta il numero di attività svolte non ha molto significato perché è una diretta conseguenza del tempo che passa (ogni giorno i dipendenti segnano le loro attività su GECO, ovviamente più il ticket rimane aperto più attività vengono svolte).

Per ogni variabile categoriale si hanno (numero di osservazioni):

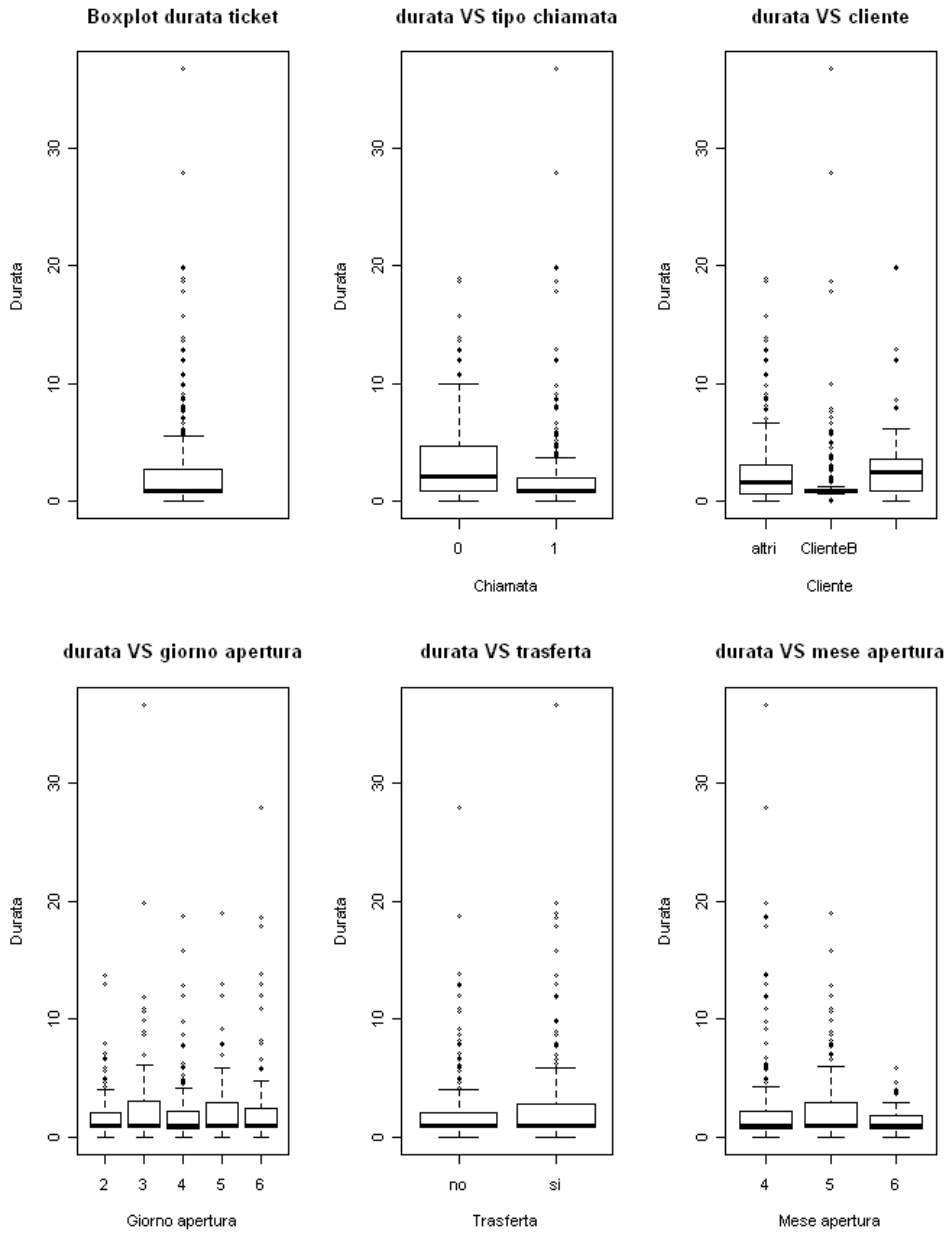
CLIENTE		**%**
ClienteB	450	50
ClienteT	216	24
Altri	229	26

MESE APERTURA		**%**
4 (Aprile)	365	41
5 (Maggio)	431	48
6 (Giugno)	99	11

CHIAMATA		**%**
0	161	18
1	734	82

GIORNO APERTURA		%
2	161	18.0
3	211	23.6
4	185	20.7
5	190	21.2
6	148	16.5

Vengono riportati i *boxplot* della durata di un ticket per evidenziarne eventuali asimmetrie nella distribuzione.

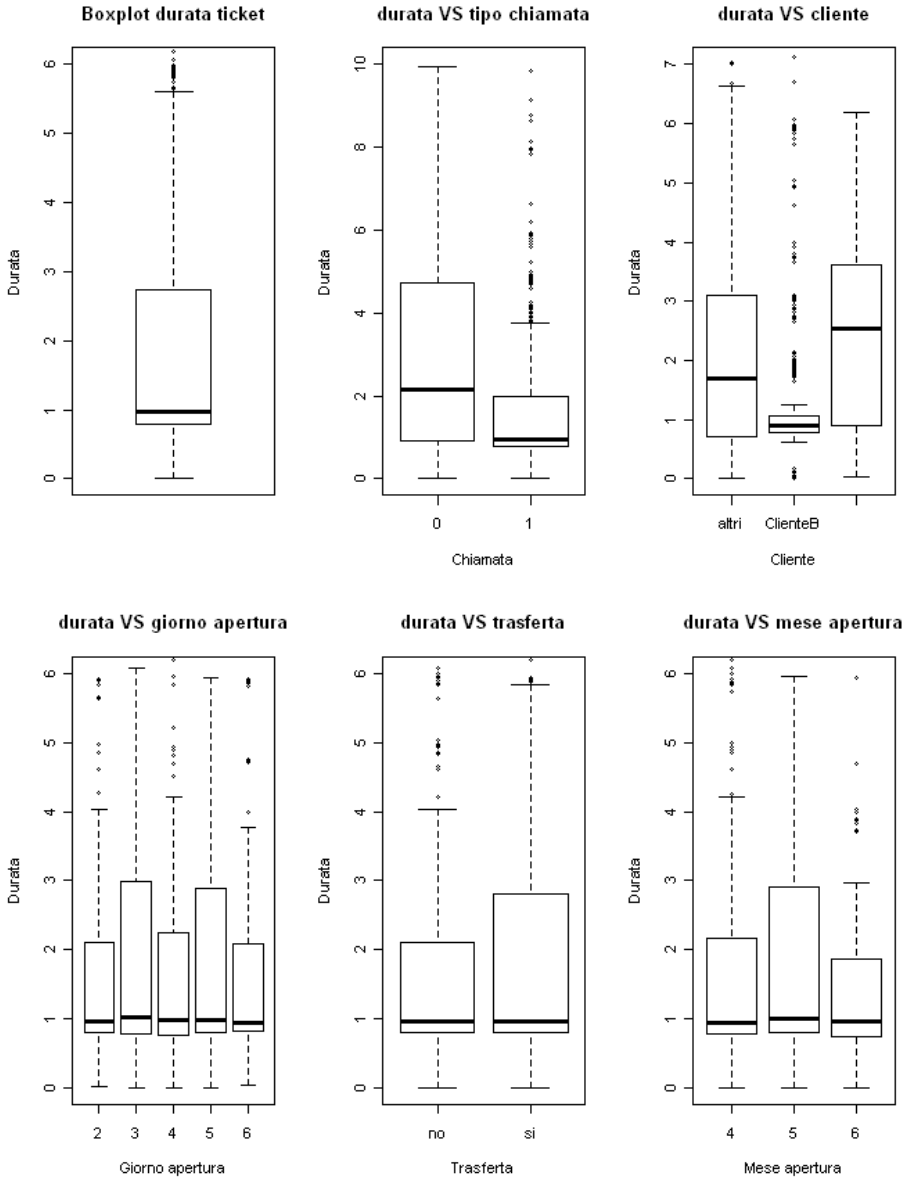


Come si vede dai grafici, la distribuzione della durata di un ticket è asimmetrica e risaltano due osservazioni, la prima corrisponde al ticket

9999 e la seconda al 9691. E' stato appurato con l'aiuto dei tecnici che queste due assistenze riguardano casi eccezionali in cui ci si occupa di un prodotto nuovo per Wintech ed una stampante molto grande e costosa, che si guasta raramente.

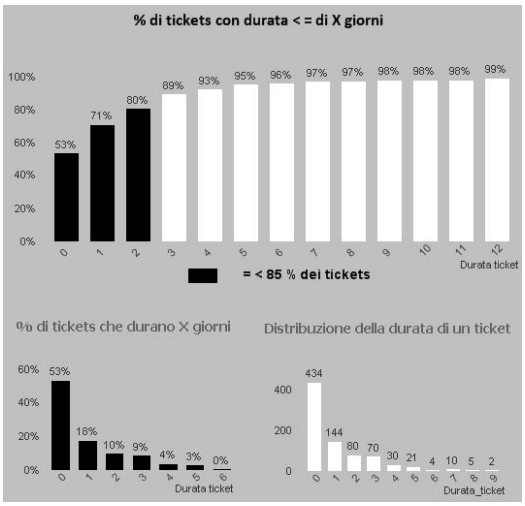
Per fare un'analisi il più corretta possibile, si decide di eliminare tali osservazioni trattandosi di casi rari e fuori dalle assistenze ordinarie.

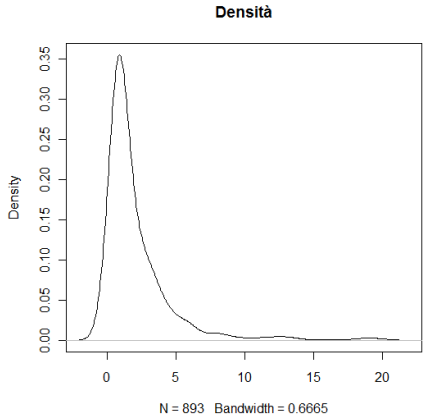
I boxplot della durata rispetto alle variabili in considerazione, ridisegnati in maniera tale da evidenziare la forma della distribuzione, sono:



Le variabili in cui si nota una sostanziale differenza in distribuzione sono la chiamata e il cliente .

I risultati descrittivi riportati anche nell'applicazione QlikView sono (considerando come unità temporale il giorno):





Data l'asimmetria della distribuzione, come indicazione sull'andamento dei ticket si è deciso di fornire all'azienda la mediana piuttosto che la media.

Le statistiche di sintesi principali valgono:

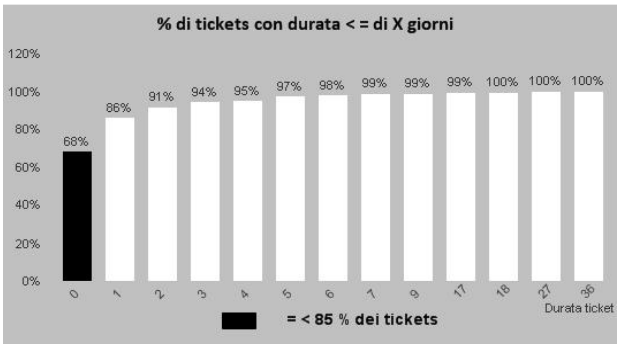
statistica	Valore
Minimo	0.002
1° quartile	0.791
Mediana	0.967
3° quartile	2.722
Massimo	19.890
Media	2.073

Al 15 giugno 2011 i dati censurati sono 77 su 895.

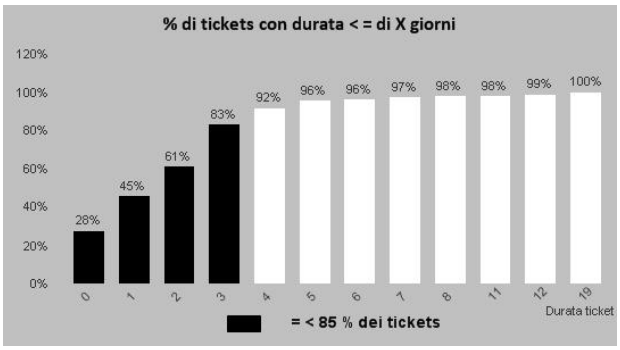
Le informazioni a nostra disposizione ci dicono che la durata mediana di un ticket per i due clienti principali è:

- Per ClienteB: poco meno di 1 giorno (0.902)
- Per ClienteT: circa 2 giorni e mezzo (2.544)

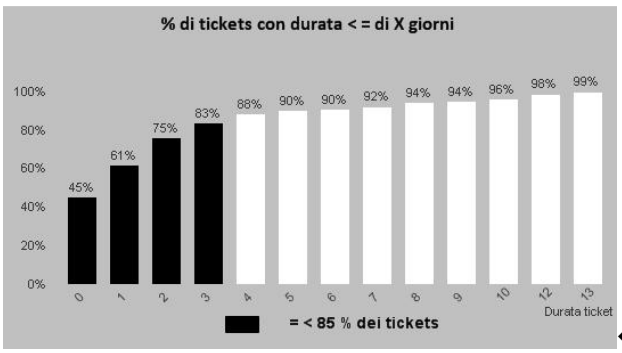
Più nel dettaglio, i dati dicono che per ClienteB l'85% dei ticket viene chiuso in meno di un giorno, mentre per ClienteT e gli altri clienti la stessa percentuale la si raggiunge in circa 3 giorni.



