Politecnico di Milano

Facoltà di Ingegneria dei Sistemi

Corso di Laurea in Ingegneria Matematica

Tesi di Laurea Specialistica

METODI STATISTICI PER L'ANALISI E L'OTTIMIZZAZIONE DEI TEMPI DI PERCORRENZA DELLE AMBULANZE NELLA PROVINCIA DI MILANO

Relatore: prof.ssa Anna Maria Paganoni
Correlatore: dott.ssa Francesca Ieva

Candidato:

Giovanni Cassarini Matr. 739361

Anno Accademico 2010-2011

Indice

Introduzione

Argomento di questo elaborato è un'analisi dei dati sui tempi di percorrenza delle ambulanze nella provincia di Milano relativi al periodo tra il 2005 e il 2008. Obiettivo di tale analisi è quello di creare un *tool* di supporto per il 118 che permetta, sulla base delle informazioni a disposizione del personale al momento del soccorso di un paziente, di stabilire quale sia il tempo di arrivo previsto in ogni ospedale e di effettuare, conseguentemente, una comparazione tra i valori stimati al fine di scegliere la destinazione più velocemente raggiungibile.

L'elaborato è suddiviso in 4 capitoli. Nel primo di essi viene riportata una descrizione dei dati a disposizione e vengono introdotte e spiegate e le covariate che sono state ritenute di interesse al fine di prevedere il tempo medio di arrivo in una data struttura ospedaliera.

Il secondo capitolo tratta invece la teoria riguardante i modelli lineari e, in particolare, quella relativa ai modelli lineari generealizzati. Tale argomento è infatti alla base della costruzione del modello di previsione proposto. Il capitolo termina con una spiegazione riguardante le tecniche di trasformazione dei dati e come esse siano state applicate al problema in esame.

Nel terzo capitolo vengono inizialmente riportati, a titolo di esempio, il procedimento per la stima dei parametri del modello e i risultati ottenuti per l'ospedale Niguarda di Milano. Successivamente viene riportato un riassunto dei risultati ottenuti ripetendo il procedimento proposto per ciascuna struttura ospedaliera oggetto dell'analisi. Il capitolo si conclude con una sezione riguardante il funzionamento e le specifiche tecniche, come i parametri in ingresso e l'interpretazione dell'output, dell'applicazione proposta.

Infine, nel quarto capitolo viene riportato il codice utilizzato per la crea-

zione del modello di previsione proposto e per l'implementazioe del *tool* di supporto al personale 118, obiettivo di questo lavoro. Per la realizzazione della parte di analisi dei dati e della parte informatica di questo elaborato è stato utilizzato il software `R` (fare riferimento a [7]). A tale sofware fa pertanto riferimento il codice riportato in questo capitolo conclusivo.

Capitolo 1

Descrizione dei Dati

I dati a disposizione per lo sviluppo di questo lavoro sono stati forniti dalla Centrale Operativa del 118 (per informazioni fare riferimento a [8]) e riguardano le chiamate effettuate nella provincia di Milano nel periodo compreso tra il 01/01/2005 e il 31/07/2008. Il database fornito contiene 1 585 654 record. Ogni "uscita" di un'ambulanza è rappresentata da due record dei quali il primo riguarda il percorso tra il punto di attesa dell'ambulanza e il luogo in cui deve avvenire il soccorso, mentre il secondo riguarda il percorso dal luogo del soccorso all'ospedale al quale l'ambulanza è stata indirizzata dagli operatori del 118. Obiettivo di questo lavoro è quello di costruire un modello di previsione che permetta di stimare il tempo medio necessario per arrivare da un qualsiasi punto della provincia di Milano a una data struttura ospedaleria, al fine di supportare la decisione degli operatori al momento di indirizzare l'ambulanza all' ospedale di destinazione. Una volta costruito tale modello sarà infatti possibile creare un'applicazione che permetta, tramite una procedura semiautomatica, di effettuare una comparazione tra i tempi di arrivo previsti per ogni ospedale ed avere quindi a disposizione un criterio decisionale di tipo sistematico per la scelta della destinazione più velocemente raggiungibile.

È importante sottolineare che la scelta dell'ospedale di destinazione per il paziente viene effettuata nel momento del soccorso dello stesso, per questa ragione, all'interno del database a disposizione, le unità statistiche che

saranno selezionate ai fini dell'analisi corrispondono ai record riguardanti il percorso dal luogo del soccorso all'ospedale di arrivo.

1.1 Ospedali

In questa sezione saranno presentati i dettagli riguardanti le strutture ospedaliere oggetto dell'analisi. Gli ospedali presenti nel database a disposizione sono raffigurati, coerentemente con la loro posizione geografica, in Figura 1.1.

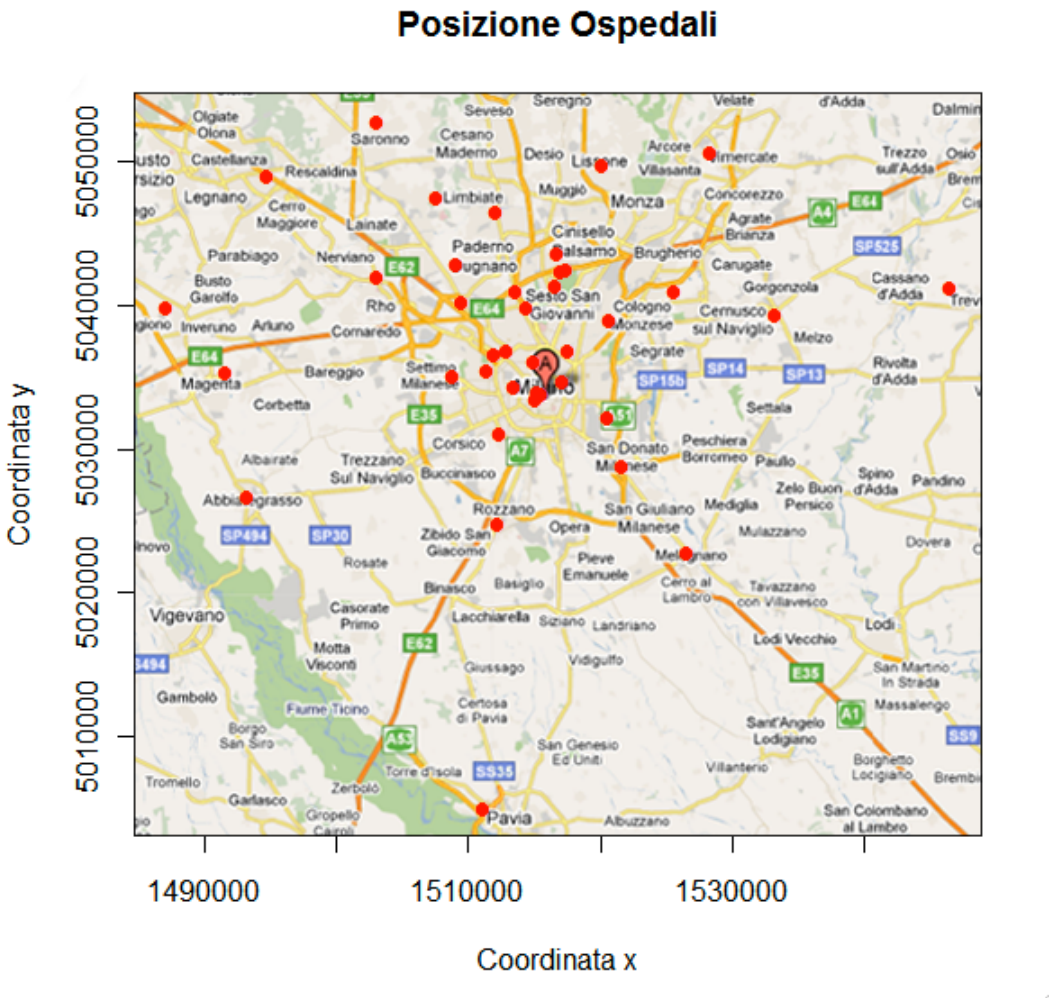


Figura 1.1: Posizione geografica delle strutture ospedaliere oggetto dell'analisi

Per ciascuna struttura ospedaliera sono inoltre riportate in Tabella 1.1 le informazioni rilevanti ai fini dell'analisi, ovvero: il nome dell'ospedale, il codice identificativo, le coordinate geografiche e il numero di arrivi totali.

Il sistema di coordinate geografiche utilizzato è quello di Gauss-Boaga. Tale sistema è del tutto comparabile a un sistema di riferimento cartesiano in cui l'origine è fissata nel punto di incontro tra il meridiano di riferimento, che è localizzato 9° a Est di Greenwich, e l'Equatore e in cui l'unità di misura è il metro. Per convenzione alla coordinata Est è sommato il valore 1 500 000. Il tema della scelta del sistema di coordinate geografiche e del passaggio da un sistema all'altro non verrà approfondito in quanto al di fuori degli scopi principali di questo lavoro. Per ulteriori approfondimenti si veda **??**.

Per chiarezza è comunque bene sottilineare che esistono software che permettono il passaggio da un sistema all'altro e in particolare a quello delle coordinate satellitari utilizzate dai comuni navigatori. Di seguito si farà comunque sempre riferimento al sistema di coordinate di Gauss-Boaga.

Ospedale	**Codice**	**Coordinata x**	**Coordinata y**	**n° arrivi**
Abbiategrasso	1	1493255	5026770	9122
Alfieri-Mangiagalli	25	1515525	5033884	2992
Bignami	26	1516673	5041435	2688
Bollate	3	1509118	5042955	6259
Cardiologico	110	1520601	5032300	3645
Cernusco sul Naviglio	9	1525654	5041081	18797
Cinisello Balsamo	10	1516785	5043685	20064
Cuggiono	12	1487080	5039920	1625
De Marchi	119	1515575	5033958	3457
Fatebenefratelli	24	1514936	5036209	46543
Gaetano Pini	27	1515186	5033538	11967

Ospedale	**Codice**	**Coordinata x**	**Coordinata y**	**n° arrivi**
Galeazzi	37	1513615	5041076	4183
Garbagnate Milanese	14	1507548	5047567	18008
Humanitas	2227	1512265	5024899	27240
Legnano	17	1494713	5049120	596
Magenta-Fornaroli	19	1491631	5035367	20857
Macedonio Melloni	32	1517153	5034839	3073
Melegnano-Predabissi	21	1526649	5022907	18667
Melzo	22	1533304	5039480	15238
Monza	41	1520229	5049857	624
Multimedica s.p.a.	2185	1517054	5042494	7650
Niguarda Ca'granda	29	1514452	5039975	54533
Paderno Dugnano	45	1512143	5046610	7999
Pavia	47	1511099	5005029	840
Policlinico	23	1515307	5033987	52028
Rho-Circolo	48	1503099	5042070	20131
Sacco	36	1509506	5040352	39197
San Carlo	35	1508886	5035217	58265
San Donato Milanese	50	1521645	5028816	12048
San Giuseppe	118	1513446	5034416	8884
San Luca	153	1511426	5035479	2140
San Paolo	34	1512375	5031095	51853
San Raffaele	33	1520727	5039052	35904
Santa Rita	112	1517583	5036895	28193
Sant'Ambrogio	144	1511942	5036706	1303
Saronno	85	1503131	5052879	18311
Sesto San Giovanni	51	1517478	5042618	12186
Treviglio-Caravaggio	53	1546496	5041279	957
Vimercate	57	1528360	5050720	643
Vittore Buzzi	28	1512971	5036928	2882

Tabella 1.1: Dati rilevanti per le strutture ospedaliere oggetto dell'analisi

1.2 Scelta del Modello

In questa sezione sarà trattato il processo di selezione delle variabili atto a identificare per ciascuna struttura ospedaliera oggetto dell'analisi un opportuno modello che metta in relazione le covariate ritenute maggiormente rilevanti con il tempo medio impiegato da un'ambulanza per arrivare da un qualsiasi punto della provincia di Milano alla struttura in questione.

All'interno del database iniziale per ogni unità statistica sono registrate le informazioni riportate in Tabella 1.2.

Nome variabile	**Descrizione**
ID Missione	Codice numerico identificativo per l'"uscita" dell'ambulanza
ID Tratta	Codice numerico identificativo del percorso seguito
DT Partenza	Data e Ora di partenza
DT Arrivo	Data e Ora di arrivo
VL Rif X, Y	Coordinate di Gauss-Boaga del punto di partenza
DS Comune	Comune nel quale viene effettuato il soccorso
DS LG Dest	Ospedale di destinazione
ID LG Dest	Codice numerico identificativo dell'ospedale di destinazione
ID Codice	Codice identificativo per la gravità delle condizioni del paziente
DS Motivo	Motivo della chiamata (incidente, caduta, aggressione...)
DS Det Motivo	Luogo del soccorso (veicolo fermo, scala, suolo...)

Tabella 1.2: Informazioni presenti nel database iniziale

Ai fini della costruzione del modello che è stato prefissato come obiettivo di questo elaborato è stata inizialmente effettuata una selezione delle variabili che sono state ritenute di maggiore interesse, ovvero quelle il cui effetto può risultare statisticamente significativo nell'influenzare il valore medio del tempo impiegato per raggiungere un dato ospedale da un qualsiasi punto nella provincia di Milano. Di seguito sono riportate le covariate selezionate.

Codice Paziente: fattore a 3 livelli (Verde, Giallo, Rosso) che rappresenta la gravità delle condizioni del paziente che deve essere trasportato. Il Codice è attribuito al momento della partenza dell'ambulanza dall'operatore del 118 sulla base del quadro clinico emerso dalla chiamata; è quindi successivamente confermato o modificato dal personale del 118 che si reca sul luogo del soccorso. Ai fini dell'analisi è stato considerato di interesse il Codice assegnato in questo secondo momento, poichè è sulla base di esso che si stabilisce l'urgenza con la quale il paziente sarà trasportato in ospedale. Per ogni unità statistica il valore assunto dal fattore Codice corrisponde alla variabile "ID Codice" presente nel database a disposizione, come riportato in Tabella 1.2.

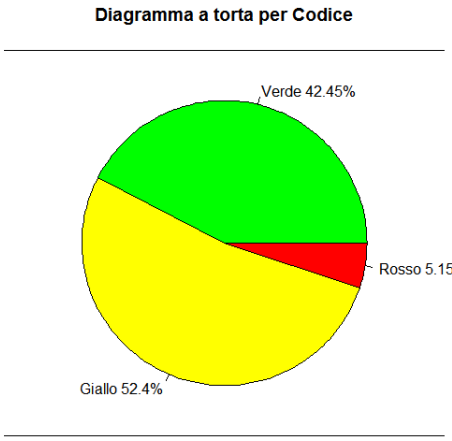
Dai grafici riportati in Figura 1.2 è possibile dedurre le seguenti osservazioni:

1. I dati relativi a pazienti con Codice Rosso hanno numerosità molto minore, in proporzione, rispetto a quelli con Codice Verde e Giallo (Figura 1.2.1).

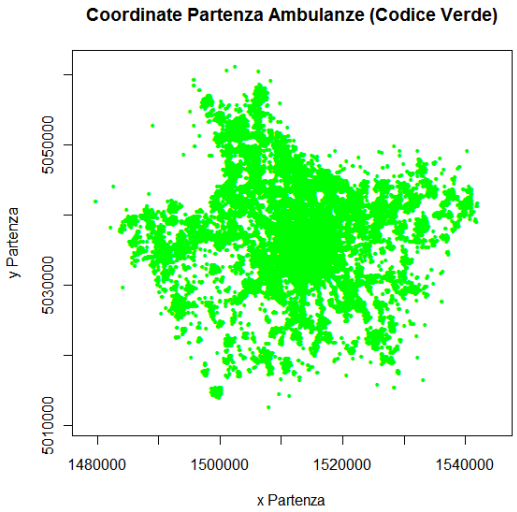
2. Le chiamate sono distribuite in modo omogeneo su tutta la zona geografica di interesse indipendentemente dal Codice assegnato dagli operatori del 118 (Figure 1.2.2, 1.2.3 e 1.2.4).

3. La distribuzione della variabile Tempo presenta una coda destra molto pesante (Figura 1.2.5). Questo fatto è spiegabile prima di tutto osservando che la variabile è solo positiva e la sua media è relativamente vicina a 0, in secondo luogo osservando che le ambulanze possono essere state inviate, per motivi che eswulano dalle sole considerazioni inerenti la distanza, a un ospedale anche molto distante dal luogo del soccorso.

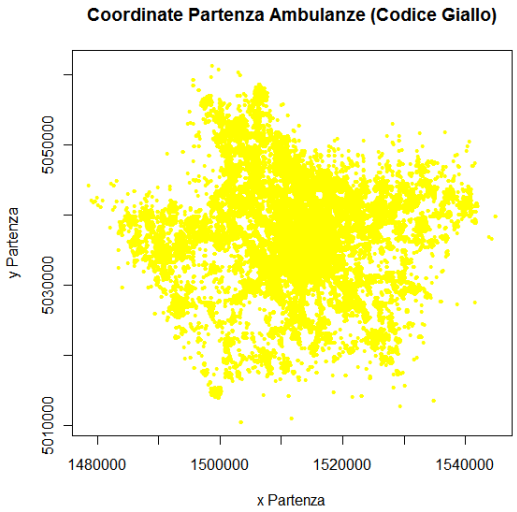
4. Come si poteva ipotizzare, il Tempo medio di trasporto in ospedale diminuisce all'aumentare della gravità delle condizioni del paziente trasportato (Figura 1.2.6). Tale supposizione viene confermata dal risultato del test non parametrico di Kruskal-Wallis per la verifica dell'ipotesi nulla che prevede che la variabile risposta (Tempo) non sia significativamente influenzata dai valori assunti dalla variabile categorica Codice (p-value $\leq 2 \cdot 10^{-16}$).



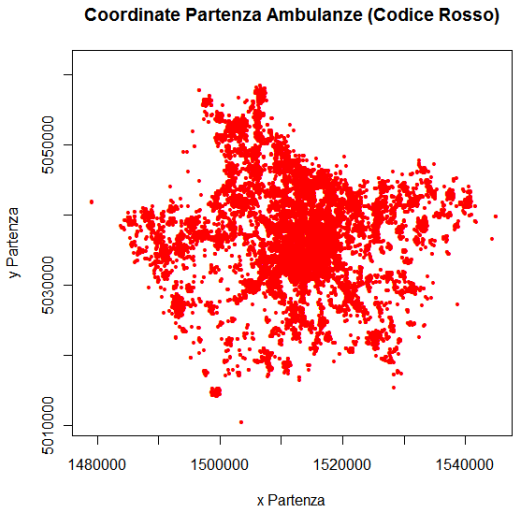
(1.2.1)



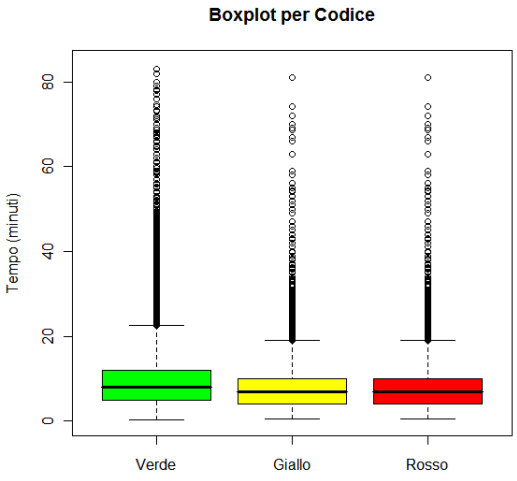
(1.2.2)



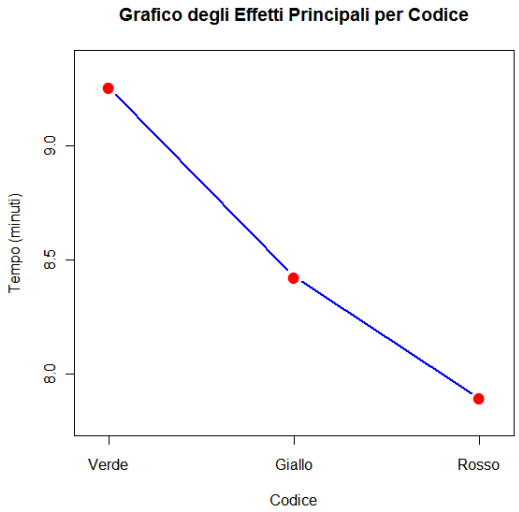
(1.2.3)



(1.2.4)



(1.2.5)



(1.2.6)

Figura 1.2: (1.2.1) Diagramma a torta per la distribuzione percentuale della variabile Codice; (1.2.2), (1.2.3), (1.2.4) Raffigurazione dei punti di partenza delle ambulanze rispettivamente per il Codice Verde, Giallo e Rosso; (1.2.5) Boxplot dei tempi intercorsi dalla partenza dell'ambulanza dal luogo del soccorso all'arrivo in ospedale rispettivamente per il Codice Verde, Giallo e Rosso; (1.2.6) Grafico degli Effetti Principali per la variabile Codice.

Tipo di Giorno: fattore a 2 livelli (Feriale, Festivo). I giorni considerati Festivi sono le Domeniche più le festività annuali riportate in Tabella 1.3. Tale distinzione, non essendo presente nel database a disposizione, è stata effettuata considerando la data di partenza delle ambulanze dal luogo del soccorso (variabile "DT Partenza" in Tabella 1.2) tramite l'utilizzo del codice `R` riportato in Sezione 4.1.

Data	Festività
1 Gennaio	Capodanno
6 Gennaio	Epifania
25 Aprile	Festa della Liberazione
1 Maggio	Festa dei lavoratori
2 Giugno	Festa della Repubblica
15 Agosto	Assunzione
8 Dicembre	Immacolata Concezione
25 Dicembre	Natale
26 Dicembre	Santo Stefano
31 Dicembre	San Silvestro

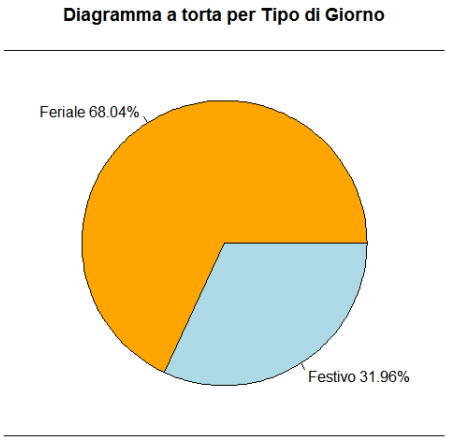
Tabella 1.3: Elenco delle festività

Dall'analisi dei grafici riportati in Figura 1.3 si osserva che:

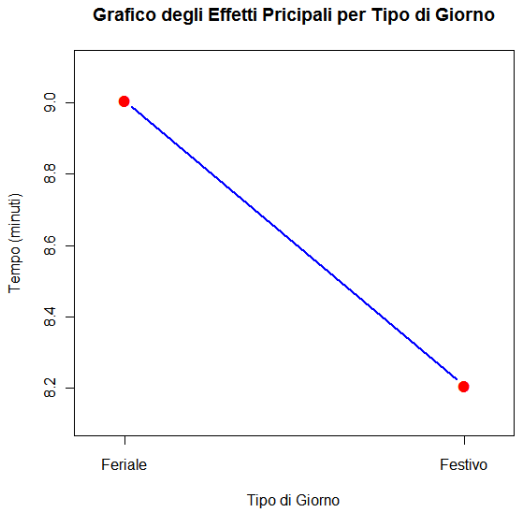
1. I dati relativi ai giorni Festivi hanno minore numerosità rispetto ai giorni Feriali (Figura 1.3.1).

2. Il Tempo medio impiegato per raggiungere un qualsiasi ospedale diminuisce se il giorno è Festivo (Figura 1.3.2). Tale supposizione è confermata dal risultato del test di Wilcoxon (p-value $\leq 2 \cdot 10^{-16}$). Questo fatto si spiega osservando che la diminuzione del traffico in seguito a una festività può comportare un aumento della velocità del trasporto in ospedale del paziente.

