UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI STATISTICA

CORSO DI LAUREA DI STATISTICA E TECNOLOGIE INFORMATICHE

TESI DI LAUREA

DURATA DELLA DEGENZA E DELLA SOPRAVVIVENZA DEI PAZIENTI AFFETTI DA SINDROME CORONARICA ACUTA

Relatore: Prof. Bruno Scarpa

Laureando: Brugnaro Luca

ANNO ACCADEMICO 2007/2008

Indice

Premessa

Ogni professionista sanitario (medico, infermiere, …) coscienzioso, durante la propria attività, si pone una serie di quesiti come: sto facendo la cosa giusta? Si potrebbe affrontare il problema in un altro modo? Ci sono modi più efficaci o semplicemente alternativi a questo trattamento? Qual è la causa del problema? Come posso migliorare?
Si trova quindi, in una condizione di dubbio, una sensazione di incertezza ed inadeguatezza, che per certi aspetti è positiva: di stimolo per cercare risposte e far sorgere quei meccanismi necessari per apprendere e per conoscere.
In ambito clinico il tutto viene riassunto ed inglobato in una strategia, una metodologia operativa per trovare le risposte ai bisogni di sapere, che nascono dalla attività clinico assistenziale chiamata Evidence Based Practice (EBP).
L'EBP è necessaria per poter formulare nel modo corretto un quesito per cui si può trovare una risposta attraverso una corretta ricerca bibliografica, ma non è sufficiente per appagare questo senso di inadeguatezza.
Infatti quando si riesce a condurre una corretta ricerca bibliografica ci si trova di fronte ad un'altra serie di difficoltà :

- gli articoli sono quasi per la loro totalità scritti in inglese,
- negli articoli vi è un costante utilizzo della statistica per giustificare le affermazioni,
- negli articoli non sempre sono garantiti standard metodologici di validità interna (ciò dipende principalmente dalla rivista),
- bisogna comprendere la loro rilevanza clinica e la loro applicabilità
- non sempre si riesce a pervenire ad una bibliografia adeguata.

Quindi anche se si dovesse pervenire a ottimali risultati di ricerca bibliografica o no, le problematiche connesse all'interpretazione dei risultati sono notevoli.
Dovere dei clinici credo diventi sempre più quello di acquisire competenze non solo specifiche del loro core professionale ma anche di approfondire competenze metodologiche di ricerca per dare risposte di salute sempre più pertinenti ed adeguate.

1 Introduzione

Scopo di questo elaborato di tesi è quello di analizzare quali sono i fattori che influenzano la durata della degenza e la sopravvivenza dei pazienti affetti da sindrome coronarica acuta (SCA) ricoverati presso l'unità di cure intensive cardiologiche (UCIC) dell'azienda Ospedaliera di Padova.

Prima di analizzare i materiali e i metodi, si ritiene opportuno dare una rapida trattazione:

- delle sindromi coronariche acute (SCA) che hanno portato al ricovero urgente dei pazienti,
- dei database relazionali, strumento impiegato preliminarmente alle analisi compiute per archiviare e gestire in modo semplice, veloce e sicuro una considerevole quantità di dati riconducibili ai pazienti sopra menzionati,
- di alcune considerazioni di statistica descrittiva dell'intera popolazione dei pazienti ricoverati in unità di cure intensive coronariche (UCIC) dell'Azienda Ospedaliera di Padova rilevati da ottobre 2006 ad ottobre 2007 con particolare attenzione ai pazienti affetti da SCA.

1.1 Le sindromi coronariche acute (SCA)

Le malattie cardiovascolari rappresentano attualmente la prima causa di morte nei paesi industrializzati e si prevede che lo diventino anche nei paesi in via di sviluppo entro il 2020(vedi Murray CJ, Lopez AD. (1997), Alternative projections of mortality and disability by cause 1990-2020: Global Burden of Disease Study. Lancet 349: 1498-504). Fra queste, la coronaropatia (CAD) rappresenta la condizione più comune, associata ad elevata mortalità e morbilità. Le presentazioni cliniche della cardiopatia ischemica comprendono l'ischemia silente, l'angina pectoris stabile e instabile, l'infarto miocardico (IM), lo scompenso cardiaco e la morte improvvisa.
In Europa, i pazienti con dolore toracico rappresentano buona parte delle ospedalizzazioni per acuti e, dal punto di vista diagnostico, risulta problematico distinguere i pazienti con sindrome coronarica acuta (SCA) da quelli con dolore toracico di sospetta origine cardiaca, soprattutto in assenza di sintomatologia e segni elettrocardiografici specifici. Nonostante la disponibilità dei moderni approcci terapeutici, l'incidenza di mortalità, IM e riospedalizzazione dei pazienti con SCA permane elevata.
È ormai accertato che le SCA, nelle loro varie forme di presentazione, condividono un substrato fisiopatologico comune. Studi anatomo-patologici, endoscopici e biologici hanno dimostrato che la rottura o l'erosione della placca aterosclerotica, su cui si sovrappongono fenomeni trombotici ed embolizzazione distale di

entità variabile determinanti ipoperfusione, costituisce il meccanismo fisiopatologico di base nella maggior parte delle SCA.
Data la pericolosità della malattia aterotrombotica, sono stati introdotti dei criteri per la stratificazione del rischio al fine di consentire al clinico di scegliere tempestivamente il miglior approccio farmacologico o interventistico.
Il sintomo primario che innesca il processo diagnostico-terapeutico è il dolore toracico, ma la classificazione dei pazienti si basa sull'ECG, tramite il quale si possono identificare due categorie di pazienti:

1. *Pazienti con dolore toracico acuto e persistente sopraslivellamento del tratto ST (>20 min).* Trattasi di SCA associata a sopraslivellamento del tratto ST (SCA-STE) e riflette generalmente un'occlusione coronarica acuta. La maggior parte di questi pazienti va incontro ad IM associato a sopraslivellamento del tratto ST (STEMI). L'obiettivo terapeutico consiste in una ricanalizzazione rapida, completa e sostenuta mediante angioplastica primaria o terapia fibrinolitica (vedi Van de Werf F, ad altri (2003), Management of acute myocardial infarction in patients presenting with ST-segment elevation. Eur Heart J 2003;24: 28-66)
2. *Pazienti con dolore toracico acuto senza persistente sopraslivellamento del tratto ST.* Trattasi del riscontro di persistente o transitorio sottoslivellamento del tratto ST, di inversione, appiattimento o pseudonormalizzazione dell'onda T, oppure di alterazioni elettrocardiografiche aspecifiche. In questi casi, la strategia iniziale è di alleviare l'ischemia e con essa la sintomatologia, di monitorare il paziente attraverso un ECG continuo e misurazioni seriate dei marker di necrosi miocardica.La diagnosi operativa di SCA senza sopraslivellamento del tratto ST (SCA-NSTE), posta alla presentazione sulla base della misurazione della troponina cardiaca, verrà successivamente diversificata in IM senza sopraslivellamento del tratto ST (NSTEMI) o angina instabile (Figura 1). In alcuni casi, si potrà escludere la CAD quale causa della sintomatologia. L'approccio terapeutico dipenderà dalla diagnosi definitiva.

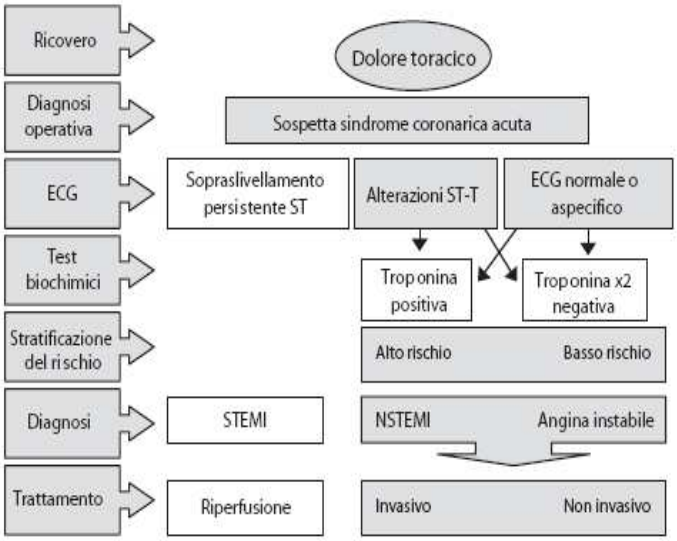


Figura 1. *Lo scenario delle sindromi coronariche acute. NSTEMI = infarto miocardico senza sopraslivellamento del tratto ST; STEMI = infarto miocardico con sopraslivellamento del tratto ST.*

I costi dell'assistenza sanitaria rappresentano una tematica sempre più importante per molti paesi e, per quanto non debbano avere ripercussioni sul processo decisionale, occorre oggigiorno operare con consapevolezza economica. Pertanto, per le opzioni terapeutiche di maggiore rilevanza viene riportato il numero dei pazienti da trattare (NNT) per prevenire un evento. L'NNT risulta l'approccio più semplice per confrontare studi di diverse dimensioni e con differenti endpoint. Ad esempio, un NNT pari a 50 per prevenire un evento fatale deve essere interpretato diversamente da un analogo NNT per risparmiare una riospedalizzazione (Cook RJ, Sackett DL.(1995), The number needed to treat: a clinically useful measure of treatment effect. BMJ; 310:452-4).

1.1.2 Epidemiologia e storia naturale

La diagnosi di SCA-NSTE è più complessa rispetto a quella di STEMI e, pertanto, è più difficile stabilirne la reale prevalenza. Inoltre, recentemente è stata introdotta una nuova definizione di IM che prevede l'utilizzo di biomarcatori di morte cellulare più sensibili e specifici (Alpert JS ad altri (2000), Infarction redefined - a consensus document of The Joint European Society of Cardiology/American College of Cardiology Committee for the redefinition of myocardial infarction. Eur Heart J; 21: 1502-13).
In questo ambito, molteplici indagini e registri hanno definito la prevalenza di SCA-NSTE in rapporto allo STEMI, riportando complessivamente un'incidenza annuale delle SCA-NSTE superiore allo STEMI. Il rapporto tra SCA-NSTE e STEMI è andato modificandosi nel tempo verso una prevalenza delle SCA-NSTE,pur in assenza di chiare motivazioni alla base ditale evoluzione, che potrebbe tuttavia essere legata ai cambiamenti degli ultimi 20 anni nel trattamento delle patologie e all'intensificarsi degli interventi di prevenzione della CAD(Bata IR, RD ed altri

(1984), Trends in the incidence of acute myocardial infarction between 1984 and1993 - The Halifax County MONICA Project. Can J Cardiol; 16: 589-95).
Dai risultati di queste indagini e registri è emersa un'incidenza annuale di circa 3 ricoveri per SCA-NSTE per 1000 abitanti. Allo stato attuale,non si dispone di una stima esatta a livello europeo,per la mancanza di un centro predisposto all'elaborazione centralizzata dei dati di statistica sanitaria. Ciononostante,si riscontra un'ampia variabilità nell'incidenza di tale patologia fra i vari paesi europei, con un'incidenza e una mortalità superiori per l'Europa centrale e orientale.
La prognosi di SCA-NSTE può essere desunta dagli studi che hanno arruolato oltre 100000 pazienti. I dati dimostrano un'incidenza di mortalità a 1 e 6 mesi più elevata nelle popolazioni incluse negli studi rispetto a quelle dei trial clinici randomizzati. La mortalità ospedaliera è maggiore nei pazienti con STEMI rispetto a quelli con SCA-NSTE (7 vs 5%), mentre a 6 mesi è assai simile per entrambe le affezioni (12 vs 13%)(vedi Savonitto S ad altri (1999), Prognostic value of the admission electrocardiogram in acute coronary syndromes. JAMA; 281: 707-13).
Il follow-up a lungo termine dei pazienti sopravvissuti ha evidenziato un'incidenza di mortalità più elevata per le SCA-NSTE rispetto alle SCA-STE, con una differenza a 4 anni di 2 volte superiore(Terkelsen Cj ed altri (2005); Mortality rates in patients with ST-elevation vs non-ST-elevation acute myocardial infarction: observations from an unselected cohort. Eur Heart J; 26: 18-26). Nell'evoluzione a medio-lungo termine, questa differenza potrebbe essere dovuta alle diverse caratteristiche dei pazienti, in ragione del fatto che i pazienti con SCA-NSTE sono più frequentemente anziani e presentano più comorbilità, in particolare diabete e insufficienza renale. Un altro motivo potrebbe essere la maggiore estensione della CAD e della vasculopatia o la presenza di fattori scatenanti quali l'infiammazione (Bahit MC ed altri (2002); Persistence of the prothrombotic state after acute coronary syndromes: implicationsfor treatment. Am Heart J; 143: 205-16).
Le implicazioni terapeutiche sono le seguenti:

- le SCA-NSTE sono più frequenti dello STEMI; a differenza dello STEMI nel quale la maggior parte degli eventi si verifica prima o immediatamente dopo la presentazione, nelle SCA-NSTE questi possono persistere anche nei successivi giorni o settimane;
- la mortalità a 6 mesi per lo STEMI e le SCA-NSTE è equiparabile. Pertanto, le strategie terapeutiche per le SCA-NSTE devono essere rivolte al trattamento tanto della fase acuta quanto a lungo termine.

1.1.3 Fisiopatologia delle SCA

L'aterosclerosi è una malattia fibroproliferativa, immuno-infiammatoria, cronica, multifocale delle arterie di grande e medio

calibro, causata principalmente da un accumulo di lipidi[]. La presenza di CAD comporta due processi distinti: da un lato, un processo costante e irreversibile che conduce, nell'arco di decenni, ad un progressivo restringimento del lume vasale (aterosclerosi), dall'altro un processo dinamico e potenzialmente reversibile che può precipitare improvvisamente in un'occlusione coronarica parziale o totale (trombosi o vasospasmo o entrambi). Pertanto, le lesioni coronariche sintomatiche contengono una miscela variabile di aterosclerosi cronica e trombosi acuta, di natura non chiaramente definibile nel singolo paziente e alla quale spesso ci si riferisce con il termine di aterotrombosi. In generale, la componente aterosclerotica è predominante nelle lesioni responsabili dell'angina stabile cronica, mentre la trombosi coronarica rappresenta la causa primaria della maggior parte delle SCA(Davies MJ. (2000); The pathophysiology of acute coronary syndromes.Heart; 83: 361-6).
Le SCA costituiscono una pericolosa manifestazione dell'aterosclerosi sollecitata dalla trombosi acuta per rottura o erosione di placca, associata o meno a vasocostrizione,che determina una riduzione repentina e critica del flusso sanguigno. Nel processo di rottura della placca, l'infiammazione gioca un ruolo determinante. Solo raramente le SCA sono di origine non aterosclerotica, come nel caso di arterite, eventi traumatici, dissecazione, tromboembolia, anomalie congenite, abuso di cocaina e complicanze del cateterismo cardiaco. Verranno approfonditi alcuni dei principali meccanismi fisiopatologici ai fini di una migliore comprensione delle strategie terapeutiche da adottare.
Evidenze cliniche e sperimentali sempre più numerose identificano nella placca instabile il meccanismo più diffuso alla base delle SCA. Nei pazienti con SCA sono stati documentati molteplici siti di rottura della placca, associata o meno a trombosi intracoronarica, ed elevati livelli dei marker sistemici di infiammazione, trombosi e coagulazione (Libby P. (2002); Inflammation in atherosclerosis. Nature; 420:868-74). In questi pazienti l'ipercolesterolemia, il fumo di sigaretta e aumentati livelli di fibrinogeno possono contribuire ad una condizione di instabilità, favorendo lo sviluppo di complicanze trombotiche.
Il concetto di instabilità diffusa ha rilevanti implicazioni terapeutiche, in quanto al di là di una procedura di rivascolarizzazione, questi pazienti necessitano di terapie sistemiche atte a stabilizzare il profilo di alto rischio che può essere fonte di ripetuti eventi ischemici.

1.1.4 Marcatori biochimici e valutazione del rischio

Recentemente sono stati valutati alcuni marcatori biochimici ai fini di un loro utilizzo nella stratificazione diagnostica e del rischio. Questi riflettono diversi aspetti fisiopatologici delle SCA, come il danno miocardico minimo, l'infiammazione e l'attivazione

piastrinica o neurormonale. Per quanto concerne la prognosi a lungo termine, anche gli indici di disfunzione ventricolare sinistra o renale o quelli per il diabete hanno un ruolo rilevante. *Marcatori di danno miocardico.* La cTnT e la cTnI sono i marker preferenziali di danno miocardico, in quanto si dimostrano più specifici e più sensibili dei tradizionali enzimi cardiaci, come la creatinchinasi (CK) o il suo isoenzima MB (CK-MB). In questo contesto, la mioglobina non è sufficientemente specifica e sensibile da consentire l'identificazione del danno cellulare miocardico e, pertanto, non ne viene raccomandato l'uso per la diagnosi routinaria e per la stratificazione del rischio (Eggers KM ed altri (2004); Diagnostic value of serial measurement of cardiac markers in patients with chest pain: limited value of adding myoglobin to troponin I for exclusion of myocardial infarction; Am Heart J; 148: 574-81).

Si ritiene che un'elevazione dei livelli di troponina rispecchi una necrosi delle cellule miocardiche irreversibile, causata da embolizzazione distale di trombi ricchi di piastrine a partenza da una placca rotta. Di conseguenza, la troponina può essere considerata un marker surrogato della formazione di trombi. Nel contesto di un quadro di ischemia miocardica (dolore toracico, alterazioni del tratto ST), secondo il Documento di Consenso ESC/ACC/AHA (Alpert JS ed altri (2000); Myocardial infarction redefined - a consensus document of…Eur Heart J; 21: 1502-13), elevati livelli di troponina depongono per una diagnosi di IM.

La troponina rappresenta il marker biochimico ottimale per predire l'outcome a breve termine (30 giorni) relativo a IM e mortalità (Antman EM ed altri (1996), Cardiac-specific troponin I levels to predict the risk of mortality in patients with acute coronary syndromes. N Engl J Med; 335: 1342-9). Il valore prognostico della misurazione della troponina è stato anche confermato a lungo termine. L'aumentato rischio che si associa al riscontro di elevati livelli di troponina è indipendente e aggiuntivo rispetto ad altri fattori di rischio, quali le alterazioni elettrocardiografiche a riposo o durante monitoraggio continuo e i marker di attività infiammatoria (Hamm CW, Braunwald E. (2000); A classification of unstable angina revisited. Circulation; 102: 118-22).

In pazienti con IM, un primo aumento della troponina nel sangue periferico si osserva dopo 3-4 ore e può persistere per un periodo fino a 2 settimane a causa della proteolisi dell'apparato contrattile come evidenziato in figura 2.