← ClienteB



← ClienteT



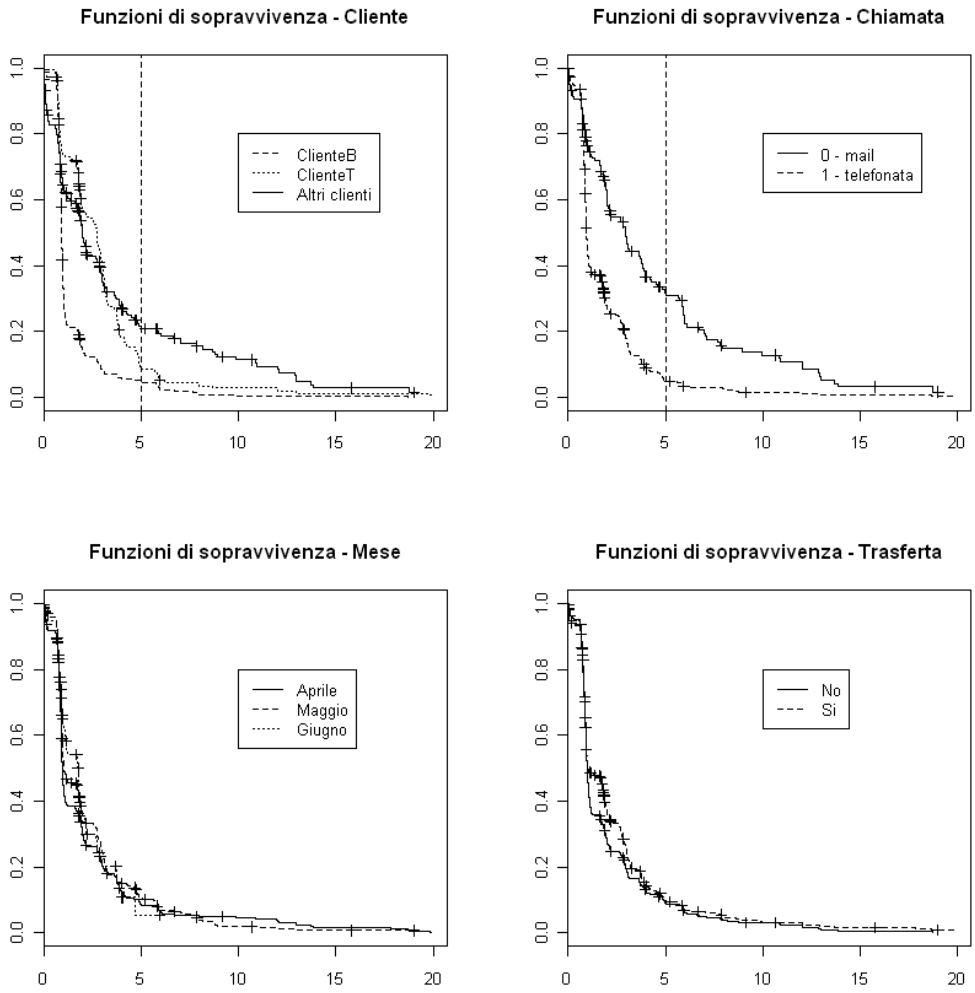
← tutti gli altri clienti

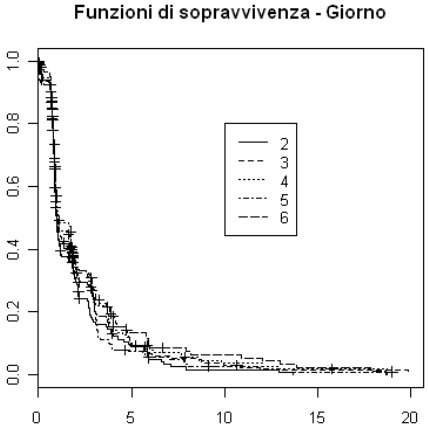
Per avere informazioni sulla probabilità che un ticket venga chiuso oltre un determinato tempo t si entra nell'ambito dell'analisi dei dati di durata

e si definisce che ogni osservazione sia rappresentata dalla tripletta $(\tau_j, d_j, Z_j(t))$ con j = 1, … ,n (n = numero di osservazioni) con:

- τ_j durata della j-esima osservazione
- d_j stato di censura della j-esima osservazione
- $Z_j(t) = (Z_{j1}(t), \dots, Z_{jp}(t))^t$ vettore delle covariate della j-esima osservazione (p = n° di parametri)

Utilizzando lo stimatore di Kaplan-Meier per la funzione di sopravvivenza (vedi paragrafo 5.1.1), che ci permette di avere la probabilità di un ticket di non essere chiuso oltre il tempo t per ogni variabile, si ottengono le curve:





Considerando il grafico relativo ai diversi clienti si vede che la probabilità che un ticket rimanga aperto più di 5 giorni è di circa

- del 5% per ClienteB
- del 10% per ClienteT
- poco più del 20% per gli altri clienti

Verifichiamo l'ipotesi nulla di omogeneità delle funzioni di sopravvivenza tra i diversi clienti per vedere se la probabilità di chiusura varia per diversi clienti.

$$H_0: \begin{cases} S_{ClienteB}(t) = S_{ClienteT}(t) \\ S_{ClienteT}(t) = S_{altri}(t) \end{cases}$$

contro l'ipotesi alternativa

H_1: $Almeno\ una\ funzione\ di\ sopravvivenza\ diversa\ dalle\ altre$

per mezzo del *test "log-rank"* (Harrington and Fleming, 1982), definito come: siano gli istanti (ordinati) di chiusura dei ticket $\tau_{(1)}, \tau_{(2)}, \dots, \tau_{(i)} \dots, \tau_{(k)}$.

$$Q_M = \sum_{j=0}^{g} \frac{O_j - A_j}{A_j}$$

con $j = 0, \dots, g$

dove:

$g + 1 = n° \ di \ popolazioni \ => 3 => g = 2$

$j = 0$ *Prima popolazione = ClienteB*

$j = 1$ *Seconda popolazione = ClienteT*

$j = 2$ *Terza popolazione = altri clienti che non sono ClienteB e ClienteT*

$m_{ji} = n°$ *di ticket chiusi al tempo* τ_i *tra i soggetti della popolazione j-esima*

r_{ji} = $n°$ *di ticket a rischio di chiusura al tempo* τ_i *tra i soggetti della popolazione j-esima*

$$m_j = m_{0i} + m_{1i} + m_{2i}$$

$$r_j = r_{0i} + r_{1i} + r_{2i}$$

$$d_{ji} = m_i * \frac{r_{ji}}{r_i} = n° \text{ di ticket chiusi attesi sotto } H_0$$

$$O_j = \sum_{i=1}^{k} m_{ji}$$

$$A_j = \sum_{i=1}^{k} m_i * \frac{r_{ji}}{r_i} = \sum_{i=1}^{k} d_{ji}$$

$Q_M \xrightarrow{d} \chi^2$ con *g* gradi di libertà (distribuzione asintotica)

Nel caso in esame la statistica test prende valore 118 con 2 gradi di libertà, ottenendo così un livello di significatività osservato prossimo allo zero. Viene quindi rifiutata l'ipotesi di omogeneità delle funzioni di sopravvivenza.

Il test log-rank assume massima potenza quando si è in una situazione di rischi proporzionali (grafico A), risulta invece inefficiente nel caso opposto (grafico B).

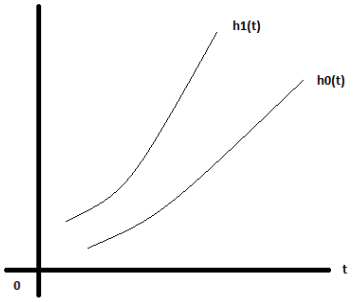


Grafico A

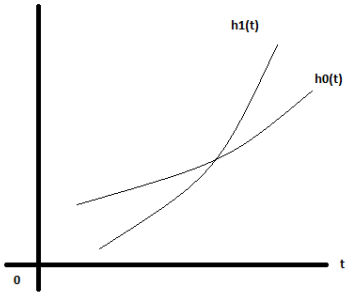


Grafico B

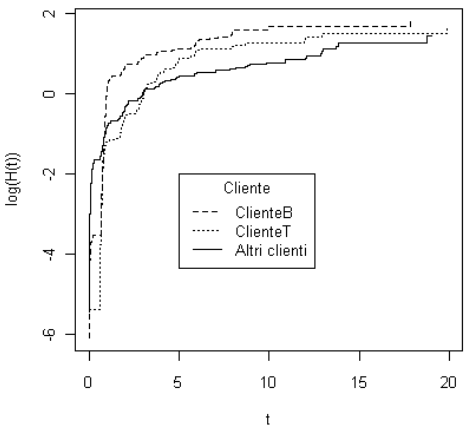
Questo perché nel grafico A la differenza tra le due funzioni di rischio rimane sempre la stessa per ogni t_i con i=1,…,k sotto H_0, ovvero $m_{1i} > d_{1i}$.

Nel grafico B, se indichiamo con t_0 il punto in cui le curve si incrociano, si ha che:

- per $t_i < t_0$ allora $m_{1i} < d_{1i}$
- per $t_i > t_0$ allora $m_{1i} > d_{1i}$

Ci sono quindi differenze positive e negative che si compensano tra di loro, il che implica un valore piccolo della statistica test che porta all'accettazione dell'ipotesi nulla anche se non è vera.

Si vedrà in dettaglio nel capitolo 5.4 come si verifica la proporzionalità dei rischi, ma si anticipa che considerando la categorizzazione dei clienti in 3 gruppi, i rischi non sono proporzionali quindi il test effettuato potrebbe non essere completamente affidabile.



5.4 Il modello di Cox

Con l'intento di verificare quali sono i fattori che influenzano la durata di un ticket, viene provato l'adattamento ai dati del modello di Cox a rischi proporzionali descritto nell'introduzione al capitolo 5, ovvero la formulazione (0) con:

$$\sum_{k=1}^{p} \beta_k Z_k = \beta_1 * clienteB + \beta_2 * clienteT +$$
$$+\beta_3 * meseApertura5 + \beta_4 * meseApertura6$$
$$+\beta_5 * giornoApertura3 + \beta_6 * giornoApertura4$$
$$+\beta_7 giornoApertura5 + \beta_8 * giornoApertura6 +$$
$$+\beta_9 * chiamata + \beta_{10} * trasferta$$

dove:

$clienteB$

$= \begin{cases} 1 \; se \; il \; cliente \; \text{è} \; ClienteB \\ 0 \; altrimenti \end{cases}$

$clienteT$

$= \begin{cases} 1 \; se \; il \; cliente \; \text{è} \; ClienteT \\ 0 \; altrimenti \end{cases}$

$meseApertura5$

$= \begin{cases} 1 \; se \; il \; mese \; di \; apertura \; \text{è} \; maggio \\ 0 \; altrimenti \end{cases}$

$meseApertura6$

$= \begin{cases} 1 \; se \; il \; mese \; di \; apertura \; \text{è} \; giugno \\ 0 \; altrimenti \end{cases}$

$giornoApertura3$

$= \begin{cases} 1 \; se \; il \; giorno \; d'apertura \; \text{è} \; martedì \\ 0 \; altrimenti \end{cases}$

$= \begin{cases} 1 \; se \; il \; giorno \; d'apertura \; \text{è} \; mercoledì \\ 0 \; altrimenti \end{cases}$

$giornoApertura5$

$= \begin{cases} 1 \; se \; il \; giorno \; d'apertura \; \text{è} \; giovedì \\ 0 \; altrimenti \end{cases}$

$giornoApertura6$

$= \begin{cases} 1 \; se \; il \; giorno \; d'apertura \; \text{è} \; venerdì \\ 0 \; altrimenti \end{cases}$

$chiamata$

$= \begin{cases} 1 \; se \; il \; ticket \; richiesto \; per \; telefono \\ 0 \; altrimenti \end{cases}$

$trasferta$

$= \begin{cases} 1 \; se \; c'\text{è} \; stata \; una \; trasferta \\ 0 \; altrimenti \end{cases}$

$giornoApertura4$
(non ci sono ticket aperti la domenica e il sabato, cioè i giorno 1 e 7.)