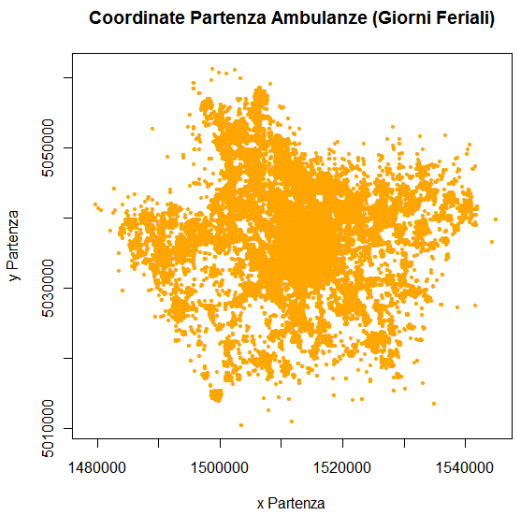
3. Le chiamate sono distribuite in modo omogeneo su tutta la zona geografica di interesse indipendentemente dal fatto che esse siano state effettuate in giorni Feriali piuttosto che Festivi (Figure 1.3.3 e 1.3.4)



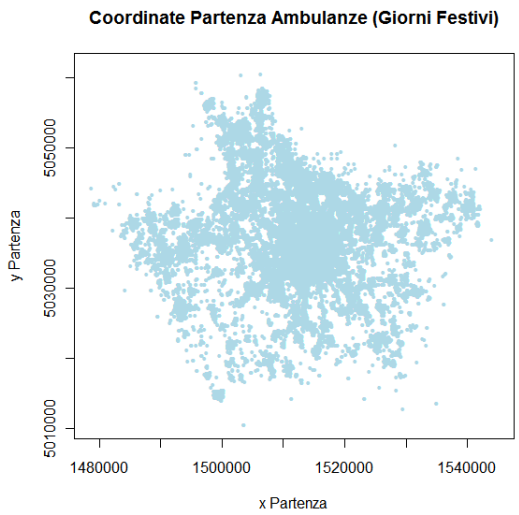
(1.3.1)



(1.3.2)



(1.3.3)



(1.3.4)

Figura 1.3: (1.3.1) Diagramma a torta per la distribuzione percentuale della variabile Tipo di Giorno; (1.3.2) Grafico degli Effetti Principali per la variabile Tipo di Giorno; (1.3.3), (1.3.4) Raffigurazione dei punti di partenza delle ambulanze rispettivamente per giorni Feriali e Festivi.

Fascia Oraria: fattore a 3 livelli (Ore di Punta, Giorno, Notte). Come nel caso precedente la motivazione alla base dell'introduzione di questa covariata risiede nel fatto che la diminuzione del traffico in determinate ore della giornata potrebbe influenzare in modo significativo il tempo medio necessario per l'arrivo in un dato ospedale. In questo caso è stato necessario effettuare una scelta per la distinzione tra le varie fasce orarie. Il riferimento adottato per eseguire tale distinzione è stato l'orario dei mezzi pubblici milanesi:

- Ore di punta: $07.00/09.00 - 17.00/20.00$
- Giorno: $09.00/17.00$
- Notte: $00.00/07.00 - 20.00/24.00$

I valori assunti dal fattore Fascia Oraria sono stati assegnati a ogni unità statistica a partire dall'ora di partenza dell'ambulanza dal luogo del soccorso (variabile "DT Partenza" in Tabella 1.2) tramite l'utilizzo del codice `R` riportato in sezione 4.1.

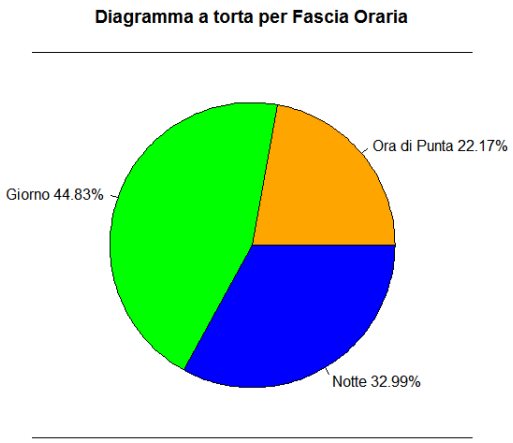
Vengono riportati in Figura 1.4 alcuni grafici finalizzati alla descrizione delle proprietà del fattore Fascia Oraria. Dall'analisi di tali grafici è possibile dedurre le seguenti osservazioni:

1. La numerosità dei dati al variare dei livelli del fattore Fascia Oraria non è costante e non è proporzionale al numero di ore dal quale i livelli stessi sono identificati (Figura 1.4.1). Come infatti è logico aspettarsi nelle ore notturne, rispetto alle ore diurne, si ha una diminuzione della frequenza delle chiamate al 118.

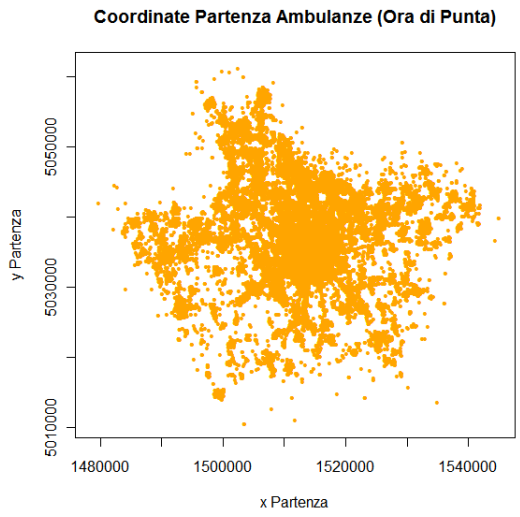
2. Le chiamate sono distribuite in modo omogeneo su tutta la zona geografica di interesse indipendentemente dal valore assunto dalla variabile Fascia Oraria (Figure 1.4.2, 1.4.3 e 1.4.4).

3. Analogamente a quanto osservato per la variabile Codice la distribuzione dela variabile Tempo presenta una coda destra molto pesante per qualsiasi livello del fattore Fascia Oraria (Figura 1.4.5).

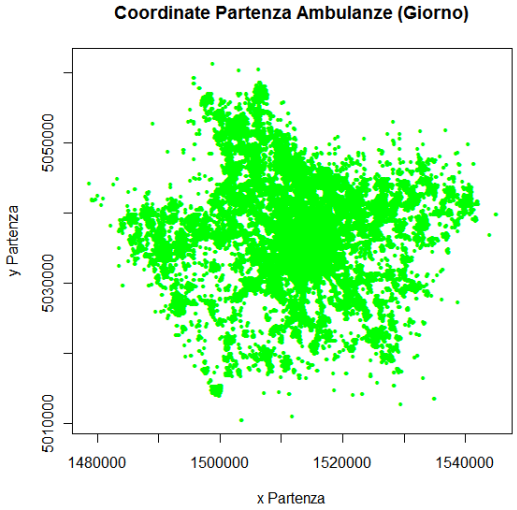
4. Il tempo medio impiegato dalle ambulanze per raggiungere una qualsiasi struttura ospedaliera diminuisce nelle ore notturne rispetto a quelle diurne e aumenta nelle ore di punta (Figura 1.4.6). Tale osservazione conferma la supposizione iniziale che ha portato all'introduzione della covariata Fascia Oraria. A questo proposito si riporta il risultato del test di Kruskal-Wallis relativo alla verifica dell'ipotesi nulla secondo la quale la variabile risposta (Tempo) non è significativamente influenzata dal fattore Fascia Oraria. Il p-value ottenuto è minore di $2 \cdot 10^{-16}$ e conferma pertanto quanto ipotizzato inizialmente.



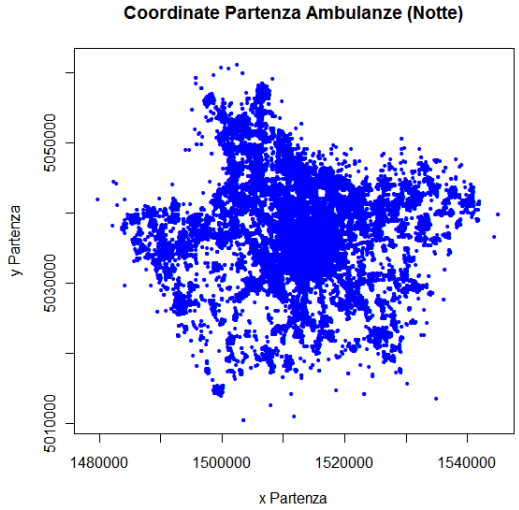
(1.4.1)



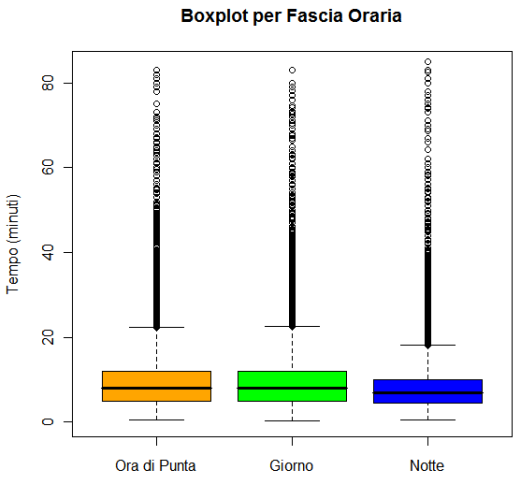
(1.4.2)



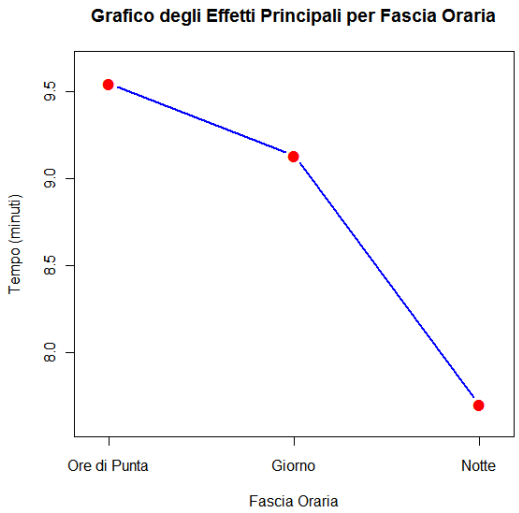
(1.4.3)



(1.4.4)



(1.4.5)



(1.4.6)

Figura 1.4: (1.4.1) Diagramma a torta per la distribuzione percentuale della variabile Fascia Oraria; (1.4.2), (1.4.3), (1.4.4) Raffigurazione dei punti di partenza delle ambulanze rispettivamente per i livelli Ore di Punta, Giorno e Notte assunti dalla variabile Fascia Oraria; (1.4.5) Boxplot dei tempi intercorsi tra la partenza dell'ambulanza dal luogo del soccorso all'arrivo in ospedale rispettivamente per i livelli Ore di Punta, Giorno e Notte assunti dalla variabile Fascia Oraria; (1.4.6) Grafico degli Effetti Principali per la variabile Fascia Oraria.

Zona: Fattore a 4 livelli (Nord-Ovest, Nord-Est, Sud-Est, Sud-Ovest). Questo fattore si riferisce alla posizione del luogo del soccorso rispetto all'ospedale al quale il paziente sarà trasportato e viene introdotto per tenere conto dell'eventuale significatività del fatto che il traffico su Milano può non essere omogeneo e, di conseguenza, può non esserlo la velocità media di un'ambulanza che proviene da una zona piuttosto che da un'altra. In altre parole, introducendo l'idea di un campo vettoriale per la rappresentazione del traffico nella zona di interesse, la covariate Zona è introdotta per spiegare la variabilità della risposta (Tempo) dovuta all'anisotropia del campo.

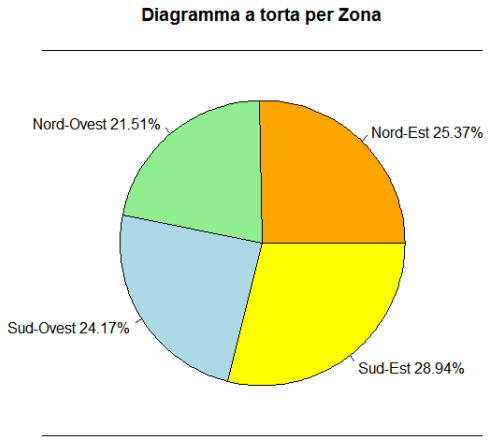
A titolo di esempio si prenda un ospedale che abbia la tangenziale a Ovest e le vie cittadine a Est, in questo caso si può immaginare che, a parità di altre condizioni, il tempo medio di arrivo a questo ospedale sia minore nelle zone dove si ha la presenza della tangenziale e che, di conseguenza, il fattore Zona possa risultare significativo.

I valori assunti dal fattore Zona sono stati assegnati a ogni unità statistica a partire dalle coordinate del punto di partenza dell'ambulanza dal luogo del soccorso (variabili "VL Rif X" "VL Rif Y" in Tabella 1.2) tramite l'utilizzo del codice `R` riportato in sezione 4.1.

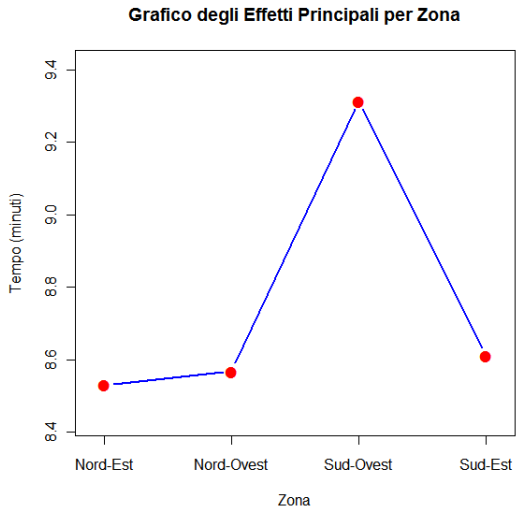
Dall'osservazione dei grafici riportati in Figura 1.5 è possibile effettuare le seguenti considerazioni:

1. Non vi sono differenze sostanziali tra le numerosità dei dati relativi a ogni livello della covariata Zona (Figura 1.5.1).

2. Il tempo medio impiegato per raggiungere una qualsiasi struttura ospedaliera non appare influenzato in modo significativo dal valore assunto dal fattore Zona (Figura 1.5.2). Questo fatto è confermato dal risultato del test di Kruskal-Wallis il cui p-value è pari a 0.4998. Nonostante ciò si è comunque deciso di includere nel modello la variabile Zona in quanto essa potrebbe risultare significativa nello spiegare la variabilità della risposta per quanto riguarda l'analisi degli ospedali considerati singolarmente.



(1.5.1)



(1.5.2)

Figura 1.5: (1.5.1) Diagramma a torta per la distribuzione percentuale della variabile Zona; (1.5.2) Grafico degli Effetti Principali per la variabile Zona.

Distanza: Covariata di tipo continuo. È naturalmente ipotizzabile che all'aumentare della distanza aumenti il tempo medio di trasporto in ospedale. Per ogni unità statistica è stata quindi stimata, tramite il sito www.tuttocitta.com, la distanza che è necessario percorrere sul tragitto cittadino che va dal luogo del soccorso all'ospedale di destinazione dell'ambulanza. Sebbene non sia possibile verificare che i percorsi effettuati dalle ambulanze siano effettivamente quelli scelti dal sito appena citato, questo procedimento garantisce di ottenere una stima adeguatamente precisa della lunghezza effettiva del tragitto in esame.

Il codice `R` che, a partire dalle coordinate geografiche del punto di partenza e di quello d'arrivo (variabili "VL Rif X", "VL Rif Y" e "DS LG Dest" del database a disposizione), permette di ottenere la stima della distanza che intercorre tra il luogo del soccorso e l'ospedale di destinazione è riportato in Sezione **??**.

1.3 Selezione dei Dati

Al fine della costruzione del modello di previsione obiettivo di questa analisi, i dati ritenuti di interesse sono quelli relativi al percorso tra il luogo dove è avvenuto il soccorso e l'ospedale al quale il paziente è stato trasportato. È inoltre risultato necessario operare una selezione dei dati eliminando inizialmente quelli privi di tutte le informazioni, considerate irrinunciabili ai fini dell'analisi, riportate nella sezione precedente. Successivamente si è resa necessaria l'eliminazione di tutti quei dati considerati privi di senso, in quanto affetti da errori di digitazione o di inserimento nel sistema da parte degli operatori del 118. Nel complesso, i dati eliminati sono quelli connotati dalle seguenti caratteristiche:

- tragitti per i quali l'orario di partenza è successivo a quello di arrivo;
- tragitti per i quali la velocità media tenuta dall'ambulanza è superiore ai 100 km/h o inferiore ai 5 km/h;
- tragitti per i quali le coordinate dell'ospedale di destinazione non corrispondono a quelle reali;
- tragitti il cui punto di partenza è molto distante dai confini della provincia di Milano. Tali dati sono presenti in numero molto basso rispetto al totale. Inoltre, se un'ambulanza viene mandata in un ospedale molto lontano dal luogo del soccorso, si può supporre che le motivazioni alla base della scelta dell'ospedale di arrivo siano altre rispetto al tempo medio di percorrenza della tratta (ad esempio la presenza di particolari strutture o équipe mediche specializzate). Non rivestirebbe pertanto particolare interesse effettuare una stima del tempo necessario a raggiungere un dato ospedale da una posizione di partenza di questo genere, in quanto in ogni caso tale stima non sarebbe utilizzata come criterio per la scelta della destinazione.

Al termine di tale processo di selezione il database consiste di 651 592 unità statistiche, ciascuna corrispondente al tragitto di un'ambulanza dal luogo del soccorso a un dato ospedale.

Prima di procedere è necessario precisare che il criterio della minimizzazione del tempo impiegato a raggiungere l'ospedale non è l'unico sulla base del quale viene effettuata la scelta della destinazione da parte degli operatori del 118. Vi sono infatti condizioni dalle quali non è possibile prescindere come, per esempio, la disponibilità dei posti letto o la presenza di reparti attrezzati per la cura del paziente che si deve trasportare. Tali elementi hanno fatto sì che all'interno del database a disposizione vi sia la presenza di dati facenti riferimento ad ambulanze che, pur partendo da una stessa zona e presentando gli stessi valori per i fattori considerati, sono state inviate verso diversi ospedali di destinazione. Questo aspetto riveste una notevole importanza in quanto ha permesso di avere le informazioni necessarie per effettuare una comparazione di carattere statistico tra i vari tempi di arrivo nelle diverse strutture ospedaliere.

Deve inoltre essere sottilineato che la procedura semiautomatica sviluppata in questo elaborato mira al supporto decisionale per ciò che concerne l'ottimizzazione dei tempi di percorrenza in funzione della distanza. Pertanto sarà necessario integrarla in unsistema evoluto, che tenga conto anche di fattori quali la disponibilità dei posti letto, il sovraccarico dei prontosoccorsi e altre variabili, tra cui quelle citate in precedenza.

Capitolo 2

Modelli Lineari

Obiettivo di questo lavoro, come detto in precedenza, è quello di costruire un modello di previsione che permetta di stimare il tempo medio impiegato da un'ambulanza a raggiungere un dato ospedale da un qualsiasi punto nella provicia di Milano. Al fine di raggiungere tale obiettivo si propone di costruire un modello lineare in cui la risposta sia la variabile Tempo e in cui le covariate siano quelle presentate nel Capitolo 1, con l'aggiunta di alcune interazioni il cui significato sarà spiegato in seguito.

2.1 Il Modello Lineare Classico

Per modello lineare si intende un modello atto a studiare la dipendenza in media di una variabile aleatoria Y da k variabili non casuali $X_1 \dots X_k$. La variabile Y si suppone di tipo quantitativo mentre le *variabili esplicative* X_j $(j = 1 \dots k)$ possono essere sia di tipo qualitativo che quantitativo. In questa sezione sarà introdotto inizialmente il metodo per la stima dei parametri di un modello con la sola presenza di variabili quantitative; sarà successivamente spiegato come generalizzarlo nel caso vi sia anche la presenza di variabili di tipo qualitativo.

Data allora una serie di n realizzazioni indipendenti $x_{1i} \dots x_{1k}$, $(i = 1 \dots n)$ delle variabili esplicative $X_1 \dots X_k$ l'obiettivo è quello di stimare i parametri della seguente relazione:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i \qquad i = 1 \ldots n$$

dove ε_i è una variabile casuale detta spesso "errore" o "perturbazione" tale che:

- $\mathbb{E}[\varepsilon_i] = 0 \qquad \forall i = 1 \ldots n$
- $\mathbb{V}\text{ar}[\varepsilon_i] = \sigma^2 \qquad \forall i = 1 \ldots n$ (omoschedasticità)
- $\mathbb{C}\text{ov}[\varepsilon_i, \varepsilon_j] = 0 \qquad \forall i, j = 1 \ldots n \quad i \neq j$ (scorrelatezza degli errori)

Queste n uguaglianze possono essere scritte in forma compatta utilizzando la seguente notazione matriciale:

$$\mathbf{X} = \begin{pmatrix} x_{01} & x_{11} & \ldots & x_{k1} \\ x_{02} & x_{12} & \ldots & x_{k2} \\ \ldots & \ldots & \ldots & \ldots \\ x_{0n} & x_{1n} & \ldots & x_{kn} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{k1} \\ 1 & x_{12} & \ldots & x_{k2} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{1n} & \ldots & x_{kn} \end{pmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_n \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_2 \\ \ldots \\ \beta_k \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{bmatrix}$$

Con tale notazione le n relazioni precedenti assumono la forma:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

dove:

- $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$
- $\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sigma^2 \mathbf{I}_n$

Al fine di stimare le $k+1$ componenti del vettore $\boldsymbol{\beta}$ due sono i metodi più frequentemente utiizzati: il metodo dei Minimi Quadrati e, nel caso in cui si conosca la distribuzione della variabile casuale ε, il metodo della Massima Verosimiglianza.

2.1.1 Metodo dei Minimi Quadrati

Sia $\hat{\boldsymbol{\beta}}$ uno stimatore di $\boldsymbol{\beta}$. Si definisce allore *vettore dei residui* il vettore:

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

Il criterio dei Minimi Quadrati prevede che gli stimatori $\beta_0 \dots \beta_k$ ottimali siano quelli che minimizzano la somma dei quadrati dei residui, data da:

$$\sum_{i=1}^{n} e_i^2 = \mathbf{e'e} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{Y'Y} + \hat{\boldsymbol{\beta}}'(\mathbf{X'X})\hat{\boldsymbol{\beta}} - 2\hat{\boldsymbol{\beta}}'\mathbf{X'Y}$$

Derivando la precedenta relazione e annullando tale derivata si ottiene:

$$\frac{\partial}{\partial \hat{\boldsymbol{\beta}}} = 2(\mathbf{X'X})\hat{\boldsymbol{\beta}} - 2\mathbf{X'Y} = \mathbf{0}$$

da cui:

$$(\mathbf{X'X})\hat{\boldsymbol{\beta}} = \mathbf{X'Y}$$

Nell'ipotesi che la matrice $\mathbf{X}$ abbia rango pieno si ha dunque l'espressione dello stimatore ai minimi quadrati $\hat{\boldsymbol{\beta}}$ per il vettore $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

Si osservi che $\hat{\boldsymbol{\beta}}$ è una funzione lineare nelle variabili $Y_1, \dots, Y_n$. Esso gode inoltre delle seguenti proprietà:

Proprietà 1 *$\hat{\boldsymbol{\beta}}$ è uno stimatore non distorto per $\boldsymbol{\beta}$.*

Infatti:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'}(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X'X})^{-1}\boldsymbol{X'\varepsilon}$$

ma $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$, quindi

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] = \mathbb{E}\left[(\boldsymbol{\beta} + (\mathbf{X'X})^{-1}\boldsymbol{X'\varepsilon})\right] = \boldsymbol{\beta} + (\mathbf{X'X})^{-1}\boldsymbol{X'}\mathbb{E}[\boldsymbol{\varepsilon}] = \boldsymbol{\beta}$$

Proprietà 2 *La matrice di varianza e covarianza dell'errore (che indicheremo con $\Sigma_{\hat{\boldsymbol{\beta}}}$) è uguale a $\sigma^2(\boldsymbol{X'X})^{-1}$*

Infatti:

$$\begin{aligned}
\Sigma_{\hat{\boldsymbol{\beta}}} &= \mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right] = \mathbb{E}\left[(\mathbf{X'X})^{-1}\boldsymbol{X'\varepsilon\varepsilon'X}(\mathbf{X'X})^{-1}\right] \\
&= (\mathbf{X'X})^{-1}\mathbf{X'}\left(\mathbb{E}[\boldsymbol{\varepsilon}\,\boldsymbol{\varepsilon'}]\right)\mathbf{X}(\mathbf{X'X})^{-1} = (\mathbf{X'X})^{-1}\mathbf{X'}(\sigma^2\mathbf{I}_n)\mathbf{X}(\mathbf{X'X})^{-1} \\
&= \sigma^2\,(\mathbf{X'X})^{-1}\mathbf{X'X}(\mathbf{X'X})^{-1} = \sigma^2\,(\mathbf{X'X})^{-1}
\end{aligned}$$

Proprietà 3 *La somma dei residui associati al modello è nulla, ovvero:*

$$\sum_{i=1}^{n} e_i = 0$$

Infatti:

$$\begin{aligned}
\frac{\partial}{\partial\hat{\beta}_0}(\mathbf{e'e}) &= \frac{\partial}{\partial\hat{\beta}_0}\left[\sum_{i=1}^{n} e_i^2\right] = \frac{\partial}{\partial\hat{\beta}_0}\left[\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \ldots - \hat{\beta}_k x_{ki})^2\right] \\
&= -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \ldots - \hat{\beta}_k x_{ki}) = -2\sum_{i=1}^{n} e_i = 0
\end{aligned}$$

Proprietà 4 $\mathbb{E}[\boldsymbol{e'e}] = (n - k - 1)\sigma^2$

Infatti si ha:

$$\begin{aligned}
\mathbf{e} &= (\boldsymbol{Y} - \boldsymbol{X\beta}) = \mathbf{Y} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'Y} \\
&= \mathbf{X\beta} + \boldsymbol{\varepsilon} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}(\mathbf{X\beta} + \boldsymbol{\varepsilon}) \\
&= \boldsymbol{\varepsilon} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}\boldsymbol{\varepsilon} \\
&= \left[\mathbf{I}_n - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}\right]\boldsymbol{\varepsilon} = \mathbf{V}\boldsymbol{\varepsilon}
\end{aligned}$$

Si può inoltre dimostrare che la matrice $\mathbf{V}$ è simmetrica e idempotente, ovvero $\mathbf{V} = \mathbf{V'} = \mathbf{V}^2$. Pertanto si ha:

$$\begin{aligned} \mathbf{e'e} &= \boldsymbol{\varepsilon}'\mathbf{V'V}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}'\mathbf{V}^2\boldsymbol{\varepsilon} \\ &= \boldsymbol{\varepsilon}'\mathbf{V}\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}'\mathbf{Q})\,\boldsymbol{\Lambda}\,(\mathbf{Q'}\boldsymbol{\varepsilon}) \end{aligned}$$

dove $\mathbf{Q}$ è una matrice ortogonale e $\boldsymbol{\Lambda}$ la matrice diagonale avente sulla diagonale principale gli autovalori della matrice $\mathbf{V}$.