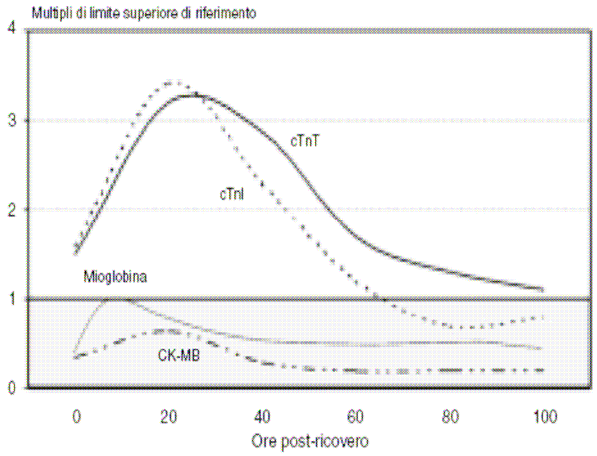


Figura 2. *Esempio di rilascio di marcatori cardiaci in un paziente con sindrome coronarica acuta senza sopraslivellamento del tratto ST (l'a*rea ombreggiata *indica il range di normalità). CK-MB = creatinchinasi-MB; cTnI = troponina cardiaca I; cTnT = troponina cardiaca T.*

1.2 Cenni sui database

In informatica, il termine database, tradotto in italiano con banca dati, base di dati (soprattutto in testi accademici) o anche base dati, indica un archivio di dati, riguardanti uno stesso argomento o più argomenti correlati tra loro, strutturato in modo tale da consentire la gestione dei dati stessi (l'inserimento, la ricerca, la cancellazione ed il loro aggiornamento) da parte di applicazioni software.
Informalmente e impropriamente, la parola "database" viene spesso usata come abbreviazione dell'espressione Database Management System (DBMS), che si riferisce a una vasta categoria di sistemi software che consentono la creazione e la manipolazione efficiente di database.
La base di dati, oltre ai dati veri e propri, deve contenere anche le informazioni sulle loro rappresentazioni e sulle relazioni che li legano. Spesso, ma non necessariamente, una base dati contiene le seguenti informazioni:

- Strutture dati che velocizzano le operazioni frequenti, tipicamente a spese di operazioni meno frequenti.
- Collegamenti con dati esterni, cioè riferimenti a file locali o remoti non facenti parte del database.
- Informazioni di sicurezza, che autorizzano solo alcuni profili utente ad eseguire alcune operazioni su alcuni tipi di dati.
- Programmi che vengono eseguiti, automaticamente o su richiesta di utenti autorizzati, per eseguire elaborazioni sui dati. Un tipico automatismo consiste nell'eseguire un programma ogni volta che viene modificato un dato di un certo tipo.

In un sistema informatico, una base di dati può essere manipolata direttamente dai programmi applicativi, interfacciandosi direttamente con il sistema operativo. Tale strategia era quella adottata universalmente fino agli anni sessanta, ed è tuttora impiegata quando i dati hanno una struttura molto semplice, o quando sono elaborati da un solo programma applicativo.
A partire dalla fine degli anni Sessanta, tuttavia, per gestire basi di dati complesse condivise da più applicazioni si sono utilizzati appositi sistemi software, detti sistemi per la gestione di basi di dati (in inglese "Database Management System" o "DBMS"). Uno dei vantaggi di questi sistemi è la possibilità di non agire direttamente sui dati, ma di vederne una rappresentazione concettuale.
La ricerca nel campo delle basi di dati studia le seguenti problematiche:

- Progettazione di basi di dati.
- Progettazione e implementazione di DBMS.
- Interpretazione (o analisi) di dati contenuti in database.

Le basi di dati spesso fanno uso di tecnologie derivate da altre branche dell'informatica. È usuale utilizzare tecniche derivate dall'intelligenza artificiale, come ad esempio il data mining, per cercare di estrarre relazioni o più in generale informazioni presenti nelle banche dati ma non immediatamente visibili.

Le basi di dati possono avere varie strutture, tipicamente, in ordine cronologico:
Gerarchica (rappresentabile tramite un albero; anni sessanta)
Reticolare (rappresentabile tramite un grafo; anni sessanta)
Relazionale (attualmente il più diffuso, rappresentabile mediante tabelle e relazioni tra esse), anni settanta
Ad oggetti (estensione alle basi di dati del paradigma "Object Oriented", tipico della programmazione a oggetti, anni ottanta).
Il formato XML, oltre che per scambi di dati su web, si sta diffondendo per la definizione di vere e proprie basi di dati. XML ha una struttura gerarchica, pare quindi un "ritorno alle origini" dei modelli di dati.
Un requisito importante di una buona base dati consiste nel non duplicare inutilmente le informazioni in essa contenute: questo è reso possibile dai gestori di database relazionali (teorizzati da Edgar F. Codd), che consentono di salvare i dati in tabelle che possono essere collegate.
La funzionalità di un database dipende in modo essenziale dalla sua progettazione: la corretta individuazione degli scopi del database e quindi delle tabelle, da definire attraverso i loro campi e le relazioni che le legano, permette poi una estrazione dei dati più veloce e, in generale, una gestione più efficiente.

1.2.1 I database relazionali

Sono il tipo di database attualmente più diffuso. I motivi di questo successo sono fondamentalmente due:

- forniscono sistemi semplici ed efficienti per rappresentare e manipolare i dati
- si basano su un modello, quello relazionale, con solide basi teoriche

Il modello relazionale è stato proposto originariamente da E.F. Codd in un ormai famoso articolo del 1970. Grazie alla sua coerenza ed usabilità, il modello è diventato negli anni '80 quello più utilizzato per la produzione di DBMS**.**
La struttura fondamentale del modello relazionale è appunto la "relazione", cioè una tabella bidimensionale costituita da righe (tuple) e colonne (attributi). Le relazioni rappresentano le entità che si ritiene essere interessanti nel database. Ogni istanza dell'entità troverà posto in una tupla della relazione, mentre gli attributi della relazione rappresenteranno le proprietà dell'entità. Ad esempio, se nel database si dovranno rappresentare delle persone, si potrà definire una relazione chiamata "Persone", i cui attributi descrivono le caratteristiche delle persone. Ciascuna tupla della relazione "Persone" rappresenterà una particolare persona (vedi figura 3)

Persone

nome	cognome	data_nascita	sesso	stato_civile
Mario	Rossi	29/03/1965	M	Coniugato
Giuseppe	Russo	15/11/1974	M	Celibe
Alessandra	Mondella	13/06/1970	F	Nubile

Fig.3

In realtà, volendo essere rigorosi, una relazione è solo la definizione della struttura della tabella, cioè il suo nome e l'elenco degli attributi che la compongono. Quando essa viene popolata con delle tuple, si parla di "istanza di relazione". Perciò la precedente figura rappresenta un'istanza della relazione persona. Una rappresentazione della definizione di tale relazione potrebbe essere la seguente:

Persone (nome, cognome, data_nascita, sesso, stato_civile)

Nel seguito si indicheranno entrambe (relazione ed istanza di relazione) con il termine "relazione", a meno che non sia chiaro dal contesto a quale accezione ci si riferisce.

Le tuple in una relazione sono un insieme nel senso matematico del termine, cioè una collezione non ordinata di elementi differenti. Per distinguere una tupla da un'altra si ricorre al concetto di "chiave primaria", cioè ad un insieme di attributi che permettono di identificare univocamente una tupla in una relazione. Naturalmente in una relazione possono esserci più combinazioni di attributi che permettono di identificare univocamente una tupla ("chiavi candidate"), ma fra queste ne verrà scelta una sola da utilizzare come chiave primaria. Gli attributi della chiave primaria non possono assumere il valore null (che significa un valore non determinato), in quanto non permetterebbero più di identificare una particolare tupla in una relazione. Questa proprietà delle relazioni e delle loro chiavi primarie va sotto il nome di integrità delle entità (entity integrity).

Spesso per ottenere una chiave primaria "economica", cioè composta da pochi attributi facilmente manipolabili, si introducono uno o più attributi fittizi, che conterranno dei codici identificativi univoci per ogni tupla della relazione.

Ogni attributo di una relazione è caratterizzato da un nome e da un dominio. Il dominio indica quali valori possono essere assunti da una colonna della relazione. Spesso un dominio viene definito attraverso la dichiarazione di un tipo per l'attributo (ad esempio dicendo che è una stringa di dieci caratteri), ma è anche possibile definire domini più complessi e precisi. Ad esempio per l'attributo "sesso" della nostra relazione "Persone" possiamo definire un dominio per cui gli unici valori validi sono 'M' e 'F'; oppure per l'attributo "data_nascita" potremmo definire un dominio per cui vengono considerate valide solo le date di nascita dopo il primo gennaio del 1960, se nel nostro database non è previsto che ci siano persone con data di nascita antecedente a quella. Il DBMS si occuperà di controllare che negli attributi delle relazioni vengano inseriti solo i valori permessi dai loro domini. Caratteristica fondamentale dei domini di un database relazionale è che siano

"atomici", cioè che i valori contenuti nelle colonne non possano essere separati in valori di domini più semplici. Più formalmente si dice che non è possibile avere attributi multi valore (multivalued). Ad esempio, se una caratteristica delle persone nel nostro database fosse anche quella di avere uno o più figli, non sarebbe possibile scrivere la relazione Persone nel seguente modo:
Persone (nome, cognome, data_nascita, sesso, stato_civile, figli)
Infatti l'attributo figli è un attributo non-atomico, sia perché una persona può avere più di un figlio, sia perché ogni figlio avrà varie caratteristiche che lo descrivono. Per rappresentare queste entità in un database relazionale bisogna definire due relazioni:
*Persone(*numero_persona, nome, cognome, data_nascita, sesso, stato_civile)*
*Figli(*numero_persona, *nome_cognome, eta, sesso)*
Nelle precedenti relazioni gli asterischi (*) indicano gli attributi che compongono le loro chiavi primarie. Si noti l'introduzione nella relazione Persone dell'attributo numero_persona, attraverso il quale si assegna a ciascuna persona un identificativo numerico univoco che viene utilizzato come chiave primaria. Queste relazioni contengono solo attributi atomici. Se una persona ha più di un figlio, essi saranno rappresentati in tuple differenti della relazione Figli. Le varie caratteristiche dei figli sono rappresentate dagli attributi della relazione Figli. Il legame fra le due relazioni è costituito dagli attributi numero_persona che compaiono in entrambe le relazioni e che permettono di assegnare ciascuna tupla della relazione figli ad una particolare tupla della relazione Persone. Più formalmente si dice che l'attributo numero_persona della relazione Figli è una chiave esterna (foreign key) verso la relazione Persone. Una chiave esterna è una combinazione di attributi di una relazione che sono chiave primaria per un'altra relazione. Una caratteristica fondamentale dei valori presenti in una chiave esterna è che, a meno che non siano null, devono corrispondere a valori esistenti nella chiave primaria della relazione a cui si riferiscono. Questa proprietà va sotto il nome di integrità referenziale (referential integrity)
Uno dei grandi vantaggi del modello relazionale è che esso definisce anche un algebra, chiamata appunto "algebra relazionale". Tutte le manipolazioni possibili sulle relazioni sono ottenibili grazie ala combinazione di cinque soli operatori: RESTRICT, PROJECT, TIMES, UNION e MINUS. Per comodità sono stati anche definiti tre operatori addizionali che comunque possono essere ottenuti applicando i soli cinque operatori fondamentali: JOIN, INTERSECT e DIVIDE. Gli operatori relazionali ricevono come argomento una relazione o un insieme di relazioni e restituiscono una singola relazione come risultato. Vediamo brevemente questi otto operatori:
RESTRICT: restituisce una relazione contenente un sottoinsieme delle tuple della relazione a cui viene applicato. Gli attributi rimangono gli stessi.
PROJECT: restituisce una relazione con un sottoinsieme degli attributi della relazione a cui viene applicato. Le tuple della

relazione risultato vengono composte dalle tuple della relazione originale in modo che continuino ad essere un insieme in senso matematico.
TIME: viene applicato a due relazioni ed effettua il prodotto cartesiano delle tuple. Ogni tupla della prima relazione viene concatenata con ogni tupla della seconda.
JOIN: vengono concatenate le tuple di due relazioni in base al valore di un insieme dei loro attibuti.
UNION: applicando questo operatore a due relazioni compatibili, se ne ottiene una contenente le tuple di entrambe le relazioni. Due relazioni sono compatibili se hanno lo stesso numero di attributi e gli attributi corrispondenti nelle due relazioni hanno lo stesso dominio.
MINUS: applicato a due relazioni compatibili, ne restituisce una terza contenente le tuple che si trovano solo nella prima relazione.
INTERSECT: applicato a due relazioni compatibili, restituisce una relazione contenente le tuple che esistono in entrambe le relazioni.
DIVIDE: applicato a due relazioni che abbiano degli attributi comuni, ne restituisce una terza contenente tutte le tuple della prima relazione che possono essere fatte corrispondere a tutti i valori della seconda relazione.
I database relazionali compiono tutte le operazioni sulle tabelle utilizzando l'algebra relazionale, anche se normalmente non permettono all'utente di utilizzarla. L'utente interagisce con il database attraverso un'interfaccia differente, il linguaggio SQL, un linguaggio dichiarativo che permette di descrivere insiemi di dati. Le istruzioni SQL vengono scomposte dal DBMS in una serie di operazioni relazionali.

1.3 Analisi descrittiva dei ricoveri dell'UCIC

L'ambito di questo studio è stata l'unità di cure intensive coronariche (UCIC) dell'Azienda Ospedaliera di Padova dove sono stati ricoverati 964 pazienti dall'1 ottobre 2006 al 30 settembre 2007 di cui 909 costituiscono nuovi ingressi e i rimanenti 55 re-ingressi.
Di questi, 658 erano di sesso maschile e 306 di sesso femminile, con un'età media rispettivamente di 64,31 e di 71,37 anni.
Per quanto riguarda i fattori di rischio cardio-vascolare: 156 pazienti (16,18%) erano affetti da obesità (13,67% del totale dei maschi e 21,56% del totale delle femmine), 341 (35,37%) presentavano familiarità per coronaropatia ed aterosclerosi (36,47% e 33%), 625 (64,83%) erano affetti da ipertensione (63,06% e 68,62%), 412 (42,73%) erano fumatori (52,43% e 21,89%), 370 (38,38%) erano affetti da dislipidemia (39,20% e 36,60%) e 226 (23,44%) da diabete mellito (22,49% e 25,49%). I pazienti dello studio presentavano quindi in media 2,21 fattori di rischio, 2,27 i soggetti di sesso maschile e 2,07 quelli di sesso femminile.
A livello anamnestico, 264 (27,38%) avevano già avuto un pregresso infarto (29,79% dei maschi e 22,22% delle femmine), 131 (13,59%) erano già stati sottoposti a procedure di interventistica coronarica (14,59% e 11,44%) ed 85 (8,82%) presentavano un bypass coronarico (9,42% e 7,52%); inoltre 45 (4,67%) avevano impiantato un ICD (5,93% e 1,96%), 31 (3,21%) un pacemaker (3,04% e 3,59%), 52 (5,39%) avevano avuto un pregresso stroke (5,77% e 4,57%), 82 (8,5%) erano affetti da insufficienza renale cronica (8,5% e 8,5%) e 64 (6,64%) da BPCO (6,38% e 7,19%).
Per ogni ricovero, si sono considerate la provenienza, il motivo dell'ingresso e la durata media della degenza. Ecco i principali dati che ne abbiamo ricavato (tab.1):

Diagnosi	Pazienti	Età media (anni)	Degenza	Decessi
S. coronarica acuta con NSTE	348 (36,10%), 35,71% M e 36,93% F	67,54 M e 71,76 F	5,78 gg	13 (3,7%)
S. coronarica acuta con STE	258 (26,76%), 28,11% M e 23,86% F	62,61 M e 74,07 F	6,14 gg	3 (1,2%)
Aritmia	120 (12,44%), 11,55% M e 14,38% F	68,60 M e 72,60 F	7,67 gg	3 (2,5%)
Scompenso cardiaco	75 (7,78%), 7,45% M e 8,50% F	64,76 M e 70,79 F	10,4 gg	17 (22,7%)
Mio/pericardite	26 (2,70%), 3,65 M e 0,65% F	35,16 M e 51,53 F	4,08 gg	0
Cardiomiopatia	21 (2,18%), 2,28% M e 1,96% F	58,27 M e 52,35 F	7,29 gg	3 (14,3%)
Valvulopatia	22 (2,28%), 1,37% M e 4,25% F	70,61 M e 73,35 F	5,36 gg	2 (9,1%)
Embolia polmonare	13 (1,35%), 0,91% M e 2,29% F	62,11 M e 69,51 F	9,15 gg	1 (7,7%)
Altro	81 (8,40%), 8,97% M e 7,19% F	63,54 M e 65,10 F	4,74 gg	2 (2,5%)

Tab.1

La durata media della degenza è stata di 6,41 giorni.

Per quanto riguarda i trasferimenti: il 65,35% dei pazienti è stato trasferito nel reparto di cardiologia, il 6,53% in un reparto di cardiochirurgia, il 15,97% in un altro reparto, il 2,07% in un altro ospedale e solo il 5,50% è stato dimesso.
I decessi sono stati 44 (il 4,56%, dei ricoveri di cui il 4,25% maschi ed il 5,23% femmine).
Vediamo ora le diagnosi di dimissione dei pazienti del nostro studio (tab.2):

IMA non-Q	335 (34,75%)	34,95%M e 34,31%F
IMA Q	228 (23,65%)	25,99%M e 18,63%F
Aritmia	123 (12,76%)	11,55%M e 15,36%F
Scompenso cardiaco	107 (11,10%)	10,33%M e 12,74%F
Altra diagnosi	72 (7,47%)	7,14%M e 8,17%F
Miocardiopatia dilatativa	58 (6,02%)	6,84%M e 4,25%F
Angina instabile	52 (5,39%)	5,32%M e 5,55%F
Mio/pericardite	30 (3,011%)	3,80%M e 1,63%F
Valvulopatia	28 (2,90%)	2,13%M e 4,57%F
Sincope	20 (2,07%)	2,28%M e 1,63%F
Dolore toracico non cardiaco	17 (1,76%)	1,98%M e 1,30%F
Embolia polmonare	14 (1,45%)	0,91%M e 2,61%F
IMA con coronarie prive di lesioni significative	13 (1,35%)	0,46%M e 3,27%F
Shock cardiogeno	11 (1,14%)	1,37%M e 0,65%F

Tab.2

1.3.1 Pazienti con sindrome coronarica acuta con sopraslivellamento del tratto ST (STE)

I pazienti con questo motivo di ricovero sono stati 258, 185 maschi e 73 femmine, con età media rispettivamente di 62,61 e di 74,07 anni. I fattori di rischio cardiovascolare sono stati in media 2,23 a paziente (2,32 nei maschi e 2 nelle femmine), e nello specifico (tab3):

SESSO	OBESITÀ		FAMILIARITA'		IPERTENSION		FUMO		DISLIPIDEMIE		DIABETE		Totali	Medie FR
F	15	21%	29	40%	49	67%	19	26%	19	26%	15	21%	73	2,00
M	27	15%	74	40%	109	59%	117	63%	69	37%	34	18%	185	2,32
Totali	42	16%	103	40%	158	61%	136	53%	88	34%	49	19%	258	2,23

Tab.3

Analizziamo la provenienza di questi pazienti: il 53,88% è giunto nella nostra UCIC inviatoci dal Pronto Soccorso, il 6,59% dal PS dell'ospedale Sant'Antonio, l'1,55% dal nostro reparto di cardiologia, il 5,43% da altri reparti del Policlinico, il 19,38% da altri ospedali ed il 13,18% dal proprio domicilio.