Il modello di Cox tra gli assunti suppone che non ci siano istanti coincidenti in cui viene chiuso il ticket, ma quando questi ultimi non sono troppo elevati si può utilizzare la *correzione di Breslow* per la verosimiglianza parziale:

$$\prod_{i=1}^{m} \frac{e^{s_i'\beta}}{\left[\sum_{v \in R_i} e^{z_v'\beta}\right]^{m_i}}$$

Dove:

$m = n° \, totale \, di \, morti$

$m_i = n° \, di \, morti \, osservate \, a \; \tau_i$

$R_i =$*classe di individui a rischio nell'istante* τ_i

$s_i =$ *somma dei regressori relativi agli individui che muoiono a* τ_i

Le stime dei coefficienti e i test alla Wald sull'ipotesi di nullità di ciascuno sono (con relativi p-value):

```
                              coef exp(coef) se(coef)      z Pr(>|z|)
CLIENTEB                   0.46332   1.58935  0.10029  4.620 3.84e-06 ***
CLIENTE                   -0.54560   0.57949  0.12528 -4.355 1.33e-05 ***
as.factor(MESE_APERTURA)5  0.05326   1.05471  0.07611  0.700   0.4840
as.factor(MESE_APERTURA)6 -0.07930   0.92376  0.13933 -0.569   0.5692
as.factor(GIORNO_APERTURA)3 -0.06894   0.93338  0.11040 -0.624   0.5323
as.factor(GIORNO_APERTURA)4 -0.15335   0.85783  0.11679 -1.313   0.1892
as.factor(GIORNO_APERTURA)5  0.05968   1.06150  0.11327  0.527   0.5983
as.factor(GIORNO_APERTURA)6 -0.20506   0.81460  0.12139 -1.689   0.0912 .
as.factor(CHIAMATA)1        0.98708   2.68337  0.11260  8.766  < 2e-16 ***
as.factor(TRASFERTA)si      0.07164   1.07427  0.07982  0.897   0.3695
```

Utilizzando una procedura *backward* per l'eliminazione delle le variabili non significative si ottiene il modello (0) con:

$$\sum_{k=1}^{3} \beta_k Z_k = \beta_1 * clienteB + \beta_2 * clienteT + \beta_3 * chiamata$$

Sfruttando il fatto che i parametri sono stime di massima verosimiglianza, ricaviamo per ognuno un intervallo di confidenza al 95% tale che $\hat{\beta} \pm z_{1-\frac{\alpha}{2}} * se(\beta)$ (Klein e Moeschberger, 2001):

```
                        coef exp(coef) se(coef)      z Pr(>|z|)
CLIENTEClienteB      0.45519   1.57647  0.09955  4.572 4.82e-06 ***
CLIENTEClienteT     -0.47328   0.62295  0.11574 -4.089 4.33e-05 ***
as.factor(CHIAMATA)1 0.95644   2.60242  0.11177  8.557  < 2e-16 ***
```

Estremo inferiore	***Stima del parametro***	***Estremo superiore***
0.2600654	β_1=0.45519	0.6503059
-0.7001361	β_2=-0.47328	-0.2464310
0.7373689	β_3=0.95644	1.1755139

Test di nullità di tutti i coefficienti	***Chi – quadro***	***con gradi di libertà:***	***p-value***
Test log-rapporto di verosimiglianza	192.8	3	0
Test alla Wald	181.9	3	0
Test score	192.3	3	0

Ora tutti i coefficienti sono significativamente diversi da zero.

Il **modello di Cox è un modello a rischi proporzionali**, ovvero l'ipotesi che sta alla base di esso è che le unità statistiche con regressori diversi abbiano funzioni di rischio proporzionali e il loro rapporto non risulti dipendente dal tempo:

$$\frac{h_q(t|Z_q)}{h_u(t|Z_u)} = \frac{h_0(t)e^{(\beta^t Z_q)}}{h_0(t)e^{(\beta^t Z_u)}} = \frac{e^{(\beta^t Z_q)}}{e^{(\beta^t Z_u)}}$$

$h_q(t|Z_q) = rischio\ della\ q-esima\ osservazone\ con\ q \in \{1, \dots, n\}$
$h_u(t|Z_u) = rischio\ della\ u-esima\ osservazone\ con\ u \in \{1, \dots, n\}$
$\forall q \neq u\ e\ q, u \in \{1, \dots, n\}$

Quindi

$$h_q(t|Z_q) = c * h_u(t|Z_u)$$

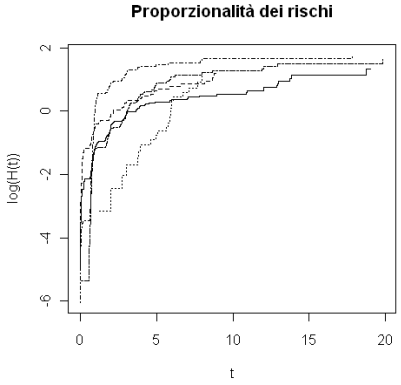
La verifica dell'assunto si effettua per via grafica sfruttando il fatto che anche $H_q(t|Z_q) = C * H_u(t|Z_u)$ e quindi

$$\log\left(H_q(t|Z_q)\right) = \log(C) + \log\left(H_u(t|Z_u)\right)$$

Ciò significa che graficamente le curve dei logaritmi dei rischi cumulati devono essere parallele.

Ora quindi vi è da verificare l'assunto di proporzionalità dei rischi per il modello stimato all'interno di tutte le classi create dalle variabili.

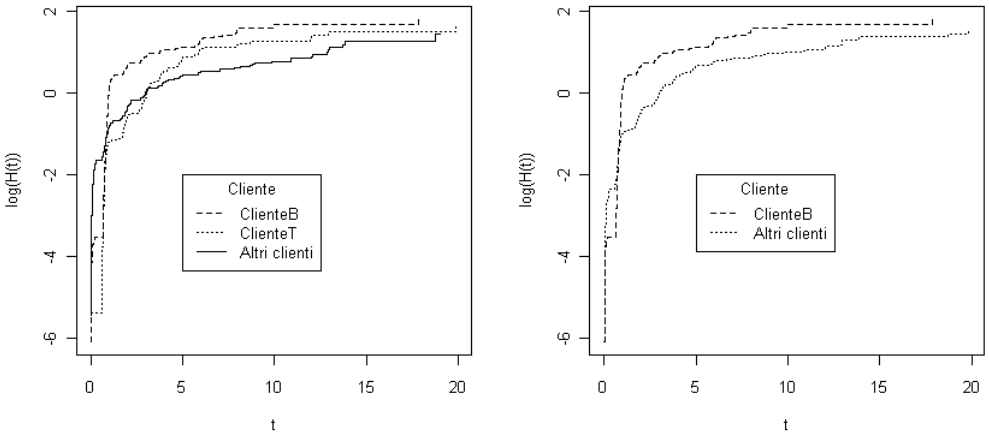
Il grafico dei logaritmi dei rischi cumulati è riportato a fianco. Come si può osservare, le curve si incrociano tra di loro, quindi l'assunto di proporzionalità non è rispettato.



Viene ora verificato se l'assunzione è rispettata almeno marginalmente dalle singole variabili in modo da riuscire ad individuare la causa della non-proporzionalità.

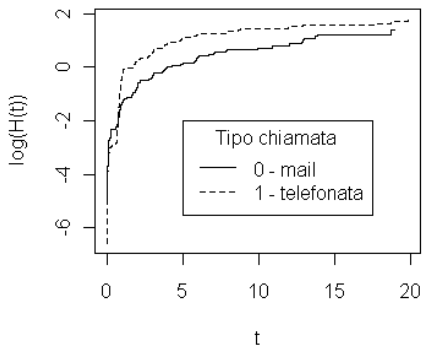
PROPORZIONALITA' PER "CLIENTE"

Il grafico sotto a sinistra rappresenta il logaritmo dei rischi cumulati considerando come variabile esplicativa solo il cliente



Le curve *ClienteT* e *altri clienti* si incrociano, per risolvere il problema si possono unire in un'unica categoria (la nuova variabile da utilizzare viene chiamata "CL"). Ridisegnando il grafico con la nuova categorizzazione si ottiene l'output di destra, la situazione è decisamente migliorata[3].

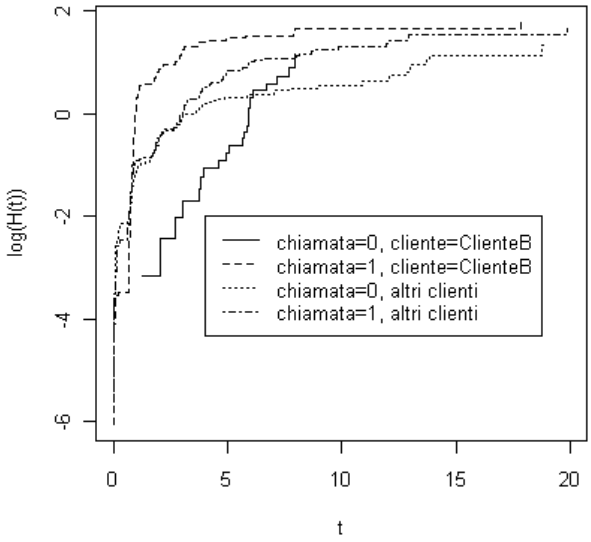
PROPORZIONALITA' PER "CHIAMATA"



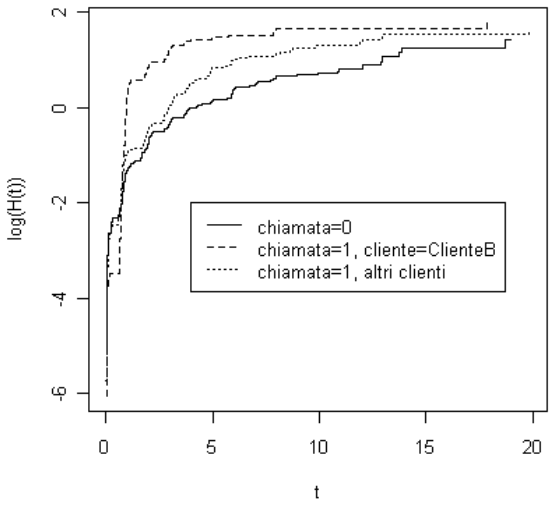
Per il tipo di contatto da parte del cliente non ci sono grossi problemi di proporzionalità.

[3] Ripetendo il test log-rank effettuato alla fine del capitolo 5.3 si ottiene $Q_M = 117$ con 1 grado di libertà e livello di significatività osservato prossimo allo zero.

Viene quindi provata la proporzionalità complessiva del modello utilizzando solo 2 categorie per il cliente



La proporzionalità non è ancora rispettata complessivamente, si è pensato quindi di unire le due categorie che hanno CHIAMATA = 0 (e-mail), e il grafico che si è prodotto è stato:



Quindi, il modello che risulta a questo punto dell'analisi e i relativi coefficienti stimati è:

$$\boldsymbol{h_j(t|Z) = h_0(t)e^{(\beta^t Z_j)} = h_0(t)e^{(\sum_{k=1}^{p} \beta_k Z_{jk})}}$$

dove

$$\boldsymbol{\sum_{k=1}^{p=2} \beta_k Z_{jk} = \beta_1 * chiamata1\_ClienteB + \beta_2 * chiamata1\_Altri}$$

Con

$$chiamata1\_ClienteB = \begin{cases} 1 \; cliente \text{ è } ClienteB \; e \; la \; richiesta \; di \; ticket \text{ è } avvenuta \; per \; telefono \\ 0 \; altrimenti \end{cases}$$

$$chiamata1\_Altri = \begin{cases} 1 \; cliente \; non \text{ è } ClienteB \; e \; la \; richiesta \; di \; ticket \text{ è } avvenuta \; per \; telefono \\ 0 \; altrimenti \end{cases}$$

I valori dei coefficienti sono:

```
          coef exp(coef) se(coef)      z Pr(>|z|)
catb1   1.3574    3.8861   0.1080 12.569  < 2e-16 ***
catta1 0.4161    1.5161   0.1077  3.865 0.000111 ***
```

Estremo inferiore	***Stima del parametro***	***Estremo superiore***
1.1457412	β_1=1.3574	1.5690743
0.2050753	β_2=0.4161	0.6271482

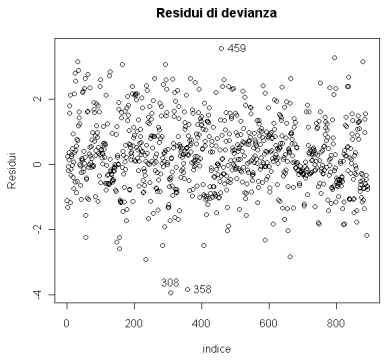
Test di nullità di tutti i coefficienti	***Chi – quadro***	***con gradi di libertà:***	***p-value***
Test log-rapporto di verosimiglianza	208.3	2	0
Test alla Wald	205.1	2	0
Test score	219.1	2	0

Per la proporzionalità dei rischi nel modello di Cox, un coefficiente positivo relativo ad una variabile categoriale significa che il rischio aumenta rispetto alla categoria di riferimento. Ad esempio, si ha che il "rischio di chiudere" un ticket ad ogni istante di tempo t è più alto per chi ha aperto il ticket telefonicamente, in particolare se chi ha chiamato è ClienteB.

Dato che abbiamo 3 categorie per le osservazioni, per evidenziare l'interpretazione dei coefficienti, il modello di Cox della facciata precedente può essere scritto tramite una semplice tabella:

categoria	**Chiama1_ClienteB**	**Chiama1_Altri**	$\mathbf{h(t\|Z)}$	**HR**	Valore HR
ClienteB, telefono	1	0	$h_0(t)e^{\beta_1}$	$e^{\beta_1} =$	3.886076
No ClienteB, telefono	0	1	$h_0(t)e^{\beta_2}$	$e^{\beta_2} =$	1.516037
e-mail	0	0	$h_0(t)$	1	1

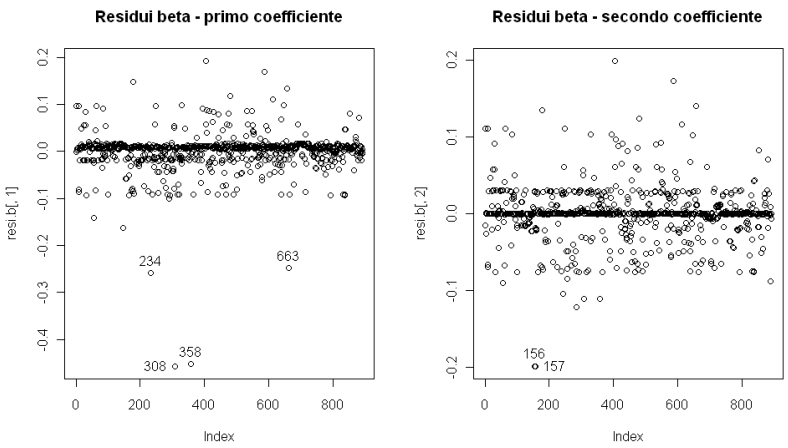
Dove HR è il rapporto con dei rischi $h(t|Z)$ con il rischio di riferimento.