Si consideri ora la trasformazione della v.a. multivariata $\mathbf{U} = \mathbf{Q'}\boldsymbol{\varepsilon}$. Sulla distribuzione di probabilità di $\mathbf{U}$ non è possibile dire nulla poichè niente è stato detto sulla distribuzione di probabilità di $\boldsymbol{\varepsilon}$. Tuttavia si può dimostrare che il vettore delle medie e la matrice di varianza e covarianza di $\mathbf{U}$ coincidono con quelli di $\boldsymbol{\varepsilon}$. Infatti:

$$\mathbb{E}\left[\mathbf{U}\right] = \mathbb{E}\left[\mathbf{Q'}\boldsymbol{\varepsilon}\right] = \mathbf{Q}'\mathbb{E}\left[\boldsymbol{\varepsilon}\right] = \mathbf{Q'0} = \mathbf{0}$$

e, considerando che per l'ortoganilità di $\mathbf{Q}$ si ha $\mathbf{Q'Q} = \mathbf{I}_n$,

$$\mathbb{E}\left[\mathbf{U'U}\right] = \mathbb{E}\left[\boldsymbol{Q'\varepsilon}\,\boldsymbol{\varepsilon'Q}\right] = \mathbf{Q'}\mathbb{E}\left[\boldsymbol{\varepsilon}\,\boldsymbol{\varepsilon'}\right]\mathbf{Q} = \mathbf{Q'}\sigma^2\mathbf{I}_n\mathbf{Q} = \sigma^2\mathbf{I}_n$$

Quindi si ha:

$$\mathbf{e'e} = (\boldsymbol{\varepsilon}'\mathbf{Q})\,\boldsymbol{\Lambda}\,(\mathbf{Q'}\boldsymbol{\varepsilon})$$

e, poichè $\mathbb{C}\mathrm{ov}\left[U_i, U_j\right] = 0$ per ogni $i \neq j$:

$$\mathbb{E}\left[\mathbf{e'e}\right] = \sum_{n}^{i=1} \lambda_i \mathbb{E}\left[U_i^2\right] = \sigma^2 \sum_{n}^{i=1} \lambda_i = \sigma^2\, tr(\boldsymbol{\Lambda}) = \sigma^2(n-k-1)$$

Si noti che questa proprietà possiede un importante corollario, ovvero:

$$\hat{S}^2 = \frac{1}{n-k-1}\mathbf{e'e} = \frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2$$

è uno stimatore non distorto di σ^2, quindi $\mathbb{E}\left[\hat{S}^2\right] = \sigma^2$.

Proprietà 5 (Teorema di Gauss-Markov) *Lo stimatore $\hat{\boldsymbol{\beta}}$ dei minimi*

quadrati è BLUE (Best Linear Unbiased Estimator), ovvero è lo stimatore lineare non distorto di $\boldsymbol{\beta}$ a varianza minima.

Ciò significa che per qualsiasi vettore $\mathbf{c} \in \mathbb{R}^{k+1}$ lo stimatore $\boldsymbol{c'\hat{\beta}}$ è il più efficiente stimatore lineare non distorto per la combinazaione lineare $\boldsymbol{c'\beta} = c_0\beta_0 + c_1\beta_1 + \ldots + c_k\beta_k$.

Dimostrazione:
Innanzitutto è facile dimostrare che $\boldsymbol{c'\hat{\beta}}$ è uno stimatore lineare di $\boldsymbol{c'\beta}$, in quanto ognuno dei β_j è una combinazione lineare delle variabili $Y_1, \ldots, Y_n$.

Inoltre $\boldsymbol{c'\hat{\beta}}$ è corretto, infatti:

$$\mathbb{E}\left[\boldsymbol{c'\hat{\beta}}\right] = \mathbf{c'}\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] = \mathbf{c'}\boldsymbol{\beta}$$

Sia ora $z = \mathbf{a'Y} = a_1Y_1 + \ldots + a_nY_n$ un qualunque altro stimatore lineare di $\mathbf{c'}\boldsymbol{\beta}$. Affinchè esso sia corretto è necessario che verifichi la condizione $\mathbb{E}[z] = \mathbf{c'}\boldsymbol{\beta}$, ma

$$\mathbb{E}[z] = \mathbf{a'}\mathbb{E}[\mathbf{Y}] = \mathbf{a'X}\boldsymbol{\beta}$$

quindi deve necessariamnte valere $\mathbf{c'} = \mathbf{X'a}$

La varianza dello stimatore z è:

$$\begin{aligned}\mathbb{V}\mathrm{ar}[z] &= \mathbb{V}\mathrm{ar}[\mathbf{a'Y}] = \mathbb{V}\mathrm{ar}[\mathbf{a'}(\boldsymbol{X\beta} + \boldsymbol{\varepsilon})] \\ &= \mathbb{V}\mathrm{ar}[\mathbf{a'}\boldsymbol{\varepsilon}] = \mathbf{a'}\mathbb{E}[\boldsymbol{\varepsilon'\varepsilon}]\,\mathbf{a} = \sigma^2\mathbf{a'a}\end{aligned}$$

Osservando inoltre che

$$\mathbb{V}\mathrm{ar}\left[\boldsymbol{c'\hat{\beta}}\right] = \boldsymbol{c'\hat{\beta}c} = \sigma^2\mathbf{a'X}(\mathbf{X'X})^{-1}\mathbf{X'a}$$

si deduce allora che la differenza tra la varianza di z e quella di $\mathbf{c'}\boldsymbol{\beta}$ vale:

$$\begin{aligned}\mathbb{V}\mathrm{ar}[z] - \mathbb{V}\mathrm{ar}[\mathbf{c'}\boldsymbol{\beta}] &= \sigma^2\mathbf{a'a} - \sigma^2\mathbf{a'X}(\mathbf{X'X})^{-1}\mathbf{X'a} = \\ &= \sigma^2\mathbf{a'}\left[\mathbf{I}_n - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}\right]\mathbf{a} = \sigma^2\mathbf{a'Va}\end{aligned}$$

Ma la matrice $\mathbf{V}$, essendo idempotente, è semidefinita positiva, e quindi si ha $\mathbf{a'Va} \geq 0$ per ogni vettore $\mathbf{a} \in \mathbb{R}^n$. Ciò comporta che $\mathbb{V}\mathrm{ar}[z] - \mathbb{V}\mathrm{ar}[\mathbf{c'}\boldsymbol{\beta}] \geq 0$,

ossia $\mathbb{V}\text{ar}\,[z] \geq \mathbb{V}\text{ar}\,[\mathbf{c'}\boldsymbol{\beta}]$. Lo stimatore $\mathbf{c'}\boldsymbol{\beta}$ è quindi più efficiente di z.

2.1.2 Metodo della Massima Verosimiglianza

Sebbene il metodo dei Minimi Quadrati sia quello classicamente utilizzato per la stima dei parametri di un modello lineare è utile introdurre anche il metodo della Massima Verosimiglianza in quanto, come sarà mostrato nella sezione successiva, esso viene utilizzato nell'ambito dei modelli lineari generalizzati. Per applicare questo metodo è necessario ipotizzare la distribuzione di probabilità degli errori, in particolare per il modello lineare "classico" si suppone:

$$\varepsilon_1 \dots \varepsilon_n \quad \text{i.i.d.} \quad N(0, \sigma^2), \quad \text{ovvero} \quad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

Sotto questa ipotesi, ricordando che $\mathbb{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$, è allora possibile ottenere l'espressione della funzione di verosimiglianza per un campione casuale di numerosità n:

$$L(\boldsymbol{\beta}, \sigma^2 | \mathbf{Y} = \mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{-\frac{n}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\}$$

da questa relazione si deduce allora l'espressione della funzione di log-verosimiglianza:

$$\ell(\boldsymbol{\beta}, \sigma^2 | \mathbf{Y} = \mathbf{y}) = -\frac{n}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \left[\mathbf{y'y} - 2\boldsymbol{\beta'}(\boldsymbol{X'X})\boldsymbol{\beta} \right]$$

derivando rispetto a $\boldsymbol{\beta}$ e uguagliando a 0 tali derivate si ottiene il sistema di equazioni:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = -2\mathbf{X'y} + 2(\mathbf{X'X})\boldsymbol{\beta} = \mathbf{0}$$

che fornisce la stima del vettore dei parametri $\boldsymbol{\beta}$ con il metodo della Massima Verosimiglianza:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

Lo stimatore di Massima Verosimiglianza nel caso in cui la distribuzione di probabilità dell'errore sia supposta essere Normale coincide quindi con lo stimatore ottenuto tramite il metodo dei Minimi Quadrati e gode, di conseguenza, delle stesse proprietà. Come sarà mostrato in seguito questo fatto non è vero in generale, infatti ipotizzando una diversa distribuzione di probabilità per l'errore i due stimatori non coincidono.

Si osservi inoltre che:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'Y} = \boldsymbol{\beta} + (\mathbf{X'X})^{-1}\boldsymbol{X'\varepsilon}$$

Quindi, essendo $\hat{\boldsymbol{\beta}}$ una funzione lineare di $\boldsymbol{\varepsilon}$, in conseguenza dell'introduzione dell'ipotesi di Normalità dell'errore si ha:

$$\hat{\boldsymbol{\beta}} \sim N_{k+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X'X})^{-1})$$

Un aspetto molto importante è quello relativo alla distribuzione della variabile aleatoria $\hat{S}^2$. Consideriamo a tale scopo le matrici $\mathbf{P} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$ e $\mathbf{V} = \mathbf{I}_n - \mathbf{P}$, le quali sono entrambe simmetriche definite positive e sono tali che $\mathbf{P} + \mathbf{V} = \mathbf{I}_n$. Quindi, ponendo $\mathbf{z} = \dfrac{1}{\sigma}\boldsymbol{\varepsilon}$, si ha:

$$\mathbf{z'z} = \mathbf{z'Pz} + \mathbf{z'Vz}$$

Si osserva che $\mathbf{z}$ è una v.a. Normale multivariata a componenti indipendenti e standardizzate. Pertanto applicando il teorema di Cochran-Fisher, e tenendo presente che $rank(V) = (n - k - 1)$ e $rank(P) = (k + 1)$, si ha che la v.a. $\mathbf{z'Vz}$ ha distribuzione chi quadrato a $(n - k - 1)$ gradi di libertà. D'altra parte si ha anche:

$$\mathbf{z'Vz} = \frac{\boldsymbol{\varepsilon' V \varepsilon}}{\sigma^2} = \frac{\mathbf{e'e}}{\sigma^2} = \frac{(n-k-1)\hat{S}^2}{\sigma^2}$$

Di conseguenza

$$\frac{(n-k-1)\hat{S}^2}{\sigma^2} \sim \chi^2_{n-k-1}$$

É inoltre possibile dimostrare che tale variabile aleatoria è indipendente da $\hat{\boldsymbol{\beta}}$, per fare ciò basta provare che le v.a. $\mathbf{e}$ e $\hat{\boldsymbol{\beta}}$ sono indipendenti. A tale scopo si osservi che $\mathbf{e} = \boldsymbol{V\varepsilon}$, perciò sia $\hat{\boldsymbol{\beta}}$ che $\mathbf{e}$ sono variabili multinormali.

Si osservi inoltre che la matrice avente come componenti le $\mathbb{C}\text{ov}\left[e_i, \hat{\beta}_j\right]$ è data da:

$$\begin{aligned}
\mathbb{E}\left[\mathbf{e}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right] &= \mathbb{E}\left[\boldsymbol{V\varepsilon}\,\boldsymbol{\varepsilon'}\boldsymbol{X}(\mathbf{X'X})^{-1}\right] \\
&= \mathbf{V}\mathbb{E}\left[\boldsymbol{\varepsilon}\,\boldsymbol{\varepsilon'}(\mathbf{X'X})^{-1}\mathbf{X'}\right] \\
&= \sigma^2\mathbf{V}(\mathbf{X'X})^{-1}\mathbf{X'} = \mathbf{0}
\end{aligned}$$

Poichè, come detto, $\hat{\boldsymbol{\beta}}$ e $\mathbf{e}$ sono variabili aleatorie multinormali, il fatto che esse siano scorrelate implica che siano indipendenti. Di conseguenza $\hat{\boldsymbol{\beta}}$ e $\hat{S}^2$ sono indipendenti.

2.1.3 Intervalli di Confidenza per $\boldsymbol{\beta}$

Come detto la variabile aleatoria $\hat{\boldsymbol{\beta}}$ ha distribuzione multinormale di media $\boldsymbol{\beta}$ e varianza $\sigma^2(\mathbf{X'X})^{-1}$. Quindi le singole componenti $\hat{\beta}_j$ $(j = 0, \ldots, k)$ di $\hat{\boldsymbol{\beta}}$ sono Normali univariate con media β_j e varianza $\sigma^2 a_{jj}$, essendo a_{jj} il j-esimo elemento della diagonale principale della della matrice $(\mathbf{X'X})^{-1}$. Si può affermare allora che:

$$\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{a_{jj}}} \sim N(0,1) \qquad\qquad j = 0, \ldots, k$$

Di norma, però, il valore di σ^2 è incognito, e quindi tale risultato non è immediatamente utilizzabile per costruire un intervallo di confidenza per β_j. Si è tuttavia visto che la variabile $(n-k-1)\hat{S}^2/\sigma^2$ ha distribuzione chi quadrato con $(n-k-1)$ gradi di libertà ed è indipendente da $\hat{\beta}_j$. Si deduce quindi che la v.a.

$$\frac{\hat{\beta}_j - \beta_j}{\hat{S}\sqrt{a_{jj}}} = \frac{\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{a_{jj}}}}{\left(\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{a_{jj}}}/(n-k-1)\right)^{1/2}}$$

è data dal rapporto tra una v.a. Normale standard e la radice quadrata di una v.a. chi quadrato divisa per i propri gradi di libertà e indipendente dalla precedente. Pertanto ha distribuzione t di Student $(n-k-1)$ gdl, cioè:

$$\frac{\hat{\beta}_j - \beta_j}{\hat{S}\sqrt{a_{jj}}} \sim t_{n-k-1}$$

Detto allora $t_{n-k-1,\alpha/2}$ il valore della v.a. t_{n-k-1} tale che $\mathbb{P}(t_{n-k-1} \geq t_{n-k-1,\alpha/2}) = \alpha/2$, è possibile ottenere l'espressione per l'intervallo di confidenza per β_j di livello $1-\alpha$:

$$\hat{\beta}_j - t_{n-k-1,\alpha/2}\hat{S}\sqrt{a_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{n-k-1,\alpha/2}\hat{S}\sqrt{a_{jj}} \qquad \forall j = 0, \ldots, k$$

2.1.4 Intervalli di Previsione

Uno dei problemi di maggior rilievo nell'analisi della regressione è indubbiamente quello della previsione si un'osservazione futura, cioè della della previsione del valore che la variabile casuale Y assumerà in corrispondenza di un dato insieme $X_{1(n+1)}, \ldots, X_{k(n+1)}$ di valori delle variabili esplicative.

La previsione "puntuale" può essere sviluppata anche senza far ricorso all'ipotesi di Normalità degli errori. Se $\hat{\boldsymbol{\beta}}$ è lo stimatore dei minimi quadrati di $\boldsymbol{\beta}$ ottenuto sulla base delle n osservazioni campionarie $Y_i, X_{1i}, \ldots X_{ki}$ (con $i = 1, \ldots, n$), allora ponendo

$$\mathbf{x}_{n+1} = \begin{bmatrix} 1 \\ x_{1(n+1)} \\ \ldots \\ x_{k(n+1)} \end{bmatrix}$$

una funzione di previsione di

$$Y_{n+1} = \beta_0 + \beta_1 x_{1(n+1)} + \ldots + \beta_k x_{k(n+1)} + \varepsilon_{n+1} = \mathbf{x'}_{n+1}\boldsymbol{\beta} + \varepsilon_{n+1}$$

è data da

$$\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{1(n+1)} + \ldots + \hat{\beta}_k x_{k(n+1)} = \mathbf{x'}_{n+1}\hat{\boldsymbol{\beta}} \tag{2.1}$$

Tale stima per l'osservazione futura Y_{n+1} possiede importanti proprietà di ottimalità in quanto è la funzione lineare in $Y_1, \ldots, Y_n$ non distorta con minimo errore quadratico medio di previsione.

Provare che la 2.1 è non distorta è molto semplice, infatti si ha:

$$\begin{aligned}
\mathbb{E}\left[\hat{Y}_{n+1} - Y_{n+1}\right] &= \mathbb{E}\left[\boldsymbol{x}_{n+1}\hat{\boldsymbol{\beta}}\right] - \mathbb{E}\left[Y_{n+1}\right] \\
&= \mathbf{x'}_{n+1}\mathbb{E}\left[\boldsymbol{\beta}\right] - \mathbf{x'}_{n+1}\boldsymbol{\beta} \\
&= \mathbf{x'}_{n+1}\hat{\boldsymbol{\beta}} - \mathbf{x'}_{n+1}\hat{\boldsymbol{\beta}} = 0
\end{aligned}$$

Inoltre il suo errore quadratico medio di previsione è:

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{Y}_{n+1} - Y_{n+1}\right)^2\right] &= \mathbb{E}\left[\left((\hat{Y}_{n+1} - \mathbf{x'}_{n+1}\boldsymbol{\beta}) - (Y_{n+1} - \mathbf{x'}_{n+1}\boldsymbol{\beta})\right)^2\right] \\
&= \mathbb{V}\mathrm{ar}\left[\hat{Y}_{n+1}\right] + \mathbb{V}\mathrm{ar}\left[Y_{n+1}\right] - 2\mathbb{C}\mathrm{ov}\left[\hat{Y}_{n+1}, Y_{n+1}\right]
\end{aligned}$$

Ma si ha anche:

1. $\mathbb{V}\mathrm{ar}\left[\hat{Y}_{n+1}\right] = \mathbb{V}\mathrm{ar}\left[\mathbf{x'}_{n+1}\hat{\boldsymbol{\beta}}\right] = \sigma^2\mathbf{x'}_{n+1}(\mathbf{X'X})^{-1}\mathbf{x}_{n+1}$
2. $\mathbb{V}\mathrm{ar}\left[Y_{n+1}\right] = \mathbb{V}\mathrm{ar}\left[\varepsilon_{n+1}\right] = \sigma^2$

inoltre

$$\begin{aligned}
\mathbb{C}\mathrm{ov}\left[\hat{Y}_{n+1}, Y_{n+1}\right] &= \mathbb{E}\left[(\hat{Y}_{n+1} - \mathbf{x'}_{n+1}\boldsymbol{\beta})(\hat{Y}_{n+1} - \mathbf{x'}_{n+1}\boldsymbol{\beta})\right] \\
&= \mathbb{E}\left[\mathbf{x'}_{n+1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\varepsilon_{n+1}\right] = \mathbb{E}\left[\mathbf{x'}_{n+1}(\mathbf{X'X})^{-1}\mathbf{X'}\boldsymbol{\varepsilon}\,\varepsilon_{n+1}\right] \\
&= \mathbf{x'}_{n+1}(\mathbf{X'X})^{-1}\mathbf{X'}\mathbb{E}\left[\boldsymbol{\varepsilon}\,\varepsilon_{n+1}\right] = 0
\end{aligned}$$

Pertanto, complessivamente, la $\hat{Y}_{n+1}$ ha un errore quadratico medio di previsione pari a:

$$\mathbb{E}\left[\left(\hat{Y}_{n+1} - Y_{n+1}\right)^2\right] = \sigma^2\left[1 + \mathbf{x}_{n+1}(\mathbf{X'X})^{-1}\mathbf{x'}_{n+1}\right]$$

Per provare infine che $\hat{Y}_{n+1}$ è la più efficiente funzione di previsione lineare corretta di Y_{n+1} consideriamo una generica funzione di previsione

$$\tilde{Y}_{n+1} = \sum_{i=1}^{n} h_i Y_i = \mathbf{h'Y} \qquad \text{con} \quad \mathbf{h} \in \mathbb{R}^n$$

La condizione di non distorsione implica che:

$$\begin{aligned} \mathbb{E}\left[\tilde{Y}_{n+1} - Y_{n+1}\right] &= \mathbb{E}\left[\tilde{Y}_{n+1}\right] - \mathbb{E}\left[Y_{n+1}\right] = \mathbf{h'}\mathbb{E}\left[\mathbf{Y}\right] - \mathbf{x'}_{n+1} = \boldsymbol{\beta} \\ &= \mathbf{h'X}\boldsymbol{\beta} - \mathbf{x'}_{n+1}\boldsymbol{\beta} = \left(\mathbf{h'X} - \mathbf{x'}_{n+1}\right)\boldsymbol{\beta} = \mathbf{0} \end{aligned}$$

da cui, necessariamente, $\mathbf{X'h} = \mathbf{x}_{n+1}$ L'errore quadratico medio di previsione vale:

$$\mathbb{E}\left[\left(\tilde{Y}_{n+1} - Y_{n+1}\right)^2\right] = \mathbb{V}\mathrm{ar}\left[\tilde{Y}_{n+1}\right] + \mathbb{V}\mathrm{ar}\left[Y_{n+1}\right] - 2\mathbb{C}\mathrm{ov}\left[\tilde{Y}_{n+1}, Y_{n+1}\right]$$

Ma dalla condizione di linearità di $\tilde{Y}_{n+1}$ si evince che:

$$\mathbb{C}\mathrm{ov}\left[\tilde{Y}_{n+1}, Y_{n+1}\right] = \mathbb{C}\mathrm{ov}\left[\mathbf{h'Y}, Y_{n+1}\right] = \sum_{i=1}^{n} h_i\, \mathbb{C}\mathrm{ov}\left[Y_i, Y_{n+1}\right] = 0$$

Inoltre dalla correttezza di $\tilde{Y}_{n+1}$ si evince che $\mathbb{E}\left[\tilde{Y}_{n+1}\right] = \mathbf{x}_{n+1}\boldsymbol{\beta}$, quindi $\tilde{Y}_{n+1}$ può essere considerato come uno stimatore corretto di $\mathbf{x}_{n+1}\boldsymbol{\beta}$. Ma anche $\hat{Y}_{n+1}$ è uno stimatore correto di $\mathbf{x}_{n+1}\boldsymbol{\beta}$, pertanto, per il teorema di Gauss-Markov, vale:

$$\mathbb{V}\mathrm{ar}\left[\hat{Y}_{n+1}\right] \leq \mathbb{V}\mathrm{ar}\left[\tilde{Y}_{n+1}\right]$$

Quindi $\hat{Y}_{n+1}$ è una funzione di previsione più efficiente di $\tilde{Y}_{n+1}$

Se gli errori sono normalmente distribuiti, è possibile costruire degli intervalli di previsione per Y_{n+1}. Infatti è facile vedere che la v.a. $\hat{Y}_{n+1} - Y_{n+1}$ ha distribuzione Normale con media 0 e varianza $\sigma^2\left[1 + \mathbf{x}_{n+1}(\mathbf{X'X})^{-1}\mathbf{x'}_{n+1}\right]$, quindi:

$$\frac{\hat{Y}_{n+1} - Y_{n+1}}{\sigma\left[1 + \mathbf{x}_{n+1}(\mathbf{X'X})^{-1}\mathbf{x'}_{n+1}\right]^{1/2}} \sim N(0,1)$$

Inoltre sia $\hat{Y}_{n+1}$ che Y_{n+1} sono indipendenti da $\hat{S}^2$ e , ricordando che

$$\frac{(n-k-1)\hat{S}^2}{\sigma^2} \sim \chi^2_{n-k-1}$$

è possibile dedurre che

$$\frac{\hat{Y}_{n+1} - Y_{n+1}}{\hat{S}\left[1 + \mathbf{x}_{n+1}(\mathbf{X'X})^{-1}\mathbf{x'}_{n+1}\right]^{1/2}} \sim t_{n-k-1}$$

Detto allora $t_{n-k-1,\alpha/2}$ il quantile di ordine $\alpha/2$ di una distribuzione t di Student con $(n-k-1)$ gradi di libertà si ha l'intervallo di previsione di livello α per Y_{n+1}:

$$\hat{Y}_{n+1} - t_{n-k-1,\alpha/2} \cdot u \leq Y_{n+1} \leq \hat{Y}_{n+1} + t_{n-k-1,\alpha/2} \cdot u$$

dove $u = \hat{S}\left[1 + \mathbf{x}_{n+1}(\mathbf{X'X})^{-1}\mathbf{x'}_{n+1}\right]^{1/2}$

2.2 Modelli Lineari Generalizzati

La classe dei modelli lineari generalizzati costituisce un'estensione del modello lineare classico. Tramite l'utilizzo di un modello lineare generalizzato è infatti possibile rimuovere l'ipotesi che le osservazioni abbiano una distribuzione Normale. L'assunzione che viene fatta è invece che la generica v.a. Y appartenga alla *Exponential Dispersion Family*, ovvero che abbia una funzione di densità esprimibile nella forma:

$$p(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{2.2}$$

dove si intende che il parametro θ (detto *parametro naturale*) sia incognito, che il parametro ϕ (detto *parametro di scala*) possa essere noto oppure incognito e che $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$ siano opportune funzioni rispettivamente di ϕ, di θ e di (y,ϕ). Si osservi che, nel caso in cui il parametro ϕ sia noto, allora la (2.2) rappresenta la funzione di densità di una v.a. appartenente alla famiglia esponenziale di parametro θ.