Di questi pazienti, 201 (77,91%) sono stati sottoposti a PTCA primaria (80,54%M e 71,23%F), 3 a PTCA facilitata e 10 a PTCA rescue. Il 26,74% presentava angina pre-infartuale all'anamnesi.
Il tempo medio tra l'insorgenza del dolore e la PTCA è stato di 4 ore e 56 minuti, 4 ore e 43 min per i maschi, 5 ore e 39 min per le femmine.
Il 19,77% presentava sopraslivellamento settale (20% maschi, 19,18% femmine), il 50% anteriore (49,19% e 52,05%),il 26,74% laterale (25,4% e 30,14%) ed il 47,29% inferiore (48,11% e 45,2%).
Vediamo i valori medi di sopraslivellamento presenti in questi pazienti: V1 sopraslivellamento medio di 1,7 mm, V2 3,2 mm, V3 3,3 mm, V4 2,7 mm, V5 2,1 mm, V6 1,9 mm, DI 1,4 mm, DII 2 mm, DIII 2,8 mm, aVL 1,7 mm, aVF 2,3 mm.
All'ingresso è stato valutato ogni paziente, che è stato quindi inserito nella classificazione di Killip; il 66,8% era in classe prima (71,98% dei maschi e 53,52% delle femmine), il 28,46% in classe seconda (24,17% e 39,44%), il 2,77% in terza(1,65% e 5,63%)e l'1,98%% in quarta (2,2% e 1,41%).
Sono stati ricoverati nella nostra Unità Coronarica 33 pazienti (12,79%, 13,51% dei maschi e 10,96% delle femmine) in seguito all'invio, da parte dei medici del 118, del tracciato elettrocardiografico per via telematica; tale tracciato è stato quindi valutato dai nostri medici i quali hanno dato indicazione di ricovero immediato senza la necessità di far transitare il paziente per il Pronto Soccorso del paziente.
I farmaci somministrati durante la degenza per pazienti con SCA STE sono stati i seguenti (tab.4):

Nitrati ev	85,60%	Clopidogrel	82,49%	Furosemide	46,30%
Eparina ev	84,82%	Beta-bloccanti	72,37%	Inotropi	16,73%
Eparina basso pm	36,58%	ACE-inibitori	64,20%	Warfarin	2,99%
ASA	95,72%	Satanici	5,45%	Statine	87,94%
Ticlopidina	3,11%	Ca-antagonisti	14,79%	Antiaritmici	17,51%
Trombolisi	5,06%	Digitale	1,80%		

Tabella 4

La trombolisi (5,95% dei maschi e 2,74% delle femmine) è stata eseguita nel 92% dei casi in un altro ospedale, e solo in un caso all'interno dell'UCIC.
Il reo-pro è stato somministrato a 108 pazienti (42,03%), 51,35% dei maschi e 17,81% delle femmine. Nel 51,51% dei casi è stato eseguito bolo in sala di emodinamica più infusione in UCIC, nel 28,28% il bolo è stato eseguito prima dell'invio del paziente in-sala di emodinamica mentre nel 20,20% è stata eseguita solo l'infusione post-procedurale.
A tutti i pazienti è stato applicato monitoraggio telemetrico e controlli periodici della pressione, della frequenza cardiaca e dei restanti parametri vitali.
Le principali aritmie rilevate sono: BAV 5,04% (4,86% dei maschi e 5,48% delle femmine), FA 7,36%(4,32% e 15,07%), TV

4,65%(4,86% e 4,11%), FV 4,26% (3,78% e 5,48%), TVNS 48,84% (49,19% e 47,94%), TPSV 6,97% (5,95% e 9,59%).
I valori medi di alcuni esami di laboratorio eseguiti sono qui riportati: acido urico 0,32 mmol/l, Lp(a) 288,48 mg/dl, omocisteina 24,15 mmol/l, PCR 21,90 mg/l. Ecco i valori medi dei lipidi nel sangue nei pazienti con dislipidemia: colesterolo totale 230 mg/dl, LDL 169 mg/dl, HDL 54 mg/dl, trigliceridi 194,36 mg/dl.
Particolare importanza è stata data ai markers di miocardio-citolisi: TnI entrata 9,55 µg/dl (10,81 nei maschi e 6,31 nelle femmine), TnI picco 76,77 µg/dl (80,32 e 67,58).I pazienti con TnI > 10 µg/l sono stati l'80,23% (78,92% maschi, 83,56% femmine).
Nei 258 pazienti con SCA STE sono state eseguite 259 CNG.
La frazione di eiezione mediamente rilevata è stata del 54,64%.
I pazienti con almeno un vaso occluso sono stati 159 (61,39%, 60,82% maschi e 63,08% femmine).
La media dei vasi malati è stata di 1,63 (1,66 nei maschi e 1,54 nelle femmine), mentre i vasi trattati sono stati 0,87 (0,88 e 0,86). Il 40,54% dei pazienti presentava un interessamento monovasale (39,17% maschi e 44,61% femmine), il 27,80% bivasale (25,77% e 33,85%) e il 20,08% trivasale (22,16% e 13,65%).
L'11,58% dei pazienti alla coronarografia presentavano delle coronarie prive di lesioni angiograficamente significative (12,89% e 7,69%).
Il 5,02% dei pazienti aveva una lesione significativa del tronco comune (5,15% e 4,61%), mentre il 25,10% presentava dei circoli collaterali (28,35% e 15,38%). Solo il 10,44% degli stent posizionati era di tipo medicato.

1.3.2 Pazienti con sindrome coronarica acuta senza sopraslivellamento del tratto ST (NSTE)

I 348 pazienti che si sono presentati con questo quadro (235 maschi e 113 femmine) avevano un'età media rispettivamente di 67,54 e di 71,76 anni.
Analizzando i fattori di rischio cardiovascolare, ne abbiamo riscontrati in media 2,57 per paziente (2,60 nei maschi e 2,49 nelle femmine)(tab.5):

SESSO	OBESITÀ		FAM.		IPERT.		FUMO		DISLIP.		DIAB		Totali	Medie FR
F	29	26%	46	41%	91	81%	25	22%	56	50%	34	30%	113	2,49
M	42	18%	88	37%	176	75%	131	56%	117	50%	58	25%	235	2,60
Totali	71	20%	134	39%	267	77%	156	45%	173	50%	92	26%	348	2,57

Tab.5

La provenienza di questi pazienti giunti in UCIC è la seguente: il 57,47% proveniva dal Pronto Soccorso, il 4,31% dal PS dell'ospedale Sant'Antonio, il 2,59% dal nostro reparto di cardiologia, il 24,42% da altri reparti del Policlinico, il 7,76% da altri ospedali ed il 3,45% dal proprio domicilio.

All'ingresso 139 pazienti (39,94%, 39,57% dei maschi e 40,71% delle femmine) presentavano elettrocardiograficamente un sottoslivellamento del tratto ST significativo.
Il 10,63% (10,21% e 11,50%) presentava invece un sopraslivellamento che si è rivelato essere non persistente, mentre lo 0,57% (0,85% dei maschi e 0% delle femmine) presentava una tachicardia ventricolare non sostenuta.
Il 48,85% dei pazienti (49,36% maschi e 47,79% femmine) presentava invece un ECG privo di alterazioni indicative di ischemia o un ECG con alterazioni elettrocardiografiche non specifiche.
Il TIMI RISK SCORE medio è risultato essere di 4,04 (4,02 nei maschi e 4,09 nelle femmine).

Le percentuali suddivise per sesso delle variabili considerate nello score sono riportate in tabella 6.

	Maschi	Femmine
Età ≥ 65 anni	62,98%	78,76%
Angina recente insorgenza	71,06%	75,22%
≥ 3 fattori di rischio	51,49%	51,33%
Aumento della troponina	81,28%	76,99%
Nota coronaropatia	43,40%	40,71%
ASA negli ultimi 7 gg	45,53%	40,71%
STsopra ≥ 0,1 mm	45,96%	45,13%

Tab.6

Secondo la classificazione Killip (noto predittore dell'insufficienza cardiaca nei pazienti con infarto miocardico acuto), il 60,22% era in classe prima (59,12% dei maschi e 66,36% delle femmine), il 34,31% in seconda (35,36% e 32,26%), il 4,38% in terza (4,42% e 4,30%) e solo l'1,09% in classe quarta (1,10% e 1,07%).

I farmaci somministrati ai pazienti con SCA NSTE sono stati qui elencati in tabella 7:

Nitrati ev	86,78%	Clopidogrel	58,91%	Furosemide	46,84%
Eparina ev	56,90%	Beta-bloccanti	67,24%	Inotropi	7,76%
Eparina basso pm	49,14%	ACE-inibitori	55,17%	Warfarin	2,79%
ASA	91,67%	Satanici	11,21%	Statine	81,61%
Ticlopidina	8,91%	Ca-antagonisti	27,59%	Antiaritmici	8,62%
Aggrastat	31,03%	Digitale	4,19%	Reo-Pro	4,88%

Tab.7

Le principali aritmie rilevate in questi pazienti sono state: BAV 2,87% (3,40% maschi e 1,77% femmine), FA 10,92% (8,94% nei maschi e 15,04%) nelle femmine, TV 3,16% (4,25% maschi e 0,88%) femmine, FV 0,86% (1,28% nei maschi e 0% nelle femmine), TVNS 20,67% (22,13% nei maschi e 17,70%) nelle femmine, TPSV 10,06% (9,79% nei maschi e 10,62% nelle femmine).

Mediamente gli esami ematochimici hanno rilevato valori medi dei lipidi nel sangue nei pazienti con dislipidemia: colesterolo totale 230 mg/dl, LDL 158 mg/dl, HDL 52 mg/dl, trigliceridi 170 mg/dl.

Particolare importanza è stata data ovviamente ai markers di miocardio-citolisi: TnI entrata 7,22 µg/dl (3,31 nei maschi e 15,11 nelle femmine), TnI picco 17,36 µg/dl (16,56 e 19,23).
I pazienti con TnI > 10 µg/l sono stati l'37,07% (28,30% maschi, 34,51% femmine).

Il tempo medio tra l'ingresso del paziente e l'esecuzione della coronarografia è stato di 1,66 giorni (1,46 per i maschi e 2,2 per le femmine); sono state eseguite 259 coronarografie (CNG).
La frazione di eiezione media rilevata è stata di 58,45% (58,35% e 58,70%). Le CNG in cui abbiamo riscontrato almeno un vaso occluso sono state 85 (32,82%, 37,30% maschi e 21,62% femmine). Il numero di vasi malati medio è stato di 1,91 (2,09 e 1,47), mentre i vasi trattati sono stati 0,63 (0,66 e 0,57). Il 23,94% dei pazienti presentavano un interessamento monovasale (21,62% dei maschi e 29,73% delle femmine), il 27,41% bivasale (28,65% e 24,32%) ed il 37,07% trivasale (42,70% e 22,97%). L'11,58% dei pazienti alla coronarografia presentavano delle coronarie prive di lesioni angiograficamente signifi-cative(7,03% dei maschi e 22,97% delle femmine). Il 10,81% dei pazienti aveva una malattia angio-graficamente significativa del tronco comune (11,89% e 8,11%), mentre il 31,66% (36,76% e 19,92%) presentava dei circoli collaterali.

L'interessamento medio delle coronarie rilevato è qui riassunto nella tabella 8:

Tronco comune	40,83%	Circonflessa	77,7%	Rami PL	64,09%
IVA	75,84%	Marginale	78,60%	IVP	71,67%
Diagonali	80,59%	Coronaria dx	75,76%	Intermedio	81,16%

Tab.8

Alla dimissione dei pazienti sono state riportate le seguenti diagnosi (tab.9):

IMA Q anteriore	11 (3,16%)	3,40%M	2,65%F
IMA Q laterale	3 (0,86%)	0,42%M	1,77%F
IMA Q inferiore	21 (6,03%)	6,38%M	5,31%F
IMA non Q	243 (69,83%),	71,91%M	65,47%F
IMA con coronarie prive di lesioni significative	6 (1,72%),	0,42%M	4,42%F
Angina instabile	47 (13,51%),	13,19%M	14,16%F
Mio/pericardite	4 (1,15%),	0,85%M	1,77%F

Dolore toracico di origine non cardiaca	10 (2,87%),	2,55%M	3,54%F
Miocardiopatia dilatativa	8 (2,30%),	2,98%M	0,88%F
Scompenso cardiaco	16 (4,60%),	4,25%M	5,31%F
Shock cardiogeno	3 (0,86%)	85%M	0,88%F
Altra diagnosi	6 (1,72%)	0,85%M	3,54%F

(Tab.9)

2 Analisi della durata della degenza

2.1 Introduzione

L'obiettivo di questo studio prospettico condotto dal 1 ottobre 2006 al 30 settembre 2007dei pazienti ricoverati (unità statistiche) in UCIC dell'AOP è l'analisi della durata della degenza utilizzando la classe dei modelli lineari generalizzati (GLM). In particolare si vogliono osservare le relazioni che intercorrono tra la durata della degenza in UCIC (unità di cure intensive cardiologiche) dell'Azienda Ospedaliera di Padova (AOP) e alcune variabili:

- antropometriche,
- dei fattori di rischio,
- bioumorali
- e dei rilievi anamnestici.

Tutta la bibliografia di riferimento inerenti i modelli lineari e i modelli linerari generalizzati compresi gli argomenti di complemento trattati in questo elaborato (sezioni da 2.3 a 2.9.2) sono ripresi da: Azzalini A. e Scarpa B. (2004), Analisi dei dati e data mining, Milano, Sprinter-Verlag Italia.

2.2 Il dataset di riferimento

I dati sono stati ricavati da un database relazionale realizzato in collaborazione con il personale medico operante presso la struttura in Microsoft Access 2000 di cui riportiamo lo schema E-R (entità relazione) (fig.).

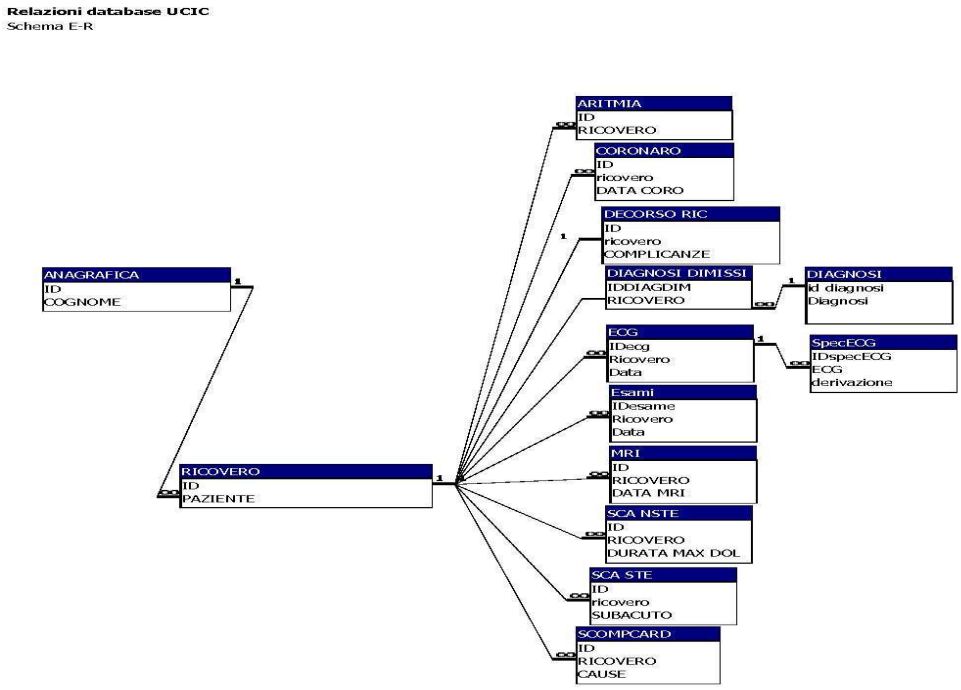


Fig.4

Per l'analisi sono state selezionate per ogni unità statistica le seguenti categorie di variabile come potenziali predittori di sopravvivenza tra tutte quelle presenti nel database:
-antropometriche:

2.3 Modelli Lineari

L'obiettivo dei modelli lineari è quello di studiare la relazione che intercorre tra le variabili che caratterizzano un fenomeno.
In particolare si assume che una variabile risposta, chiamata Y, sia legata linearmente ad una o più variabili esplicative fissate, chiamate $X_1,..., X_k$
La dipendenza lineare di Y rispetto alle variabili esplicative viene introdotta assumendo che la media della variabile di risposta sia una combinazione lineare delle variabili esplicative con $ß_1, ..., ß_k$ incogniti.
Il valore osservato della variabile risposta è quindi composto da due termini, ovvero

$$Y = r(x_1, \ldots, x_n) + \epsilon = \sum_{j=1}^{k} \beta_j x_j + \epsilon \qquad (1.1)$$

in cui il termine r $(x_1,..., x_n)$, che rappresenta la combinazione lineare delle variabili esplicative, `e detta componente sistematica, mentre ε è detta componente accidentale o di errore. Quest'ultima rappresenta lo scostamento di natura casuale tra Y e r $(x_1,..., x_n)$.
Poichè μ = E(Y) è dato solo dalla componente sistematica dovrà essere E(ε) = 0.
Supponiamo ora di essere in presenza di n osservazioni $Y_1,..., Y_n$ e assumiamo che

$$Y_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

con i = 1,..., n e dove Y_i `e la i -esima componente della variabile di risposta, x_{i1} rappresenta l'i - esimo valore della variabile esplicativa x_1, e così via. Si può riscrivere tutto in forma compatta come:
$Y = Xß + ε$
dove $X = (x_{ij})$ è detta matrice di regressione mentre i ß sono chiamati parametri di regressione.
In questa classe di modelli, detti modelli lineari, si assume che $E(ε) = 0$ e $Var[ε]=s^2I$. Assumendo inoltre che $Y \sim N_n[Xß, σ^2I]$ il modello viene detto lineare normale.

2.4 Carenze dei Modelli Lineari

I modelli lineari presentano delle carenze, facciamo alcuni esempi
Può succedere che la relazione sia del tipo (1.1) ma con r(·) decisamente non lineare nei parametri.
Anche quando non si sa com'è r(·), dalla natura del fenomeno si è in grado di escludere a priori una relazione di tipo lineare.

La varianza del termine di errore e quindi anche della variabile risposta è stata posta costante (ipotesi del secondo ordine) mentre spesso si riscontra empiricamente che ciò non è vero.
Nei modelli lineari si assume che la distribuzione della variabile risposta sia normale (ipotesi di normalità) ma spesso ciò non si riscontra nella realtà. Si trasforma la variabile risposta in una nuova variabile con distribuzione normale. Questo metodo però non si può applicare sempre; in particolare quando Y è una variabile discreta può risultare problematico o anche impossibile trasformare Y in modo da ottenere una variabile normale.