Quindi il rischio che sia chiuso un ticket di ClienteB quando è stato aperto per telefono è 4 volte maggiore rispetto ad un ticket aperto per e-mail.

A sinistra è riportato un grafico con rappresentati i residui di devianza, per la ricerca dei punti anomali. Non c'è nessun punto eccessivamente distaccato dalla nuvola complessiva.

Per la ricerca di eventuali punti leva vengono usati i residui β, ottenuti togliendo una per volta le osservazioni, stimando nuovamente il modello senza di esse e facendo la differenza tra le due stime dei coefficienti.



Dai grafici in alto risaltano 6 possibili punti leva, e dall'analisi della storia di ognuno risulta che si tratta di ticket che riguardano

- malintesi tra la Wintech e il cliente
- ticket che sono stati sospesi per molto tempo, cioè sono stati chiusi e poi riaperti dopo un lungo periodo, ciò può avvenire quando si attende un pezzo di ricambio e questo è in ritardo.
- articoli nuovi e che quindi comportano un maggiore dispendio di tempo sia per la riparazione che l'arrivo del materiale di ricambio
- cliente particolarmente lontano geograficamente e conseguenti lunghi periodo di consegna di un ricambio

Avendo avuto dai tecnici Wintech la spiegazione che si tratta di casi rari e fuori dall'ordinario, si prova a stimare il modello senza tali osservazioni, i risultati delle stime dei coefficienti che si ottengono sono:

```
                      coef exp(coef) se(coef)      z Pr(>|z|)
as.factor(ct)b1  1.6158    5.0321   0.1163 13.889  < 2e-16 ***
as.factor(ct)ta1 0.5415    1.7186   0.1109  4.885 1.04e-06 ***
```

	Stime "vecchie" – non effettuando alcuna modifica al dataset	***Standard error***	***Stime "nuove" – togliento i punti segnalati***	***Standard error***
$\boldsymbol{\beta_1}$	1.3574	0.1080	1.6158	0.1163
$\boldsymbol{\beta_2}$	0.4161	0.1077	0.5415	0.1109

Estremo inferiore	***Stima del parametro***	***Estremo superiore***
1.3878011	β_1=1.6158292	1.8438573
0.3242416	β_2=0.5415335	0.7588254

Categoria	**Chiama1_clienteB**	**Chiama1_Altri**	$\mathbf{h(t\|Z)}$	**HR**	Valore HR
ClienteB, telefono	1	0	$h_0(t)e^{\beta_1}$	e^{β_1} =	5.0321
No ClienteB, telefono	0	1	$h_0(t)e^{\beta_2}$	e^{β_2} =	1.7185
e-mail	0	0	$h_0(t)$	1	1

I punti segnalati non fanno variare di molto la stima dei parametri, quindi non si tratta di punti leva ma al più di casi anomali. Aumenta invece il distacco tra il rischio di riferimento (ticket aperti per e-mail) e ticket aperti telefonicamente per ClienteB.
Per dare una valutazione su quale dei due modelli spiega meglio i dati si passa alla bontà di adattamento del modello ai dati.

Per valutare le bontà di adattamento del modello si usano i residui di Cox Snell (Cox e Snell 1968) i quali si comportano come un campione censurato proveniente da un'esponenziale di parametro 1 quando il modello spiega bene i dati.
Per spiegare il comportamento di questo tipo di residui, siano X e U variabili casuali continue dove

- F(x) funzione di ripartizione di X
- U = F(x) e u una sua realizzazione.

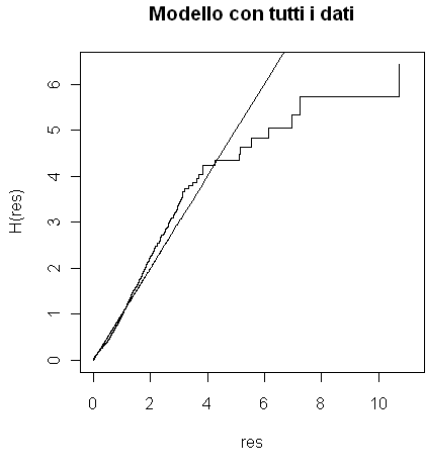
Si ottenga la funzione di ripartizione di U:
$F_U(u) = \Pr(U \leq u) = \Pr\,(F(x) \leq u)$ con $u\epsilon(0,1)$
$= \Pr\left(F^{-1}\big(F(x)\big) \leq F^{-1}(u)\right) = \Pr\big(x \leq F^{-1}(u)\big) = F\big(F^{-1}(u)\big) = u$

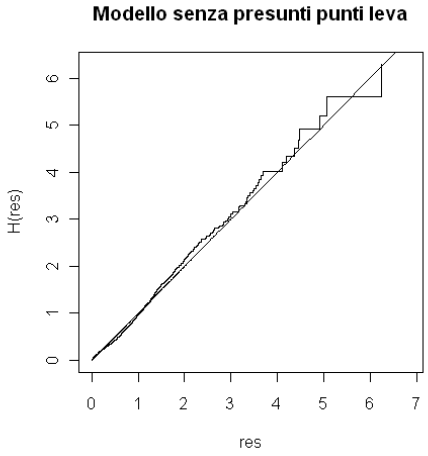
Si definisce V come: $V = H(x) = -\log\big(S(x)\big) = -\log\big(1 - F(x)\big) =$
$-\log(1-u)$

Quindi, la funzione di ripartizione di V è:
$F_V(v) = \Pr(V \leq v) = \Pr(H(x) \leq v) = \Pr(-\log(1-u) \leq v)$
$= \Pr(1 - u \geq e^{-v}) =$
$= \Pr(u \leq 1 - e^{-v}) = F_u(1 - e^{-v}) = 1 - e^{-v}$
che è la funzione di ripartizione di un esponenziale di parametro 1.

Quindi $V \sim Esponenziale(1)$ e vengono definiti *residui di Cox-Snell* le quantità: $\boldsymbol{V_j = \widehat{H_0}(T_j) * e^{(\sum_{k=1}^{p} \beta_k Z_{jk})}}$ con j=1, … , n.

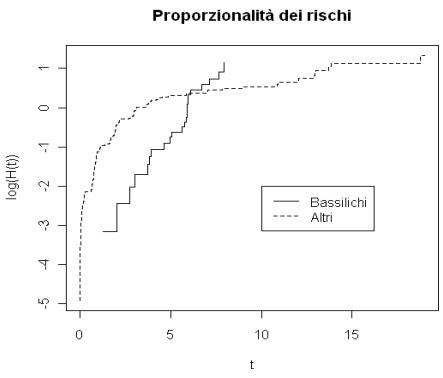
Nei grafici sotto riportati sono rappresentati i rischi cumulati dei residui di Cox-Snell e la retta che li attraversa rappresenta il rischio cumulato di un campione censurato proveniente da esponenziale di parametro 1.





Dal primo al secondo modello i residui migliorano considerevolmente, quindi si decide di mantenere il modello che non considera le 6 osservazioni anomale, dato che queste rappresentano casi eccezionali. Si potrebbe costruire un modello migliore se si avesse a disposizione una variabile che indichi quale è la tipologia del ticket.

A riprova del fatto che se un ticket è stato aperto via mail non c'è alcuna differenza tra clienti diversi (i clienti sono sempre divisi in due categorie, ClienteB e tutti gli altri) proviamo ad effettuare un test log-rank solo per i dati che riguardano le richieste aperte via e-mail (chiamata = 0).
La statistica test assume valore 1.4 con 1 grado di libertà, a cui corrisponde un p-value di 0.236. Viene quindi accetta l'ipotesi nulla dell'uguaglianza delle funzioni di sopravvivenza nei due gruppi.



Bisogna fare attenzione che la proporzionalità all'interno dei clienti che hanno fatto richiesta del ticket per e-mail non è completamente rispettata, quindi il test potrebbe

anche fallire, anche se la motivazione della mancata proporzionalità è quella che in una categoria ci sono solo 24 osservazioni.

Viene ora presentato anche un modello con l'interazione tra il cliente suddiviso in due categorie e la via di comunicazione per la richiesta del ticket, anche se bisogna far presente che la proporzionalità in questo caso è rispettata solo marginalmente

MODELLO CON INTERAZIONE TRA CLIENTE E TIPO DI CHIAMATA

Viene ora valutata l'interazione tra la modalità di richiesta di un ticket e il cliente.

L'output con la stima dei coefficienti del modello di Cox con la suddetta interazione (con i dati senza valori anomali) è il seguente:

```
                          coef exp(coef) se(coef)      z Pr(>|z|)
as.factor(cl)ta          0.3417    1.4073   0.2294  1.489    0.136
as.factor(db$CHIAMATA)1  1.8830    6.5735   0.2191  8.596  < 2e-16 ***
(cl)ta:(db$CHIAMATA)1   -1.4187    0.2420   0.2470 -5.743  9.3e-09 ***
```

Intervalli di confidenza:

Estremo inferiore	***Stima del parametro***	***Estremo superiore***
-0.1079968	β_1=0.3416601	0.7913170
1.4536933	β_2=1.8830412	2.3123890
-1.9028432	β_3=-1.4186768	-0.9345104

L'interazione risulta significativa, quindi verrà mantenuta all'interno del modello.

Da notare è che l'intervallo di confidenza al 95% per β_1 comprende lo zero e valori negativi, il che può implicare la non significatività del parametro o una riduzione dei tempi di risoluzione, invece che un aumento, come indicato dalla stima di massima verosimiglianza del parametro.

Interpretazione dei coefficienti:

categoria	**No clienteB**	**Chiamata1**	**Interazione**	$\mathbf{h(t\|Z)}$	**HR**	Valore HR
ClienteB, telefono	0	1	0	$h_0(t)e^{\beta_2}$	$e^{\beta_2} =$	6.5735
No ClienteB, telefono	1	1	1	$h_0(t)e^{\beta_1+\beta_2+\beta_3}$	$e^{\beta_1+\beta_2+\beta_3} =$	0.8236
ClienteB, e-mail	0	0	0	$h_0(t)$	1	1
No ClienteB, e-mail	1	0	0	$h_0(t)e^{\beta_1}$	e^{β_1}	1.40773

Test per la nullità di tutti i coefficienti:

Test	***Chi – quandro***	***con gradi di libertà:***	***p-value***
Test log-rapporto di verosimiglianza	257.9	3	0
Test alla Wald	242.6	3	0
Test score	265.1	3	0

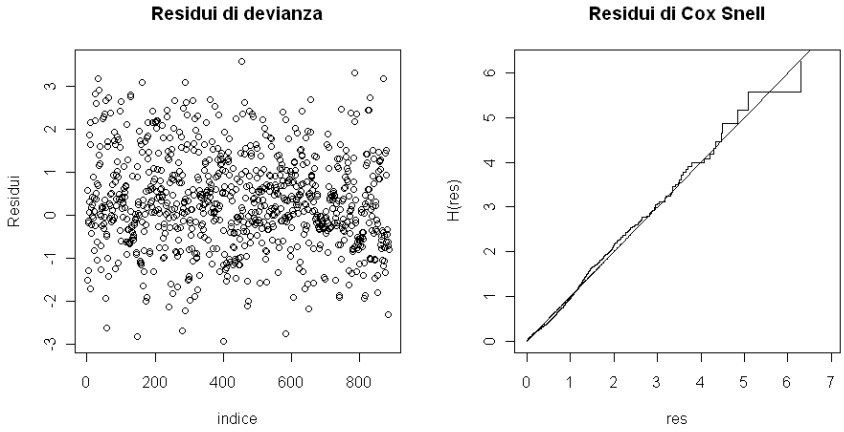
Riguardo all'interpretazione dei coefficienti abbiamo l'informazione in più che se è stato richiesto un ticket per telefono da clienti diversi da ClienteB il "rischio di chiudere" un ticket diminuisce (quindi dura di più) rispetto a se il cliente è ClienteB e se la richiesta è stata fatta per e-mail.

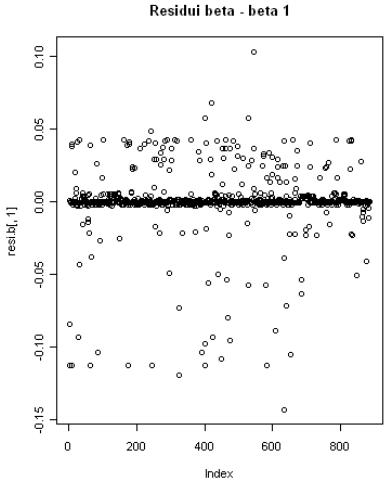
Formalmente, il modello per spiegare la durata di un ticket risulta la (0) del paragrafo 5.1 con

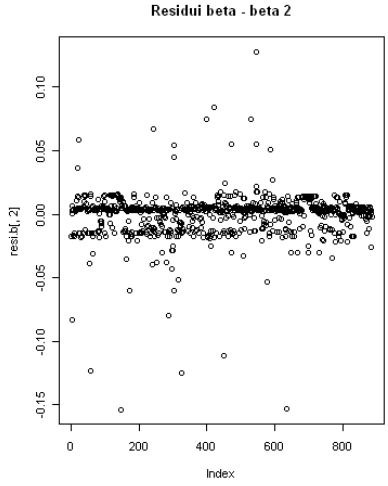
$$\sum_{k=1}^{p} \beta_k Z_k = \beta_1 * clienteNoClienteB + \beta_2 * Chiamata1 + \beta_3 * clienteNoClienteB * Chiamata1$$

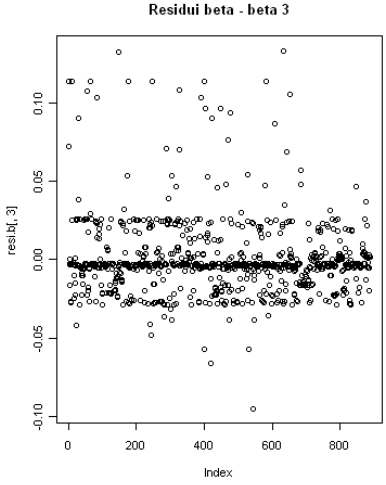
$$clienteClienteB_j = \begin{cases} 1\ cliente\ NON\ \text{è}\ ClienteB \\ 0\ altrimenti \end{cases}$$

$$Chiamata1 = \begin{cases} 1\ se\ la\ chiamata\ \text{è}\ avvenuta\ per\ telefono \\ 0\ altrimenti \end{cases}$$









I residui di devianza non mostrano outlier e dai residui di Cox Snell si osserva che il modello si adatta bene ai dati. Non risaltano possibili punti leva .