Con opportuni accorgimenti è possibile calcolare in forma generale il valore atteso e la varianza di una generica v.a. Y con funzione di densità del tipo (2.2). È infatti noto che sotto opportune ipotesi di regolarità, sempre soddisfatte dalle distribuzioni appartenenti alla famiglia esponenziale, valgono le seguenti relazioni:

$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\left(\ln p(Y;\theta,\phi)\right)\right] = 0 \tag{2.3}$$

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\left(\ln p(Y;\theta,\phi)\right)\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\left(\ln p(Y;\theta,\phi)\right)\right] \tag{2.4}$$

Dalla 2.3 si ottiene:

$$\begin{aligned}\mathbb{E}\left[\frac{\partial}{\partial\theta}\left(\ln p(Y;\theta,\phi)\right)\right] &= \mathbb{E}\left[\frac{\partial}{\partial\theta}\left(\frac{Y\theta - b(\theta)}{a(\phi)} + c(Y,\phi)\right)\right] \\ &= \mathbb{E}\left[\frac{Y - b'(\theta)}{a(\phi)}\right] = \frac{\mathbb{E}[Y] - b'(\theta)}{a(\phi)} = 0\end{aligned}$$

$$\text{Da cui } \mathbb{E}[Y] = b'(\theta)$$

Dalle 2.4 e 2.5 si ottiene:

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\left(\ln p(Y;\theta,\phi)\right)\right)^2\right] &= \mathbb{E}\left[\left(\frac{Y - b'(\theta)}{a(\phi)}\right)^2\right] \\ &= \frac{1}{a(\phi)^2}\mathbb{E}\left[\left(Y - b'(\theta)\right)^2\right] = \frac{1}{a(\phi)^2}\mathbb{V}\text{ar}\left[Y\right]\end{aligned}$$

$$\begin{aligned}-\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\left(\ln p(Y;\theta,\phi)\right)\right] &= -\mathbb{E}\left[\frac{\partial}{\partial\theta}\left(\frac{Y - b'(\theta)}{a(\phi)}\right)\right] \\ &= \mathbb{E}\left[\frac{b''(\theta)}{a(\phi)}\right] = \frac{b''(\theta)}{a(\phi)}\end{aligned}$$

$$\text{Da cui } \mathbb{V}\text{ar}[Y] = a(\phi)b''(\theta)$$

Si ha quindi che se Y è una v.a. con funzione di densità del tipo (2.2) allora valgono:

$$\mathbb{E}[Y] = b'(\theta) \tag{2.5}$$

$$\mathbb{V}\text{ar}[Y] = a(\phi)b''(\theta) \tag{2.6}$$

La (2.5) mette in evidenza che il valore atteso di Y dipende da θ ma non da ϕ. La (2.6) mostra invece come la varianza di Y possa essere espressa come il prodotto di due funzioni, una dipendente solo da ϕ e l'altra solo da θ. In particolare la varianza di Y dipende da θ, e quindi dal valore atteso di Y, solo tramite la funzione $b''(\theta)$, che è detta *funzione di varianza*. Il parametro ϕ dal quale dipende (eventualmente) la varianza ma non la media è detto *parametro di dispersione*.

2.2.1 Formulazione del Modello

Siano $y_1 \dots y_n$ n osservazioni provenienti dalle v.a. indipendenti $Y_1 \dots Y_n$. La generica variabile aleatoria Y_i $(i = 1 \dots n)$ ha funzione di densità del tipo (2.2) con parametri θ_i (incognito) e ϕ (noto o incognito):

$$p(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y, \phi)\right\} \qquad i = 1 \dots n \tag{2.7}$$

Si osservi che il parametro θ_i può variare da un'osservazione all'altra mentre il paramtro ϕ è sempre costante per tutte le osservazioni. Da quanto precedentemente ottenuto si deduce:

$$\mathbb{E}[Y_i] = b'(\theta_i) \qquad i = 1 \dots n \tag{2.8}$$

$$\mathbb{V}\text{ar}[Y_i] = a(\phi)b''(\theta_i) \qquad i = 1 \dots n \tag{2.9}$$

Dalla (2.8) si deduce che le v.a. $Y_i \dots Y_n$ possono avere in generale medie diverse. Dalla (2.9) risulta invece che esse hanno tutte la stessa varianza se e solo se la funzione di varianza $b''(\cdot)$ è costante in θ_i; in questo caso si è

in condizioni di omoschedasticità. Se, al contrario, la funzione di varianza non è costante in θ_i le v.a. $Y_i \dots Y_n$ hanno differenti varianze, cioè sono eteroschedastiche.

Detto allora $\mathbf{Y} = [Y_1 \dots Y_n]'$ definiamo:

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{Y}] = \begin{bmatrix} b'(\theta_1) \\ \dots \\ b'(\theta_n) \end{bmatrix} \quad \boldsymbol{\Sigma} = \mathbb{V}\text{ar}[Y] = a(\phi) \begin{pmatrix} b''(\theta_1) & 0 & \dots & 0 \\ 0 & b''(\theta_2) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & b'(\theta_n) \end{pmatrix}$$

L'idea alla base del modello lineare classico è quello di esprimere il generico valore medio μ_i come funzione lineare delle k variabile esplicative, ovvero $\mu_i = \mathbf{x'}_{i\cdot}\boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}$. In un modello generalizzato il valore medio μ_i viene invece espresso come funzione non necessariamente lineare di $\mathbf{x}'_{i\cdot}\boldsymbol{\beta}$. In particolare:

$$\mu_i = f(\mathbf{x}'_{i\cdot}\boldsymbol{\beta}) = f(\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}) \qquad i = 1 \dots n$$

Assumendo che f sia invertibile e ponendo $g = f^{-1}$ si ottiene:

$$g(\mu_i) = \mathbf{x}'_{i\cdot}\boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} \qquad i = 1 \dots n$$

La funzione g è detta *funzione link*, essa determina la "scala" nella quale un incremento di una variabile esplicativa determina un incremento proporzionale di $g(\mu_i)$. Una tipologia particolare di funzioni link sono i *link canonici*. Nel caso in cui la funzione $b'(\cdot)$ sia invertibile g è detta link canonico se vale l'uguaglianza

$$g(\cdot) = b'^{-1}(\cdot)$$

In questo caso si ha

$$\theta_i = b'^{-1}(\mu_i) = b'^{-1}[g^{-1}(\mathbf{x}'_{i\cdot}\boldsymbol{\beta})] = \mathbf{x}'_{i\cdot}\boldsymbol{\beta} \qquad i = 1 \dots n$$

Quindi se g è un link canonico i parametri $\theta_1 \dots \theta_n$ sono esprimibili tramite

una funzione lineare nei parametri $\beta_0 \dots \beta_k$.

Va comunque osservato che, sebbene l'utilizzo di funzioni link canoniche faciliti (nel senso che sarà spiegato in seguito) le procedure di inferenza sui parametri del modello, non vi è alcuna ragione particolare per la quale esse dovrebbero essere utilizzate. Pertanto sono molto spesso impiegati modelli lineari generalizzati in cui la funzione link non è quella canonica. In questo senso una famiglia di funzioni link molto spesso utilizzata è quella delle *funzioni di potenza*, definite come:

$$g(\mu_i) = \begin{cases} \frac{\mu_i^\lambda - 1}{\lambda} & \text{se } \lambda \neq 0 \\ \ln \mu_i & \text{se } \lambda = 0 \end{cases}$$

2.2.2 Stima dei Parametri del Modello

Solitamente quando si tratta il problema della stima dei parametri incogniti $\beta_0, \dots, \beta_k$ e ϕ di un modello lineare generalizzato il criterio di ottimalità utilizzato è quello della massimizzazione della funzione di verosimiglianza. La stima dei parametri $\beta_0, \dots, \beta_k$ e, nel caso in cui esso sia incognito, del parametro ϕ avverrà dunque utilizzando questo metodo.

Dalla (2.7) si deduce che la funzione di verosimiglianza associata al problema in esame è:

$$\begin{aligned} L(\boldsymbol{\theta}, \phi) &= \prod_{i=1}^{n} \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\} \\ &= \exp\left\{ \frac{1}{a(\phi)} \sum_{i=1}^{n} [y_i\theta_i - b(\theta_i)] + \sum_{i=1}^{n} c(y_i, \phi) \right\} \end{aligned}$$

Da ciò discende che la funzione di log-verosimiglianza è:

$$l(\boldsymbol{\theta}, \phi) = \ln L(\boldsymbol{\theta}, \phi) = \frac{1}{a(\phi)} \sum_{i=1}^{n} [y_i\theta_i - b(\theta_i)] + \sum_{i=1}^{n} c(y_i, \phi) \tag{2.10}$$

Per calcolare gli stimatori in esame è quindi necessario derivare la funzione

$l(\boldsymbol{\theta}, \phi)$ rispetto a $\beta_0, \ldots, \beta_k$ e ϕ, tenendo conto del fatto che $\theta_1 \ldots \theta_n$ sono anch'esse funzioni di $\beta_0, \ldots, \beta_k$. Va ricordato infatti che, per quanto prima spiegato, si ha:

$$\theta_i = b'^{-1}(\mu_i) = b'^{-1}[f(\mathbf{x}'_{i\cdot}\boldsymbol{\beta})] = h(\mathbf{x}'_{i\cdot}\boldsymbol{\beta}) \qquad i = 1 \ldots n$$

Si osservi che nel caso in cui $g(\cdot)$ fosse la funzione link canonica si avrebbe $\theta_i = \mathbf{x}'_{i\cdot}\boldsymbol{\beta}$. In generale invece, posto $\eta_i = \mathbf{x}'_{i\cdot}\boldsymbol{\beta}$ si ha:

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l}{\partial \mu_i}\left(\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}\right) = \sum_{i=1}^{n} \frac{\partial l}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}x_{ji}$$

Inoltre vale:

$$\frac{\partial l}{\partial \mu_i} = \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i} = \left(\frac{y_i - b'(\theta_i)}{a(\phi)}\right)\frac{\partial \theta_i}{\partial \mu_i} = \frac{y_i - \mu_i}{a(\phi)b''(\theta_i)} = \frac{y_i - \mu_i}{\mathbb{V}\mathrm{ar}[Y_i]}$$

Quindi dalle due precedenti relazioni si ottiene:

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathbb{V}\mathrm{ar}[Y_i]}\frac{\partial \mu_i}{\partial \eta_i}x_{ji} \qquad j = 0 \ldots k$$

Queste $k + 1$ derivate possono essere espresse in forma più compatta ricorrendo a una notazione vettoriale. Infatti ponendo

$$\frac{\boldsymbol{\partial l}}{\boldsymbol{\partial \beta}} = \begin{bmatrix} \frac{\partial l}{\partial \beta_0} \\ \ldots \\ \frac{\partial l}{\partial \beta_k} \end{bmatrix} \quad \mathbf{M} = \begin{pmatrix} \frac{\partial \mu_1}{\partial \eta_1} & 0 & \ldots & 0 \\ 0 & \frac{\partial \mu_2}{\partial \eta_2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \frac{\partial \mu_n}{\partial \eta_n} \end{pmatrix} \quad \mathbf{V} = \frac{1}{a(\phi)}\boldsymbol{\Sigma}$$

è facile verificare che:

$$\frac{\boldsymbol{\partial l}}{\boldsymbol{\partial \beta}} = \frac{1}{a(\phi)}\mathbf{X}'\mathbf{M}\mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu})$$

Lo stimatore di massima verosimiglianza cercato si ottiene quindi risolvendo il sistema:

$$\mathbf{X}'\mathbf{M}\mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}) = \mathbf{0} \tag{2.11}$$

Per quanto riguarda la stima del parametro ϕ, invece, si ha:

$$\begin{aligned}\frac{\partial l}{\partial \phi} &= \left[\sum_{i=1}^{n}\left[y_i\theta_i - b(\theta_i)\right]\right]\frac{d}{d\phi}\left(\frac{1}{a(\phi)}\right) + \sum_{i=1}^{n}\frac{\partial c(y_i,\phi)}{\partial \phi} \\ &= -\frac{a'(\phi)}{a(\phi)^2}\sum_{i=1}^{n}\left[y_i\theta_i - b(\theta_i)\right] + \sum_{i=1}^{n}\frac{\partial c(y_i,\phi)}{\partial \phi}\end{aligned}$$

La stima di massima verosimiglianza di ϕ si ottiene sostituendo a ogni θ_i il suo valore stimato (che può essere calcolato indipendentemente da ϕ come sarà mostrato in seguito) $\hat{\theta}_i = b'^{-1}(f(\mathbf{x}_{i\cdot}'\hat{\boldsymbol{\beta}}))$ e poi annullando $\partial l/\partial\phi$. Si ha in tal modo l'equazione:

$$-\frac{a'(\phi)}{a(\phi)^2}\sum_{i=1}^{n}\left[y_i\hat{\theta}_i - b(\hat{\theta}_i)\right] + \sum_{i=1}^{n}\frac{\partial c(y_i,\phi)}{\partial \phi} = 0 \tag{2.12}$$

Sfortunatamente, a parte pochi casi particolari (sostanzialmente il solo modello lineare classico), le equazioni (2.11) e (2.12) non sono lineari in $\boldsymbol{\beta}$ e non possiedono una soluzione esplicita. Risulta pertanto necessario ricorrere ad un metodo di tipo numerico. Il software `R` utilizza l'algoritmo *Scoring* di Fisher, il cui funzionamento viene quindi di seguito riportato.

2.2.3 L'algoritmo *Scoring* di Fisher

Consideriamo il caso generale in cui si abbiano n realizzazioni $x_1, \ldots, x_n$ di una certa variabile aleatoria X con densità $p(x;\boldsymbol{\vartheta})$ e in cui l'obiettivo sia quello di determinare la stima di massima verosimiglianza per il vettore di parametri $\boldsymbol{\vartheta} = [\vartheta_1, \ldots, \vartheta_k]'$. Indicando con $l(\vartheta_1 \ldots, \vartheta_k) = l(\boldsymbol{\vartheta})$ la funzione di log-verosimiglianza è necessario risolvere il sistema di k equazioni in k incognite:

$$\frac{\boldsymbol{\partial l}}{\boldsymbol{\partial \vartheta}} = \mathbf{0}$$

L'idea alla base dell'algoritmo Scoring di Fisher è quella di sviluppare in serie di Taylor centrata nel punto $\boldsymbol{\vartheta_0} = [\vartheta_{10}, \ldots, \vartheta_{k0}]'$ la funzione di log-verosimiglianza $l(\boldsymbol{\vartheta})$, ottenendo così anche un'approssimazione per il gradiente della stessa. A tale scopo si indichino con:

- $\nabla l(\boldsymbol{\vartheta_0})$ il gradiente della funzione di log-verosimiglianza calcolato nel punto $\boldsymbol{\theta} = \boldsymbol{\vartheta_0}$

- $\mathbf{H}(\boldsymbol{\vartheta_0})$ la matrice Hessiana della funzione di log-verosimiglianza calcolata nel punto $\boldsymbol{\vartheta} = \boldsymbol{\vartheta_0}$

Si osservi che $\nabla l(\boldsymbol{\vartheta_0})$ e $\mathbf{H}(\boldsymbol{\vartheta_0})$ sono rispettivamente un vettore e una matrice aleatori, in quanto dipendono dalle realizzazioni $x_1, \ldots, x_n$ della variabile aleatoria X.

Arrestando lo sviluppo in serie di Taylor al secondo ordine si ottiene:

$$l(\boldsymbol{\vartheta}) \simeq l(\boldsymbol{\vartheta_0}) + (\boldsymbol{\vartheta} - \boldsymbol{\vartheta_0})' \nabla l(\boldsymbol{\vartheta_0}) + \frac{1}{2}(\boldsymbol{\vartheta} - \boldsymbol{\vartheta_0})' \mathbf{H}(\boldsymbol{\vartheta_0})(\boldsymbol{\vartheta} - \boldsymbol{\vartheta_0})$$

e derivando entrambi i membri rispetto a $\boldsymbol{\vartheta}$:

$$\frac{\boldsymbol{\partial l}}{\boldsymbol{\partial \vartheta}} \simeq \nabla l(\boldsymbol{\vartheta_0}) + \mathbf{H}(\boldsymbol{\vartheta_0})(\boldsymbol{\vartheta} - \boldsymbol{\vartheta_0})$$

Si osserva che nel punto $\hat{\boldsymbol{\vartheta}}$ soluzione del problema in esame si ha $\nabla l(\hat{\boldsymbol{\vartheta}}) = \mathbf{0}$, quindi valutando entrambi i membri della precedente equazione nel punto $\boldsymbol{\vartheta} = \hat{\boldsymbol{\vartheta}}$ si ottiene:

$$\nabla l(\boldsymbol{\vartheta_0}) + \mathbf{H}(\boldsymbol{\vartheta_0})(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta_0}) \simeq \mathbf{0}$$

da cui si deduce l'approssimazione per $\hat{\boldsymbol{\vartheta}}$:

$$\hat{\boldsymbol{\vartheta}} \simeq \boldsymbol{\vartheta_0} - \mathbf{H}(\boldsymbol{\vartheta_0})^{-1} \nabla(\boldsymbol{\vartheta_0})$$

Poichè, come detto, la matrice Hessiana $\mathbf{H}(\boldsymbol{\vartheta_0})$ è aleatoria al fine di calcolarne il valore si sceglie di utilizzare, al posto delle derivate seconde di $l(\boldsymbol{\vartheta})$, il loro valore atteso. La matrice avente per elementi le quantità

$$-\mathbb{E}\left[\frac{\partial^2 l}{\partial \vartheta_j \partial \vartheta_h}\right] \qquad j, h = 1, \ldots, k$$

è detta matrice di informazione di Fisher ed è indicata con il simbolo $\mathcal{I}_n(\boldsymbol{\vartheta})$.

A partire da un valore iniziale $\boldsymbol{\vartheta_0}$ al passo i-esimo si calcola:

$$\boldsymbol{\vartheta}^{(i)} = \boldsymbol{\vartheta}^{(i-1)} + \mathcal{I}_n(\boldsymbol{\vartheta}^{(i-1)})^{-1}\nabla l(\boldsymbol{\vartheta}^{(i-1)})$$

Il metodo si arresta quando, fissato $\epsilon > 0$ "sufficientemente piccolo", si ha:

$$|\vartheta_j^{(i)} - \vartheta_j^{(i-1)}| < \epsilon \qquad \text{per ogni } j = 1, \ldots, k$$

Ritornando al caso specifico della stima dei parametri $\beta_0, \ldots, \beta_k$ e ϕ per un modello lineare generalizzato risulta a questo punto giustificabile l'affermazione precedente secondo la quale la stima del parametro ϕ possa avvenire in seguito alla stima dei parametri $\beta_0, \ldots, \beta_k$. Si osserva infatti come la stima di questi ultimi sia indipendentemente da ϕ. Applicando l'algoritmo con $\boldsymbol{\vartheta} = [\boldsymbol{\beta}'|\phi]'$ si avrebbe infatti:

$$\boldsymbol{\beta}^{(i)} = \boldsymbol{\beta}^{(i-1)} + \mathcal{I}_n(\boldsymbol{\beta}^{(i-1)}, \phi)^{-1}\nabla l(\boldsymbol{\beta}^{(i-1)}, \phi)$$

Calcolando esplicitamente le derivate seconde della funzione di log-verosimiglianza si ha:

$$\begin{aligned}
\frac{\partial^2 l}{\partial\beta_j\partial\beta_t} &= \frac{\partial}{\partial\beta_t}\left(\frac{\partial l}{\partial\beta_j}\right) = \frac{\partial}{\partial\beta_t}\left[\sum_{i=1}^{n}\frac{(y_i-\mu_i)}{\mathbb{V}\mathrm{ar}[Y_i]}\frac{\partial\mu_i}{\partial\eta_i}x_{ji}\right] \\
&= \sum_{i=1}^{n}\frac{\partial}{\partial\beta_t}\left[\frac{(y_i-\mu_i)}{\mathbb{V}\mathrm{ar}[Y_i]}\frac{\partial\mu_i}{\partial\eta_i}x_{ji}\right]x_{ji} \\
&= \sum_{i=1}^{n}\left[\frac{\partial(y_i-\mu_i)}{\partial\beta_t}\left(\frac{1}{\mathbb{V}\mathrm{ar}[Y_i]}\frac{\partial\mu_i}{\partial\eta_i}\right)x_{ji}\right] + \sum_{i=1}^{n}(y_i-\mu_i)\frac{\partial}{\partial\beta_t}\left(\frac{1}{\mathbb{V}\mathrm{ar}[Y_i]}\frac{\partial\mu_i}{\partial\eta_i}\right)x_{ji}
\end{aligned}$$

Tenendo presente che

$$\frac{\partial(y_i-\mu_i)}{\partial\beta_t} = -\frac{\partial\mu_i}{\partial\beta_t} = -\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_t} = -\frac{\partial\mu_i}{\partial\eta_i}x_{ti}$$

si ottiene

$$\frac{\partial^2 l}{\partial\beta_j\partial\beta_t} = -\sum_{i=1}^{n}\frac{1}{\mathbb{V}\mathrm{ar}[Y_i]}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 x_{ji}x_{ti} + \sum_{i=1}^{n}(y_i-\mu_i)\left[\frac{\partial}{\partial\beta_t}\left(\frac{1}{\mathbb{V}\mathrm{ar}[Y_i]}\frac{\partial\mu_i}{\partial\eta_i}\right)\right]x_{ji}$$

Ricordando inoltre che $\mathbb{V}\text{ar}[Y_i] = a(\phi)b''(\theta_i)$ si deduce che:

$$\begin{aligned}
\left[\mathcal{I}_n(\boldsymbol{\beta},\phi)\right]_{j,t} &= -\mathbb{E}\left[\frac{\partial^2 l}{\partial\beta_j\partial\beta_t}\right] = \frac{1}{a(\phi)}\sum_{i=1}^{n}\frac{1}{b''(\theta_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 x_{ji}x_{ti} \\
\left[\nabla l(\boldsymbol{\beta},\phi)\right]_j &= \frac{1}{a(\phi)}\sum_{i=1}^{n}\frac{y_i-\mu_i}{b''(\theta_i)}\frac{\partial\mu_i}{\partial\eta_i}x_{ji}
\end{aligned}$$

Di conseguenza la stima dei parametri incogniti$\boldsymbol{\beta}^i$ ottenuta tramite l'algoritmo Scoring di Fisher è data da $\boldsymbol{\beta}^{(i)} = \boldsymbol{\beta}^{(i-1)} + \mathcal{I}_n(\boldsymbol{\beta}^{(i-1)},\phi)^{-1}\nabla l(\boldsymbol{\beta}^{(i-1)},\phi)$ e risulta indipendente dal parametro ϕ.