2.5 Un sottoinsieme della Famiglia Esponenziale

Introduciamo ora una classe di distribuzioni di probabilità che permetterà di generalizzare l'assunzione di normalità tipica dei modelli lineari.
Tale categoria di distribuzioni di probabilità è un sottoinsieme delle famiglie esponenziali che, ricordiamo, hanno densità

$$f(y) = q(y) \exp\left\{\sum \mu_i(\theta)\, t_i(y) - \tau(\theta)\right\} \tag{1.2}$$

Data l'importanza che tale classe di distribuzioni di probabilità avrà nel nostro studio, ne introduciamo una notazione specifica. Per una variabile continua reale Y diremo che

$$Y \sim EF\left(b(\theta), \frac{\psi}{\omega}\right)$$

se Y ha funzione di densità del tipo

$$f(y) = \exp\left\{\left(\frac{\omega}{\psi}\left(y\theta - b(\theta)\right)\right) + c(y, \psi)\right\} \tag{1.3}$$

dove θ e ψ sono dei parametri scalari ignoti, ω è una costante nota , e b(·) e c(·) sono funzioni note la cui scelta individua una particolare distribuzione di probabilità. Per ogni particolare scelta di ψ, che è detto parametro di dispersione, la (1.3) costituisce una famiglia esponenziale di parametro θ, ma bisogna sottolineare che la (1.3) non sempre appartiene alla famiglia esponenziale nel caso in cui sia θ che ψ varino simultaneamente. Si consideri per ora ψ fissato. La (1.3) la si può scrivere come

$$f(y) = \exp\left\{\left(\frac{\theta}{\psi} y\omega - b(\theta)\frac{\omega}{\psi} + c(y;\psi)\right)\right\} \tag{1.4}$$

$$= \exp\left\{\left(\frac{\theta}{\psi}\omega y - b(\theta)\frac{\omega}{\psi}\right)\right\} \exp\{c(y;\psi)\} \tag{1.5}$$

Quindi facendo un parallelo con la (1.2) con i=1 otteniamo:

$$\left.\begin{aligned} t(y) &= y \\ \mu(\theta) &= \frac{\theta}{\psi}\omega \\ q(y) &= \exp\{c(y;\psi)\} \\ \tau(\theta) &= b(\theta)\frac{\omega}{\psi} \end{aligned}\right| \qquad \begin{aligned} &(1.6)\\ &(1.7)\\ &(1.8)\\ &(1.9) \end{aligned}$$

Si vuole calcolare la media e la varianza di Y.
Per potelo fare dobbiamo per prima cosa determilarle nel caso di una famiglia esponeziale.
La f(y) essendo una distribuzione di probabilità avrà:

$$\left.\int f(y,\theta)dy = 1,\right|$$

inoltre f(.) è sufficientemente regolare da poter giustificare la derivazione sotto segno di integrale e quindi che:

$$0 = \frac{d}{d\theta}1 = \frac{d}{d\theta}\int_y f(y;\theta)dy = \int_y \frac{d}{d\theta}f(y;\theta)dy.$$

Osserviamo allora che la derivata prima della mia f(·) è

$$\begin{aligned} \frac{d}{d\theta}q(y)exp\left(\mu(\theta)t(y)-\tau(\theta)\right) &= q(y)\left(\mu'(\theta)t(y)-\tau'(\theta)\right)exp\left(\mu(\theta)t(y)-\tau(\theta)\right) \\ &= f(y;\theta)\left(\mu'(\theta)t(y)-\tau'(\theta)\right) \end{aligned}$$

Quindi otteniamo

$$\int_y f(y,\theta)\left(\left(\mu'(\theta)t(y)-\tau'(\theta)\right)dy = 0\right.$$

Essendo μ'(θ) e τ'(θ) costanti all'interno degli intervalli si avrà

$$\mu'(\theta)\int_y f(y;\theta)t(y)dy = \tau'(\theta)\int_y f(y;\theta)dy$$

In definitiva:

$$\left.\mathbb{E}\{t(y)\} = \frac{\tau'(\theta)}{\mu'(\theta)}\right| \qquad (1.10)$$

Per il calcolo della varianza osserviamo innanzi tutto che

$$0 = \int_y \frac{d}{d\theta}\left[f(y,\theta)\left(\left(\mu'(\theta)t(y)-\tau'(\theta)\right)\right]dy \qquad (1.11)$$

dove

$$\left.\begin{aligned} f'(y;\theta) &= q(y)\left(\mu'(\theta)t(y)-\tau'(\theta)\right)exp\left(\mu(\theta)-\tau(\theta)\right) \\ &= f(y;\theta)\left(\mu'(\theta)t(y)-\tau'(\theta)\right) \end{aligned}\right|$$

ora si può calcolare esplicitamente al (1.11)

$$0 = \int_y f'(y;\theta)\left(\mu'(\theta)t(y) - \tau'(\theta)\right) + f(y;\theta)\left(\mu''(\theta)t(y) - \tau''(\theta)\right) dy$$

$$= \int_y f(y;\theta)\left(\mu'(\theta)t(y) - \tau'(\theta)\right)^2 + f(y,\theta)\left(\mu''(\theta)t(y) - \tau''(\theta)\right) dy$$

L'integrale del primo addendo è il momento secondo della variabile continua $\{\mu'(\theta)t(y) - \tau^0(\theta)\}$; tale variabile ha valor medio nullo, in base alle
relazioni ottenute precedentemente e quindi il momento secondo coincide con la varianza si può quindi scrivere

$$Var\left\{\mu'(\theta)t(y) - \tau'(\theta)\right\} = -\int_y f(y)\left(\mu''(\theta)t(y) - \tau''(\theta)\right) dy \quad (1.12)$$

$$= -\mathbb{E}\left\{\mu''(\theta)t(y) - \tau''(\theta)\right\} \quad (1.13)$$

$$= \frac{\mu'(\theta)\tau''(\theta) - \mu''(\theta)\tau'(\theta)}{\mu'(\theta)} \quad (1.14)$$

e in definitiva si ottiene

$$Var\left\{t(y)\right\} = \frac{\mu'(\theta)\tau''(\theta) - \mu''(\theta)\tau'(\theta)}{\mu'(\theta)^3}. \quad (1.15)$$

A questo punto osservando che nella formula (1.3) la statistica t(y) assume in realtà proprio il valore y si può concludere che se Y ~ EF (b(θ), ψ/ω)

$$\mathbb{E}(y) = \frac{\tau'(\theta)}{\mu'(\theta)} = b'(\theta)$$

$$Var(y) = \frac{\mu'(\theta)\tau''(\theta) - \mu''(\theta)\tau'(\theta)}{\mu'(\theta)^3} = b''(\theta)\frac{\psi}{\omega}$$

che è quello che si voleva determinare. Procedendo allo stesso modo con derivate di ordine superiore si possono ottenere espressione dei momenti superiori della v.c Y . Si vede peraltro come la (1.15), e ancor più le espressioni dei momenti superiori, si semplificano notevolmente se μ(θ) = θ, si parla in tal caso di parametrizzazione naturale della famiglia esponenziale.
Se Y è una osservazione da una v.c Gamma(v, v/λ), il parametro naturale, è rappresentato dal $-v^{-1}$. Concludiamo osservando che per le successive elaborazioni algebriche si porrà

$$\mu = b'(\theta) \quad (1.16)$$

$$V(\mu) = b''(\theta)\big|_{\theta = b'^{-1}(\mu)} \quad (1.17)$$

2.6 Modelli Lineari Generalizzati (GLM)

I Modelli Lineari Generalizzati sono una naturale estensione dei modelli lineari.

Si considera il caso in cui la funzione r(·), introdotta nella formula (1) non sia lineare e le variabili non siano normali. Questa nuova classe di modelli non è molto ampia da un punto di vista strettamente matematico ma è sufficientemente flessibile da incorporare un gran numero di situazioni rilevanti per le applicazioni pratiche. Inoltre la classe dei GLM ha anche il pregio di permettere una trattazione unificata di una serie di modelli specifici, che prima dell'introduzione di questa classe erano trattati come casi singoli e non come sottocasi di un modello generale.
Per definire questa nuova classe di modelli si riconsiderano gli elementi caratteristici dei modelli lineari.
Per la generica unità i-esima poniamo $\eta_i = x_i^T$ ß dove x_i^T è la i-esima riga della matrice X per i = 1,…, n. Tale quantità incognita verrà chiamata predittore lineare.
Ricordando che nei modelli lineari normali veniva ipotizzato che $Y_i \sim N(\mu; \sigma^2)$ dove la relazione tra μ_i e il predittore lineare η_i era rappresentata dalla funzione identità, la classe dei GLM si ottiene estendendo la formulazione precedente in due direzioni:
si pone Yi non strettemente Normale ma $Y_i \sim EF(b(\theta_i),\Psi/\omega_i)$ tale che $b'(\theta_i) = \mu i$;
si prendono in considerazione altre forme di legame tra il predittore lineare η_i e il valor medio μ_i, ovvero si ipotizza $g(\mu_i) = \eta_i$.

Sintetizzando il tutto si afferma che un GLM è caratterizzato dai seguenti

$$Y_i \sim EF\left(b(\theta_i), \frac{\psi}{\omega_i}\right) \tag{2.1}$$

$$g(\mu_i) = \eta_i \tag{2.2}$$

$$\eta_i = x_i^T \beta \tag{2.3}$$

con $b'(\theta_i) = \mu_i$. Più analiticamente un GLM è caratterizzato dai seguenti elementi:
le osservazioni $y_1, \ldots, y_n$ sono tratte da variabili continue $Y_1, \ldots, Y_n$ tra loro indipendenti;
ciascuna Y_i ha distribuzione del tipo $EF(b(\theta_i), \psi/\omega_i)$ con $E(Y_i) = \mu_i = b'(\theta_i)$ per i = 1, ..., n;
esiste una funzione $g(\cdot)$ tale per cui $g(\mu_i) = x_i^T$ ß,dove x_i è un vettore di costanti e ß un vettore di parametri;
le funzioni g(μ), b(θ) e c(y; ψ) e il parametro di dispersione ψ sono comuni a tutte le Y_i, mentre il fattore peso ω può variare da individuo a individuo.

2.7 Verosimiglianza e Informazione di Fisher

Date le osservazioni campionarie $y_1, \ldots, y_n$ si vuole procedere a fare inferenza sui parametri ß e ψ con particolare interesse per ß poichè determina la relazione tra le variabili esplicative e la media μ.

Sia p la dimensione di ß e X = (x_{ij}) la matrice n × p con i-esima riga x_i^T. Essendo tutte le componenti indipendenti, si ha che la log-verosimiglianza è

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} \left(\frac{\omega_i \left(y_i \theta_i - b(\theta_i) \right)}{\psi} + c_i(y_i, \psi) \right) \\ &= \sum_{i=1}^{n} l_i(\beta) \end{aligned}$$

Per ottenere le equazioni di verosiglianza si calcola

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \tag{2.4}$$

$$\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{\frac{\psi}{\omega_i}} \tag{2.5}$$

ed essendo μ = b'(θ) e Var(Y_i) = b''(θ) Ψ/ω_i

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{\omega_i}{\psi} Var(Y_i)$$

e poiché η = x_i^Tβ

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$$

In definitiva le equazioni della verosimiglianza per β sono

$$\sum_{i=1}^{n} \frac{(y_i - \eta_i) x_{ij}}{Var\{Y_i\}} \frac{\partial \mu_i}{\partial \eta_i} = 0$$

Per ottenere l'informazione di Fisher si considerano le derivate seconde di l_i(β) ricavando

$$\begin{aligned} -\mathbb{E}\left\{ \frac{\partial^2 l}{\partial \beta_i \partial \beta_k} \right\} &= \mathbb{E}\left\{ \frac{\partial l_i}{\partial \beta_j} \frac{\partial l_i}{\partial \beta_k} \right\} \\ &= \mathbb{E}\left\{ \left(\frac{(Y_i - \mu_i) x_{ij}}{var\{Y_i\}} \frac{\partial \mu_i}{\partial \eta_i} \right) \left(\frac{(Y_i - \mu_i) x_{ik}}{var\{Y_i\}} \frac{\partial \mu_i}{\partial \eta_i} \right) \right\} \\ &= \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \frac{x_{ij} x_{ik}}{Var\{Y_i\}^2} \mathbb{E}\left\{ (Y_i - \mu_i)^2 \right\} \\ &= \frac{x_{ij} x_{ik}}{Var\{Y_i\}} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \end{aligned}$$

Quindi la matrice dell'informazione attesa ha elemento (j,k)-esimo

$$\sum_{i=1}^{n} \mathbb{E}\left\{ \frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k} \right\}$$

ovvero in forma matriciale

$$I(\beta) = X^T \widetilde{\mathcal{W}} X$$

dove

$$\widetilde{\mathcal{W}} = \begin{pmatrix} \widetilde{w_1} & 0 & \cdots & 0 \\ 0 & \widetilde{w_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \widetilde{w_n} \end{pmatrix}$$

avendo posto

$$\widetilde{w_i} = \frac{1}{var\{Y_i\}} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$$

2.8 Legami Canonici e Statistiche Sufficienti

La g(μ) si può scegliere in modo arbitrario in quanto non è soggetta a particolari restrizioni. Ad esempio si può scegliere in modo tale che θ_i coincida con η_i, parametro naturale della famiglia esponenziale ovvero: $g(\mu_i) = \theta_i$

Questo particolare legame è definito legame canonico. In questo caso si verifica che

$$\begin{aligned} l(\beta) &= \frac{1}{\psi} \left\{ \sum_{i=1}^{n} \omega_i \left(y_i \theta_i - b\left(\theta_i\right) \right) \right\} + \sum_{i=1}^{n} c_i(y_i, \psi) \\ &= \frac{1}{\psi} \left\{ \sum_{i=1}^{n} \omega_i \left(y_i {x_i}^T \beta - b({x_i}^T \beta) \right) \right\} + \sum_{i=1}^{n} c_i(y_i, \psi) \\ &= \frac{1}{\psi} \left\{ \left(\sum_{i=1}^{n} \omega_i y_i x_i \right)^T \beta - \sum_{i=1}^{n} \omega_i b\left(x_i^T \beta \right) \right\} + \sum_{i=1}^{n} c_i\left(y_i, \psi\right) \end{aligned}$$

che mostra che (Σi ωiyixi) è una statistica sufficiente per β nel caso che il parametro Ψ sia assente oppure noto.

Se Ψ è ignoto, ma la verosimiglianza è ancora distribuita esponenzialmente, la (Σi ωiyixi) è comunque una componente della statistica sufficiente minimale. Avendo posto $g(\mu i) = \theta_i$ dà luogo anche ad altri vantaggi .

Per quanto riguarda le derivate della log-verosimiglianza si ha che

$$\frac{d\mu_i}{d\eta_i} = \frac{d\mu_i}{d\theta_i} = \frac{db'(\theta_i)}{d\theta_i} = b''(\theta_i)$$

e quindi

$$\frac{\partial l_i}{\partial \beta_j} = \frac{(y_i - \mu_i)x_{ij}}{var\{Y_i\}} b''(\theta_i) = \frac{\omega_i (y_i - \mu_i) x_{ij}}{\psi}$$

ed implica che

$$\sum_i y_i x_{ij} = \sum_i x_{ij} \hat{\mu}_i$$

nel caso che $\omega_i = 1$
In notazione matriciale:

$$X^T y = X^T \hat{\mu}$$

indicando con $\hat{\mu}_i$ i valori di μ_i orrispondenti alla SMV $\hat{\beta}$di β. Si supponga che una colonna di X sia I_n allora $X^T y = X^T \hat{\mu}$ ci dice che il totale dei valori osservati y è uguale al totale dei valori interpolati $\hat{\mu}$ e un'uguaglianza analoga vale per le altre colonne di X.
Per quando riguarda invece l'informazione di Fischer derivando si ottiene

$$\frac{\partial^2 l}{\partial \beta_i \beta_k} = \mathbb{E}\left\{ \frac{\partial^2 l}{\partial \beta_i \beta_k} \right\}$$

Quindi l'informazione attesa e osservata coincidono.
Riportiamo una tabella in cui, tra sono anche rappresentati i legami canonici relativi ad alcune distribuzioni.

Distribuzione	Normale $N(\mu, \sigma^2)$	Binomiale/m $Bin(m, \mu)/m$	Gamma $G(\omega, \omega/\mu)$
Supporto	$(-\infty, \infty)$	$\{0, 1/m, ..., 1\}$	$(0, \infty)$
ψ	σ^2	1	ω^-1
ω	1	m	1
$b(\theta)$	$\theta^2/2$	$\log\left(1 + e^\theta\right)$	$-\log(-\theta)$
$c(y;\psi)$	$-\frac{1}{2}\left(\frac{y^2}{\psi} + \log(2\pi\psi)\right)$	$\log\begin{pmatrix} m \\ my \end{pmatrix}$	$\frac{\log(\omega y)^\omega}{\log y \log \Gamma(\omega)}$
$\mu(\theta)$	θ	$e^\theta/1 + e^\theta$	$-\frac{1}{\theta}$
Legame canonico	Identità	logit	reciproco
$V(\mu)$	1	$\mu(1-\mu)$	μ^2

Tabella 2.1: Elementi caratteristici di alcune distribuzioni

2.9 Adeguatezza dei Modelli

2.9.1 Devianza

Si considera il problema di confrontare due GLM. Siano M_1 e M_2 due modelli distinti con la condizione che M_2 incluso in M_1, si parla

quindi di modelli annidati. In particolare si prendono due modelli lineari generalizzati, e su M_2 si impongono dei vincoli supplementari sui parametri del predittore lineare. Ovvero se M_1 è un modello contenente p_1 parametri e M_2 un modello con p_2 parametri, si impongono dei vincoli del tipo $g_i(ß) = 0$ per i = 1, ..., p_2-p_1 come metodo per confrontare i modelli annidati si utilizza quello del rapporto di massima verosimiglianza. Nei modelli lineari la seguente quantità

$$Q\left(\hat{\beta}\right) = ||y - \hat{\mu}||^2$$

viene chiamata Devianza e si può dimostrare che il test del rapporto di verosimiglianza è funzione della devianza associata a ciascuno dei modelli.
Infatti la verosimiglianza dipende dai dati solo attraverso D, come si vede scrivendo

$$l\left(\hat{\beta}\right) = c - \frac{n}{2}\log\left(\sigma^2\right) - \frac{D}{2\sigma^2}.$$

Si definisce modello saturo quello in cui le stime di μ_i coincidono con y_i, cosa che è possibile con un modello contenente n parametri.
Tale modello non è di utilità pratica, ma serve come termine di confronto per il modello effettivamente in esame. Tecnicamente esso serve a liberare la log-verosimiglianza dalle costanti arbitratrie.
Se si confrontasse la log-verosimiglianza del modello in questione con quello saturo, con $\tilde{\mu}_i = y_i$, si ottiene che il rapporto di verosimiglianza tra i due modelli, saturo e annidato, vale

$$-2\left\{l\left(\hat{\beta}\right) - l\left(\tilde{\beta}\right)\right\} = \frac{D}{\sigma^2}$$

ed è pari alla devianza stessa, a meno di una costante. Inoltre per confrontare due modelli annidati, il rapporto di verosimiglianza diventa

$$W = -2\left\{l\left(\hat{\beta}_2\right) - l\left(\hat{\beta}_1\right)\right\} = \frac{D_2 - D_1}{\sigma^2}$$

e cio è sostanzialmente la differenza delle devianze a meno di una costante σ^2 che abbiamo supposto nota.
Si può ora estendere il concetto di devianza nell'ambito del GLM. Il modello saturo rimane invariato e si indica con $\tilde{\theta}_i$ il corrispondente valore di θ_i.
Risulta allora che

$$W(y) = -2\left[l\left(\hat{\beta}\right) - l\left(\tilde{\beta}\right)\right]$$

$$= -2\sum_i \frac{\omega_i}{\psi}\left[\left(y_i\hat{\theta}_i - b\left(\hat{\theta}_i\right)\right) - \left(y_i\tilde{\theta}_i - b\left(\tilde{\theta}_i\right)\right)\right]$$

$$= \frac{\sum_i d_i}{\psi}$$

Dove d_i è il contributo della i-esima osservazione alla devianza; la quantita risultante è detta devianza normalizzata. Nel caso dei modelli annidati

$$\frac{D\left(y;\hat{\mu_2}\right) - D\left(y;\hat{\mu_1}\right)}{\psi} \overset{d}{\longmapsto} \chi^2_{p_1-p_2}$$

per n → ∞ e se p1 e p2 sono costanti rispetto a n.

2.9.2 Residui

I residui costituiscono uno strumento grazie al quale si può valutare informalmente l'adeguatezza di un modello lineare.
Si estende il concetto di residuo agli GLM nel modo che segue. Si prendono in esame

$$e_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V\left(\mu_i\right)/\omega_i}}$$

definito come residuo di Pearson

$$e_i^D = sign(y_i - \hat{\mu})\sqrt{d_i}$$

detto residuo di devianza

$$e_i^A = \frac{A(y_i) - A(\mu_i)}{A'(\mu_i)\sqrt{V\left(\mu_i\right)}}$$

in cui

$$A(x) = \int \frac{1}{V(x)^{1/3}} dx$$

è scelto in modo tale da rendere la loro distribuzione normale.

2.9.3 Criterio Informativo di Akaike

Il criterio informativo di Akaike (AIC), sviluppato da Hirotsugu Akaike sotto il nome di "un criterio di informazioni" (AIC) nel 1971 e proposto dallo stesso Akaike nel 1974, è una misura della bontà di un modello statistico per stimare la distribuzione dei dati in esame.
Si basa sul concetto di entropia. L'AIC è un modo operativo di controbilanciare la bonta del modello confrontando la sua complessità in termini di numero di parametri e la sua capacità di adattarsi ai dati.