5.5 Modello di Cox con effetto frailty

Nel dataset è presente un'ulteriore variabile finora non presa in considerazione, ovvero il dipendente che ha risolto il ticket. Attraverso quest'ultima informazione si vanno a formare tanti piccoli sottogruppi a cui può essere associato un "*effetto frailty*", cioè un effetto casuale inosservabile, condiviso dai ticket appartenenti allo stesso sottogruppo, cioè risolti dallo stesso dipendente. L'idea di base è che alcuni gruppi di ticket abbiano un rischio più alto di altri ad essere chiusi entro un tempo t per il fatto che se ne è occupato un determinato dipendente (questo spiega il termine inglese, ci sono ticket "più fragili"). Un esempio più concreto lo si trova nell'ambito dei dati di sopravvivenza dove l'effetto frailty è usato per considerare diverse filiate in cui le madri trasmettono, per costituzione, una robustezza diversa ai figli.

Una via per la stima dei coefficienti di un modello di Cox che considera effetti casuali fa utilizzo della classe dei "modelli penalizzati" (Therneau e Grambsch, 2000). Dato che questa procedura è implementata nelle librerie di R (survival e coxme) che si stanno utilizzando per l'analisi, nel paragrafo seguente viene effettuata una piccola introduzione ai modelli penalizzati.

5.5.1 Modelli con effetto frailty

Sia τ una variabile casuale e *t* una sua realizzazione. Si consideri il modello di Cox a rischi proporzionali scritto nella seguente maniera:

$$\boldsymbol{h_{ki}(t) = h_0(t) Z_i e^{x'_{ik}\beta}}$$

β = *vettore dei p parametri degli effetti fissi* (p x 1)

x_{ik} = *vettore delle covariate del ticket k-esimo del dipendente i-esimo* (p x 1)

Z_i = *effetto frailty per il dipendente i-esimo con i=1, …, q*

T_{ik} =*k-esima durata di chiusura più piccola del ticket del dipendente i-esimo k=1, …, n_i dove n_i è la numerosità del sottogruppo di ticket del dipendente i-esimo*

Tale formulazione è utilizzata da McGilchrist e Aisbett (1991).

5.5.2 Modelli penalizzati

Si assuma che ogni ticket j-esimo, con j=1, ... ,n , sia membro di un solo gruppo i, con i=1,…,q e sia τ una variabile casuale e t una sua realizzazione. Il modello di Cox può essere scritto come:

$$h_j(t) = h_0(t)e^{(X_j\beta + Z_j\omega)}$$

con

X_j *e* Z_j = *righe j-esime della matrice delle covariate degli effetti fissi,* X_{nxp}*, e degli effetti casuali,* Z_{nxq}

X_{nxp} = *matrice delle covariate relative ai p effetti fissi*

Z_{nxq} = *matrice delle covariate relative ai q effetti casuali*

$Z_{ji} = \begin{cases} 1\ se\ il\ soggetto\ j-esimo\ appartine\ al\ gruppo\ i-esimo \\ 0\ altrimenti \end{cases}$

β = *vettore dei p coefficienti per gli effetti fissi*
$(relativi\ alle\ covariate\ stimate\ anche\ nel\ classico\ modello\ di\ Cox)$

ω = *vettore contenente i q effetti casuali che hanno densità* $p(\omega;\theta)$

La stima dei parametri avviene attraverso la massimizzazione della log-verosimiglianza parziale penalizzata (= PPL = penalized partial log-likelihood), calcolata facendo la differenza tra l'usuale log-verosimiglianza parziale di Cox (D. R. Cox, 1972) e la densità di ω.

$$PPL = PL(\beta, \omega; dati) - p(\omega;\theta)$$

Dove:

$PL(\beta, \omega; dati)$ = *usuale log-verosimiglianza parziale del modello di Cox*

$p(\omega;\theta)$ = *funzione di penalità = funzione che "penalizza" i valori "meno desiderabili" di* ω.

Per il procedimento di stima dei parametri si veda Therneau e Grambsh, 2000.

5.5.3 Modello "shared frailty"

Se si considera quindi il modello descritto nel paragrafo 5.5.2, si può implementare un modello di Cox penalizzato con funzione di penalità $p(\omega) = \frac{1}{2\theta}\sum_{k=1}^{n}\omega_k^2$ (dove n è numero totale di ticket). Tale modello è equivalente ad un modello a effetti casuali con distribuzione normale, descritti da McGilchrist e Aisbett (1991), dove il parametro θ è la varianza degli ω_j. Una valore elevato di θ riflette un alto grado di eterogeneità tra i gruppi e una forte associazione all'interno degli stessi.

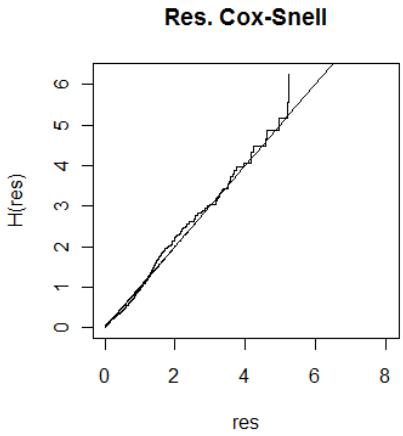
Stimando un modello di Cox con effetto frailty del dipendente i risultati sono i seguenti:

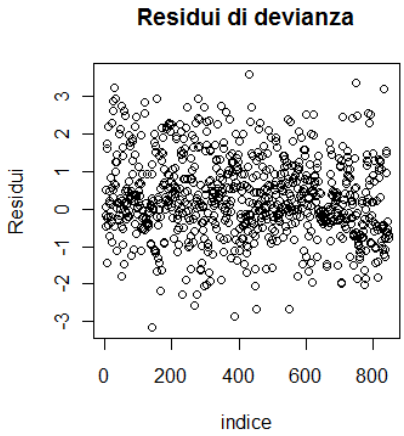
```
                       coef exp(coef)  se(coef)     z       p
as.factor(ct2)b1  1.7876316  5.975284 0.1449962 12.33 0.00000
as.factor(ct2)ta1 0.5096637  1.664731 0.1353299  3.77 0.00017
```

e la variabilità dell'effetto casuale vale **0.129**.

I coefficienti stimati per ogni impiegato sono:

Dipendente1	**Dipendente2**	**Dipendente3**	**Dipendente4**
0.54196035	0.63562273	-0.14308287	0.44687501
Dipendente5	**Dipendente6**	**Dipendente7**	**Dipendente8**
0.06540874	-0.06412256	-0.19409623	-0.16508570
Dipendente9	**Dipendente10**	**Dipendente11**	**Dipendente12**
-0.29154140	-0.27270543	-0.19092647	-0.20505363
Dipendente13	**Dipendente14**		
-0.09412778	-0.06912477		





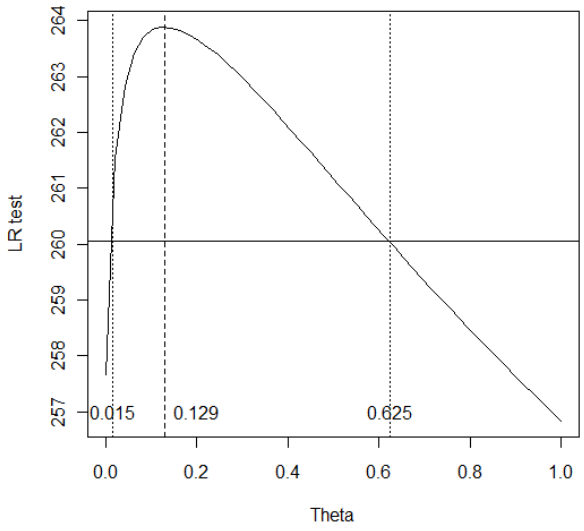
INTERVALLO DI CONFIDENZA PER β

Un intervallo di confidenza al 95% per β è:

Estremo inferiore	*Stima del parametro*	*Estremo superiore*
1.477058	β_1=1.7876316	2.045443
0.2380859	β_2=0.5096637	0.7685791

INTERVALLO DI CONFIDENZA PER ϑ

Per ottenere un intervallo di confidenza per θ si utilizza la log-verosimiglianza profilo tracciata facendo variare il parametro di interesse.

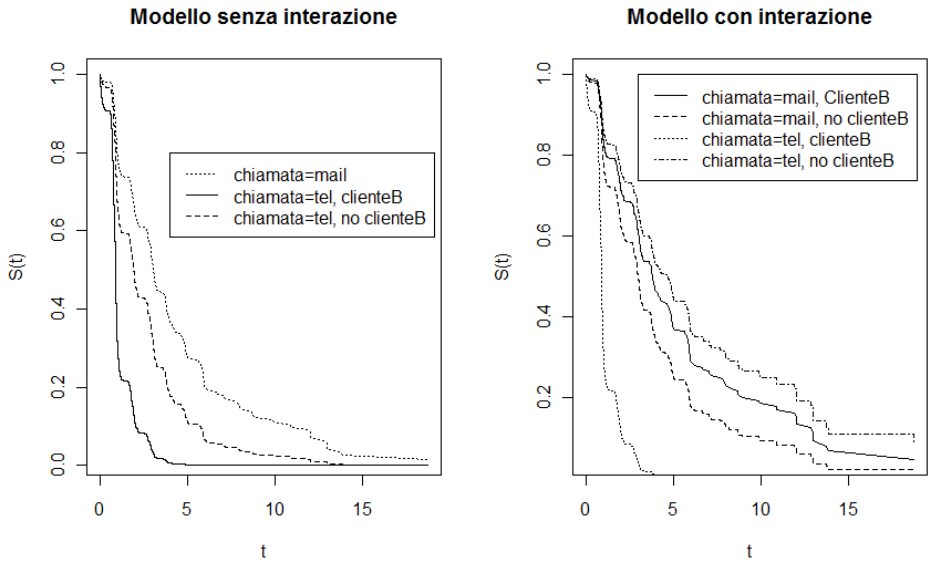


Quindi, un intervallo di confidenza al 95% per θ è (0.015, 0.625). E' importante che tale intervallo non contenga lo zero, in quanto se ci fosse vorrebbe dire che potrebbe essere che l'effetto frailty non è presente nel campione.

Il risultato del test e l'intervallo di confidenza ottenuti per θ permettono quindi di affermare che tra un dipendente e l'altro vi è differenza nei tempi di risoluzione.

6 Conclusioni

Le curve di sopravvivenza predette dal modello di Cox con e senza interazione tra le variabili sono:



Vediamo per esempio qual è la probabilità che un ticket rimanga aperto più di 5 giorni, per il modello senza interazione

Chiamata via mail	Chiamata telefonica e clienteB	Chiamata telefonica e no clienteB
25%	0%	10%

e per il modello con interazione

Chiamata via mail e clienteB	Chiamata via mail e no clienteB	Chiamata telefonica e clienteB	Chiamata telefonica e no clienteB
40%	25%	0%	40%

Come si nota dalle percentuali riportate nelle tabelle, i due modelli sono discordanti quindi si decide di non tenere in considerazione il secondo modello, dato che non è rispettato l'assunto di proporzionalità dei rischi.

Per quando riguarda il modello che tiene conto di effetti casuali, per verificare se il parametro θ è significativo si utilizza il test log-rapporto di

verosimiglianza per modelli annidati, che serve per saggiare l'ipotesi nulla di nullità del coefficiente del modello frailty.
Il modello completo è quello che comprende gli effetti frailty e il modello ridotto è quello senza effetti.

LVMC = log-verosimiglianza modello completo = modello con effetto frailty
LVMR = log-verosimiglianza modello ridotto = modello senza effetto frailty
gradi di libertà = n° di variabili riconducibili agli effetti casuali = 1
Test = 2*(LVMC-LVMR) $\xrightarrow{d} \chi^2$ con 1 grado di libertà.

LVMC	LVMR	Test	Gradi di libertà	p-value
-4328.467	-4331.780	6.626895	1	0.01

La verosimiglianza del modello completo è una verosimiglianza integrata sui termini frailty.

Si rifiuta quindi l'ipotesi nulla di considerare il modello senza effetto frailty con un livello di α osservato <0.05, di conseguenza si conclude che il dipendente che ha risolto il ticket influisce sulla durata dello stesso.