2.3 Trasformazioni dei dati

Accade molto spesso che, nel tentativo di costruire un modello lineare, si vada incontro alla impossibiltà di vedere verificata l'assunzione di Normalità per la variabile risposta. Al fine di ovviare a questo problema la procedura più frequentemente utilizzata è quella di effettuare una trasformazione dei dati. Tra tutte le possibili trasformazioni una classe particolarmente appropriata, e utilizzata qualora i dati siano positivi (come nel caso in esame), è quella che va sotto il nome di trasformazioni di Box-Cox. Queste ultime sono trasformazioni di tipo potenza. Data una serie di realizzazioni $x_1,\ldots,x_n$ esse sono definite come segue:

$$y_i = \begin{cases} \frac{x_i^\lambda - 1}{\lambda} & \text{se } \lambda \neq 0 \\ \ln x_i & \text{se } \lambda = 0 \end{cases}$$

Il parametro λ che identifica la trasformazione è scelto in modo tale da massimizzare la funzione di verosimiglianza dei dati $\mathbf{x} = (x_1 \ldots x_n)$ a partire dall'ipotesi che i dati trasformati $\mathbf{y} = (y_1 \ldots y_n)$ seguano una distribuzione di probabilità Normale di media μ e varianza σ^2. In particolare si ha:

$$f_Y(y|\mu,\sigma^2) = L(\mu,\sigma^2|y) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$$

$$L(\mu, \sigma^2|\mathbf{y}) = \prod_{i=1}^{n} L(\mu, \sigma^2|y_i) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}$$

Inoltre vale la relazione $f_X(x) = f_Y(y) \cdot \left|\frac{dy}{dx}\right| = f_Y(y) \cdot \left|x^{\lambda-1}\right|$, dalla quale si ottiene:

$$L(\mu, \sigma^2|\mathbf{x}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\} \cdot \prod_{i=1}^{n}|x_i|^{\lambda-1}$$

Poichè i valori dei parametri μ e σ^2 non sono noti essi vengono approssimati tramite i rispettivi stimatori di massima verosimiglianza

$$\hat{\mu} = \overline{y}_n = \frac{\sum_{i=1}^{n} y_i}{n}$$
$$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n}(y_i-\overline{y}_n)^2}{n-1}$$

Osservando inoltre che la massimizzazione della funzione di verosimiglianza equivale a quella della funzione di log-verosimiglianza, la trasformata di Box-Cox seleziona il parametro λ tale da massimizzare:

$$\ell(\hat{\mu}, \hat{\sigma}^2|\mathbf{x}) = \ln L(\hat{\mu}, \hat{\sigma}^2|\mathbf{x}) = -\frac{n}{2}\ln 2\pi\hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(y_i-\hat{\mu})^2 + (\lambda-1)\sum_{i=1}^{n}\ln|x_i|$$

Questo procedimento non garantisce tuttavia che i dati trasformati siano effettivamente distribuiti Normalmente. La trasformazione di Box-Cox ottiene infatti il miglior risultato possibile per una trasformazione di tipo potenza, ma non è comunque garantito che la potenza sia una trasformazione ottimale al fine del recupero della normalità dei dati. In generale quindi il

risultato della trasformazione di Box-Cox non assicura la verifica dell'ipotesi di Normalità.

Nel costruire il modello di previsione obiettivo di questo lavoro si incorre esattamente in questo problema. Ciò è probabilmente dovuto all'asimmetria della distribuzione dela variabile risposta (Tempo), che vede la presenza di una coda destra molto pesante. Al fine di determinare una soluzione più soddisfacente (in termini della verifica delle ipotesi sulla distribuzione che sono alla base del modello) si propone allora l'implementazione di una trasformata di tipo potenza il cui obiettivo sia quello, tramite un procedimento analogo a quello di Box-Cox, di ottenere per la variabile risposta una distribuzione di probabilità Gamma di parametri α, β. Le motivazioni che hanno portato alla scelta di tale distribuzione sono le seguenti:

- La distribuzione Gamma risponde alla necessità di avere un supporto positivo, dal momento che la variabile osservata in questione è un tempo;
- La distribuzione Gamma ammette una coda destra più pesante rispetto alla distribuzione Normale
- La distribuzione Gamma appartiene comunuque alla famiglia esponenziale, fatto che consente di sfruttare le proprietà e i procedimenti esposti in precedenza

Una volta ottenuto questo risultato sarà allora possibile costruire un modello lineare generalizzato in cui la variabile risposta sia distribuita come una Gamma, la cui funzione di densità soddisfi quindi la condizione (2.2).

Si noti che una trasformazione di tipo Box-Cox, nel caso in cui il parametro λ ottimale sia diverso da 0, è del tutto equivalente alla trasformazione $y = x^{\lambda}$, infatti una trasformazione lineare dei dati non comporta nessuna differenza sostanziale in quella che sarà la significatività dei parametri del modello che si sta costruendo. Detto ciò, poiché l'interesse è quello di definire una trasformazione che renda la variabile risposta distribuita come una Gamma, si è scelto di implementare tale trasformazione secondo la relazione

$y = x^{\lambda}$ in modo da garantire, per ogni λ, la positività di tutti i valori ottenuti. Seguendo il procedimento mostrato in precedenza si ha:

$$f_Y(y|\alpha,\beta) = L(\alpha,\beta|y) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}$$

$$L(\alpha,\beta|\mathbf{y}) = \prod_{i=1}^{n} L(\alpha,\beta|y_i) = \left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\right)^{n} \left(\prod_{i=1}^{n} y_i\right)^{\alpha-1} \cdot \exp\left\{-\beta \sum_{i=1}^{n} y_i\right\}$$

Poiché vale la relazione $f_X(x) = f_Y(y) \cdot \left|\frac{dy}{dx}\right| = f_Y(y) \cdot |\lambda x^{\lambda-1}|$ si ottiene:

$$L(\alpha,\beta|\mathbf{x}) = \left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\right)^{n} \left(\prod_{i=1}^{n} y_i\right)^{\alpha-1} \cdot \exp\left\{-\beta \sum_{i=1}^{n} y_i\right\} \cdot |\lambda|^n \prod_{i=1}^{n} |x_i|^{\lambda-1}$$

Come nel caso precedente i parametri α e β della distribuzione di riferimento non sono noti e dovono, di conseguenza, essere stimati. Le stime utilizzate sono le seguenti:

$$\hat{\alpha} = \frac{\overline{y}_n^2}{s^2} \qquad\qquad \hat{\beta} = \frac{\overline{y}_n}{s^2}$$

Il parametro λ ottimale sarà allora quello che massimizza la funzione di log-verosimiglianza

$$\ell(\hat{\alpha},\hat{\beta}|\mathbf{x}) = n \ln \frac{\hat{\beta}^{\hat{\alpha}}}{\Gamma(\hat{\alpha})} + (\hat{\alpha}-1) \sum_{i=1}^{n} \ln y_i - \hat{\beta} \sum_{i=1}^{n} y_i + n \ln|\lambda| + (\lambda-1) \sum_{i=1}^{n} \ln|x_i|$$

Capitolo 3

Analisi dei Dati

In questo capitolo saranno esposti inizialmente i criteri che hanno portato alla scelta delle covariate per il modello di previsione e una spiegazione delle stesse; successivamente saranno riportati, a titolo di esempio, il procedimento seguito e i risultati ottenuti per una struttura ospedaliera di riferimento (il Niguarda di Milano); infine sarà presentato un riassunto dei risultati ottenuti e alcuni esempi sul funzionamento del pacchetto `R` creato.

Il modello che si è scelto di implementare è un modello di tipo lineare in cui la variabile risposta sia il Tempo e in cui le covariate siano le seguenti:

- Codice Paziente: fattore a 3 livelli (Verde, Giallo, Rosso)
- Tipo di Giorno: fattore a 2 livelli (Feriale, Festivo)
- Fascia Oraria: fattore a 3 livelli (Ore di Punta, Giorno, Notte)
- Zona: fattore a 4 livelli (Nord-Ovest, Nord-Est, Sud-Est, Sud-Ovest)
- Distanza

Oltre a questi fattori, la cui scelta è già stata motivata nel Capitolo 1, nell'ambito dell'analisi è stato ritenuto opportuno aggiungere, affinchè il modello colga e descriva meglio la variabilità del fenomeno in esame, le seguenti interazioni a due fattori:

- Interazione Tipo di Giorno / Fascia Oraria: è ipotizzabile che in un giorno festivo, durante il quale il flusso del traffico è sicuramente meno consistente rispetto a quello di un giorno feriale, l'effetto del fattore Fascia Oraria possa essere meno significativo. In particolare è ragionevole pensare che in un giorno festivo il tempo medio impiegato per raggiungere un dato ospedale sia influenzato in modo minore dalla fascia oraria.

- Interazione Distanza / Zona: questa covariata ha lo scopo di descrivere il flusso del traffico al variare della velocità di percorrenza delle strade nella provincia di Milano. Questa interazione potrà ad esempio risultare significativa nel caso in cui in una determinata zona vi sia le presenza di strade a scorrimento veloce. In questo caso infatti saranno considerati significativiti sia il fattore principale Zona, poichè in media il tempo di arrivo in ospedale sarà inferiore rispetto ad altre zone, sia l'interazione Distanza / Zona, in quanto il tempo medio di percorrenza sarà influenzato in modo minore dalla distanza.

L'analisi è stata effettuata inizialmente trasformando la variabile risposta in modo tale che essa soddisfi le ipotesi sulla distribuzione dell'errore casuale che sono alla base del modello. In particolare, poiché la trasformazione di Box-Cox non ha garantito risultati soddisfacenti in termini di p-value del test di normalità di Anderson-Darling, si è scelto di trasformare i dati per "adattarli" a una distribuzione di probabilità di tipo $\Gamma(\alpha, \beta)$ (secondo quanto esposto nella sezione 2.3) e sfruttare poi la teoria dei modelli lineari generalizzati. Sebbene questa sia la motivazione principale che ha portato a questa scelta è anche necessario sottilineare che la verifica delle ipotesi alla base del modello non è stato l'unico criterio adottato per stabilire la bontà dell'analisi. Una volta identificato il modello migliore per ogni ospedale sia in termini di buon adattamento, tramite l'analisi dei residui e il calcolo dell'AIC, sia in termini di verifica delle ipotesi sulla distribuzione dei dati, al fine di valutarne le capacità predittive è stato calcolato via cross-validazione l'errore medio di previsione. Anche sotto questo aspetto il modello migliore risulta quello nel quale la variabile risposta è stata trasformata in una Gamma.

Dal momento che ai fini della costruzione del modello si è reso necessario l'inserimento di covariate sia di tipo continuo che di tipo categorico si è scelto di trattare queste ultime introducendo variabili *dummy* per ognuno dei livelli dei fattori qualitativi tenuti in considerazione. Il modello del quale ci si propone di stimare i parametri risulta pertanto della forma:

$$\begin{aligned}
\mathbb{E}\left[\text{Tempo}_i^{\lambda}\right] &= \beta_0 + \beta_1 \cdot \text{Giallo}_i + \beta_2 \cdot \text{Rosso}_i + \beta_3 \cdot \text{OraDiPunta}_i \\
&+ \beta_4 \cdot \text{Notte}_i + \beta_5 \cdot \text{Festivo}_i + \beta_6 \cdot \text{NordOvest}_i + \beta_7 \cdot \text{SudOvest}_i \\
&+ \beta_8 \cdot \text{SudEst}_i + \beta_9 \cdot \text{Distanza}_i + \beta_{10} \cdot \text{OraDiPunta}_i \cdot \text{Festivo}_i \\
&+ \beta_{11} \cdot \text{Notte}_i \cdot \text{Festivo}_i + \beta_{12} \cdot \text{Distanza}_i \cdot \text{NordOvest}_i \\
&+ \beta_{13} \cdot \text{Distanza}_i \cdot \text{SudOvest}_i + \beta_{14} \cdot \text{Distanza}_i \cdot \text{SudEst}_i
\end{aligned}$$

dove:

- le variabili Giallo, Rosso, OraDiPunta, Notte, Festivo, NordOvest, SudOvest, SudEst sono di tipo dummy, ovvero variabili dicotomiche che assumono valore 1 nel caso in cui il dato in questione soddisfi la condizione che dà il nome alla variabile e assumono valore 0 altrimenti
- λ è il valore ottimale ottenuto dalla trasformazione dei dati

In questo modello l'espressione $\beta_0 + \beta_9 \cdot \text{Distanza}$ rappresenta, fissata una certa distanza, il tempo medio di arrivo in un dato ospedale nel caso in cui al paziente che viene soccorso sia assegnato un codice Verde, la chiamata sia stata effettuata in un giorno Feriale durante la fascia oraria che abbiamo definito con Giorno e dal quadrante di Nord-Est rispetto all'ospedale.

Il valore assunto dagli altri coefficienti β_i rappresenta la differenza nella stima del tempo medio impiegato per raggiungere l'ospedale che si ha quando la covariata i-esima è "attiva". Ad esempio il coefficiente β_2 rappresenta la differenza nella stima del tempo medio impiegato nel caso in cui al paziente trasportato sia assegnato codice Rosso invece che Verde: ci si aspetta quindi che, in generale, il coefficiente β_2 sia negativo.

Nella sezione seguente viene riportato un esempio significativo, relativo all'ospedale Niguarda di Milano, nel quale viene esposto nel dettaglio il

procedimento utilizzato per la costruzione del modello di previsione. Tale procedimento è esemplificativo di quello adottato per ogni struttura ospedaliera oggetto dell'analisi. I risultati relativi ad ogni ospedale sono riportati nella sezione 3.2.

3.1 Procedimento per la stima dei parametri del modello

Per la costruzione del modello di previsione viene effettuata un'analisi in 3 passi:

1. Trasformazione della variabile risposta e descrizione qualitativa dei dati
2. Stima dei parametri tramite un modello lineare generalizzato
3. Verifica delle ipotesi e della capacità di previsione

3.1.1 Analisi preliminare dei dati

Il primo passo per la costruzione del modello di previsione consiste nella trasformazione della variabile risposta (Tempo) in modo tale che la distribuzione di probabilità empirica relativa a questa variabile sia adattata a una distribuzione $\Gamma(\alpha, \beta)$. Applicando la funzione il cui codice `R` è riportato nella sezione 4.2 si ottiene:

```
> trasformata_gamma(tempo)
[1] 0.26
```

La variabile risposta per il modello sarà dunque $\text{Tempo}^{0.26}$.

In Figura 3.1 si osserva un istogramma normalizzato della variabile Tempo prima e dopo la trasformazione. Come si può notare la distribuzione empirica della variabile trasformata approssima in modo evidentemente migliore la distribuzione di una Gamma. Tale supposizione è confermata dal p-value del test Chi Quadrato, che risulta pari a 0.34783. Sebbene questo valore non sia molto alto esso può comunque essere considerato accettabile, soprattutto in

considerazione del fatto che applicando la trasformazione classica di Box-Cox il p-value del test di normalità della distribuzione è inferiore a $2 \cdot 10^{-16}$.

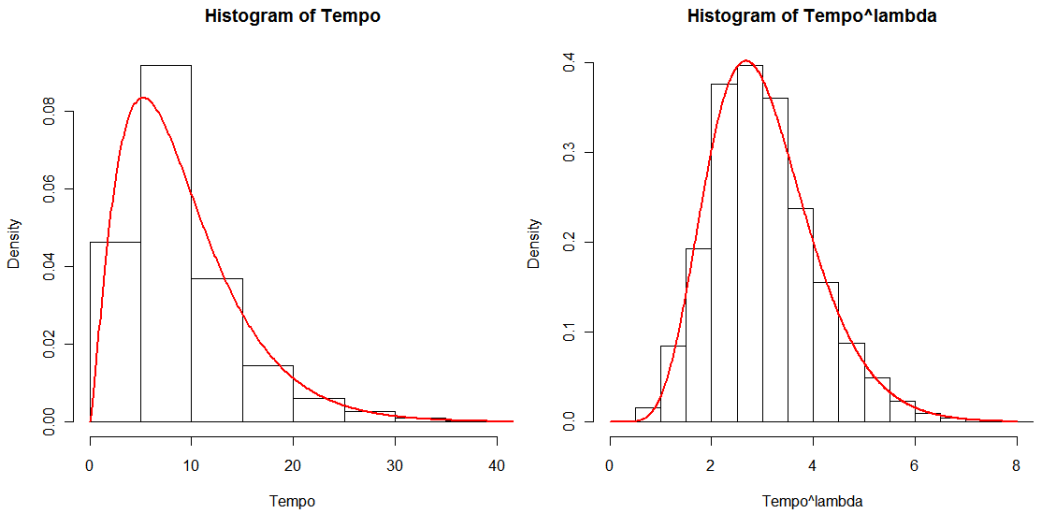
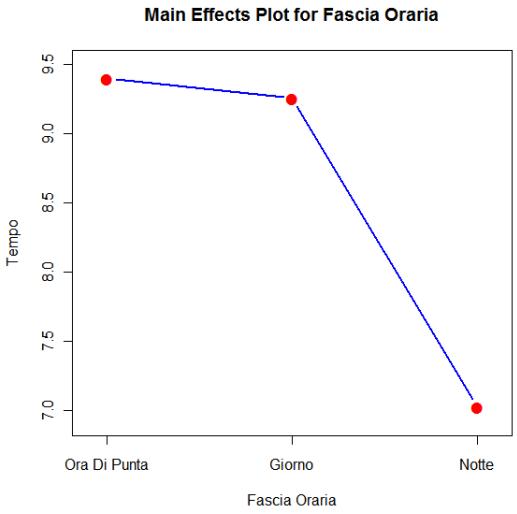


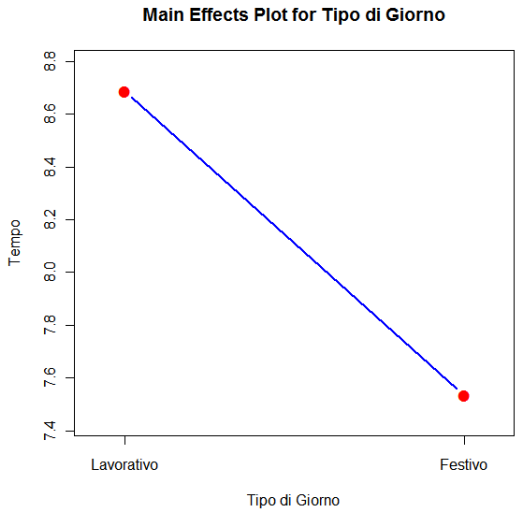
Figura 3.1: Istogramma normalizzato per la variabile Tempo ripettivamente prima e dopo la trasformazione.

In Figura 3.2 vengono riportati i grafici riguardanti l'effetto dei fattori principali. In ognuno di questi grafici il punto in rosso rappresenta la media della variabile risposta in corrispondenza di ciascun livello del fattore considerato. L'analisi di questi grafici conferma quelle che sono le supposizioni iniziali riguardo al ruolo dei fattori principali, si osserva infatti che il tempo medio impiegato per raggiungere l'ospedale:

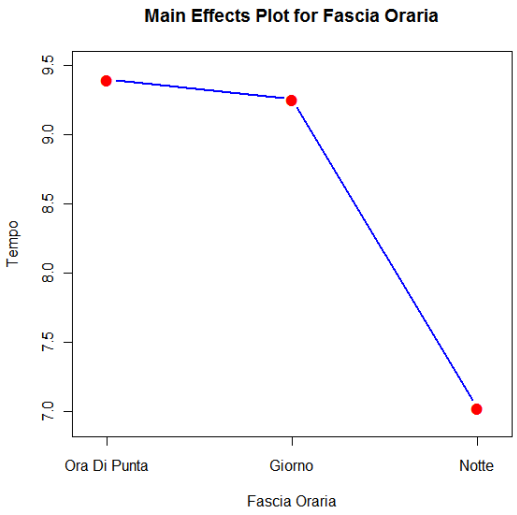
1. diminuisce con l'aumentare della gravità del paziente trasportato
2. diminuisce nei gioni festivi rispetto ai giorni feriali
3. aumenta nelle ore di punta e diminuisce notevolmente nelle ore notturne
4. ha un andamento casuale al variare del quadrante di partenza



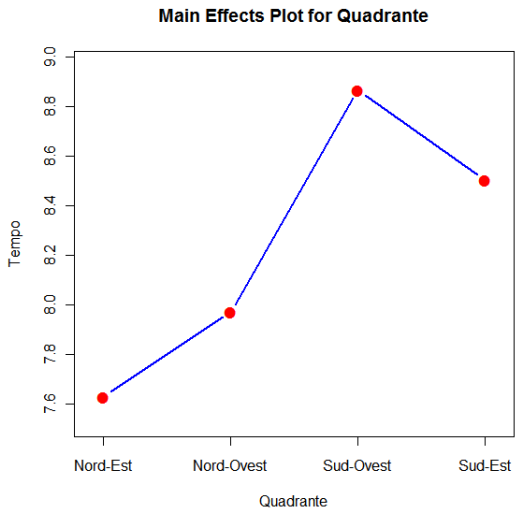
(3.2.1)



(3.2.2)



(3.2.3)



(3.2.4)

Figura 3.2: Grafici degli effetti principali rispettivamente per le variabili Codice (3.2.1), Tipo di Giorno (3.2.2), Fascia Oraria (3.2.3) e Zona (3.2.4)

In Figura 3.3 viene riportato un grafico relativo alla possibile significatività dell'interazione tra le variabili Tipo Di Giorno e Fascia Oraria.

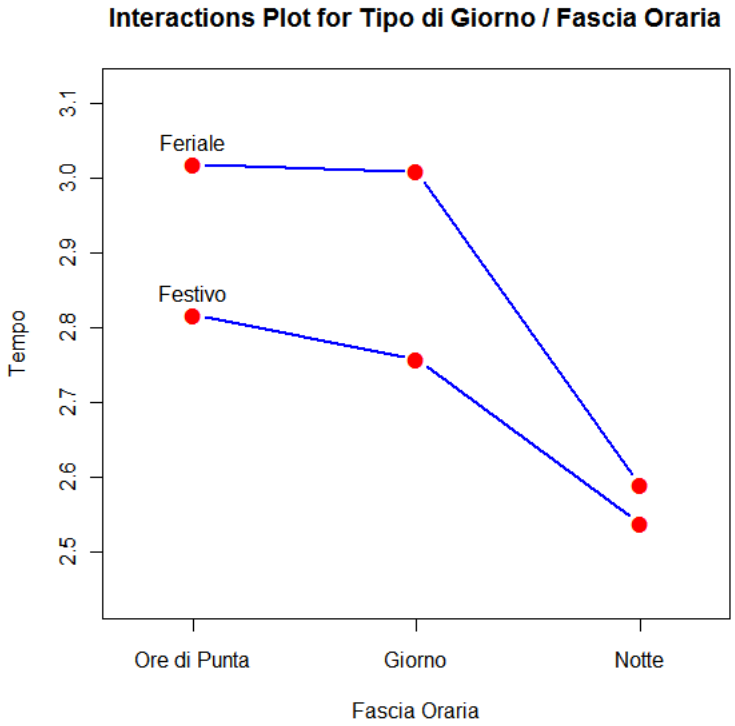


Figura 3.3: Grafico dell'interazione tra le variabili Tipo di Giorno e Fascia Oraria

Come si può osservare durante la fascia oraria notturna si ha un notevole decremento della differenza tra il tempo medio di trasporto all'ospedale tra i giorni lavorativi e i giorni festivi. Anche questo grafico sembra conferma quindi le supposizioni iniziali e giustifica ulteriormente l'introduzione nel modello dell'interazione tra i fattori Tipo Di Giorno e Fascia Oraria.