Nel caso generale, l'AIC è

$$AIC = 2k - 2\ln(L)$$

dove k è il numero di parametri del modello statistico, e L è la funzione di massima verosimiglianza.
Oltre a ciò, si presume che gli errori del modello siano normalmente e indipendentemente distribuiti. Supponiamo che n sia il numero di osservazioni e RSS definito come

$$RSS = \sum_{i=1}^{n} \hat{\epsilon}_i^2,$$

somma dei quadrati degli errori.
Quindi AIC si può scrivere

$$AIC = 2k + n[\ln(2\pi RSS/n) + 1]\,.$$

Più piccolo è il valore di AIC più il modello giudica buono il modello. Tale criterio penalizza i modelli con un elevato numero di parametri e premia quelli che hanno residui piccoli (RSS).

3 Un'applicazione dei GLM in campo biomedico: analisi della durata della degenza in unità di cure intensive cardiologiche (UCIC)

3.1 Descrizione dell'insieme di dati

In questo capitolo verrà presentata un'applicazione reale in cui la classe dei modelli lineari generalizzati viene utilizzata per modellare un insieme di dati di natura biomedica.
Va inoltre considerata che la durata della degenza è una delle variabili particolarmente osservate in ambito clinico di terapia intensiva soprattutto per motivi economici: infatti tali strutture complesse hanno costi sensibilmente più elevati per posto letto rispetto alle altre strutture ospedaliere (degenze, ambulatori).
Nella fattispecie abbiamo a disposizione un campione di 585 pazienti nuovi ricoverati (no-reingressi) nell'unità operativa di cure intensive cardiologiche (UCIC) dell'Azienda Ospedaliera di Padova (AOP) affetti da sindrome coronarica acuta (SCA): malattia cardiovascolare nota con severità di prognosi in funzione del dilatarsi del ritardo delle prime cure.

Su ogni unità sono state rilevate le seguenti classi di variabili:

- **antropometriche**: età, peso, altezza, body mass index,
- **fattori di rischio** : fumo, obesità, famigliarità, diabete, ipertensione, dilipidemie
- **anamnestiche**: presenza di defibrillatori impiantabili sottocutanei, pacemaker, patologie, presenza di bypass aortocorarici,…
- **bioumorali**: TnIng, PCR, Colesterolo (totale, LDL, HDL),…

La nostra popolazione è quindi composta da individui di sesso maschile o femminile, che codificheremo con M o F, ad ognuno dei quali è stata diagnosticata una SCA (STE o NSTE), alla quale è stata riconosciuta o no una serie di fattori di rischio (famigliarità, obesità, ipertensione, fumatore, diabete), sottoposta ad una serie di esami bioumorali e alla quale è stata redatta congrue note anamnestiche ed è stata ricoverata per un certo numero di giorni (GGDEG) presso l'UCIC non superiori ai 25 giorni.
Consideriamo quindi un dataset in cui la variabile dipendente che misura la durata della degenza (GGDEG) non possa assumere valori superiori a 25.
Tale scelta di porre un blocco ai giorni di degenza risponde a due ordini di fattori:

- pazienti con caratteristiche cliniche e gestionali definibili come particolari (componente esogena),
- complicanze incorse durante il ricovero (componente endogena).

Operando il filtro ai dati otteniamo un nuovo dataset di 577 pazienti (solo 8 esclusi dal totale del campione precedentemente considerato).

3.2 Analisi bivariata della durata della degenza dei pazienti con SCA in UCIC

Analizziamo ora la durata della degenza in funzione delle singole variabili: fattori di rischio, BMI, ICD, numero di vasi coronarici malati e valori ematici di Troponina I. Per vedere se sussiste qualche relazione tra le variabili elencate.

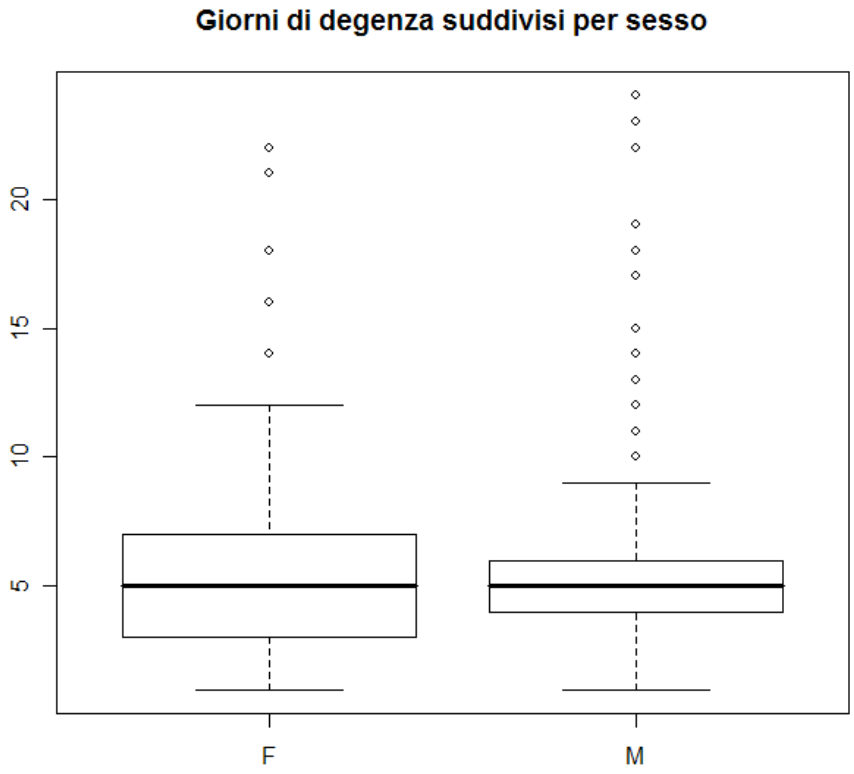


Fig.5

```
> wilcox.test(GGDEG ~ SESSO)
Wilcoxon rank sum test with continuity correction
data: GGDEG by SESSO
W = 36167, p-value = 0.2057
alternative hypothesis: true location shift is
not equal to 0
```

Dalla fig.5 e dal test di Wilcoxon non sembrano esserci significative differenze ($p>0.05$) nella durata della degenza tra maschi e femmine. In entrambe le distribuzioni vi sono comunque numerosi valori che si pongono “lontani” dai valori medi della distribuzione.

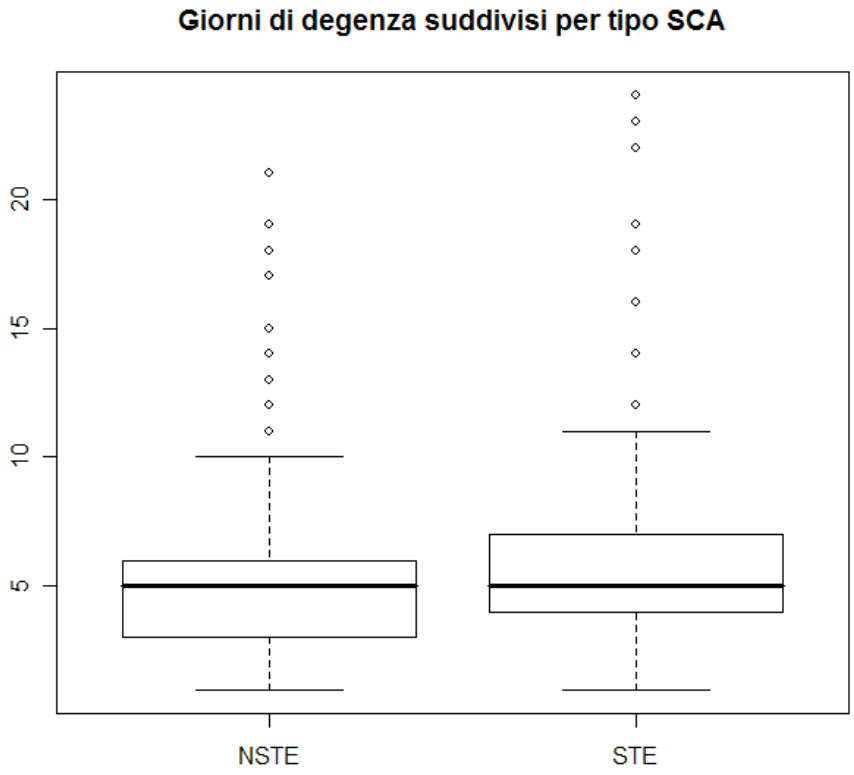


Fig.6

```
> wilcox.test(GGDEG ~ WHY)
Wilcoxon rank sum test with continuity correction
data: GGDEG by WHY
W = 36466, p-value = 0.1810
alternative hypothesis: true location shift is
not equal to 0
```

Dalla fig.6 e dal test di Wilcoxon, si conviene che non vi sia una significativa differenza (p>0.05) tra la durata della degenza dei pazienti con STE da quelli con NSTe SCA.

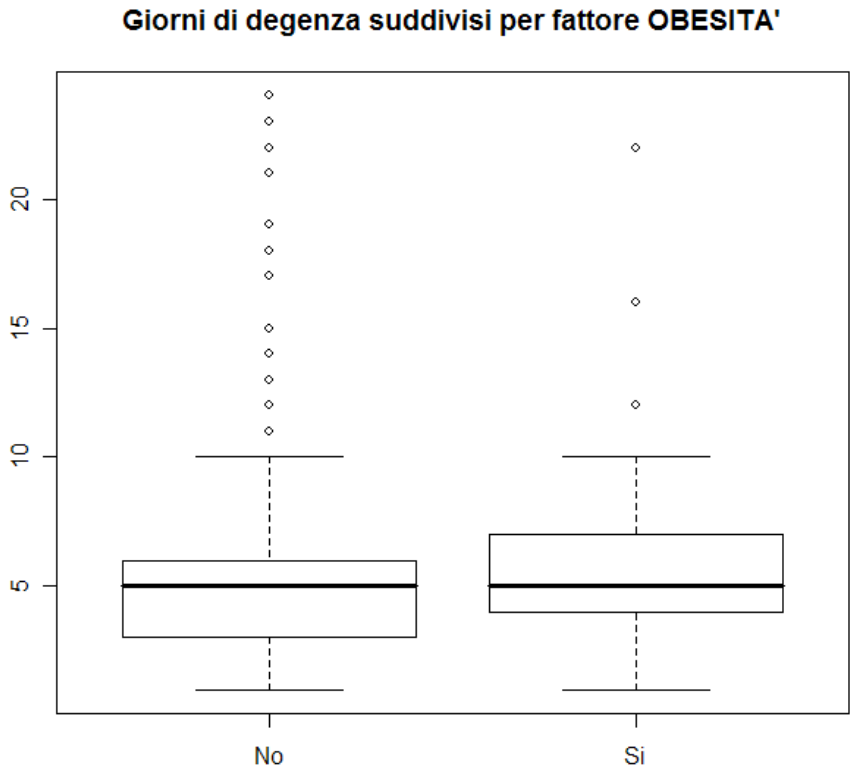


Fig.7

```
> wilcox.test(GGDEG ~ OBE)

Wilcoxon rank sum test with continuity correction

data: GGDEG by OBE

W = 20368, p-value = 0.0688
alternative hypothesis: true location shift is
not equal to 0
```

La fig.7 e il test di wilcoxon suggeriscono che vi sia una dubbia non significatività (p-limite=0.05) tra la durata della degenza dei pazienti obesi da quelli non obesi.
Nella pratica clinica comunque si nota una certa differenza per quanto riguarda la ripresa delle attività di vita quotidiane da parte dei soggetti con obesità patologica (BMI>30) che richiedono quindi un maggior carico assistenziale.

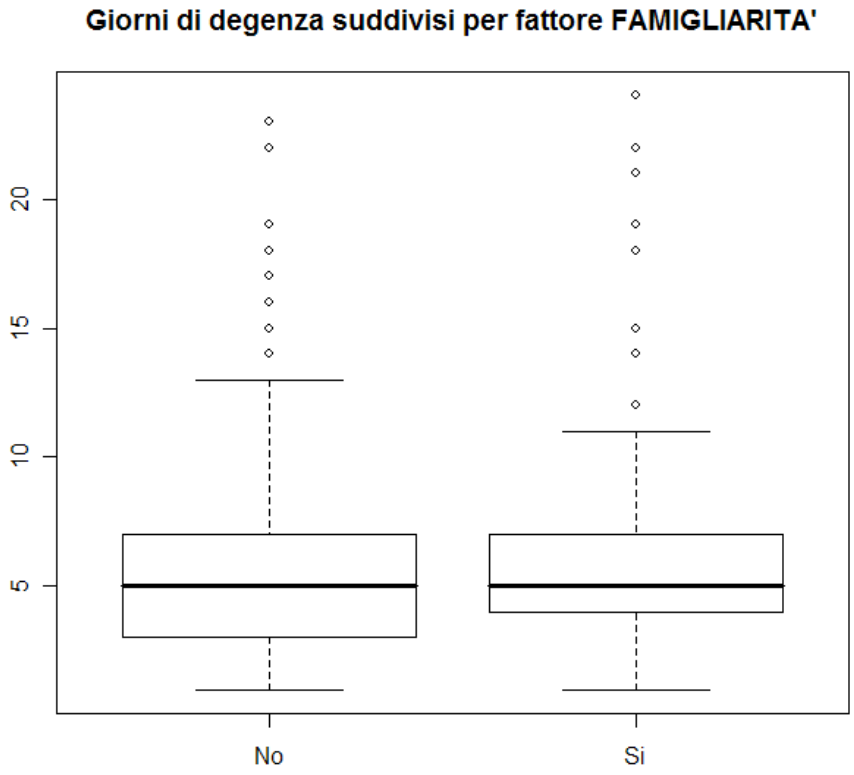


Fig.8

```
> wilcox.test(GGDEG ~ FAM)
Wilcoxon rank sum test with continuity correction
data: GGDEG by FAM
W = 38029, p-value = 0.8838
alternative hypothesis: true location shift is
not equal to 0
```

La fig.8 e il test sono suggestivi nel far considerare la famigliarità non un fattore significativo nel determinare la durata della degenza ($p>0.05$).

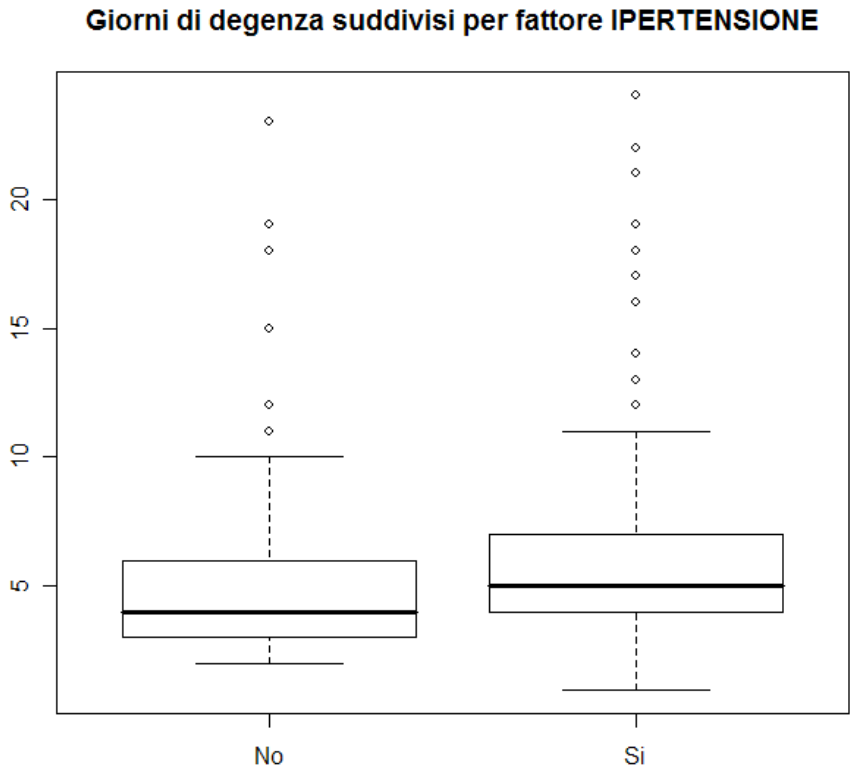


Fig.9

```
> wilcox.test(GGDEG ~ IXT)
Wilcoxon rank sum test with continuity correction
data: GGDEG by IXT
W = 30558, p-value = 0.0836
alternative hypothesis: true location shift is
not equal to 0
```

La fig.9 e il test di Wilcoxon suggeriscono che l'ipertensione non sia un fattore significativo nel determinare la durata della degenza ($p>0.05$).

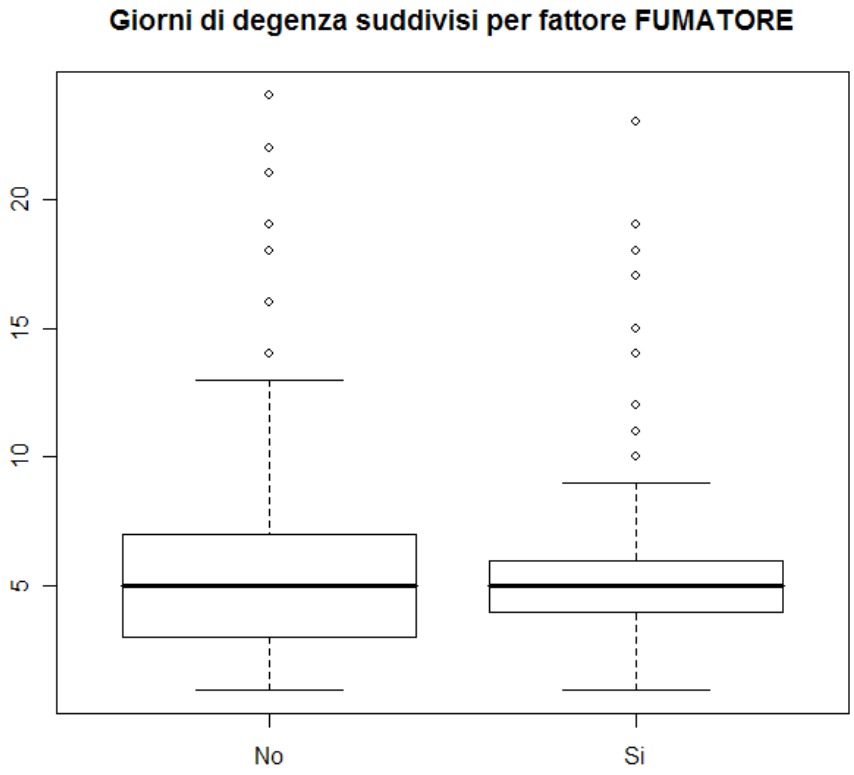


Fig.10

```
> wilcox.test(GGDEG ~ FUMO)
Wilcoxon rank sum test with continuity correction
data: GGDEG by FUMO
W = 42078.5, p-value = 0.2178
alternative hypothesis: true location shift is
not equal to 0
```

La fig.10 e il test di Wilcoxon convengono nel non considerare il fattore di rischio “fumatore” significativo nel determinare la durata della degenza (p>0.05).