Confrontiamo infine i modelli mediante il criterio di informazione di Akaike (AIC, Akaike 1974) che penalizza la massima log-verosimiglianza con il numero di parametri portando alla scelta del modello con indice più basso.

$$AIC = -2 * PPL + penalità$$

PPL = massima log-verosimiglianza parziale
"Penalità" = 2 volte il numero di parametri del modello nel caso del modello di Cox con e senza interazione; i gradi di libertà del modello penalizzato nel caso del modello con effetto frailty.

Indicando i 3 modelli considerati con
M1 = modello di Cox semplice
M2 = modello di Cox con interazione tra cliente e chiamata
M3 = modello di Cox con effetto frailty del dipendente

	M1	**M2**	**M3**
Log verosimiglianza parziale	-4581.689	-4580.506	-4312.713
N° di parametri o gradi di libertà	p = 2	p = 3	Gdl = 11.846 (p = 3)
Penalità	4	6	23.692 (6)
AIC	9167.378	9167.012	8649.11736 (8631.426)

I gradi di libertà del modello penalizzato si calcolano:

$$gdl = (p + q) - Tr[GH^{-1}]$$

Dove

p = n° di parametri relativi agli effetti fissi

q = n° di parametri relativi agli effetti casuali

$G = \begin{bmatrix} 0 & 0 \\ 0 & -g'' \end{bmatrix}$ con g = derivata seconda della funzione di penalità (nel caso in esame la gaussiana, vedi paragrafo 5.5.3)

$H = I + G$ con I = matrice di informazione del modello di Cox usuale

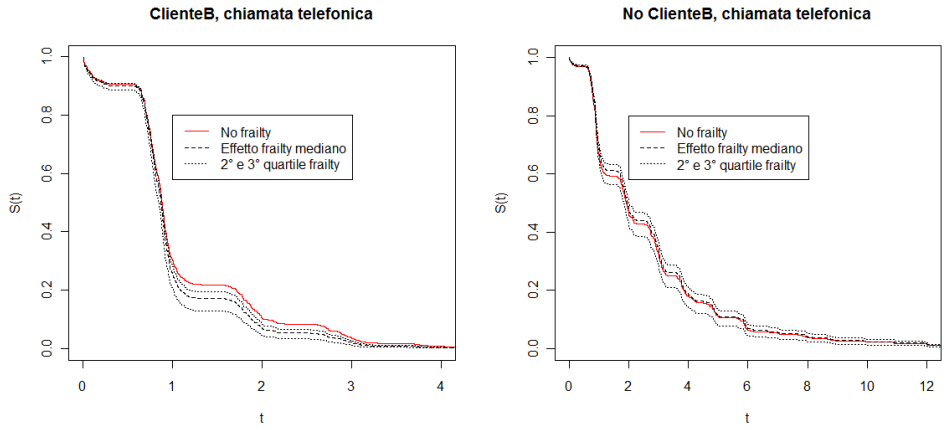
Per i dettagli si veda (Therneau e Grambsch, 2000).

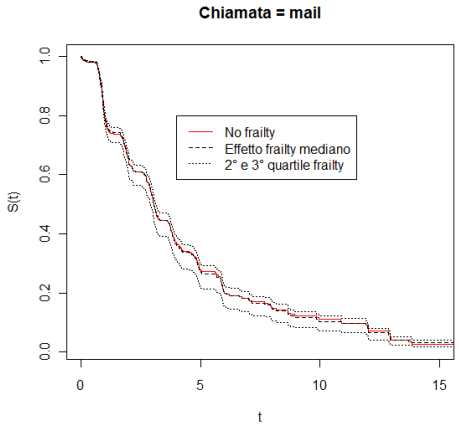
Tra parentesi è riportato anche il valore considerando i gradi di libertà come nei due modelli precedenti, ovvero il numero di parametri.

Nonostante i gradi di libertà siano in numero superiore per il modello con effetto frailty, secondo il criterio di Akaike tra i 3 modelli quest'ultimo è il migliore, il che significa che effettivamente vi è differenza tra un dipendente e l'altro, dovuto forse anche al fatto che alcuni tecnici risolvono solo problemi specifici (di cui però non abbiamo una classificazione).

Per quanto riguarda i due modelli senza effetto del dipendente l'indice AIC è molto simile, quindi si predilige il primo modello perché rispetta le assunzioni e perché stima meno parametri.

Le curve di sopravvivenza messe a confronto graficamente dal modello con e senza frailty sono (per il modello con effetto frailty viene disegnato il frailty mediano, il secondo e il terzo quartile):





Nel primo grafico si nota un certo distaccamento tra i due modelli; nel secondo e nel terzo grafico la curva relativa al frailty mediano è abbastanza coincidente con quella relativa al modello che non considera il dipendente.

APPENDICE – PRINCIPALI COMANDI ESEGUITI IN R PER EFFETTUARE L'ANALISI DELLA DURATA DI UN TICKET

LETTURA DEI DATI E CARICAMENTO DELLA LIBRERIA survival()

```
> library(survival)
Carico il pacchetto richiesto: splines
> dat <- read.csv('dati6-2_DurataInGiorniLavorativi.csv', header =
T, sep=';')
> head(dat)
> attach(dat)

table(PRIORITA, CHIAMATA)
# N.B.: kayako automaticamente assegna priorità 2 ai contatti
# telefonici e 1 ai contatti via mail, quindi non consideriamo la
# variabile priorità (di cui i tecnici non fanno utilizzo)
```

ANALISI ESPLORATIVE

```
# Quanti dati censurati ci sono:
> table(STATO)
STATO
  0   1
 77 818
> table(CLIENTE)
CLIENTE
   altri ClienteB ClienteT
     229      450      216
> table(MESE_APERTURA)
MESE_APERTURA
  4   5   6
365 431  99
> table(GIORNO_APERTURA)
GIORNO_APERTURA
  2   3   4   5   6
161 211 185 190 148
> table(CHIAMATA)
CHIAMATA
  0   1
161 734
> length(CLIENTE)  #895
[1] 895
> length(CLIENTE[CLIENTE=='ClienteT'])/length(CLIENTE)
[1] 0.2413408
> length(CLIENTE[CLIENTE=='ClienteB'])/length(CLIENTE)
[1] 0.5027933


# Statistiche di sintesi di DURATA
> summary(DURATA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0020  0.7925  0.9690  2.1400  2.7280 36.6800

# Box plot della durata dei ticket VS le variabili del dataset
par(mfrow=c(2,3))
boxplot(dat$DURATA, main='Boxplot durata ticket', ylab='Durata')
plot(dat$DURATA~as.factor(dat$CHIAMATA), main='durata VS tipo
      chiamata', xlab='Chiamata', ylab='Durata')
plot(dat$DURATA~dat$CLIENTE, main='durata VS cliente',
      xlab='Cliente', ylab='Durata')
plot(dat$DURATA~as.factor(dat$GIORNO_APERTURA), main='durata VS
      giorno apertura', xlab='Giorno apertura', ylab='Durata')
plot(dat$DURATA~as.factor(dat$TRASFERTA), main='durata VS
```

```
      trasferta', xlab='Trasferta', ylab='Durata')
plot(dat$DURATA~as.factor(dat$MESE_APERTURA), main='durata VS mese
      apertura', xlab='Mese apertura', ylab='Durata')
par(mfrow=c(1,1))
# Identificazione ed eliminazione dei due valori anomali che si
# osservano nei boxplot.

Ticketid[DURATA==27.951]  #9999
Ticketid[DURATA==36.684]  #9691

d <- dat[DURATA!= 27.951 & DURATA!=36.684,]
d

detach(dat)
attach(d)
```

FUNZIONE DI SOPRAVVIVENZA

```
# per CLIENTE
surv <- survfit(Surv(DURATA,STATO)~CLIENTE)
plot(surv, main='Funzioni di sopravvivenza - Cliente', lty=1:3)
legenda <- c('ClienteB','ClienteT','Altri clienti')
legend(10,0.8,legend=legenda, lty=c(2,3,1))
abline(v=5, lty=2)
# per CHIAMATA
surv <- survfit(Surv(DURATA,STATO)~as.factor(CHIAMATA))
plot(surv, main='Funzioni di sopravvivenza - Chiamata', lty=1:2)
legenda <- c('0 - mail','1 - telefonata')
legend(10,0.8,legend=legenda, lty=1:2)
abline(v=5, lty=2)
# per GIORNO_APERTURA
surv <- survfit(Surv(DURATA,STATO)~as.factor(GIORNO_APERTURA))
plot(surv, main='Funzioni di sopravvivenza - Giorno', lty=1:5)
legenda <- c(2,3,4,5,6)
legend(10,0.8,legend=legenda, lty=1:5)
# per MESE_APERTURA
surv <- survfit(Surv(DURATA,STATO)~as.factor(MESE_APERTURA))
plot(surv, main='Funzioni di sopravvivenza - Mese', lty=1:3)
legenda <- c('Aprile','Maggio','Giugno')
legend(10,0.8,legend=legenda, lty=1:3)
# per la TRASFERTA
surv <- survfit(Surv(DURATA,STATO)~TRASFERTA)
plot(surv, main='Funzioni di sopravvivenza - Trasferta', lty=1:3)
legenda <- c('No','Si')
legend(10,0.8,legend=legenda, lty=1:2)
```