L'analisi di questi grafici non può comunque essere considerata esaustiva, non è infatti garantito che le variazioni tra le medie dei tempi di percorrenza osservate siano statisticamente significative. A questa conclusione si potrà giungere soltanto dall'analisi dei p-value dei test di significatività delle singole covariate che verrà effettuata nella sezione seguente.

3.1.2 Stima dei parametri del modello

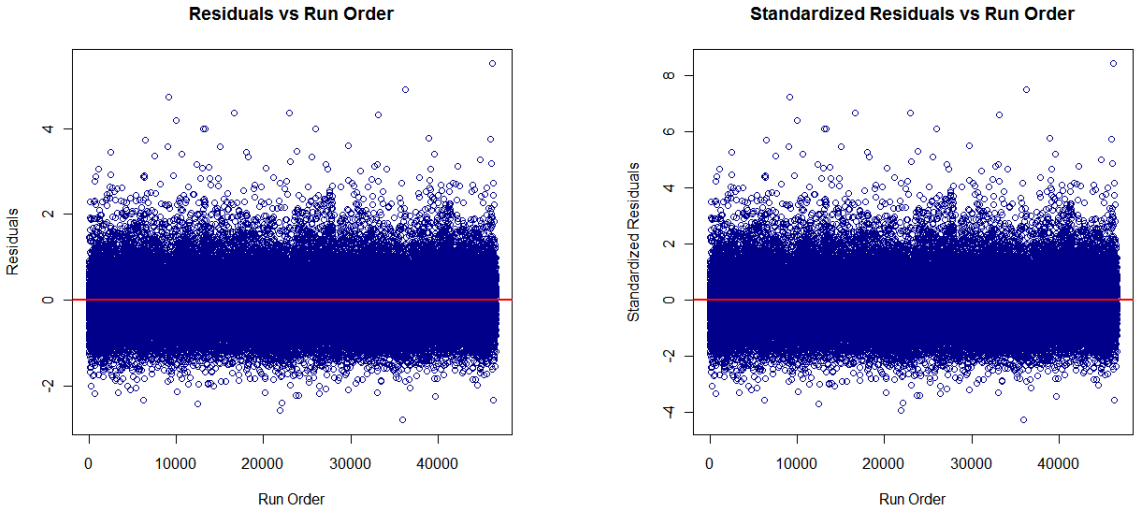
Si riporta di seguito l'output associato al modello relativo alla struttura ospedaliera Niguarda di Milano, ottenuto mediante la chiamata del comando `glm()` del software `R`.

```
Call:
glm(formula = tempo_lambda ~ giallo + rosso + ora_di_punta +
    notte + festivo + q_n_o + q_s_o + q_s_e + distanza + ora_di_punta:festivo +
    notte:festivo + q_n_o:distanza + q_s_o:distanza + q_s_e:distanza,
    family = Gamma(link = "identity"), maxit = 400)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-0.3562452  -0.0386650   0.0003385   0.0375786   0.2982107

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.252e+00  1.662e-03 753.298  < 2e-16 ***
giallo               -2.149e-02  7.321e-04 -29.353  < 2e-16 ***
rosso                -4.820e-02  1.520e-03 -31.711  < 2e-16 ***
ora_di_punta          1.391e-02  1.025e-03  13.568  < 2e-16 ***
notte                -3.662e-02  8.753e-04 -41.834  < 2e-16 ***
festivo              -2.619e-02  1.463e-03 -17.899  < 2e-16 ***
q_n_o                -3.061e-02  2.285e-03 -13.396  < 2e-16 ***
q_s_o                 2.716e-02  2.194e-03  12.377  < 2e-16 ***
q_s_e                -2.339e-02  1.932e-03 -12.106  < 2e-16 ***
distanza              3.455e-05  3.681e-07  93.872  < 2e-16 ***
ora_di_punta:festivo -8.921e-03  2.535e-03  -3.519 0.000433 ***
notte:festivo         1.872e-02  2.070e-03   9.041  < 2e-16 ***
q_n_o:distanza       -6.675e-06  4.847e-07 -13.773  < 2e-16 ***
q_s_o:distanza       -7.221e-06  4.678e-07 -15.435  < 2e-16 ***
q_s_e:distanza        7.896e-06  4.612e-07  17.119  < 2e-16 ***

(Dispersion parameter for Gamma family taken to be 0.003627001)

    Null deviance: 412.03  on 54016  degrees of freedom
Residual deviance: 197.43  on 54002  degrees of freedom
AIC: -116728
Number of Fisher Scoring iterations: 4
```

Come si può osservare, ai fini della spiegazione della variabilità della risposta, tutti i fattori possono essere considerati altamente significativi.

3.1.3 Verifica della bontà del modello

Posto che l'adattamento della distribuzione empirica dei dati a una distribuzione di tipo Gamma è già stato verificato e che per la costruzione di un modello lineare generalizzato non è necessaria l'ipotesi di omoschedasticità dei dati, per verificare la bontà del modello si è scelto innanzitutto di analizzarne i residui. A tale scopo in Figura 3.5 sono riportati i grafici rispettivamente dei residui e dei residui standardizzati. Tali grafici presentano in ascissa l'indice che identifica la posizione del residuo all'interno del dataset, detto *Run Order*.



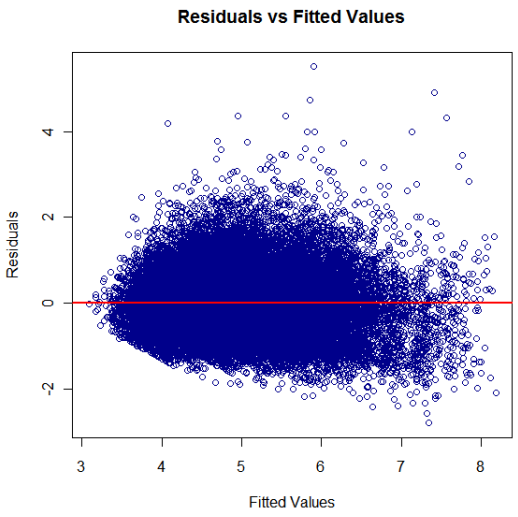
(3.4.1) (3.4.2)

Figura 3.4: (3.4.1) Residui vs Run Order; (3.4.2) Residui standardizzati vs Run Order

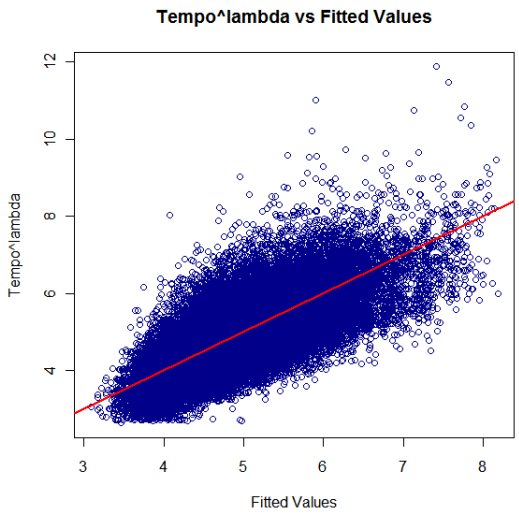
Si osserva che i residui e i residui standardizzati sono distribuiti abbastanza uniformemente attorno al valore 0. Per quanto riguarda i residui standardizzati si può notare la presenza di valori superiori, in modulo, a quelli che sono considerati i limiti di accettabilità. Da un'analisi più approfondita si verifica che i residui standardizzati aventi valore assoluto maggiore

di 4 sono 122 su un totale di 54017 (circa lo 0.22%); i residui standardizzati aventi valore assoluto maggiore di 3 sono invece 486 su un totale di 54017 (circa lo 0.9%). Si può quindi concludere che sebbene la presenza di residui con valore assoluto così alto non sia una buona caratteristica del modello la loro presenza non può essere considerata anomala, visto il numero così basso (in proporzione) con cui essi appaiono. Si osserva inoltre che la maggior parte dei residui con valore assoluto elevato sono positivi, si deduce quindi la presenza di dati, che potrebbero essere considerati outliers, che sono probabilmente relativi a trasporti di pazienti avvenuti molto più lentamente di quanto non avvengano in media. Poichè però non si hanno informazioni in merito alle motivazioni che potrebbero aver causato dati così anomali si è scelto di non escluderli dal modello.

In Figura 3.5.1 è riportato il grafico dei residui in funzione dei valori stimati dal modello. In Figura 3.5.2 è invece riportato il grafico dei valori della variabile risposta in funzione dei valori stimati.



(3.5.1)



(3.5.2)

Figura 3.5: (3.5.1) Grafico dei residui in funzione dei valori stimati; (3.5.2) Grafico della variabile risposta in funzione dei valori stimati

Anche sotto questo aspetto sembra essere assicurata una buona capacità descrittiva del modello. Nel primo di essi si osserva che, escludendo valori i valori temporali più bassi per i quali la varianza dei residui appare inferiore, la distribuzione di questi ultimi non è dipendente dai valori stimati. Nel secondo grafico si può invece notare come i valori stimati approssimino bene l'andamento della variabile risposta. L'unico aspetto negativo può essere il fatto che il modello sovrastima, seppur non in maniera troppo evidente, i tempi medi di arrivo in ospedale quando questi sono molto elevati.

Poichè l'obiettivo finale non è quello di descrivere i dati bensì di poterli prevedere si è scelto di utilizzare come indice per la bontà del modello l'errore medio di previsione calcolato via cross-validazione. A tale scopo viene costruita una partizione casuale del dataset in cui ogni gruppo contiene l'1% dei dati. Per ognuno di questi gruppi viene in seguito calcolato l'errore medio di previsione sulla base del modello i cui parametri sono stimati utilizzando il restante 99% dei dati a disposizione. Per l'ospedale Niguarda si ottiene una stima dell'errore medio di previsione pari a 2 minuti e 45 secondi, che equivale a un errore medio percentuale pari al 33.55%. Tale risultato è comparabile a quello ottenuto per le altre strutture ospedaliere (si veda la sezione 3.2) ed è comunque migliore di quello che si otterrebbe scegliendo di implementare un modello lineare "classico", ovvero con errore casuale distribuito Normalmente

3.2 Riassunto dei risultati ottenuti

In questa sezione sono riportati i risultati ottenuti per ogni strutture ospedaliera oggetto dell'analisi. Deve essere sottolineato che per gli ospedali di Treviglio, Legnano, Pavia e Monza, a causa della scarsa numerosità dei dati a disposizione, non è stato possibile stimare i coefficienti relativi ai livelli del fattore Zona. Tali coefficienti sono quindi stati posti uguali a 0 a priori.

Ospedale	**35**	**9**	**23**	**34**	**19**
$\boldsymbol{\lambda}$	0.64	1.2	0.53	0.86	1.39
$\hat{\boldsymbol{\beta}}_{\mathbf{0}}$	1.74 e+00	2.31 e+00	1.53 e+00	1.23 e+00	1.35 e+00
$\hat{\boldsymbol{\beta}}_{\mathbf{1}}$	-1.96 e-01	-5.90 e-01	-1.62 e-01	-2.81 e-01	-1.18 e+00
$\hat{\boldsymbol{\beta}}_{\mathbf{2}}$	-4.91 e-01	-1.70 e+00	-4.06 e-01	-6.63 e-01	-2.02 e+00
$\hat{\boldsymbol{\beta}}_{\mathbf{3}}$	1.50 e-01	1.21 e+00	-3.73 e-02	3.22 e-01	6.08 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{4}}$	-4.09 e-01	-8.45 e-01	-4.61 e-01	-6.52 e-01	-1.54 e+00
$\hat{\boldsymbol{\beta}}_{\mathbf{5}}$	-2.64 e-01	-4.43 e-01	-4.00 e-01	-4.25 e-01	-1.37 e+00
$\hat{\boldsymbol{\beta}}_{\mathbf{6}}$	-1.54 e-02	-1.74 e-00	4.42 e-01	1.10 e-00	3.86 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{7}}$	4.07 e-01	-1.37 e+00	4.53 e-02	1.07 e-00	3.12 e+00
$\hat{\boldsymbol{\beta}}_{\mathbf{8}}$	-1.14 e-01	-7.03 e-01	6.34 e-01	1.50 e-01	2.76 e+00
$\hat{\boldsymbol{\beta}}_{\mathbf{9}}$	6.00 e-04	2.45 e-03	4.34 e-04	1.14 e-03	4.17 e-03
$\hat{\boldsymbol{\beta}}_{\mathbf{10}}$	-1.49 e-01	-1.14 e-00	5.98 e-02	-2.39 e-01	-5.84 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{11}}$	2.29 e-01	4.10 e-01	3.51 e-01	3.93 e-01	1.13 e+00
$\hat{\boldsymbol{\beta}}_{\mathbf{12}}$	-1.38 e-04	-1.08 e-05	-4.62 e-05	1.98 e-05	5.57 e-04
$\hat{\boldsymbol{\beta}}_{\mathbf{13}}$	-1.77 e-04	5.08 e-04	-1.68 e-05	-2.87 e-04	-3.99 e-04
$\hat{\boldsymbol{\beta}}_{\mathbf{14}}$	-2.16 e-06	1.39 e-04	-9.24 e-05	-3.12 e-04	-2.48 e-04
AIC	142968	101371	112865	210870	146253
Errore	2.22	2.01	2.91	2.49	2.35
Errore (%)	33.91%	30.58%	41.64%	37.59%	31.65%

Tabella 3.1: Risultati ottenuti per gli ospedali San Carlo, Cernusco Sul Naviglio, Policlinico, San Paolo, Magente

Ospedale	**37**	**48**	**21**	**2227**	**24**
$\boldsymbol{\lambda}$	0.7	1.02	1.06	0.5	0.54
$\hat{\boldsymbol{\beta}}_{\mathbf{0}}$	1.97 e+00	1.63e+00	2.46 e+00	1.93 e+00	1.71 e+00
$\hat{\boldsymbol{\beta}}_{\mathbf{1}}$	-1.19 e-01	-3.70 e-01	-8.44 e-01	-1.12 e-01	-1.12 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{2}}$	-7.87 e-01	-1.07 e-00	-1.38 e+00	-2.84 e-01	-4.17 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{3}}$	3.51 e-01	3.56 e-01	2.71 e-01	6.66 e-02	1.13 e-02
$\hat{\boldsymbol{\beta}}_{\mathbf{4}}$	-4.53 e-01	-6.49 e-01	-7.23 e-01	-9.17 e-02	-3.83 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{5}}$	-3.80 e-01	-4.97 e-01	-1.32 e-01	-4.51 e-02	-2.98 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{6}}$	7.38 e-01	-3.12 e-01	5.82 e-01	3.81 e-02	1.40 e-02
$\hat{\boldsymbol{\beta}}_{\mathbf{7}}$	9.48 e-01	4.55 e-01	-1.68 e-01	-2.31 e-01	3.98 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{8}}$	2.25 e-01	1.66 e-00	-4.63 e-01	-2.00 e-01	1.97 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{9}}$	7.49 e-04	1.95 e-03	1.20 e-03	2.03 e-04	5.68 e-04
$\hat{\boldsymbol{\beta}}_{\mathbf{10}}$	-3.16 e-01	-2.96 e-01	-7.66 e-01	-5.23 e-02	4.05 e-02
$\hat{\boldsymbol{\beta}}_{\mathbf{11}}$	1.97 e-01	2.92 e-01	1.93 e-02	5.51 e-02	2.62 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{12}}$	-1.86 e-04	-3.58 e-04	1.82 e-04	4.76 e-06	-9.60 e-05
$\hat{\boldsymbol{\beta}}_{\mathbf{13}}$	-1.21 e-04	-4.15 e-04	5.07 e-04	-4.54 e-05	-1.35 e-04
$\hat{\boldsymbol{\beta}}_{\mathbf{14}}$	7.01 e-05	-3.77 e-04	1.16 e-04	3.97 e-06	-8.66 e-05
AIC	14412	94372	96080	37420	95601
Errore	3.41	2.16	2.55	2.61	2.72
Errore (%)	33.79%	34.72%	31.59%	28.01%	40.40%

Tabella 3.2: Risultati ottenuti per gli ospedali Galeazzi, Rho, Melegnano, Humanitas, Fatebenefratelli

Ospedale	**112**	**14**	**29**	**33**	**85**
$\boldsymbol{\lambda}$	0.46	0.17	0.15	0.23	1.1
$\hat{\boldsymbol{\beta}}_{\mathbf{0}}$	1.66 e+00	1.28 e+00	1.25 e+00	1.48 e+00	2.46 e+00
$\hat{\boldsymbol{\beta}}_{\mathbf{1}}$	-9.44 e-02	-2.81 e-02	-2.15 e-02	-4.15 e-02	-6.26 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{2}}$	-2.46 e-01	-4.16 e-02	-4.82 e-02	-8.86 e-02	-1.02 e-00
$\hat{\boldsymbol{\beta}}_{\mathbf{3}}$	4.11 e-02	6.48 e-03	1.39 e-02	2.73 e-02	6.17 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{4}}$	-2.25 e-01	-1.08 e-02	-3.66 e-02	-5.18 e-02	-5.69 -01
$\hat{\boldsymbol{\beta}}_{\mathbf{5}}$	-1.83 e-01	-7.19 e-03	-2.62 e-02	-4.14 e-02	-3.70 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{6}}$	-1.37 e-02	-1.81 e-02	-3.06 e-02	-3.11 e-03	5.50 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{7}}$	-4.08 e-02	8.86 e-03	2.71 e-02	-3.09 e-02	5.53 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{8}}$	1.40 e-02	-9.66 e-03	-2.34 e-02	-7.03 e-03	-1.14 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{9}}$	3.49 e-04	3.72 e-05	3.46 e-05	4.01 e-05	1.53 e-03
$\hat{\boldsymbol{\beta}}_{\mathbf{10}}$	-2.83 e-03	-1.85 e-03	-8.92 e-03	-2.79 e-02	-4.44 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{11}}$	1.73 e-01	1.09 e-02	1.87 e-02	3.24 e-02	3.82 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{12}}$	6.26 e-05	-6.65 e-07	-6.68 e-06	9.63 e-06	5.15 e-04
$\hat{\boldsymbol{\beta}}_{\mathbf{13}}$	-3.23 e-06	-4.77 e-06	-7.22 e-06	1.21 e-05	5.44 e-04
$\hat{\boldsymbol{\beta}}_{\mathbf{14}}$	-3.24 e-05	4.49 e-07	7.90 e-06	1.88 e-06	2.70 e-04
AIC	37112	-41027	-116727	-350066	91459
Errore	2.30	2.14	2.76	2.87	1.94
Errore (%)	41.20%	25.75%	33.55%	31.50%	31.28%

Tabella 3.3: Risultati ottenuti per gli ospedali Santa Rita, Garbagnate, Niguarda, San Raffaele, Saronno

Ospedale	**45**	**118**	**10**	**36**	**50**
$\boldsymbol{\lambda}$	0.3	0.6	0.3	0.34	0.66
$\hat{\boldsymbol{\beta}}_{\mathbf{0}}$	1.47 e+00	1.84 e+00	1.79 e+00	1.10 e+00	2.43 e+00
$\hat{\boldsymbol{\beta}}_{\mathbf{1}}$	-4.03 e-02	-1.57 e-01	-3.31 e-02	-5.43 e-02	-1.45 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{2}}$	-9.62 e-02	-3.43 e-01	-1.12 e-02	-1.05 e-01	-4.49 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{3}}$	1.73 e-02	-6.06 e-02	3.47 e-02	4.04 e-02	2.48 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{4}}$	-6.44 e-02	-4.99 e-01	-7.75 e-02	-1.08 e-1	-1.10 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{5}}$	-3.72 e-02	-4.21 e-01	-4.14 e-02	-6.73 e-02	-7.74 e-02
$\hat{\boldsymbol{\beta}}_{\mathbf{6}}$	4.10 e-02	2.56 e-01	2.65 e-01	2.31 e-01	-4.45 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{7}}$	-5.33 e-02	2.05 e-01	-3.34 e-01	9.10 e-01	-7.91 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{8}}$	-1.08 e-01	-5.80 e-02	-3.37 e-01	-5.34 e-01	-5.80 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{9}}$	8.19 e-05	6.50 e-04	2.28 e-05	9.81 e-05	3.79 e-04
$\hat{\boldsymbol{\beta}}_{\mathbf{10}}$	-5.13 e-03	1.78 e-01	-4.54 e-02	-3.98 e-02	-2.70 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{11}}$	1.05 e-02	3.29 e-01	2.35 e-02	4.66 e-02	2.70 e-03
$\hat{\boldsymbol{\beta}}_{\mathbf{12}}$	2.9 e-05	-2.43 e-06	1.05 e-04	-1.38 e-05	1.21 e-04
$\hat{\boldsymbol{\beta}}_{\mathbf{13}}$	3.42 e-05	-1.31 e-04	1.16 e-04	-7.57 e-05	1.45 e-04
$\hat{\boldsymbol{\beta}}_{\mathbf{14}}$	4.26 e-05	-6.11 e-05	1.84 e-04	3.42 e-05	1.07 e-04
AIC	-3029	21815	-4301	8832	31650
Errore	1.88	2.49	1.98	2.45	2.32
Errore (%)	32.07%	40.75%	30.85%	33.13%	36.37%

Tabella 3.4: Risultati ottenuti per gli ospedali Paderno Dugnano, San Giuseppe, Cinisello Balsamo, Sacco, San Donato Milanese

Ospedale	**22**	**110**	**25**	**51**	**2185**
$\boldsymbol{\lambda}$	0.64	0.32	0.25	0.29	0.38
$\hat{\boldsymbol{\beta}}_{\mathbf{0}}$	2.03 e+00	1.46 e+00	2.12 e+00	1.65 e+00	1.78 e+00
$\hat{\boldsymbol{\beta}}_{\mathbf{1}}$	-3.18 e-01	-1.87 e-01	-1.96 e-01	-1.85 e-01	-2.25 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{2}}$	-8.29 e-01	-2.88 e-01	-2.52 e-01	-5.03 e-01	-5.84 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{3}}$	3.27 e-01	-9.23 e-04	-5.06 e-02	2.07 e-01	1.52 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{4}}$	-3.95 e-01	-1.38 e-01	-2.87 e-01	-3.90 e-01	-2.60 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{5}}$	-1.76 e-01	-1.50 e-01	-1.79 e-01	-1.28 e-01	-1.76 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{6}}$	-4.16 e-02	2.09 e-01	1.09 e-01	1.86 e-01	1.44 e+00
$\hat{\boldsymbol{\beta}}_{\mathbf{7}}$	8.50 e-01	6.79 e-01	8.03 e-02	2.05 e-01	6.50 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{8}}$	9.11 e-01	2.97 e-01	8.03 e-02	1.41 e-01	2.02 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{9}}$	9.76 e-04	2.87 e-04	1.95 e-04	1.10 e-03	6.67 e-04
$\hat{\boldsymbol{\beta}}_{\mathbf{10}}$	-4.17 e-01	6.83 e-02	3.54 e-02	-3.57 e-01	3.30 e-03
$\hat{\boldsymbol{\beta}}_{\mathbf{11}}$	-1.16 e-01	7.05 e-02	2.06 e-01	1.36 e-01	1.63 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{12}}$	2.12 e-04	1.01 e-04	-8.76 e-06	-9.82 e-05	-3.26 e-04
$\hat{\boldsymbol{\beta}}_{\mathbf{13}}$	5.49 e-05	-4.08 e-05	-2.37 e-05	5.06 e-05	-1.41 e-04
$\hat{\boldsymbol{\beta}}_{\mathbf{14}}$	1.15 e-04	-8.59 e-06	-3.07 e-05	1.37 e-04	3.68 e-05
AIC	65791	8461	4295	42340	21190
Errore	2.13	2.63	3.84	2.23	2.69
Errore (%)	29.83%	35.20%	36.27%	41.83%	34.74%