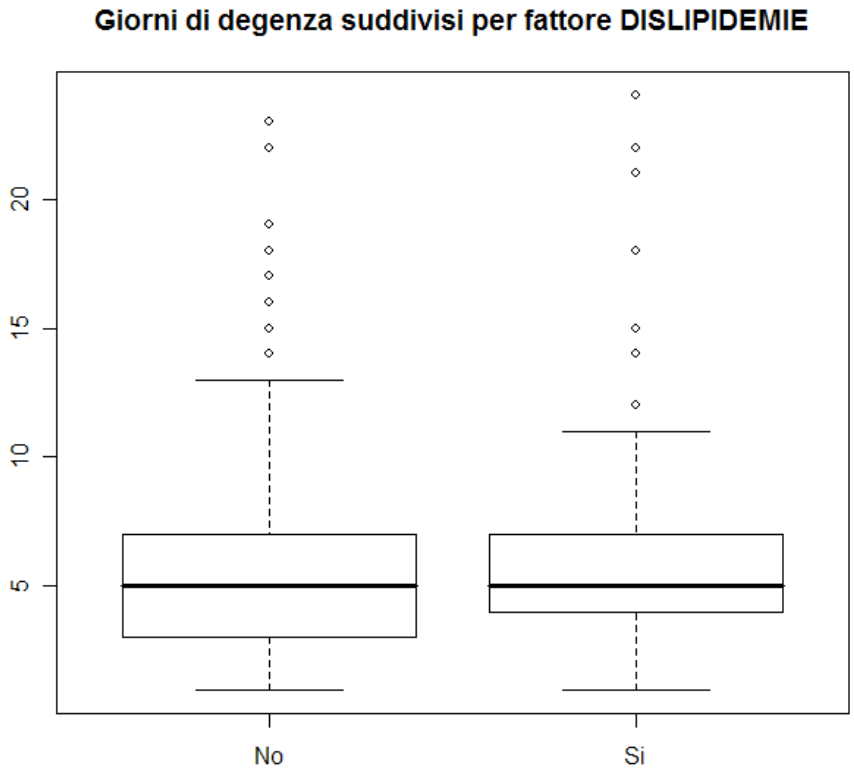


Fig.11

```
> wilcox.test(GGDEG ~ DIS)
Wilcoxon rank sum test with continuity correction
data: GGDEG by DIS
W = 39847.5, p-value = 0.6403
alternative hypothesis: true location shift is
not equal to 0
```

I boxplot (fig.11) e il Wilcoxon test suggeriscono che il fattore di rischio “dislipidemie” non sia significativo nel determinare la durata della degenza (p>0.05).

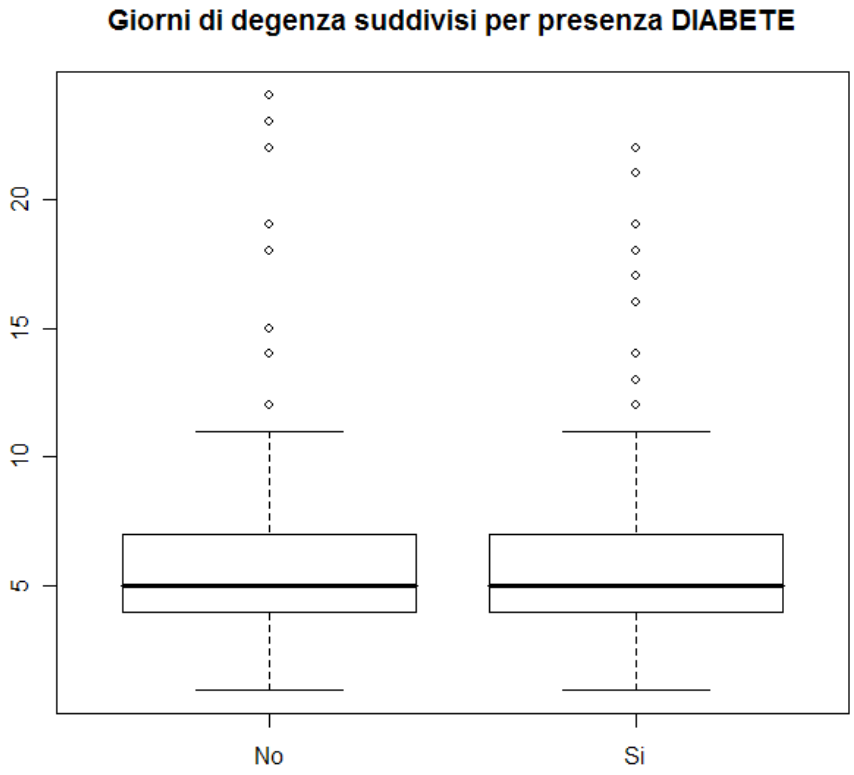


Fig.12

```
> wilcox.test(GGDEG ~ DIAB)
Wilcoxon rank sum test with continuity correction
data: GGDEG by DIAB
W = 26047.5, p-value = 0.3321
alternative hypothesis: true location shift is
not equal to 0
```

I boxplot (fig.12) e il test di Wilcoxon convengono che il fattore di rischio “diabete” non sia un fattore significativo (p>0.05) nel determinare la durata della degenza. Nonostante ciò, va considerato il fatto che la presenza della malattia diabetica nel paziente affetto da SCA comporta un aggravio del carico assistenziale dovuto all’esecuzione di stick glicemici seriati e controllo dell’infusione di insulina rapida con pompa a siringa come previsto dal protocollo glicemico vigente in reparto.

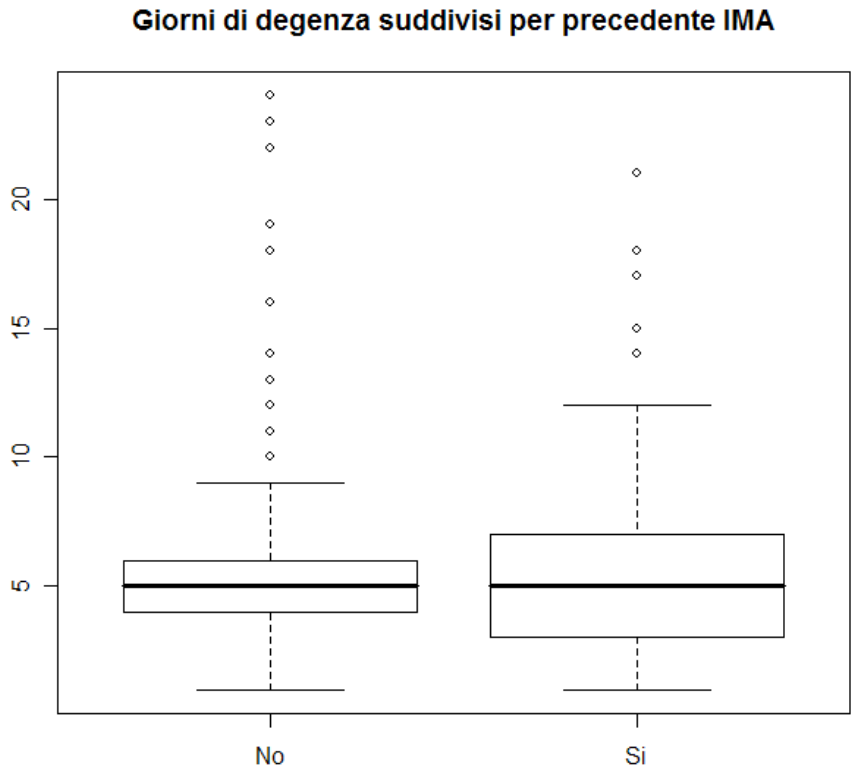


Fig.13

```
> wilcox.test(GGDEG ~ PreIMA)
Wilcoxon rank sum test with continuity correction
data: GGDEG by PreIMA
W = 32359, p-value = 0.1752
alternative hypothesis: true location shift is
not equal to 0
```

Il fattore pregresso infarto del miocardio da considerazioni sul grafico (fig.13) e sul test di Wilcoxon, non sembra essere un fattore significativo (p>0.05) nel determinare la durata della degenza..

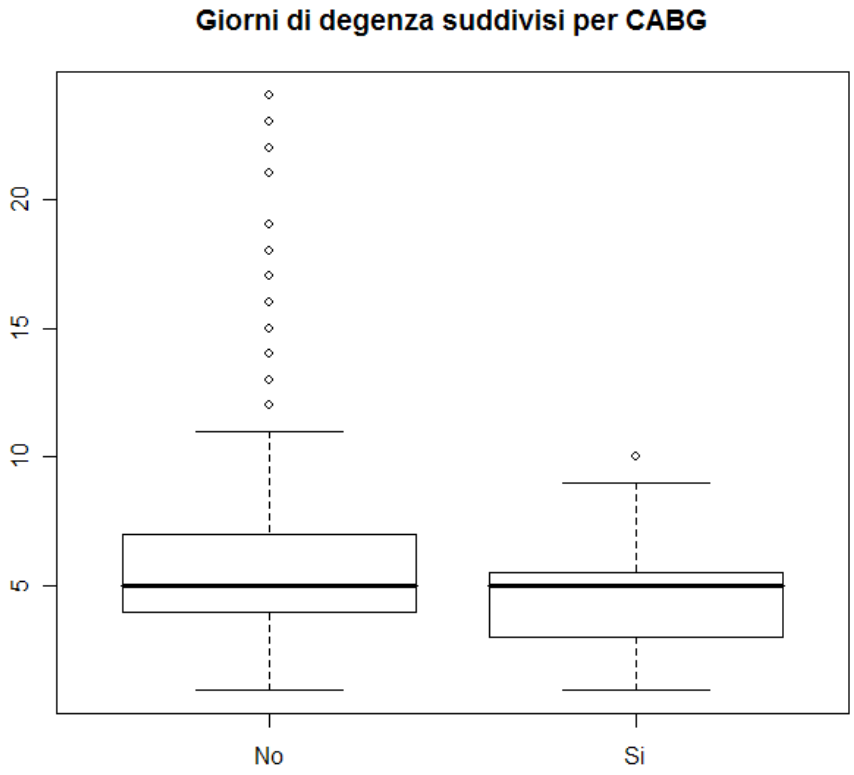


Fig.14

```
> wilcox.test(GGDEG ~ PreCABG)
Wilcoxon rank sum test with continuity correction
data: GGDEG by PreCABG
W = 11226, p-value = 0.3088
alternative hypothesis: true location shift is
not equal to 0
```

Il fattore pregresso bypass aortocoronarico da considerazioni sul grafico (fig.14) e sul rispettivo test di Wilcoxon, non sembra essere un fattore significativo (p>0.05) nel determinare la durata della degenza..

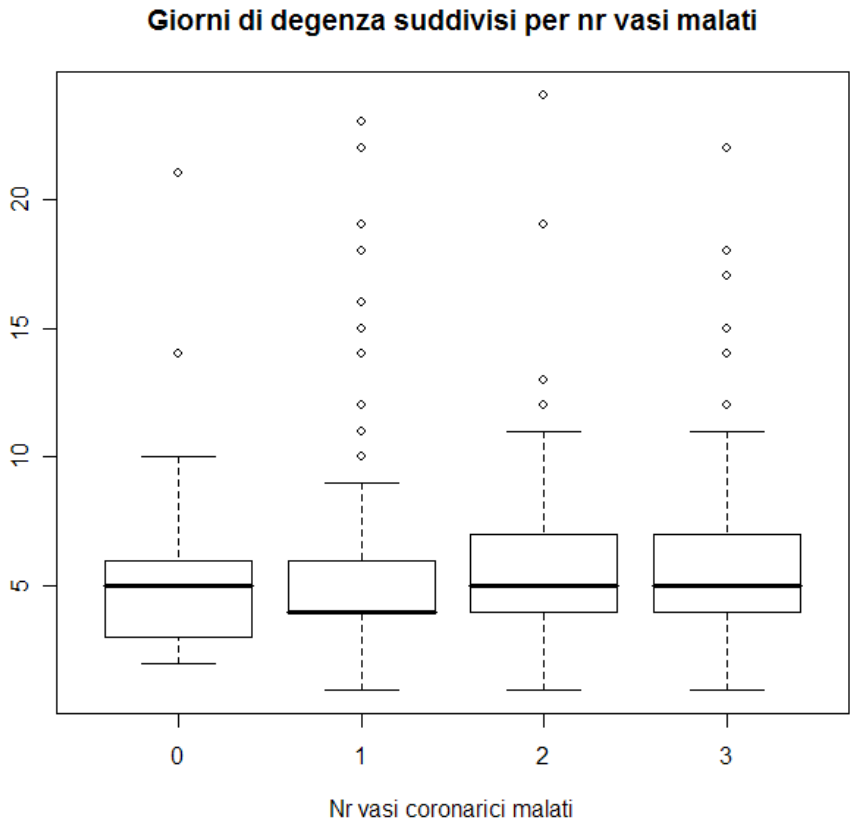


Fig.15

```
> kruskal.test(GGDEG~NrVasiMal)

        Kruskal-Wallis rank sum test

data:  GGDEG by NrVasiMal
Kruskal-Wallis chi-squared = 14.6281, df = 3, p-
value = 0.002164
```

Il test risulta significativo quindi risulta una differente distribuzione nella durata della degenza in base al numero dei vasi coronarici affetti da malattia aterosclerotica considerata critica dall'esame coronarografico. Probabilmente questa differenza è da attribuirsi alle indicazioni terapeutiche diverse a cui i vari gruppi di pazienti sono suddivisi: sono più probabili le indicazioni di by pass aortocoronarico nei pazienti affetti da malattia di due o più vasi rispetto a quelli con patologia critica focalizzata su un solo vaso coronario.

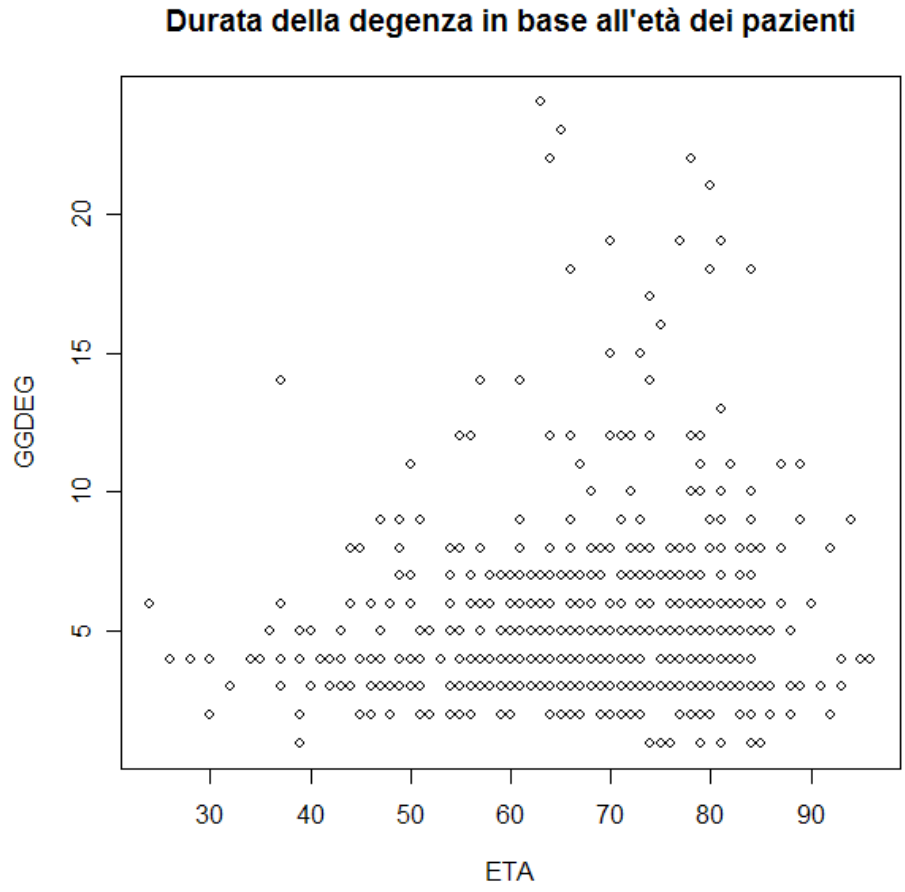


Fig.16

Per quanto riguarda l'analisi dei giorni di degenza in funzione dell'età del paziente ricoverato, risulta un lieve aumento della durata della degenza ma un chiaro aumento della variabilità del fenomeno durata della degenza in funzione dell'età. Tradotto in numeri:

```
> list(var(ETA,GGDEG),cor(ETA,GGDEG))

[1] 5.217377

[1] 0.1201915
```

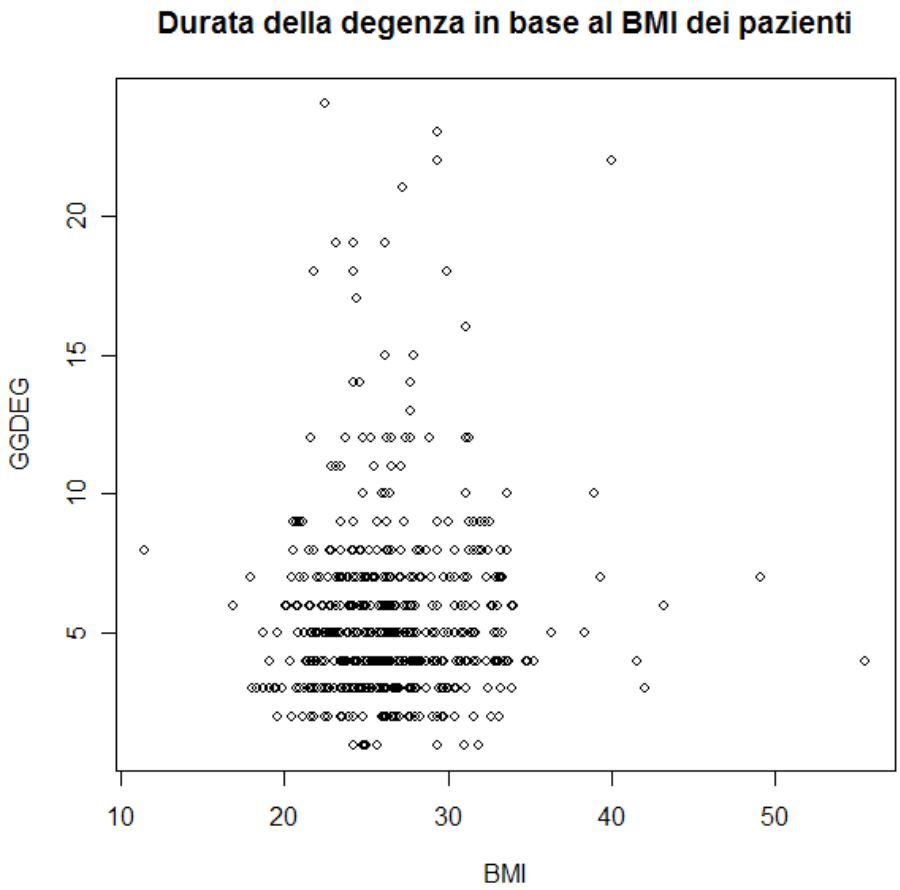


Fig.17

Difficile dare una interpretazione della durata della degenza in funzione del BMI (fig.17). La correlazione lineare esigua ci suggerisce che non vi sia correlazione lineare tra durata della degenza e i valori del BMI:

```
> list(var(BMI,GGDEG),cor(BMI,GGDEG))
[1] 0.6847428
[1] 0.04986358
```

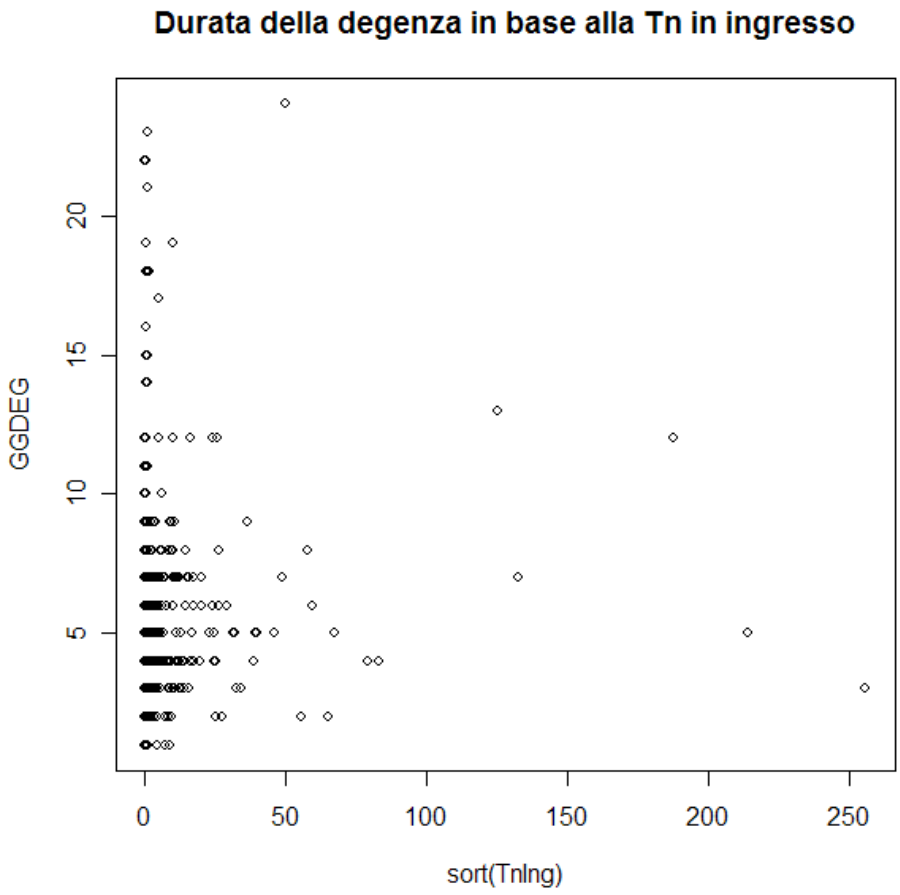


Fig.18

Dal grafico della distribuzione dei giorni degenza in base ai valori della troponina non vi sembra essere una evidente relazione. È però pensabile che vi sia una qualche relazione positiva tra le due variabili. In effetti esaminando la correlazione risulta in R:

```
> cor(TnIng,GGDEG)

[1] 0.1435665
```

modesta ma presente.