PROCESSO DI ELIMINAZIONE DELLE VARIABILI NON SIGNIFICATIVE

```
> fit.cox <- coxph(Surv(DURATA,STATO)~CLIENTE +
as.factor(MESE_APERTURA)
+ + as.factor(GIORNO_APERTURA)+ as.factor(CHIAMATA)
+ + as.factor(TRASFERTA), method='breslow')
> summary(fit.cox)
Call:
coxph(formula = Surv(DURATA, STATO) ~ CLIENTE +
as.factor(MESE_APERTURA) +
    as.factor(GIORNO_APERTURA) + as.factor(CHIAMATA) +
as.factor(TRASFERTA),
    method = "breslow")
  n= 893

                              coef exp(coef) se(coef)      z Pr(>|z|)
CLIENTEClienteB             0.46332   1.58935  0.10029  4.620 3.84e-06 ***
CLIENTEClienteT            -0.54560   0.57949  0.12528 -4.355 1.33e-05 ***
as.factor(MESE_APERTURA)5   0.05326   1.05471  0.07611  0.700   0.4840
as.factor(MESE_APERTURA)6  -0.07930   0.92376  0.13933 -0.569   0.5692
as.factor(GIORNO_APERTURA)3 -0.06894   0.93338  0.11040 -0.624   0.5323
as.factor(GIORNO_APERTURA)4 -0.15335   0.85783  0.11679 -1.313   0.1892
```

```
as.factor(GIORNO_APERTURA)5  0.05968   1.06150  0.11327  0.527   0.5983
as.factor(GIORNO_APERTURA)6 -0.20506   0.81460  0.12139 -1.689   0.0912 .
as.factor(CHIAMATA)1         0.98708   2.68337  0.11260  8.766  < 2e-16 ***
as.factor(TRASFERTA)si       0.07164   1.07427  0.07982  0.897   0.3695
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                          exp(coef) exp(-coef) lower .95 upper .95
CLIENTEClienteB              1.5893     0.6292    1.3057    1.9346
CLIENTEClienteT              0.5795     1.7256    0.4533    0.7408
as.factor(MESE_APERTURA)5    1.0547     0.9481    0.9086    1.2244
as.factor(MESE_APERTURA)6    0.9238     1.0825    0.7030    1.2138
as.factor(GIORNO_APERTURA)3  0.9334     1.0714    0.7518    1.1589
as.factor(GIORNO_APERTURA)4  0.8578     1.1657    0.6823    1.0785
as.factor(GIORNO_APERTURA)5  1.0615     0.9421    0.8502    1.3254
as.factor(GIORNO_APERTURA)6  0.8146     1.2276    0.6421    1.0334
as.factor(CHIAMATA)1         2.6834     0.3727    2.1520    3.3460
as.factor(TRASFERTA)si       1.0743     0.9309    0.9187    1.2562

Rsquare= 0.203   (max possible= 1 )
Likelihood ratio test= 202.4  on 10 df,   p=0
Wald test            = 189.7  on 10 df,   p=0
Score (logrank) test = 200.5  on 10 df,   p=0

> fit.cox <- coxph(Surv(DURATA,STATO)~CLIENTE
+ + as.factor(GIORNO_APERTURA)+ as.factor(CHIAMATA)
+ + as.factor(TRASFERTA), method='breslow')
> summary(fit.cox)
Call:
coxph(formula = Surv(DURATA, STATO) ~ CLIENTE +
as.factor(GIORNO_APERTURA) +
    as.factor(CHIAMATA) + as.factor(TRASFERTA), method = "breslow")

  n= 893

                             coef exp(coef) se(coef)      z Pr(>|z|)
CLIENTEClienteB             0.45825   1.58130  0.09992  4.586 4.51e-06 ***
CLIENTEClienteT            -0.54444   0.58017  0.12528 -4.346 1.39e-05 ***
as.factor(GIORNO_APERTURA)3 -0.07119   0.93128  0.11035 -0.645   0.5188
as.factor(GIORNO_APERTURA)4 -0.17589   0.83871  0.11471 -1.533   0.1252
as.factor(GIORNO_APERTURA)5  0.06970   1.07218  0.11284  0.618   0.5368
as.factor(GIORNO_APERTURA)6 -0.20795   0.81224  0.12138 -1.713   0.0867 .
as.factor(CHIAMATA)1         0.97642   2.65493  0.11214  8.707  < 2e-16 ***
as.factor(TRASFERTA)si       0.06878   1.07120  0.07966  0.863   0.3879
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                          exp(coef) exp(-coef) lower .95 upper .95
CLIENTEClienteB              1.5813     0.6324    1.3001    1.9234
CLIENTEClienteT              0.5802     1.7236    0.4539    0.7416
as.factor(GIORNO_APERTURA)3  0.9313     1.0738    0.7502    1.1562
as.factor(GIORNO_APERTURA)4  0.8387     1.1923    0.6698    1.0501
as.factor(GIORNO_APERTURA)5  1.0722     0.9327    0.8594    1.3376
as.factor(GIORNO_APERTURA)6  0.8122     1.2312    0.6403    1.0304
as.factor(CHIAMATA)1         2.6549     0.3767    2.1311    3.3075
as.factor(TRASFERTA)si       1.0712     0.9335    0.9164    1.2522

Rsquare= 0.202   (max possible= 1 )
Likelihood ratio test= 201.2  on 8 df,   p=0
Wald test            = 189.0  on 8 df,   p=0
Score (logrank) test = 199.6  on 8 df,   p=0

> fit.cox <- coxph(Surv(DURATA,STATO)~CLIENTE
+ + as.factor(GIORNO_APERTURA)+ as.factor(CHIAMATA)
+ , method='breslow')
> summary(fit.cox)
Call:
coxph(formula = Surv(DURATA, STATO) ~ CLIENTE +
as.factor(GIORNO_APERTURA) +
    as.factor(CHIAMATA), method = "breslow")

  n= 893

                             coef exp(coef) se(coef)      z Pr(>|z|)
CLIENTEClienteB             0.46323   1.58920  0.09961  4.651 3.31e-06 ***
CLIENTEClienteT            -0.50701   0.60230  0.11748 -4.316 1.59e-05 ***
as.factor(GIORNO_APERTURA)3 -0.06962   0.93274  0.11033 -0.631   0.5280
as.factor(GIORNO_APERTURA)4 -0.16758   0.84571  0.11430 -1.466   0.1426
```

```
as.factor(GIORNO_APERTURA)5  0.07120   1.07380  0.11280  0.631   0.5279
as.factor(GIORNO_APERTURA)6 -0.20088   0.81801  0.12108 -1.659   0.0971 .
as.factor(CHIAMATA)1         0.97274   2.64517  0.11186  8.696  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                          exp(coef) exp(-coef) lower .95 upper .95
CLIENTEClienteB              1.5892     0.6292    1.3074    1.9318
CLIENTEClienteT              0.6023     1.6603    0.4784    0.7582
as.factor(GIORNO_APERTURA)3  0.9327     1.0721    0.7514    1.1579
as.factor(GIORNO_APERTURA)4  0.8457     1.1824    0.6760    1.0581
as.factor(GIORNO_APERTURA)5  1.0738     0.9313    0.8608    1.3395
as.factor(GIORNO_APERTURA)6  0.8180     1.2225    0.6452    1.0371
as.factor(CHIAMATA)1         2.6452     0.3780    2.1244    3.2936

Rsquare= 0.201   (max possible= 1 )
Likelihood ratio test= 200.5  on 7 df,   p=0
Wald test            = 188.1  on 7 df,   p=0
Score (logrank) test = 198.8  on 7 df,   p=0

>
> fit.cox <- coxph(Surv(DURATA,STATO)~CLIENTE + as.factor(CHIAMATA)
+ , method='breslow')
> summary(fit.cox)
Call:
coxph(formula = Surv(DURATA, STATO) ~ CLIENTE +
as.factor(CHIAMATA),
    method = "breslow")

  n= 893

                       coef exp(coef) se(coef)      z Pr(>|z|)
CLIENTEClienteB       0.45519   1.57647  0.09955  4.572 4.82e-06 ***
CLIENTEClienteT      -0.47328   0.62295  0.11574 -4.089 4.33e-05 ***
as.factor(CHIAMATA)1  0.95644   2.60242  0.11177  8.557  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                      exp(coef) exp(-coef) lower .95 upper .95
CLIENTEClienteB           1.576     0.6343    1.2970    1.9161
CLIENTEClienteT           0.623     1.6053    0.4965    0.7816
as.factor(CHIAMATA)1      2.602     0.3843    2.0904    3.2398

Rsquare= 0.194   (max possible= 1 )
Likelihood ratio test= 192.8  on 3 df,   p=0
Wald test            = 181.9  on 3 df,   p=0
Score (logrank) test = 192.3  on 3 df,   p=0

# Intervalli di confidenza per i parametri
> se <- sqrt(c(fit.cox$var[1,1], fit.cox$var[2,2],
fit.cox$var[3,3]))
> u <- fit.cox$coef+1.96*se
> l <- fit.cox$coef-1.96*se
> data.frame(l,fit.cox$coef,u)
                                l fit.cox.coef          u
CLIENTEClienteB        0.2600654    0.4551856  0.6503059
CLIENTEClienteT       -0.7001361   -0.4732835 -0.2464310
as.factor(CHIAMATA)1   0.7373689    0.9564414  1.1755139


# Verifica della proporzionalità per "CLIENTE" e "CHIAMATA" assieme

surv.pr <- survfit(Surv(DURATA,STATO)~CLIENTE +
as.factor(CHIAMATA))
surv.pr$strata

s <- c(rep(1,125),rep(2,90),rep(3,24),rep(4,314),rep(5,1),
rep(6,205))
plot(surv.pr$time, log(-log(surv.pr$surv)), type='n',
ylab='log(H(t))', xlab='t', main='Proporzionalità dei rischi')
lines(surv.pr$time[s==1], log(-log(surv.pr$surv[s==1])), type='s',
lty=1)
```

```
lines(surv.pr$time[s==2], log(-log(surv.pr$surv[s==2])), type='s',
lty=2)#, col=2)
lines(surv.pr$time[s==3], log(-log(surv.pr$surv[s==3])), type='s',
lty=3)#, col=3)   #verde
lines(surv.pr$time[s==4], log(-log(surv.pr$surv[s==4])), type='s',
lty=4)#, col=4)   #blu
lines(surv.pr$time[s==5], log(-log(surv.pr$surv[s==5])), type='s',
lty=5)#, col=5)   #azzurra
lines(surv.pr$time[s==6], log(-log(surv.pr$surv[s==6])), type='s',
lty=6)#, col=6)   #rosa


# Verifica della proporzionalità per "CLIENTE"
surv.c <- survfit(Surv(DURATA,STATO)~CLIENTE)
surv.c$strata

s <- c(rep('a',209), rep('b',338), rep('t', 206))
plot(surv.c$time, log(-log(surv.c$surv)), type='n', xlab='t',
ylab='log(H(t))')
lines(surv.c$time[s=='a'], log(-log(surv.c$surv[s=='a'])),
type='s', lty=1)
lines(surv.c$time[s=='b'], log(-log(surv.c$surv[s=='b'])),
type='s', lty=2)
lines(surv.c$time[s=='t'], log(-log(surv.c$surv[s=='t'])),
type='s', lty=3)
legenda <- c('ClienteB','ClienteT','Altri clienti')
legend(5,-2,legend=legenda, lty=c(2,3,1), title='Cliente')

# Creazione di una nuova variabile per riclassificare in solo 2
categorie
# il cliente
CL <- rep(NA,length(CLIENTE))
for(i in 1:length(CLIENTE))
{
 if(CLIENTE[i]=='ClienteT' | CLIENTE[i]=='altri')
       CL[i] <- 'ta'
 else CL[i] <- 'b'
}

# Verifica della proporzionalità per "CL"
surv.c <- survfit(Surv(DURATA,STATO)~CL)
surv.c$strata

s <- c(rep('b',338), rep('ta', 405))
plot(surv.c$time, log(-log(surv.c$surv)), type='n', xlab='t',
ylab='log(H(t))')
lines(surv.c$time[s=='b'], log(-log(surv.c$surv[s=='b'])),
type='s', lty=2)
lines(surv.c$time[s=='ta'], log(-log(surv.c$surv[s=='ta'])),
type='s', lty=3)
legenda <- c('ClienteB','Altri clienti')
legend(5,-2,legend=legenda, lty=c(2,3), title='Cliente')

#Test log-rank
> survdiff(Surv(DURATA,STATO)~CL)
Call:
survdiff(formula = Surv(DURATA, STATO) ~ CL)

        N Observed Expected (O-E)^2/E (O-E)^2/V
CL=b  448      433      290      70.0       117
CL=ta 445      383      526      38.7       117

 Chisq= 117  on 1 degrees of freedom, p= 0

# Verifica della proporzionalità per "CHIAMATA"
surv.pr <- survfit(Surv(DURATA,STATO)~as.factor(CHIAMATA))
surv.pr$strata
```

```
s <- c(rep(0,149),rep(1,561))
plot(surv.pr$time, log(-log(surv.pr$surv)), type='n',
ylab='log(H(t))', xlab='t')
lines(surv.pr$time[s==0], log(-log(surv.pr$surv[s==0])), type='s',
lty=1)
lines(surv.pr$time[s==1], log(-log(surv.pr$surv[s==1])), type='s',
lty=2)
legenda <- c('0 - mail','1 - telefonata')
legend(5,-2,legend=legenda, lty=c(1,2), title='Tipo chiamata')

# Verifica della proporzionalità per "CL" e  "CHIAMATA" assieme
surv.pr <- survfit(Surv(DURATA,STATO)~ CL + as.factor(CHIAMATA))
surv.pr$strata

s <- c(rep(1,24),rep(2,314),rep(3,126),rep(4,294))
plot(surv.pr$time, log(-log(surv.pr$surv)), type='n',
ylab='log(H(t))', xlab='t')
lines(surv.pr$time[s==1], log(-log(surv.pr$surv[s==1])), type='s',
lty=1)        #nero
lines(surv.pr$time[s==2], log(-log(surv.pr$surv[s==2])), type='s',
lty=2)#, col=2) #rosso
lines(surv.pr$time[s==3], log(-log(surv.pr$surv[s==3])), type='s',
lty=3)#, col=3) #verde
lines(surv.pr$time[s==4], log(-log(surv.pr$surv[s==4])), type='s',
lty=4)#, col=4) #blu
legenda <- c('chiamata=0, cliente=ClienteB','chiamata=1,
cliente=ClienteB','chiamata=0, altri clienti', 'chiamata=1, altri
clienti')
legend(4,-2,legend=legenda, lty=1:4)

#Proviamo a trattare congiuntamente le categorie 1 e 3 (3=tutti
clienti con chiamata = 0)
cat <- rep(NA,length(CL))
for(i in 1:length(CL))
{
 if(CL[i]=='b' & CHIAMATA[i]==1) cat[i] <- 'b1'
 if(CL[i]=='ta' & CHIAMATA[i]==1) cat[i] <- 'ta1'
 if(CHIAMATA[i]==0) cat[i] <- '0'
}

surv.cat <- survfit(Surv(DURATA,STATO)~ cat)
surv.cat$strata

s <- c(rep(0,149),rep('b1',314),rep('ta1',292))
plot(surv.pr$time, log(-log(surv.pr$surv)), type='n',
ylab='log(H(t))', xlab='t')
lines(surv.cat$time[s==0], log(-log(surv.cat$surv[s==0])),
type='s', lty=1)
lines(surv.cat$time[s=='b1'], log(-log(surv.cat$surv[s=='b1'])),
type='s', lty=2)#, col=2)
lines(surv.cat$time[s=='ta1'], log(-log(surv.cat$surv[s=='ta1'])),
type='s', lty=3)#, col=3)
legenda <- c('chiamata=0','chiamata=1,
cliente=ClienteB','chiamata=1, altri clienti')
legend(4,-2,legend=legenda, lty=1:3)

# Modello finale
# ovvero con le categorie:
# - chiamata 1, clienteB
# - chiamata 1, no clienteB
# - chiamata 0, tutti i clienti

> fit.cox <- coxph(Surv(DURATA,STATO)~as.factor(cat),
+  method='breslow')
> summary(fit.cox)
Call:
coxph(formula = Surv(DURATA, STATO) ~ as.factor(cat), method =
"breslow")
```

```
  n= 893

                    coef exp(coef) se(coef)      z Pr(>|z|)
as.factor(cat)b1  1.3574    3.8861   0.1080 12.569  < 2e-16 ***
as.factor(cat)ta1 0.4161    1.5161   0.1077  3.865 0.000111 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                  exp(coef) exp(-coef) lower .95 upper .95
as.factor(cat)b1      3.886     0.2573     3.145     4.802
as.factor(cat)ta1     1.516     0.6596     1.228     1.872

Rsquare= 0.208   (max possible= 1 )
Likelihood ratio test= 208.3  on 2 df,   p=0
Wald test            = 205.1  on 2 df,   p=0
Score (logrank) test = 219.1  on 2 df,   p=0

# Intervallo di confidenza
> se <- sqrt(c(fit.cox$var[1,1], fit.cox$var[2,2]))
> u <- fit.cox$coef+1.96*se
> l <- fit.cox$coef-1.96*se
>
> data.frame(l,fit.cox$coef,u)
                         l fit.cox.coef         u
as.factor(cat)b1  1.1457412    1.3574077 1.5690743
as.factor(cat)ta1 0.2050753    0.4161117 0.6271482

# Residui di devianza – ricerca dati anomali
resi.d <- residuals(fit.cox, type='dev')
plot(resi.d, main='Residui di devianza', xlab='indice',
ylab='Residui')
identify(resi.d)
# Residui beta – ricerca punti leva
resi.b <- residuals(fit.cox, type='dfbetas')
head(resi.b)
par(mfrow=c(1,2))
plot(resi.b[,1], main='Residui beta - primo coefficiente')
identify(resi.b[,1])
plot(resi.b[,2], main='Residui beta - secondo coefficiente')
identify(resi.b[,2])

# Residui di Cox Snell – adattamento del modello ai dati
resi.m <- residuals(fit.cox, type='mart')
resi.cs <- STATO-resi.m
s.res <- survfit(Surv(resi.cs, STATO)~1)
plot(s.res$time, -log(s.res$surv), type='s', main='Modello con
tutti i dati',
             xlab='res', ylab='H(res)')
lines(s.res$time,s.res$time)

# Eliminazione punti evidenziati dalla ricerca dei punti leva che
# si sono evidenziati come outlier
i <- 1:length(cat)
db <- d[-c(148,375,219,188,694,695),]
cl <- CL[-c(148,375,219,188,694,695)]
ct <- cat[-c(148,375,219,188,694,695)]

# Stima del modello senza punti leva evidenziati
> fit.cox <- coxph(Surv(db$DURATA,db$STATO)~as.factor(ct),
+  method='breslow')
> summary(fit.cox)
Call:
coxph(formula = Surv(db$DURATA, db$STATO) ~ as.factor(ct), method =
"breslow")

  n= 887
```

```
                    coef exp(coef) se(coef)      z Pr(>|z|)
as.factor(ct)b1  1.3543    3.8740   0.1081 12.523  < 2e-16 ***
as.factor(ct)ta1 0.4141    1.5130   0.1077  3.844 0.000121 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                 exp(coef) exp(-coef) lower .95 upper .95
as.factor(ct)b1      3.874     0.2581     3.134     4.789
as.factor(ct)ta1     1.513     0.6609     1.225     1.869

Rsquare= 0.207   (max possible= 1 )
Likelihood ratio test= 206.0  on 2 df,   p=0
Wald test            = 203.2  on 2 df,   p=0
Score (logrank) test = 217  on 2 df,   p=0

# Per visualizzare i residui di Cox Snell e gli intervalli di
# confidenza per i parametri, utilizzare il codice del modello
# precedente
```