Tabella 3.5: Risultati ottenuti per gli ospedali Melzo, Cardiologico, Alfieri, Sesto San Giovanni, Multimedica

Ospedale	**3**	**32**	**1**	**12**	**27**
$\boldsymbol{\lambda}$	0.43	0.46	0.51	0.61	0.9
$\hat{\boldsymbol{\beta}}_{\mathbf{0}}$	1.73 e+00	1.85 e+00	1.76 e+00	2.11 e+00	2.35 e+00
$\hat{\boldsymbol{\beta}}_{\mathbf{1}}$	-8.64 e-02	-1.13 e-01	-8.51 e-02	-8.74 e-02	4.53 e-02
$\hat{\boldsymbol{\beta}}_{\mathbf{2}}$	-2.19 e-01	-1.08 e-01	-2.29 e-01	-2.73 e-01	1.81 e+00
$\hat{\boldsymbol{\beta}}_{\mathbf{3}}$	4.81 e-02	1.52 e-03	3.67 e-02	5.87 e-02	2.85 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{4}}$	-1.53 e-01	-2.17 e-01	-1.29 e-01	2.18 e-02	-1.02 e+00
$\hat{\boldsymbol{\beta}}_{\mathbf{5}}$	-5.87 e-02	-2.14 e-01	-4.04 e-02	1.02 e-01	-1.04 e+00
$\hat{\boldsymbol{\beta}}_{\mathbf{6}}$	-1.11 e-01	1.15 e-01	4.62 e-03	-5.49 e-01	-1.03 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{7}}$	-4.16 e-02	1.44 e-01	-3.66 e-01	-1.13 e+00	1.70 e-00
$\hat{\boldsymbol{\beta}}_{\mathbf{8}}$	-1.92 e-01	-4.25 e-02	-7.57 e-03	-9.68 e-01	-2.88 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{9}}$	2.05 e-04	2.92 e-04	2.19 e-04	3.26 e-04	2.01 e-03
$\hat{\boldsymbol{\beta}}_{\mathbf{10}}$	-7.60 e-02	-1.14 e-02	-5.61 e-02	-2.05 e-01	-1.13 e-02
$\hat{\boldsymbol{\beta}}_{\mathbf{11}}$	-4.28 e-02	2.37 e-01	4.88 e-02	-9.57 e-02	7.54 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{12}}$	1.13 e-05	-5.41 e-06	5.76 e-05	-2.38 e-05	4.62 e-05
$\hat{\boldsymbol{\beta}}_{\mathbf{13}}$	-2.38 e-05	-6.21 e-05	1.95 e-04	1.62 e-04	4.36 e-05
$\hat{\boldsymbol{\beta}}_{\mathbf{14}}$	5.37 e-05	-3.68 e-05	-6.41 e-06	1.18 e-04	-1.82 e-04
AIC	6696	4590	13696	3798	58298
Errore	2.29	2.79	2.14	1.45	3.17
Errore (%)	38.42%	41.15%	32.34%	31.72%	32.99%

Tabella 3.6: Risultati ottenuti per gli ospedali Bollate, Macedonio Melloni, Abbiategrasso, Cuggiono, Gaetano Pini

Ospedale	**153**	**26**	**119**	**53**	**28**
$\boldsymbol{\lambda}$	0.5	0.57	0.66	0.21	0.58
$\hat{\boldsymbol{\beta}}_{\mathbf{0}}$	1.65 e+00	1.87 e+00	2.76 e+00	1.54 e+00	1.92 e+00
$\hat{\boldsymbol{\beta}}_{\mathbf{1}}$	-2.38 e-01	-8.96 e-02	-2.11 e-01	-4.24 e-02	-2.48 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{2}}$	-3.29 e-01	4.30 e-01	-3.77 e-01	-8.80 e-02	-3.90 e-02
$\hat{\boldsymbol{\beta}}_{\mathbf{3}}$	-3.89 e-02	2.18 e-01	5.94 e+00	3.32 e-02	6.43 e-02
$\hat{\boldsymbol{\beta}}_{\mathbf{4}}$	-1.66 e-01	-1.50 e-01	-7.00 e-01	-5.55 e-02	-3.41 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{5}}$	-2.01 e-01	-1.30 e-02	-5.02 e-01	-4.11 e-02	-2.50 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{6}}$	3.44 e-01	-2.27 e-01	4.23 e-01	-	3.24 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{7}}$	3.41 e-01	-1.51 e-01	3.18 e-02	-	7.70 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{8}}$	2.13 e-01	-3.16 e-01	3.27 e-01	-	2.52 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{9}}$	4.56 e-04	4.49 e-04	6.27 e-04	2.57 e-05	5.00 e-04
$\hat{\boldsymbol{\beta}}_{\mathbf{10}}$	1.85 e-01	-1.78 e-01	-8.81 e-02	-3.82 e-02	-2.17 e-02
$\hat{\boldsymbol{\beta}}_{\mathbf{11}}$	1.92 e-01	3.26 e-02	2.90 e-01	5.01 e-02	1.98 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{12}}$	-2.11 e-04	5.93 e-05	-1.06 e-04	-	-1.31 e-04
$\hat{\boldsymbol{\beta}}_{\mathbf{13}}$	-1.89 e-04	5.37 e-05	-7.34 e-05	-	-1.39 e-04
$\hat{\boldsymbol{\beta}}_{\mathbf{14}}$	-1.15 e-04	7.53 e-05	-1.31 e-04	-	-4.03 e-05
AIC	3744	6285	11326	-1577	6966
Errore	2.48	2.93	3.46	3.22	2.95
Errore (%)	42.29%	33.97%	35.95%	24.54%	36.88%

Tabella 3.7: Risultati ottenuti per gli ospedali San Luca, Bignami, De Marchi, Treviglio, Buzzi

Ospedale	**17**	**57**	**47**	**41**	**144**
$\boldsymbol{\lambda}$	0.61	0.52	0.25	0.28	0.33
$\hat{\boldsymbol{\beta}}_{\mathbf{0}}$	3.86 e+00	1.77 e+00	1.70 e+00	1.93 e+00	1.48 e+00
$\hat{\boldsymbol{\beta}}_{\mathbf{1}}$	-7.77 e-01	-6.02 e-02	-1.16 e-02	-1.48 e-01	-1.61 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{2}}$	-9.54 e-01	-1.40 e-01	-7.02 e-03	-2.30 e-01	-2.19 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{3}}$	2.47 e-01	5.41 e-02	2.69 e-02	4.89 e-02	3.15 e-02
$\hat{\boldsymbol{\beta}}_{\mathbf{4}}$	4.93 e-02	-4.42 e-02	-2.39 e-02	-8.35 e-02	-3.85 e-02
$\hat{\boldsymbol{\beta}}_{\mathbf{5}}$	6.01 e-03	8.79 e-03	-9.56 e-03	-7.65 e-02	-3.30 e-02
$\hat{\boldsymbol{\beta}}_{\mathbf{6}}$	-	-7.96 e-02	-	-	2.11 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{7}}$	-	7.04 e-02	-	-	2.64 e-01
$\hat{\boldsymbol{\beta}}_{\mathbf{8}}$	-	5.45 e-02	-	-	8.09 e-02
$\hat{\boldsymbol{\beta}}_{\mathbf{9}}$	2.26 e-04	7.01 e-05	2.25 e-05	4.80 e-05	2.37 e-04
$\hat{\boldsymbol{\beta}}_{\mathbf{10}}$	-1.45 e-01	-6.18 e-02	-1.46 e-02	-1.95 e-02	-3.95 e-02
$\hat{\boldsymbol{\beta}}_{\mathbf{11}}$	-2.23 e-01	-3.62 e-03	-1.38 e-02	9.99 e-02	-2.53 e-02
$\hat{\boldsymbol{\beta}}_{\mathbf{12}}$	-	-8.38 e-06	-	-	-9.91 e-05
$\hat{\boldsymbol{\beta}}_{\mathbf{13}}$	-	-1.99 e-05	-	-	-9.96 e-05
$\hat{\boldsymbol{\beta}}_{\mathbf{14}}$	-	-1.4 e-05	-	-	-4.56 e-05
AIC	1748	-279	-1019	-51	1119
Errore	4.24	2.94	3.97	5.29	2.73
Errore (%)	27.27%	21.85%	19.07%	29.26%	42.12%

Tabella 3.8: Risultati ottenuti per gli ospedali Legnano, Vimercate, Pavia, Monza, Sant'Ambrogio

3.3 Metodo di utilizzo dell'applicazione

L'applicazione proposta consiste in una funzione `R` a cui è stato assegnato il nome `get.prediction118`. In questa sezione sarà riportata inizialmente una spiegazione dettagliata riguardante i parametri che tale funzione richiede in ingresso, successivamente sarà mostrato come interpretare l'output ottenuto. Il codice utilizzato per l'implementazione è invece riportato nella sezione 4.4.

3.3.1 Parametri in Ingresso

`codice` : codice rappresentativo della gravità delle condizioni del paziente trasportato.

Valore di default : `NULL`

Valori ammessi : `verde`, `giallo`, `rosso`

Tipologia di variabile : `character`

`indirizzo` : indirizzo del luogo del soccorso.

Valore di dafault : `NULL`

Valori ammessi : Qualsiasi stringa di caratteri rappresentante un indirizzo. Per il riconoscimento dell'indirizzo e il calcolo della distanza da ogni ospedale l'applicazione fa uso del motore di ricerca `GoogleMaps`. In particolare:

- L'immissione di parole quali `via` o `piazza` è facoltativo.
- Se non viene specificato il numero civico o ne viene specificato uno non presente nella via selezionata (per esempio perchè troppo alto) l'applicazione prende come riferimento il numero civico centrale della via stessa.
- Se vi è ambiguità nell'indirizzo digitato (per esempio perchè non viene specificato il Comune di appartenenza della via) l'applicazione restituisce il messaggio
  `Errore: specificare meglio l'indirizzo`

- Se l'indirizzo digitato è errato viene restituito il messaggio `Errore: indirizzo non riconosciuto`

Tipologia di variabile : `character`

`anno` : anno nel quale si vuole effettuare la previsione.

Valore di default : `thisyear`

Valori ammessi : qualsiasi anno successivo al 2008

Tipologia di variabile : `numeric`

`mese` : mese dell'anno nel quale si vuole effettuare la previsione.

Valore di default : `thismonth`

Valori ammessi : {1,2,...,12}

Tipologia di variabile : `numeric`

`giorno` : giorno del mese nel quale si vuole effettuare la previsione.

Valore di default : `today`

Valori ammessi : {1,2,...,31}. Nel caso il valore scelto non sia coerente con il mese selezionato l'applicazione restituisce il messaggio: `Errore: data non corretta`

Tipologia di variabile : `numeric`

`sort` : tipologia di ordinamento che si vuole utilizzare per visualizzare l'output.

Valore di default : `tempo`: output ordinato in modo crescente secondo il tempo previsto di arrivo in ogni ospedale

Valori ammessi :

- `distanza`: output ordinato in modo crescente secondo la distanza del luogo del soccorso da ogni ospedale

- `ospedali`: output ordinato in ordine alfabetico secondo il nome dell'ospedale

Tipologia di variabile : `character`

`graph` : parametro che permette di scegliere se visualizzare o meno i grafici in output

Valore di default : `OFF`

Valori ammessi : `ON`

Tipologia di variabile : `character`

N.B: I parametri in ingresso di tipo `character` devono essere scritti tra virgolette.

3.3.2 Output

L'output della funzione consiste in data frame all'interno del quale ogni riga corrisponde a una struttura ospedaliera. Per ciascuna di queste strutture sono memorizzate le seguenti informazioni:

`codice` : codice identificativo dell'ospedale (si veda Tabella 1.1)

`fit` : tempo di arrivo previsto

`lwr` : estremo inferiore dell'intervallo di previsione

`upr` : estremo superiore dell'intervallo di previsione

`distanza` : distanza dell'ospedale dal luogo del soccorso

`ospedale` : nome dell'ospedale

Inoltre, nel caso in cui al parametro `graph` sia stato assegnato il valore `ON`, la funzione produce i grafici riportati nelle Figure 3.6, 3.7 e 3.8. Tali grafici sono relativi al comando

```
get.prediction118("rosso","piazza duomo milano",graph="ON")
```

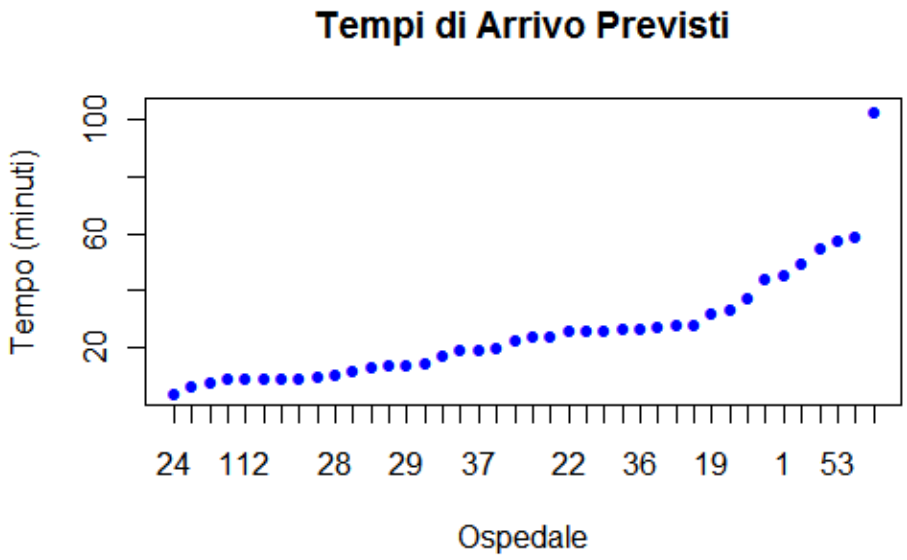


Figura 3.6: Tempi di arrivo previsti in ogni ospedale riportati in ordine crescente

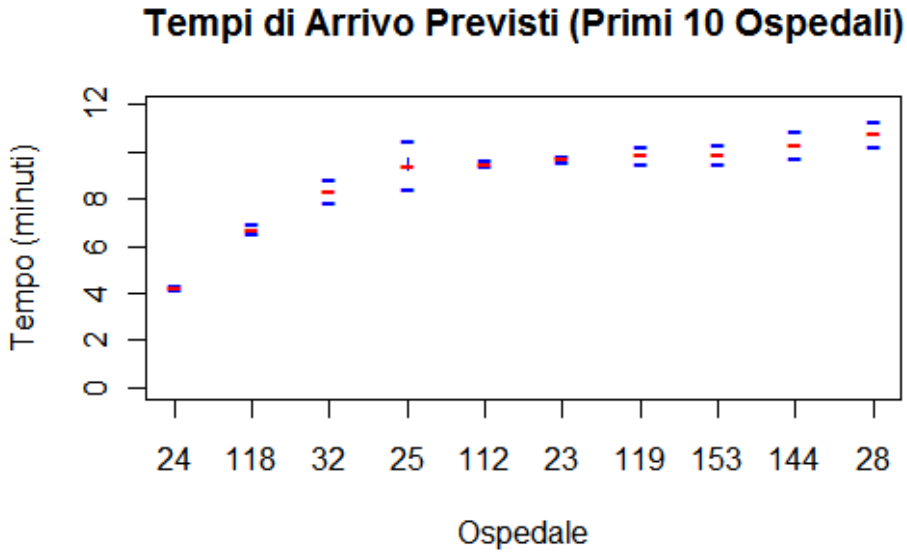


Figura 3.7: Intervalli di previsione per i 10 ospedali con il più basso tempo di arrivo stimato.

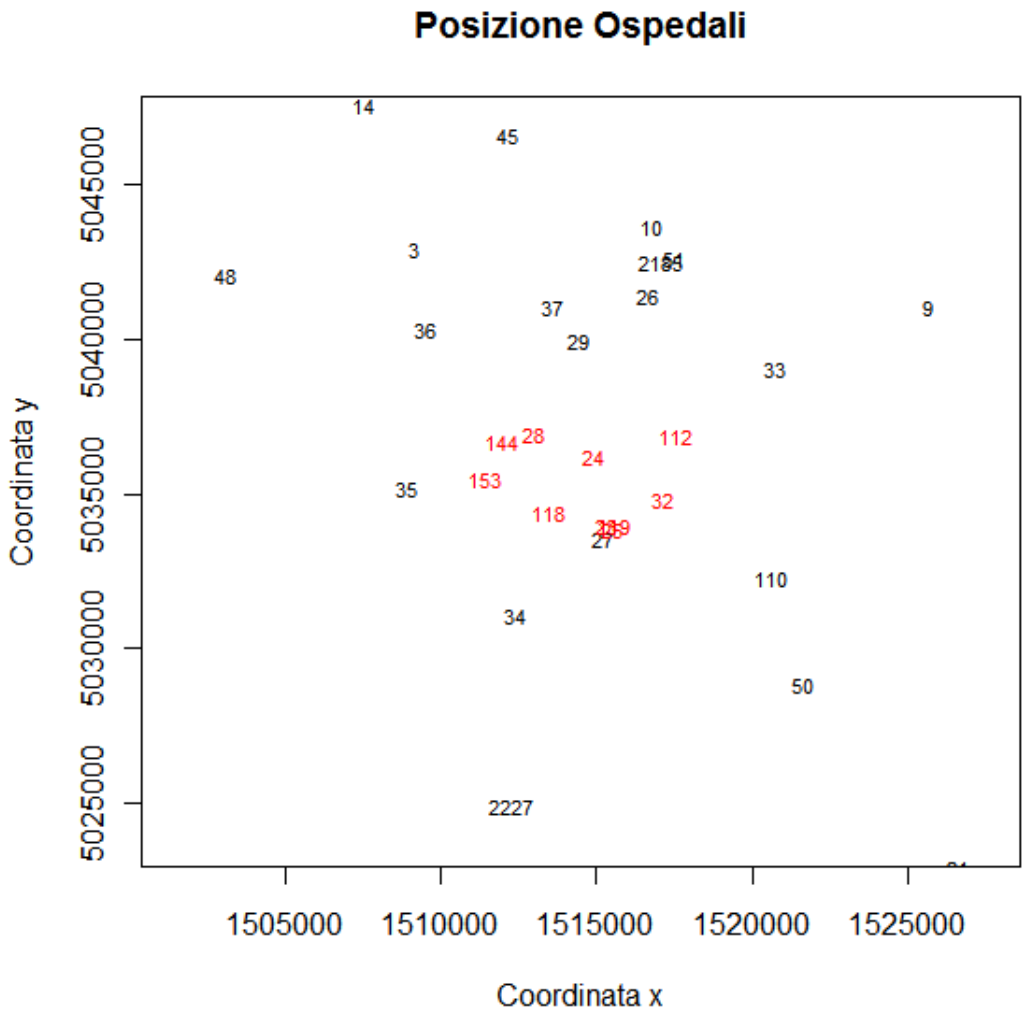


Figura 3.8: Posizione geografica degli ospedali. In rosso sono riportati i 10 ospedali con il più basso tempo di arrivo stimato.

Capitolo 4

Codice R

In questo capitolo vengono riportati e descritti gli script `R` utilizzati per la realizzazione dell'elaborato. In particolare in sezione 4.4 è riportato il codice utilizzato per l'implementazione dell'applicazione proposta.

4.1 Definizione delle Covariate

In questa sezione viene riportato il codice relativo alla creazione delle covariate costruite ad hoc per l'analisi e non presenti in forma esplicita all'interno del database a disposizione. Il dataset al quale nello script è assegnato il nome "`data`" viene interpretato, secondo le convenzioni utilizzate da `R`, come un *data frame* costituito da 723 214 righe (unità statistiche) e 12 colonne (covariate). Il nome delle covariate presenti, ordinate in modo crescente secondo la numerazione delle colonne, è riportato di seguito:

```
> attributes(data)$names
 [1] "id_missione"
 [2] "codice"
 [3] "data_partenza"
 [4] "ora_partenza"
 [5] "data_arrivo"
 [6] "ora arrivo"
 [7] "codice_ospedale"
 [8] "x_partenza"
 [9] "y_partenza"
 [10] "long_partenza"
```

```
 [11] "lat_partenza"
 [12] "comune_partenza"
```

Nello script presentato si fa inoltre utilizzo del data frame `ospedali` in cui sono memorizzate per ogni riga (rappresentante una data struttura ospedaliera) le seguenti informazioni:

```
> attributes(ospedali)$names
 [1] "codice_ospedale"
 [2] "x_ospedale"
 [3] "y_ospedale"
 [4] "n_arrivi"
 [5] "long_ospedale"
 [6] "lat_ospedale"
 [7] "comune_ospedale"
```

Lo script seguente permette inoltre di aggiungere al file `data` le covariate

```
> attributes(data)$names[c(13:18)]
 [13] "giorno"
 [14] "mese"
 [15] "anno"
 [16] "ora_partenza_min"
 [17] "ora_arrivo_min"
 [18] "tempo"
```

dove le variabili n° `[14], [15] e [16]` sono espresse in minuti.

```
n=dim(data)[1]

giorno=rep(0,n)
mese=rep(0,n)
anno=rep(0,n)
ora_partenza_min=rep(0,n)
ora_arrivo_min=rep(0,n)
tempo=rep(0,n)

for(i in 1:n)
{
ora_part_char=as.character(data[i,5])
ora_arr_char=as.character(data[i,7])
```

```
ore_p=as.numeric(substr(ora_part_char,start=1,stop=2))
min_p=as.numeric(substr(ora_part_char,start=4,stop=5))
sec_p=as.numeric(substr(ora_part_char,start=7,stop=8))
ora_partenza_min[i]=ore_p*60+min_p+(sec_p/60)

ore_a=as.numeric(substr(ora_arr_char,start=1,stop=2))
min_a=as.numeric(substr(ora_arr_char,start=4,stop=5))
sec_a=as.numeric(substr(ora_arr_char,start=7,stop=8))
ora_arrivo_min[i]=ore_a*60+min_a+(sec_a/60)

tempo[i]=ora_arrivo[i]-ora_partenza[i]
if(tempo[i]<0){ tempo[i]=tempo[i]+24*60 }

gma_char_part=as.character(data[i,4])
gma_char_arr=as.character(data[i,6])

giorno[i]=as.numeric(substr(gma_char_arr,start=1,stop=2))
mese[i]=as.numeric(substr(gma_char_arr,start=4,stop=5))
anno[i]=as.numeric(substr(gma_char_arr,start=7,stop=8))
}

data=cbind(data,giorno,mese,anno,ora_partenza_min,ora_arrivo_min,tempo)
```

Di seguito si riporta lo script tramite il quale, utilizzando la classe `POSIXlt` per la gestione dei calendari, viene creata la covariata `tipo_giorno` assegnando valore 1 ai giorni festivi e 0 a quelli feriali. Si riporta inoltre il codice per l'implementazione delle variabili `fascia_oraria` e `zona`.

```
# --- Tipo di Giorno

# Feriale <---> 0
# Festivo <---> 1

tipo_giorno=rep(0,n)

for(i in seq(1:n))
{
date_char=paste(as.character(data[i,15]),as.character(data[i,14]),
    as.character(data[i,13]),sep="/")
```

```
if(as.POSIXlt(date_char)$wday==0)
{ tipo_giorno[i]=1 }
}

tipo_giorno[which( (data[,2]==1) & (data[,3]==1) )]=1
tipo_giorno[which( (data[,2]==6) & (data[,3]==1) )]=1
tipo_giorno[which( (data[,2]==25) & (data[,3]==4) )]=1
tipo_giorno[which( (data[,2]==1) & (data[,3]==5) )]=1
tipo_giorno[which( (data[,2]==2) & (data[,3]==6) )]=1
tipo_giorno[which( (data[,2]==15) & (data[,3]==8) )]=1
tipo_giorno[which( (data[,2]==1) & (data[,3]==11) )]=1
tipo_giorno[which( (data[,2]==25) & (data[,3]==12) )]=1
tipo_giorno[which( (data[,2]==26) & (data[,3]==12) )]=1
tipo_giorno[which( (data[,2]==7) & (data[,3]==12) )]=1
tipo_giorno[which( (data[,2]==8) & (data[,3]==12) )]=1


# --- Fascia Oraria

# Ora di Punta <---> 0
# Giorno <---> 1
# Notte <---> 2

fascia_oraria=rep(0,n)

fascia_oraria[which(data[,16]>540 & data[,16]<1020)]=1
fascia_oraria[which(data[,16]<420 | data[,16]>1200)]=2


# --- Zona

# Nord-Est <---> 0
# Nord-Ovest <---> 1
# Sud-Ovest <---> 2
# Sud-Est <---> 3

zona=rep(0,n)

zona[which(data[,8]<ospedali[,2]&data[,9]>=ospedali[,3])]=1
```

```
zona[which(data[,8]<ospedali[,2]&data[,9]<ospedali[,3])]=2
zona[which(data[,8]<ospedali[,2]&data[,9]>=ospedali[,3])]=3

cbind(data,tipo_giorno,fascia,zona)
```