È invece evidente che gran parte dei pazienti viene ricoverata con valori di troponina I all'ingresso inferiori a 10 ng/ml.

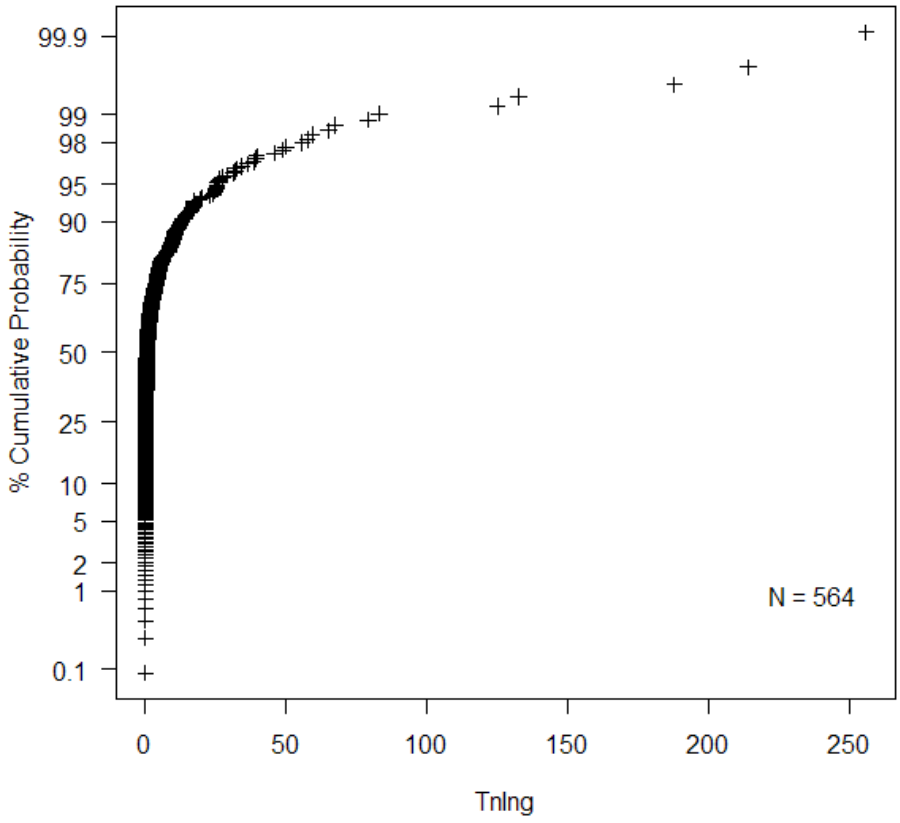


Fig.19

```
> length(TnIng[TnIng<10])

[1] 490

> 490/564

[1] 0.8687943
```

L'87% del campione risulta avere Troponina I all'ingresso dosabile a livello ematico inferiore ai 10 mcg/l.

Dall'analisi di varianza e correlazione lineare risulta che la Troponina ingresso da sola non può spiegare la durata della degenza e che vi è una modesta correlazione lineare tra la durata della degenza e i livelli di Troponina I rilevati dal prelievo ematico all'ingresso.

```
> list(var(TnIng,GGDEG),cor(TnIng,GGDEG))

[[1]]

[1] 9.500877

[]

[1] 0.1435665
```

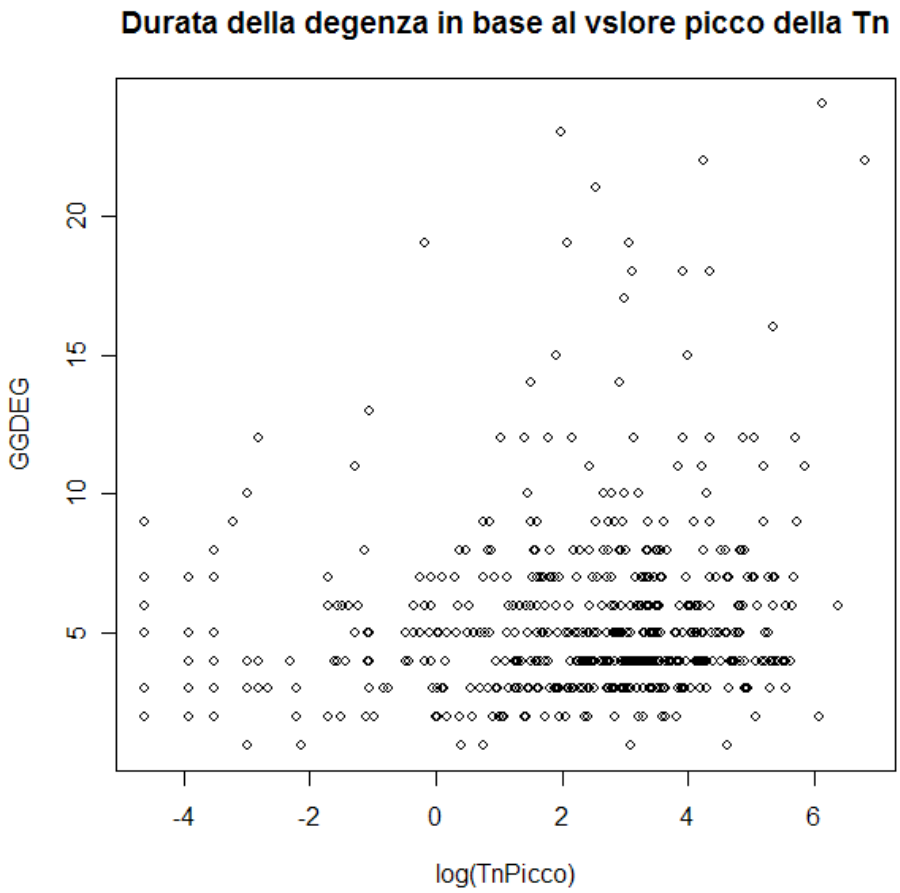


Fig.20

Dal grafico (fig.20) e dai dati:

```
> cor(TnPicco,GGDEG))
[1] 0.2284158
```

Risulta una modesta correlazione positiva tra i valori del picco della troponina dosata a livello ematico e la durata della degenza dei pazienti in UCIC.
Questa variabile (TnPicco) verrà successivamente non considerata nelle analisi multivariate perché considerata un predittore tardivo (disponibile il più delle volte dopo alcuni giorni dal momento del ricovero del paziente in UCIC).

3.3 Scelta del GLM

Nello specifico, considerando il limitato numero di eventi avversi e la standardizzazione delle cure, scopo dell'analisi è studiare la relazione che intercorre tra la quantità giorni di degenza in UCIC e le variabili esogene al contesto UCIC. In particolare, dato che il supporto della variabile risposta Y è il semi-asse positivo, e la natura dei dati (vedi figg.21-22):

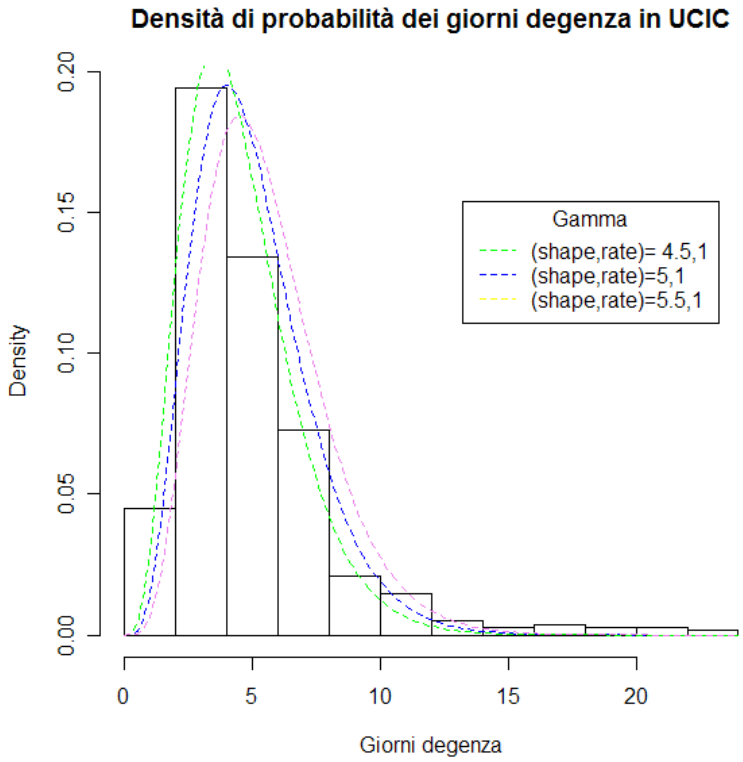


Fig.21

assumiamo che tale variabile abbia distribuzione di tipo gamma che ha densità

$$
\begin{aligned}
f(y;\nu,\lambda) &= \frac{\lambda^{\nu} y^{\nu-1} e^{-\lambda y}}{\Gamma(\nu)} \\
&= \exp\left\{\nu log(\lambda) + (\nu-1) log(y) - \lambda y - log(\Gamma(\nu))\right\} \\
&= \exp\left\{\nu\left(log(\lambda) + \left(\frac{\nu-1}{\nu}\right) log(y) - \frac{\lambda}{\nu} y - \frac{log(\Gamma(\nu))}{\nu}\right)\right\}
\end{aligned}
$$

Per la nostra applicazione iniziamo scegliendo una funzione di legame di tipo logaritmico ovvero quella funzione che esplicita la relazione tra predittore lineare e il valore atteso della distribuzione $g(\mu_i) = x_i^T ß$.

Per una verifica anche grafica di come i dati relativi ai giorni degenza dei pazienti ricoverati in UCIC seguano una distribuzione di tipo gamma piuttosto che una normale di seguito sono riportati i grafici quantile-quantile del logaritmo dei giorni di degenza rispettivamente con una distribuzione normale e una gamma (fig.)

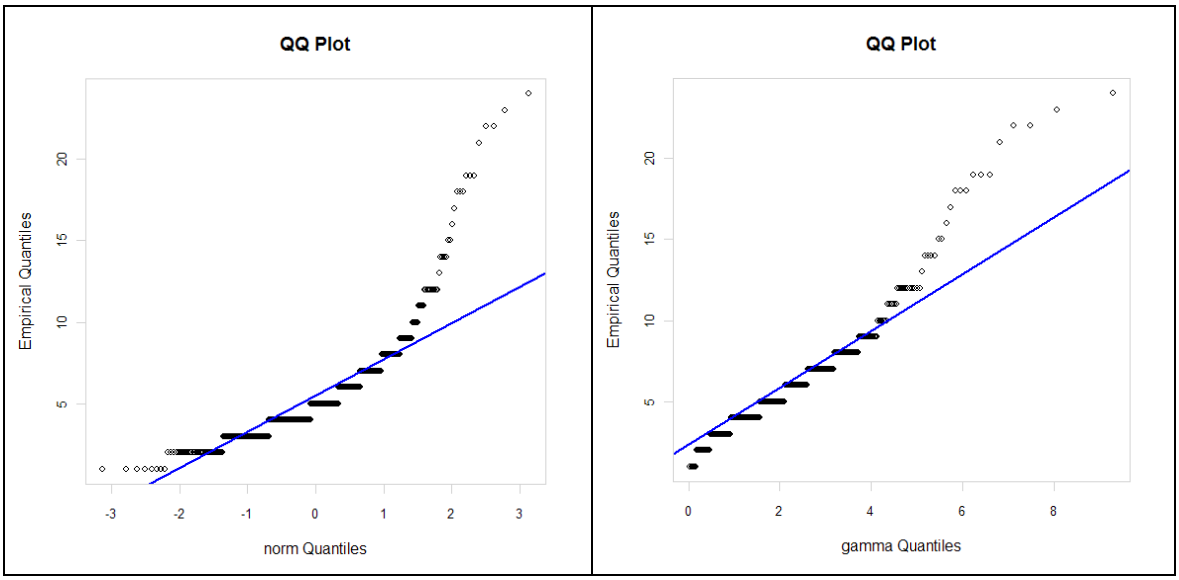


Fig.22

Ponendo θ = - λ/v abbiamo che la densità della v.a Gamma può essere riscritta nel modo seguente

$$\begin{aligned} f(y;\nu,\lambda) &= \exp\left\{\nu\left(log(-\theta\nu)+\left(\frac{\nu-1}{\nu}\right)log(y)+\theta y-\frac{log(\Gamma(\nu))}{\nu}\right)\right\} \\ &= \exp\{\nu(\theta y+log(-\theta\nu))+(\nu-1)log(y)-log(\Gamma(\nu))\} \\ &= \exp\{\nu(\theta y+log(-\theta)+log(\nu))+(\nu-1)log(y)-log(\Gamma(\nu))\} \\ &= \exp\{\nu(\theta y+log(-\theta))+\nu log(\nu)+\nu log(y)-log(y)-log(\Gamma(\nu))\} \\ &= \exp\{\nu(\theta y+log(-\theta))+\nu log(\nu y)-log(y)-log(\Gamma(\nu))\} \end{aligned}$$

per cui ponendo

$$\begin{aligned} \psi &= \nu^{-1} \\ \omega &= 1 \\ b(\theta) &= -log(-\theta) \\ c(y;\psi) &= \nu log(\nu y)-log(y)-log(\Gamma(\nu)) \end{aligned}$$

risulta che il nostro modello può essere scritto nel modo seguente

$$Y_i \sim EF(-log(-\theta_i),\psi)$$

La relazione che in base al nostro modello lega il valore atteso della quantità dei giorni di degenza dell'i-esimo soggetto rispetto alle covariate è dunque la seguente:

log (E [Yi]) = $ß_1 + ß_2x_{i2} + ß_3x_{i3} + ß_4x_{i4} + ß_5x_{i5} + ß_6x_{i6} + \ldots + ß_{20}x_{i20} + ß_{21}x_{i21}$

dove $ß_1$ rappresenta l'intercetta della relazione tra il logaritmo del valore atteso di Y e le covariate quando sono nulle,

x_{i2} indica il sesso del paziente (0 = femmina, 1 = maschio),

x_{i3} indica il valore del body mass index (Kg/m^2),
x_{i4} indica la presenza del fattore di rischio obesità (0 = non obeso, 1 = obeso)
x_{i5} indica la presenza del fattore di rischio famigliarità (0 = assenza, 1 = presenza)
x_{i6} indica la presenza del fattore di rischio ipertensione (0 = assenza, 1 = presenza)
x_{i7} indica la presenza del fattore di rischio fumatore (0 = assenza, 1 = presenza)
x_{i8} indica la presenza di dislipidemie (0 = assenza, 1 = presenza)
x_{i9} indica la presenza della patologia metabolica diabete (0 = assenza, 1 = presenza)
x_{i10} indica la presenza di insufficienza renale cronica (IRC) (0 =assenza, 1 = presenza)
x_{i11} indica la presenza di broncopneumopatie croniche ostruttive (BPCO) (0 = assenza, 1 = presenza)
x_{i12} indica la diagnosi d'ingresso (WHY) (defibrillatore cardiaco impiantabile) (0 = assenza, 1 = presenza)
x_{i13} indica un precedente infarto del miocardio (PreIMA) (0 = assenza, 1 = presenza)
x_{i14} indica la presenza di precedenti attacchi di angina pectoris (PreAng) (0 = assenza, 1 = presenza)
x_{i15} indica precedenti procedure interventistiche coronariche (PrePTCA) (0 = assenza, 1 = presenza)
x_{i16} indica precedenti bypass aortocoronarici (PreCABG) (0 = assenza, 1 = presenza)
x_{i17} indica la presenza di ICD (defibrillatore cardiaco impiantabile) (0 = assenza, 1 = presenza)
x_{i18} indica la presenza di angina precedente alla SCA (0 = assenza, 1 = presenza)
x_{i19} indica precedente ictus cerebrale (PreStroke) (0 = assenza, 1 = presenza)
x_{i20} indica il numero dei vasi coronarici affetti da placche aterosclerotiche critiche
x_{i21} indica i valore di Troponina I (ng/mL) all'ingresso in UCIC.

Come primo obiettivo stimiamo i coefficienti ß. Tali parametri, infatti, definiscono la relazione che intercorre tra il numero di giorni di degenza e le covariate. Tra i vari pacchetti statistici che permettono di stimare un modello lineare generalizzato abbiamo scelto di utilizzare (www.R-project.org). In particolare avendo chiamato (dati) la matrice le cui colonne rappresentano le variabili descritte in precedenza, possiamo stimare il modello richiesto attraverso la sintassi

```
glm(GGDEG~ETA+SESSO+BMI+OBE+FAM+IXT+FUMO+DIS+DIAB
+IRC+BPCO+WHY+PreIMA+PreAng+PrePTCA+PreCABG+ICD+A
ngPreIMA+PreSTROKE+NrVasiMal+TnIng,
family=gamma(link=log),data=dati)
```

la formula

```
GGDEG~ETA+SESSO+BMI+OBE+FAM+IXT+FUMO+DIS+DIAB+IRC
+BPCO+WHY+PreIMA+PreAng+PrePTCA+PreCABG+ICD+AngPr
eIMA+PreSTROKE+NrVasiMal+TnIng
```

indica che si vuole ottenere delle informazioni sul numero dei giorni di degenza(GGDEG) in funzione all'età(ETA), al sesso(SESSO), al body mass index(BMI), ai fattori di rischio (obesità(OBE), famigliarità(FAM), ipertensione(IXT), fumo(FUMO), dislipidemie(DIS), diabete (DIAB)), … e delle rimanenti covariiate del modello sopra esplicitato (vedi pagina precedente).