MODELLO DI COX CON INTERAZIONE TRA CLIENTE E CHIAMATA

```
> fit.cox.int <-
coxph(Surv(db$DURATA,db$STATO)~as.factor(cl)*as.factor(db$CHIAMATA)
,
+ method='breslow')
> summary(fit.cox.int)
Call:
coxph(formula = Surv(db$DURATA, db$STATO) ~ as.factor(cl) *
as.factor(db$CHIAMATA),
    method = "breslow")

  n= 887

                                        coef exp(coef) se(coef)      z
as.factor(cl)ta                       0.3417    1.4073   0.2294  1.489
as.factor(db$CHIAMATA)1               1.8830    6.5735   0.2191  8.596
as.factor(cl)ta:as.factor(db$CHIAMATA)1 -1.4187    0.2420   0.2470 -5.743
                                      Pr(>|z|)
as.factor(cl)ta                          0.136
as.factor(db$CHIAMATA)1                < 2e-16 ***
as.factor(cl)ta:as.factor(db$CHIAMATA)1  9.3e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                                      exp(coef) exp(-coef) lower .95
as.factor(cl)ta                          1.4073     0.7106    0.8976
as.factor(db$CHIAMATA)1                  6.5735     0.1521    4.2789
as.factor(cl)ta:as.factor(db$CHIAMATA)1    0.2420     4.1316    0.1491
                                        upper .95
as.factor(cl)ta                            2.2063
as.factor(db$CHIAMATA)1                   10.0984
as.factor(cl)ta:as.factor(db$CHIAMATA)1    0.3928

Rsquare= 0.252   (max possible= 1 )
Likelihood ratio test= 257.9  on 3 df,   p=0
Wald test            = 242.6  on 3 df,   p=0
Score (logrank) test = 265.1  on 3 df,   p=0
```

MODELLO DI COX CON EFFETTO FRAILTY

```
# Bisogna eliminare dal dataset le osservazioni che non hanno
# impiegato
db2 <- db[db$NOME_IMPIEGATO!='-',]
ct2 <- ct[db$NOME_IMPIEGATO!='-']
cl2 <- cl[db$NOME_IMPIEGATO!='-']

> fit.coxme.f <- coxme(Surv(db2$DURATA,db2$STATO)~as.factor(ct2),
ties='breslow', random=~1|db2$NOME_IMPIEGATO)
> fit.coxme.f
```

```
Cox mixed-effects model fit by maximum likelihood

  events, n = 773, 847
  Iterations= 15 79
                    NULL Integrated Penalized
Log-likelihood -4460.409  -4328.467 -4317.537

                   Chisq    df p    AIC    BIC
Integrated loglik 263.88  3.00 0 257.88 243.93
 Penalized loglik 285.74 11.32 0 263.10 210.46

Model:  Surv(db2$DURATA, db2$STATO) ~ as.factor(ct2)
Fixed coefficients
                       coef exp(coef)  se(coef)     z       p
as.factor(ct2)b1  1.7876316  5.975284 0.1449962 12.33 0.00000
as.factor(ct2)ta1 0.5096637  1.664731 0.1353299  3.77 0.00017

Random effects
 Group              Variable  Std Dev   Variance
 db2.NOME_IMPIEGATO Intercept 0.3589109 0.1288170

# Per avere tutti i tipi di residui utilizziamo coxph() ponendo il
# parametro theta dell'effetto frailty uguale a quello ottenuto con
# coxme()

# Residui di Cox Snell
> fit.cox.f <- coxph(Surv(db2$DURATA,db2$STATO)~as.factor(ct2)+
+   frailty(db2$NOME_IMPIEGATO, dist='gaussian', theta=0.1288170),
method='breslow')
> summary(fit.cox.f)
Call:
coxph(formula = Surv(db2$DURATA, db2$STATO) ~ as.factor(ct2) +
    frailty(db2$NOME_IMPIEGATO, dist = "gaussian", theta =
0.128817),
    method = "breslow")

  n= 847
                          coef  se(coef) se2   Chisq DF   p
as.factor(ct2)b1          1.788 0.144    0.138 155.2  1.0 0.0e+00
as.factor(ct2)ta1         0.509 0.140    0.136  13.2  1.0 2.8e-04
frailty(db2$NOME_IMPIEGAT                       35.7 10.0 9.8e-05

                  exp(coef) exp(-coef) lower .95 upper .95
as.factor(ct2)b1       5.98      0.167      4.51      7.92
as.factor(ct2)ta1      1.66      0.601      1.26      2.19

Iterations: 1 outer, 5 Newton-Raphson
     Variance of random effect= 0.129
Degrees of freedom for terms=  1.8 10.0
Rsquare= 0.294   (max possible= 1 )
Likelihood ratio test= 295  on 11.8 df,   p=0
Wald test            = 166  on 11.8 df,   p=0

> resi.m <- residuals(fit.cox.f, type='mart')
> resi.cs <- db2$STATO-resi.m
> s.res <- survfit(Surv(resi.cs, db2$STATO)~1)
> plot(s.res$time, -log(s.res$surv), type='s', main='Res. Cox-
Snell',
+ xlab='res', ylab='H(res)')
> lines(s.res$time,s.res$time)

# Residui di devianza
> resi.d <- residuals(fit.cox.f, type='dev')
> plot(resi.d, main='Residui di devianza', xlab='indice',
ylab='Residui')
> identify(resi.d)
```

```
# Test log-rapporto di verosimiglianza per verificare la
# significatività dell'effetto frailty

> #Modello completo:
> fit.coxme.f
> #Modello ridotto
> fit.coxph <- coxph(Surv(db2$DURATA,db2$STATO)~as.factor(ct2),
method='breslow')

> fit.coxme.f$loglik[1:2]
      NULL Integrated
 -4460.409  -4328.467

# Statistica test:
> ll <- fit.coxme.f$loglik[1:2] - fit.coxph$loglik

# I gradi di libertà della distribuzione della statistica test è il
# numero di effetti casuali, in questo caso c'è solo l'effetto del
# dipendente.

> ll
      NULL Integrated
  0.000000   3.313447

> 1-pchisq(2*ll[2],1)
Integrated
0.01004502
> #Effetti frailty
> fit.coxme.f$frail
# Non vengono riportati i nomi dei dipendenti per motivi di
privacy, per i
# valori si veda la tesi
> #Intervalli di confidenza per beta
> u1 <- fit.cox.f$coef[1]+1.96*0.1449962
> u2 <- fit.cox.f$coef[2]+1.96*0.1353299
> l1 <- fit.cox.f$coef[1]-1.96*0.1449962
> l2 <- fit.cox.f$coef[2]-1.96*0.1353299

> u1
as.factor(ct2)b1
        2.071841
> u2
as.factor(ct2)ta1
        0.7746392
> l1
as.factor(ct2)b1
        1.503455
> l2
as.factor(ct2)ta1
         0.244146

# Previsioni modello senza interazione e senza effetto frailty
## Per usare survfit bisogna ricordarsi di fare un attach dei dati
detach(d)
da.usare <- data.frame(db$DURATA,db$STATO,as.factor(ct))
colnames(da.usare) <- c('t','s','c')
attach(da.usare)
fit.cox <- coxph(Surv(t,s)~c, data=da.usare,
              method='breslow')
summary(fit.cox)

rb0 <- basehaz(fit.cox, centered=F)

plot(rb0$time, exp(-rb0$hazard), type='l',lty=3, main='Modello
senza interazione', ylab='S(t)', xlab='t')
lines(rb0$time, exp(-5.0321*rb0$hazard), type='s', lty=1)
lines(rb0$time, exp(-1.7186*rb0$hazard), type='s', lty=2)
```

```
legenda <-
c('chiamata=mail','chiamata=tel,’ClienteB','chiamata=tel, no
clienteB')
legend(4,0.8,legend=legenda, lty=c(3,1,2))

# Previsioni con interazione e senza effetto frailty
detach(da.usare)
da.usare <-
data.frame(db$DURATA,db$STATO,as.factor(cl),as.factor(db$CHIAMATA))
colnames(da.usare) <- c('t','s','c','ch')
attach(da.usare)

fit.cox.int <- coxph(Surv(t,s)~as.factor(c)*as.factor(ch),
                     method='breslow', data=da.usare)
summary(fit.cox.int)

rb0 <- basehaz(fit.cox.int, centered=F)

plot(rb0$time, exp(-rb0$hazard), type='l',lty=1, main='Modello con
interazione', ylab='S(t)', xlab='t')  #funzione di sopravvivenza di
base, b0
lines(rb0$time, exp(-1.4073*rb0$hazard), type='s', lty=2)
lines(rb0$time, exp(-6.5735*rb0$hazard), type='s', lty=3)
lines(rb0$time, exp(-0.8236*rb0$hazard), type='s', lty=4)

legenda <- c('chiamata=mail, ClienteB','chiamata=mail, no
clienteB','chiamata=tel, clienteB','chiamata=tel, no clienteB')
legend(3,1,legend=legenda, lty=1:4)

#Curve di sopravvivenza con effetto frailty
```

Riferimenti bibliografici

Atzeni, Ceri, Paraboschi, Torlone (2002). *Basi di Dati: modelli e linguaggi di interrogazione.* McGraw-Hill.

Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika9I* 69, 553-566.

Terry M.Therneau and Patricia M. Grambsch (2000), *Modelling Survival Data – extending the Cox Model*, Springer .

John P. Klein e Melvin L. Moeshberger (1997), *Survival Analysis - Techniques for Censored and Truncated Data*, Springer

D. R. Cox (1972). Regression models and life-tables (with discussion), *Journal of Royal Statistical Society B*, pp. 187-220

C. A. McGilchrist e C. W. Aisbett, Regression with Frailty in Survival Analysis, *Biometrics 47*, pp. 461-466.

L. Pace e A. Salvan (2001), *Introduzione alla statistica – II Inferenza, Verosimiglianza, Modelli*, Cedam

A. Azzalini e B. Scarpa (2009), *Analisi dei dati e data mining*, Springer.

T. Di Fonzo e F. Lisi (2001), *Complementi di statistica economica. Analisi delle serie storiche uni variate*, Cleup Editrice

Altro materiale di sussidio alla tesi

Dispensa “Cox Proportional-Hazards Regression for Survival Data – Appendix to an R and S-PLUS Companion to Applied Regression”, 2002, John Fox .

Manuale di riferimento della libreria “survival” di R, 2011, T. Therneau.

Materiale didattico del corso “Analisi dei dati di durata”, Anno Accademico 2010/2011, tenuto dal Professor Adimari, Professore presso la Facoltà di Scienze Statistiche – Università degli Studi di Padova.

Materiale didattico del corso di “Analisi delle serie temporali”, Anno Accademico 2009/2010, tenuto dai Professori Masarotto e Capizzi, Professori presso la Facoltà di Scienze Statistiche – Università degli Studi di Padova.
Materiale didattico del corso di “Statistica computazionale”, Anno Accademico 2009/2010, tenuto dal Professor Masarotto, Facoltà di Scienze Statistiche – Università degli Studi di Padova.

Materiale didattico del corso di Statistica Medica tenuto dalle Professoressa Laura Ventura, Professoressa presso la Facoltà di Scienze Statistiche – Università degli Studi di Padova, seminario sui dati di sopravvivenza tenuto dalla Professoressa Stefania Galimberti, Professoressa presso la Facoltà di Medicina e Chirurgia – Università Bicocca di Milano.