Di seguito viene riportato il codice `R` utilizzato per l'implementazione della funzione `get_distance(x)`. Essa prende come unico parametro di ingresso un'unità statistica del file `data` e restituisce in output la distanza (in metri) che è necessario percorrere sul tragitto cittadino che collega il punto di partenza dell'ambulanza all'ospedale di destinazione. Tale distanza viene calcolata tramite il servizio offerto dal sito www.tuttocitta.com nel seguente modo:

Step 1 : creazione di una stringa di testo rappresentante l'url a cui è necessario connettersi per calcolare il percorso desiderato;

Step 2 : lettura, tramite il comando `ReadLines()`, del codice html relativo alla pagina internet associata all'url creato nel punto 1;

Step 3 : ricerca e memorizzazione dell'informazione cercata (lunghezza del percorso).

```
# --- Step 1

get_distance=function(x)
{
comune_partenza=as.character(x[,12])
comune_arrivo=as.character(ospedali[which(ospedali[,1]==x[,7]),6])

distance=-1

i_partenza=which(comune_partenza==str_char[,1])
i_arrivo=which(comune_arrivo==str_char[,1])

if(length(i_partenza)==0){distance="errore"}

codice_partenza=as.character(str_char[i_partenza,2])
```

```
provincia_partenza=as.character(str_char[i_partenza,3])
str1_partenza=as.character(str_char[i_partenza,4])
str2_partenza=as.character(str_char[i_partenza,5])
long_partenza=paste(substr(as.character(x[,10]),1,1),
    substr(as.character(x[,10]),3,nchar(as.character(x[,10]))),sep=".")
lat_partenza=paste(substr(as.character(x[,11]),1,2),
    substr(as.character(x[,11]),4,nchar(as.character(x[,11]))),sep=".")

codice_arrivo=as.character(str_char[i_arrivo,2])
provincia_arrivo=as.character(str_char[i_arrivo,3])
str1_arrivo=as.character(str_char[i_arrivo,4])
str2_arrivo=as.character(str_char[i_arrivo,5])
long_arrivo=paste(substr(as.character(ospedali[which(ospedali[,1]==x[,7])
    ,5]),1,1),substr(as.character(ospedali[which(ospedali[,1]==x[,7]),5])
    ,3,nchar(as.character(ospedali[which(ospedali[,1]==x[,7]),5]))),sep=".")
lat_arrivo=paste(substr(as.character(ospedali[which(ospedali[,1]==x[,7])
    ,6]),1,2),substr(as.character(ospedali[which(ospedali[,1]==x[,7]),6])
    ,4,nchar(as.character(ospedali[which(ospedali[,1]==x[,7]),6]))),sep=".")

url=paste("http://www.tuttocitta.com/tcol/percorsi/",str1_partenza,"-",
    str1_arrivo,"?cb=0&op=rs&cx1=",long_partenza,"&cy1=",lat_partenza,
    "&dv1=",str2_partenza,provincia_partenza,",Italia&ind1=&cx2=",
    long_arrivo,"&cy2=",lat_arrivo,"&dv2=",str2_arrivo,provincia_arrivo,",
    Italia&ind2=&tp=cbd&ccd1=",codice_partenza,"&ccd2=",codice_arrivo,
    "&lpr=",substr(provincia_partenza,2,3),"&cre=7&lcn=",str2_partenza,
    "&poi=0000&mtp=1&me=0&ae=0&te=",sep="")
```

Prima di proseguire si deve sottilineare che per la generazione della stringa di caratteri che rappresenta l'url desiderato è necessario avere a disposizione il data frame `str_char`. Tale data frame è stato appositamente creato e, per ogni comune trattato, deve contenere le seguenti informazioni:

`comune` : stringa che deve coincidere con la colonna `[12]` relativa al dato passato in input (ad esempio "`CERNUSCO_SUL_NAVIGLIO`")

`codice` : codice numerico (ma interpretato come carattere) identificativo del comune (ad esempio "`29990`")

`provincia` : provincia alla quale appartiene il comune (ad esempio "`(MI)`")

`str1` : stringa rappresentante il nome del comune. All'interno di essa gli spazi devono essere sostituiti da "`%20`" (ad esempio "`cernusco%20sul%20naviglio`")

`str2` : stringa rappresentante il nome del comune. All'interno di essa non devono esservi spazi. (ad esempio "`cernuscosulnaviglio`")

Per quanto riguarda il secondo passo del procedimento, il comando `ReadLines(url,n)` importa, in formato di testo, il codice html della pagina internet identificata dal parametro `url`. Il parametro `n` indica quante righe di tale codice si vogliono importare; ponendo `n=-1` si sceglie di importare l'intero codice.

```
# --- Step 2

text=readLines(url,n=-1)
```

Il terzo passo del procedimento permette di cercare all'interno del file di testo ottenuto al passo 2 l'informazione richiesta, ovvero la lunghezza del percorso, e di memorizzarla come valore numerico.

```
# --- Step 3

if( text[135]=="<span class=\"valore\"/>" ){distance="errore"}

if( (substr(text[135],nchar(text[135])-11,nchar(text[135])-11)==",")
    &(distance!="errore") )
{distance=paste(substr(text[135],22,nchar(text[135])-12),substr(text[135],
    nchar(text[135])-10,nchar(text[135])-10),sep=".")}

if( (substr(text[135],nchar(text[135])-11,nchar(text[135])-11)!=",")&
    (distance!="errore") )
{distance=substr(text[135],22,nchar(text[135])-10)}

if( (substr(text[135],nchar(text[135])-8,nchar(text[135])-8)=="k")&
    (distance!="errore") )
{ distance=as.character(as.numeric(distance)*1000) }

distance
}
```

```
distanza=rep(0,n)
for(i in seq(1:n))
{ distanza[i]=get_distance(data[i,]) }

data=cbind(data,tipogiorno,fascia,zona,distanza)
```

4.2 Trasformazioni dei Dati

In questa sezione viene riportato il codice R utilizzato per implementare, secondo quanto spiegato nella sezione 2.3, una trasformata di tipo potenza per la variabile risposta in modo tale che essa si adatti a una distribuzione Gamma.

```
trasformata_gamma=function(x)
{
n=length(x)

t=NULL
z=NULL

for(lambda in seq(0.05,5,0.01))
{

y=x^lambda

alpha=((mean(y))^2)/var(y)
beta=(mean(y))/var(y)

L=n*log((beta^alpha)/(gamma(alpha)))+(alpha-1)*sum(log(y))-beta*sum(y)+
    n*log(abs(lambda))+(lambda-1)*sum(log(x))

t=c(t,L)
z=c(z,lambda)
}

lambda_max=z[which(t==max(t))]
lambda_max
}
```

4.3 Stima dei Parametri del Modello

In questa sezione viene riportato il codice R utilizzato per la stima dei parametri del modello. Come è stato fatto per il capitolo 3.1, tale codice si riferisce, atitolo di esempio, all'ospedale Fatebenfratelli di Milano. Il procedimento è stato poi ripetuto per ogni struttura ospedaliera oggetto dell'analisi.

```
osp=24
data=dati[which(data[,7]==osp),]

n=dim(data)[1]

# --- Variabile risposta

tempo=data[,12]

# --- Creazione di variabili dummy per le covariate di tipo quantitativo

giallo=rep(0,n)
giallo[which(data[,13]==2)]=1
rosso=rep(0,n)
rosso[which(data[,13]==3)]=1
ora_di_punta=rep(0,n)
ora_di_punta[which(data[,15]==0)]=1
notte=rep(0,n)
notte[which(data[,15]==2)]=1
festivo=rep(0,n)
festivo[which(data[,14]==3)]=1
q_n_o=rep(0,n)
q_n_o[which(data[,5]<data[,8]&data[,6]>=data[,9])]=1
q_s_o=rep(0,n)
q_s_o[which(data[,5]<data[,8]&data[,6]<data[,9])]=1
q_s_e=rep(0,n)
q_s_e[which(data[,5]>=data[,8]&data[,6]<data[,9])]=1

# --- Distanza

distanza=data[,20]
```

```
# --- Trasformazione dei dati

lambda=trasformata_gamma(tempo)
tempo_lambda=tempo^lambda

# --- Stima dei parametri

reg=glm( tempo_lambda ~ giallo + rosso + ora_di_punta + notte + festivo +
    q_n_o + q_s_o + q_s_e + distanza + ora_di_punta:festivo +
    notte:festivo + q_n_o:distanza + q_s_o:distanza + q_s_e:distanza,
    family=Gamma(link="identity"),maxit=400 )

fatebenfratelli24=reg
```

Viene di seguito riportato lo script utilizzato per stimare l'errore medio di previsione associato al modello. Tale stima viene effettuata creando una partizione casuale del dataset in 100 gruppi. L'errore medio di previsione viene successivamente calcolato via cross-validazione confrontando i risultati ottenuti con il modello costruito utilizzando il 99% dei dati a disposizione con i valori della variabile risposta assunti dai dati appartenenti al restante 1% del totale.

```
data=cbind(data,giallo,rosso,ora_di_punta,notte,festivo,q_n_o,q_s_o,
    q_s_e,tempo_lambda)

x=runif(n,0,1)
indici_cv=sort(x,index.return=TRUE)$ix

predizione_cv=rep(0,n)

formula = tempo_lambda_cv ~ giallo_cv + rosso_cv + ora_di_punta_cv +
    notte_cv + festivo_cv + q_n_o_cv + q_s_o_cv + q_s_e_cv + distanza_cv +
    ora_di_punta_cv:festivo_cv + notte_cv:festivo_cv +
    distanza_cv:q_n_o_cv + distanza_cv:q_s_o_cv + distanza_cv:q_s_e_cv

dim_test=floor(n/100)

for(i in 0:(floor(n/dim_test-1)))
{
```

```
if(i<(floor(n/dim_test-1)))
{
data_cv=data[which((indici_cv<=(dim_test*i))|(indici_cv>(dim_test*(i+1)))),]
data_test=data[which((indici_cv>(dim_test*i))&(indici_cv<=dim_test*(i+1))),]
}
if(i==(floor(n/dim_test-1)))
{
data_cv=data[which( (indici_cv<=(dim_test*i)) ),]
data_test=data[which( (indici_cv>(dim_test*i)) ),]
}

tempo_lambda_cv=data_cv[,29]
giallo_cv=data_cv[,21]
rosso_cv=data_cv[,22]
ora_di_punta_cv=data_cv[,23]
notte_cv=data_cv[,24]
festivo_cv=data_cv[,25]
q_n_o_cv=data_cv[,26]
q_s_o_cv=data_cv[,27]
q_s_e_cv=data_cv[,28]
distanza_cv=data_cv[,20]

new=data.frame(giallo_cv=data_test[,21],rosso_cv=data_test[,22],
    ora_di_punta_cv=data_test[,23],notte_cv=data_test[,24],
    festivo_cv=data_test[,25],q_n_o_cv=data_test[,26],q_s_o_cv=
    data_test[,27],q_s_e_cv=data_test[,28],distanza_cv=data_test[,20])

reg=glm(formula,family=Gamma(link="identity"),maxit=400)

prediction=predict.glm(reg,interval="prediction",newdata=new,level=0.95,
    type="response",se.fit=TRUE)

if(i<(floor(n/dim_test-1)))
{
predizione_cv[which((indici_cv>(dim_test*i))&(indici_cv<=dim_test*(i+1)))]
    =prediction$fit
}
if(i==(floor(n/dim_test-1)))
{ predizione_cv[which((indici_cv>(dim_test*i)))]=prediction$fit } }
```

```
errore=abs(tempo-predizione_cv^(1/lambda))
errore_medio=mean(errore)
errore_medio_percentuale=mean(errore/tempo)
```

4.4 Funzione get.prediction118

Viene di seguito riportato il codice dell'applicazione proposta. Per le specifiche tecniche riguardanti il significato dei parametri in ingresso e l'interpretazione dell'output si faccia riferimento alla sezione 3.3.

```
get.prediction118=function(codice,indirizzo,giorno="today",mese="thismonth",
    anno="thisyear",ora="now",sort="tempo",graph="OFF")
{
if(is.character(giorno)==TRUE&giorno!="today")
{
print("Errore: il parametro 'giorno' può assumere solo valori numerici")
return()
}

if(is.character(mese)==TRUE&mese!="thismonth")
{
print("Errore: il parametro 'mese' può assumere solo valori numerici")
return()
}

if(sort!="tempo"&sort!="distanza"&sort!="ospedale")
{
print("                                                                    ")
print("  Errore: il parametro 'sort' può assumere solo i valori 'tempo',
    'distanza','ospedale'  ")
print("                                                                    ")
return()
}

if(graph!="ON"&graph!="OFF")
{
print("                                                                    ")
```

```
print("Errore: il parametro 'graph' può assumere solo i valori 'ON','OFF'")
print("                                                                    ")
return()
}

if(codice!="giallo"&codice!="rosso"&codice!="verde")
{
print("                                                                    ")
print("  Errore: il parametro 'codice' può assumere solo i valori 'verde',
    'giallo','rosso'  ")
print("                                                                    ")
return()
}

if(ora!="now"&(ora<0|ora>23))
{
print("  Errore: il parametro 'ora' deve essere compreso tra 0 e 23  ")
return()
}

if(giorno!="today"&(giorno<1|giorno>31))
{
print("  Errore: il parametro 'giorno' deve essere compreso tra 1 e 31  ")
return()
}

if(mese!="thismonth"&(mese<1|mese>12))
{
print("  Errore: il parametro 'mese' deve essere compreso tra 1 e 12  ")
return()
}

giallo=0
rosso=0
if(codice=="giallo"){giallo=1}
if(codice=="rosso"){rosso=1}

orapartenza=0
oradipunta=0
```

```
notte=0
if(ora=="now")
{ orapartenza=as.numeric(substr(as.character(Sys.time()),start=12,stop=13))
    *60+as.numeric(substr(as.character(Sys.time()),start=15,stop=16)) }
if(ora!="now"&as.numeric(ora<10))
{
orapartenza=as.numeric(substr(as.character(ora),start=1,stop=1))*60+
    as.numeric(substr(as.character(ora),start=3,stop=4))
}
if(ora!="now"&as.numeric(ora>=10))
{
orapartenza=as.numeric(substr(as.character(ora),start=1,stop=2))*60+
    as.numeric(substr(as.character(ora),start=4,stop=5))
}
if((orapartenza>=420&orapartenza<=540)|(orapartenza>=1020&orapartenza<=1200))
{ oradipunta=1 }
if((orapartenza>=0&orapartenza<420)|(orapartenza>1200&orapartenza<1440))
{ notte=1 }

festivo=0
if( giorno!="today" & mese!="thismonth" & anno!="thisyear" )
{ data=paste(as.character(anno),as.character(mese),
    as.character(giorno),sep="/") }
if( giorno=="today" & mese=="thismonth" & anno=="thisyear" )
{
giorno=as.numeric(substr(as.character(Sys.time()),start=9,stop=10))
mese=as.numeric(substr(as.character(Sys.time()),start=6,stop=7))
anno=as.numeric(substr(as.character(Sys.time()),start=1,stop=4))
data=as.character(Sys.Date())
}
if(as.POSIXlt(data)$wday==0)
{ festivo=1 }
if( (giorno==1&mese==1)|(giorno==6&mese==1)|(giorno==25&mese==4)|
    (giorno==1&mese==5)|(giorno==2&mese==6)|(giorno==15&mese==8)|
    (giorno==1&mese==11)|(giorno==7&mese==12)|(giorno==8&mese==12)|
    (giorno==25&mese==12)|(giorno==26&mese==12) )
{ festivo=1 }

lat=0
```

```
long=0
for(i in seq(1:nchar(indirizzo)))
{
if( substr(indirizzo,i,i)==" " )
{ indirizzo=paste(substr(indirizzo,1,(i-1)),"+",substr(indirizzo,(i+1)
    ,nchar(indirizzo)),sep="") }
}

url=paste("http://maps.googleapis.com/maps/api/geocode/xml?address=",
    indirizzo,"&sensor=false",sep="")
url=readLines(url,n=100)

for(j in seq(1:20))
{
if(substr(url[j],2,21)=="<status>ZERORESULTS")
{
print("Errore: indirizzo non riconosciuto")
return()
}
if(j==20){break()}
}

warning=0
comune=0
for(j in seq(1:20))
{
if((substr(url[j],3,23)=="<type>locality</type>")&(substr(url[j+1],3,24)
    =="<type>political</type>"))
{
warning=1
comune=substr(url[j+4],15,(nchar(url[j+4])-12))
}
if(j==20){break()}
}

for(i in seq(25,100))
{
if( substr(url[i],4,12)=="<location" )
{
```

```
lat=as.numeric(substr(url[i+1],start=10,stop=19))
long=as.numeric(substr(url[i+2],start=10,stop=18))
break()
}}

fit=NULL
upr=NULL
lwr=NULL
distanzavect=NULL

# --- Questa parte viene ripetuta per ogni ospedale

i=which(ospedali[,1]==35)
longosp=ospedali[i,5]
latosp=ospedali[i,6]

qno=0
qso=0
qse=0
if(long<longosp&lat>=latosp){qno=1}
if(long<longosp&lat<latosp){qso=1}
if(long>=longosp&lat<latosp){qse=1}
distanza=ottienidistanza(lat,long,latosp,longosp)
distanzavect=c(distanzavect,distanza)
new=data.frame(giallo=giallo,rosso=rosso,oradipunta=oradipunta,
    notte=notte,festivo=festivo,qno=qno,qso=qso,qse=qse,distanza=distanza)
prediction=predict.glm(sancarlo35linear,interval="prediction",newdata=new,
    level=0.95,type="response",se.fit=TRUE)
fit=c(fit,(prediction$fit-trasf[i,2])^(1/trasf[i,3]))
upr=c(upr,(prediction$fit-trasf[i,2]+prediction$se.fit)^(1/trasf[i,3]))
lwr=c(lwr,(prediction$fit-trasf[i,2]-prediction$se.fit)^(1/trasf[i,3]))

# --- --- --- --- --- --- --- ---

ospedale=c("San Carlo","Cernusco","Policlinico","San Paolo","Magenta",
    "Galeazzi","Rho","Melegnano","Humanitas","Fatebenefratelli",
    "Santa Rita","Garbagnate","Niguarda","San Raffaele","Saronno","Paderno",
    "San Giuseppe","Cinisello","Sacco","San Donato","Melzo","Cardiologico",
    "Alfieri","Sesto","Humanitas","Bollate","Melloni","Abbiategrasso",
```

```
    "Cuggiono","Pini","San Luca","Bignami","De Marchi","Treviglio","Buzzi",
    "Legnano","Vimercate","Pavia","Monza","Sant'Ambrogio")
codosp=ospedali[,1]
distanza=distanzavect

ordinetempi=sort(fit,index.return=TRUE)$ix
ordinedistanze=sort(distanzavect,index.return=TRUE)$ix
ordinenomi=sort(ospedale,index.return=TRUE)$ix

x=as.data.frame(cbind(ospedale,codosp,fit,lwr,upr,distanza))

if(sort=="tempo")
{y=x[ordinetempi,]}
if(sort=="distanza")
{y=x[ordinedistanze,]}
if(sort=="ospedale")
{y=x[ordinenomi,]}
attributes(y)$row.names=seq(1:40)

if(graph=="ON")
{

coordx=1487080+(long-8.8346)*(59416/0.760748)
coordy=5005029+(lat-45.19782)*(47850/0.43078)
windows()
plot(ospedali[,2],ospedali[,3],type='n',main="Posizione Ospedali",
    xlab="Coordinata x",ylab="Coordinata y",xlim=
    c(min(ospedali[ordinetempi[c(1:10)],2])-5000,
    max(ospedali[ordinetempi[c(1:10)],2])+5000),
    ylim=c(min(ospedali[ordinetempi[c(1:10)],3])-5000,
    max(ospedali[ordinetempi[c(1:10)],3])+5000))
text(ospedali[ordinetempi[c(11:40)],2],ospedali[ordinetempi[c(11:40)],3]
    ,ospedali[ordinetempi[c(11:40)],1],cex=0.7)
text(ospedali[ordinetempi[c(1:10)],2],ospedali[ordinetempi[c(1:10)],3],
    ospedali[ordinetempi[c(1:10)],1],col='red',cex=0.7)
text(coordx,coordy,"x",cex=1,col='blue')

windows()
plot(fit[ordinetempi],main="Tempi di Arrivo Previsti",xaxt='n',
```

```
    xlab="Ospedale",ylab="Tempo (minuti)",type='n')
text(fit[ordinetempi],".",cex=1.5,col='blue')
axis(side=1,at=seq(1,40),labels=ospedali[ordinetempi,1])

windows()
plot(fit[ordinetempi[c(1:10)]],main="Tempi di Arrivo Previsti (Primi
    10 Ospedali)",xaxt='n',xlab="Ospedale",ylab="Tempo (minuti)",
    type='n',ylim=c(0,upr[ordinetempi[10]]+0.5))
for(i in seq(1:10))
{
points(c(i,i),c(lwr[ordinetempi[i]],upr[ordinetempi[i]]),type='b',col='blue')
points(c(i,i),c(lwr[ordinetempi[i]],upr[ordinetempi[i]]),col='white')
}
text(lwr[ordinetempi[c(1:10)]],"-",col='blue',cex=1.5)
text(upr[ordinetempi[c(1:10)]],"-",col='blue',cex=1.5)
axis(side=1,at=seq(1,10),labels=ospedali[ordinetempi[c(1:10)],1])
text(fit[ordinetempi[c(1:10)]],"-",col='red',cex=1.5)
}

if(warning==1)
{
print("Attenzione: Il percorso è atato calcolato")
print(paste("a partire dal centro di ",comune,sep=""))
}
print(y)
return(y)
}
```

La funzione `get.prediction118` richiama, per il calcolo della distanza tra il luogo dove avviene il soccorso e ogni struttura ospedaliera, la funzione `get.distance2`, il cui codice è riportato di seguito.

```
ottienidistanza=function(lat,long,latosp,longosp)
{
url=paste("http://maps.googleapis.com/maps/api/distancematrix/xml?origins=
    ",as.character(lat),",",as.character(long),"&","destinations=",
    as.character(latosp),",",as.character(longosp),"&sensor=false",sep="")
url=readLines(url,25)
for(j in seq(1,20))
{
```

```
if(substr(url[j],4,23)=="<status>ZERORESULTS")
{
print(paste("errore: non riesco a calcolare la distanza per l'ospedale",
    as.character(ospedali[which(ospedali[,5]==longosp&ospedali[,6]==
    latosp),1]),sep=" "))
break()
}
if(substr(url[j],4,13)=="<distance>")
{
distanza=as.numeric(substr(url[j+1],12,(nchar(url[j+1])-8)))
break()
}}
distanza
}
```

Bibliografia

[1] Richard A. Johnson, Dean W. Wichern - Applied Multivariate Statistical Analysis - Sixth Edition - Prentice Hall - 2007

[2] Douglas C. Montgomery - Design and Analysis of Experiments - McGraw-Hill - 2005

[3] Ornello Vitali - Statistica Per Le Scienze Applicate - Cacucci Editore - 1993

[4] Sanford Weisberg - Applied Linear Regression - Second Edition - Wiley Series in Probabiliy and Mathematical Statistics - 1985

[5] G. E. P. Box, D. R. Cox - An analysis of Transformations - Journal of the Royal Statistical Society. Series B (Methodological), Vol. 26, No. 2., pp. 211-252 - 1964

[6] http://geomatica.como.polimi.it

[7] R: A Language and Environment for Statistical Computing - R Development Core Team - R Foundation for Statistical Computing, Vienna, Austria - 2008 - http://www.R-project.org

[8] Centrale Operativa del 118 di Milano - Azienda Ospedaliera Niguarda Ca' Granda - Piazza Ospedale Maggiore, 3, Milano - www.118milano.it