Un estratto dell'output di R è il seguente:

```
> summary(fit_gamma_log_1)

Call:
glm(formula = GGDEG ~ ETA + SESSO + BMI + OBE + FAM + IXT +
FUMO + DIS + DIAB + IRC + BPCO + WHY + PreIMA + PreAng +
PrePTCA + PreCABG + ICD + AngPreIMA + PreSTROKE + NrVasiMal
+ TnIng, family = Gamma(link = log), data = dati)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.221663   0.290811   4.201 3.24e-05 ***
ETA          0.006335   0.002204   2.874  0.00426 **
SESSOM      -0.086567   0.065318  -1.325  0.18579
BMI          0.002030   0.009356   0.217  0.82830
OBESi       -0.037174   0.098908  -0.376  0.70722
FAMSi       -0.005065   0.053738  -0.094  0.92495
IXTSi        0.001906   0.059344   0.032  0.97440
FUMOSi      -0.017717   0.057596  -0.308  0.75853
DISSi       -0.054333   0.056139  -0.968  0.33368
DIABSi       0.072055   0.070848   1.017  0.30972
IRCSi        0.104732   0.179107   0.585  0.55903
BPCOSi       0.042596   0.121938   0.349  0.72702
WHYSTE       0.031717   0.055942   0.567  0.57104
PreIMASi     0.011042   0.079557   0.139  0.88968
PreAngSi     0.018525   0.066715   0.278  0.78140
PrePTCASi    0.016651   0.092396   0.180  0.85708
PreCABGSi   -0.126006   0.124518  -1.012  0.31214
ICDSi       -0.356974   0.563001  -0.634  0.52639
AngPreIMASi -0.030633   0.061160  -0.501  0.61673
PreSTROKESi  0.057058   0.158169   0.361  0.71847
NrVasiMal    0.049714   0.030169   1.648  0.10013
TnIng        0.003630   0.001236   2.936  0.00350 **
---
```

Sulla colonna Estimate sono riportati i valori delle stime dei coeffcienti di regressione β.
I dati relativi all'intercetta sono riferiti alla popolazione di riferimento che è costituita dalle donne, prive di fattori di rischio, con SCA NSTE e di valori nulli di troponina I ematica. Il valore 1.222 quindi indica la stima del logaritmo del numero medio di giorni di degenza in questa popolazione quando le rimanenti variabili sono nulle (dato che dal punto di vista clinico non ha nessuna rilevanza).

Il valore stimato del regressore riferito all'età è 0.006 ed indica che se un paziente assume valore n per l'età essa apporterà un incremento pari a 0.006*n al calcolo del log naturale della stima durata della degenza.
Analogo discorso vale per la stima dei rimanenti repressori.
L'output di R riporta anche i risultati del test di ipotesi $H_0 : ß_i = 0$ contro $H_1 : ß_i \neq 0$ sui coefficienti di regressione. Tali risultati sono riportati sulla colonna chiamata t value e nell'ultima colonna vengono riportati i valori p.
Ricordiamo che se il valore p è piccolo tenderemo ad accettare l'ipotesi H_1 rifiutando quindi l'ipotesi nulla $H_0 : ß_i = 0$.

Analizziamo ora la devianza del modello che se piccola è indice di un buon adattamento ai dati. Ricordiamo che la devianza normalizzata è data

$$D_N = 2\sum_{j=1}^{n}\left\{logf\left(y_j;\tilde{\eta}_j,\hat{\psi}\right) - logf\left(y_j;\hat{\eta}_j,\hat{\psi}\right)\right\}$$

mentre la devianza è data da

$$D = \hat{\psi}\left\{2\sum_{j=1}^{n}\left[logf\left(y_j;\tilde{\eta}_j,\hat{\psi}\right) - logf\left(y_j;\hat{\eta}_j,\hat{\psi}\right)\right]\right\}$$

Nel nostro caso abbiamo che la devianza è 107,6

Ora si procede alla realizzazione di altri tre modelli lineari generalizzati sullo stesso dataset per valutare quale sia il GLM che maggiormente si adatta alla nostra distribuzione:

- famiglia gamma con funzione legame identità,
- famiglia gaussiana con funzione legame logaritmico,
- famiglia gaussiana con funzione legame identità.

Per confrontare tra loro i vari modelli realizzati. A tal fine verranno fatte alcune considerazioni basate sull'analisi dei grafici riportati nelle fig.re 23-27 e confrontando i corrispettivi AIC (Akaike information criterion).

Grafici di diagnostica per il GLM gamma e legame logaritmico

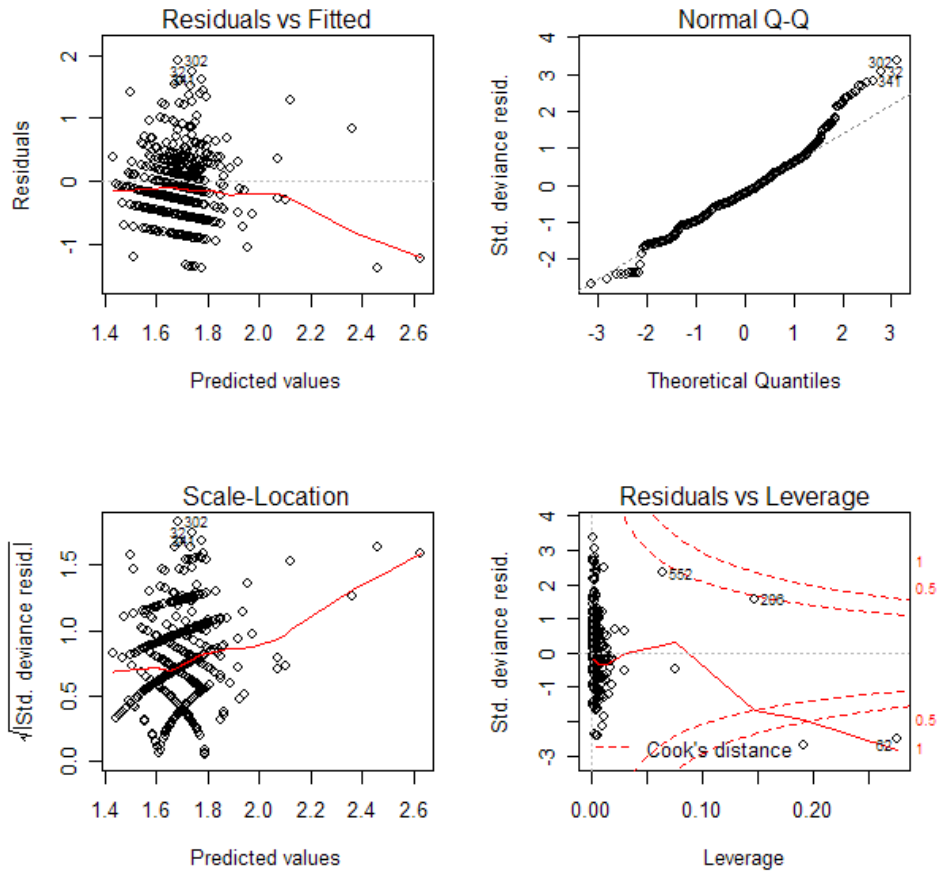


Fig.23

Per quanto riguarda il primo modello (gamma con legame logaritmico fig.23) risulta qualche problematica per i valori estremi della distribuzione e la maggior parte dei residui risulta di trascurabile entità (compresi tra ±1).

Grafici di diagnostica per il GLM gamma e legame identità

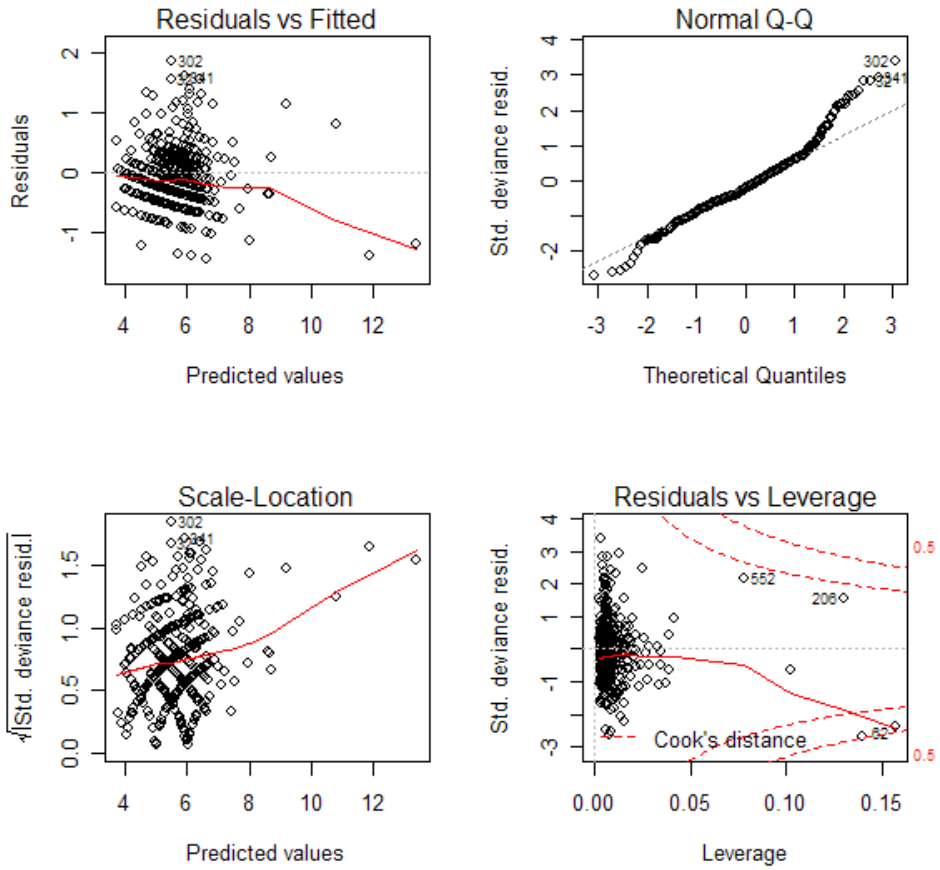


Fig.24

Valgono anche per questo modello (gamma con legame canonico fig.24) le considerazioni fatte a quello precedente (lo stesso modello ma con funzione di legame logaritmico). Si nota un sensibile miglioramento per quel che riguarda il grafico dei residui su punti leva.

Grafici di diagnostica per il GLM normale e legame canonico

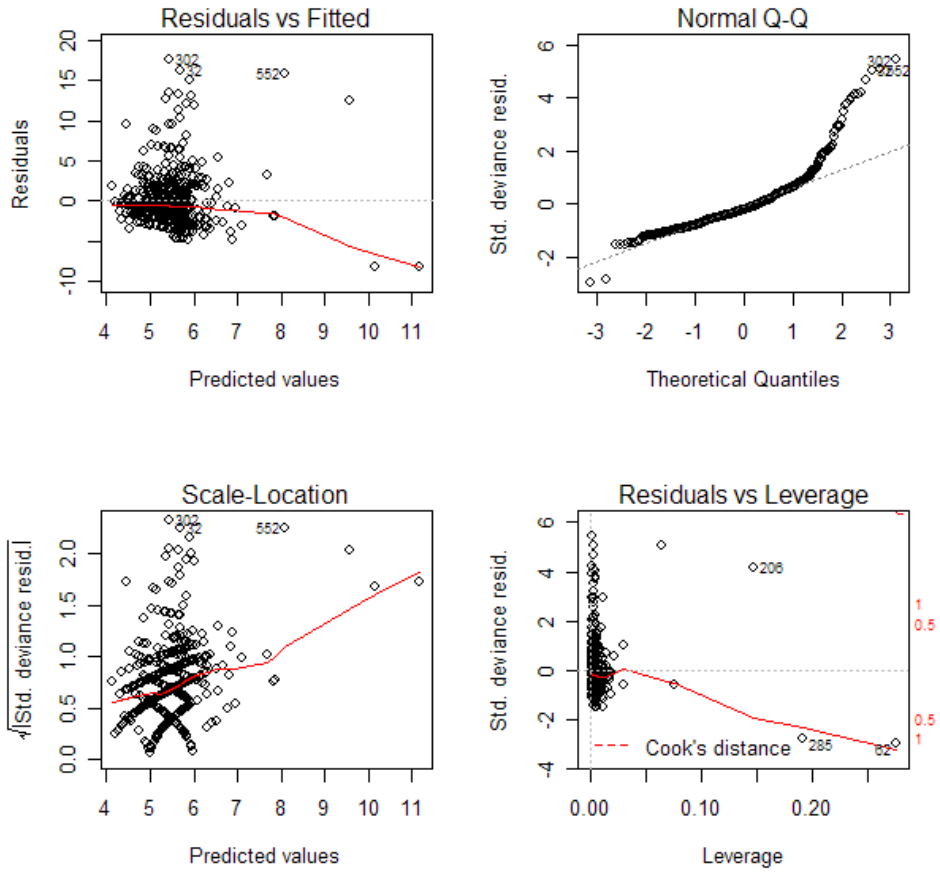


Fig.25

Per quanto riguarda questo modello (normale con legame la funzione identità fig.25) risulta qualche differenza rispetto ai precedenti per avere sicuramente una peggiore capacità di adattarsi ai valori grandi di durata della degenza (vedi fig. grafico Normal Q-Q) e per avere una considerevole quantità di residui di entità inacettabile (le stime risultano per molti casi “lontane” dai veri dati anche per quantità maggiori di 5 giorni).

Grafici di diagnostica per il GLM normale e funzione di legame logaritmica

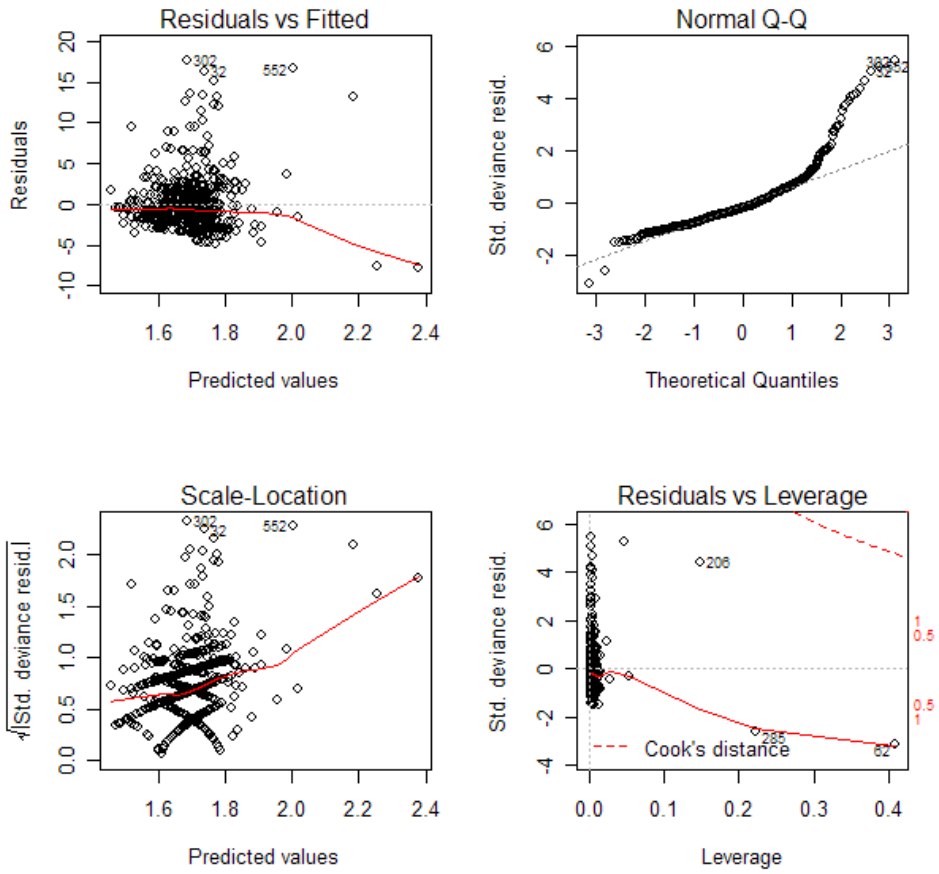


Fig.26

Per quanto riguarda questo modello (normale con legame la funzione logaritmica fig.26) vi sono considerazioni analoghe a quello normale con legame la funzione identità: residui e Q-Q plot non soddisfacenti.

Per maggiore chiarezza vediamo nel dettaglio dei residui (fig. 27) di come i quattro modelli lineari generalizzati differiscono tra loro.

Grafici dei residui dei quattro modelli considerati

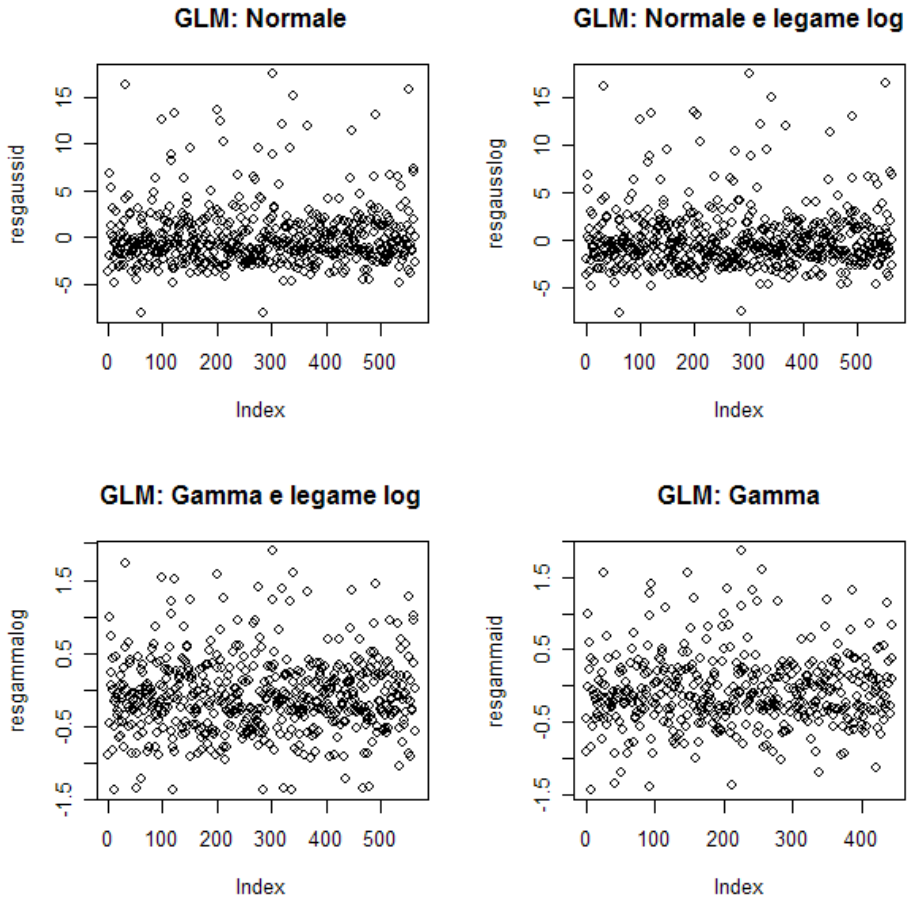


Fig.27

Senza farci confondere del cambio di scala operato nei quattro grafici, dall'analisi dei residui risulta che per la stima della durata della degenza sono preferibili i modelli con famiglia di distribuzione gamma rispetto a quelli normali e tra i due modelli con distribuzione gamma quello con funzione di legame identità non dovuta a sostanziali differenze nei residui ma a una sua maggior semplicità.

Verificando ora anche l'AIC dei quattro modelli testati ed abbiamo:

- 2650 per il modello gamma con legame logaritmico,
- 2649 per il modello gamma con legame identità,
- 2930 per il modello normale con legame identità,
- 2932 per il modello normale con legame logaritmico.

Anche il criterio di Akaike da una chiara preferenza al modello lineare generalizzato gamma con funzione di legame identità.

3.4 Conclusioni

Dalle analisi inferenziali operate con i GLM delle famiglie (gamma e normale) sul dataset dei dati relativi ai ricoveri dei pazienti affetti da SCA in UCIC dell'Azienda Ospedaliera di Padova risulta che un

modello con distribuzione gamma e legame canonico sia in base a considerazioni sui residui, sui grafici Q-Q e sui risultati dell'AIC.
Inoltre da semplici computazioni di sotto riportate si ottiene:

```
> resgammaid2<-residuals(fitgammaid2)
> scarti<-abs(resgammaid2)
> length(scarti)
[1] 564
> length(scarti[scarti<1])
[1] 534
> 534/564
[1] 0.9468085
```

che il modello considerato (GLM: gamma con legame canonico e con i dati in nostro possesso), per circa il 95% commette errori di stima inferiori ad 1 giorno.
Qui di seguito viene esplicitato tale modello dove risulta che la durata della degenza dei pazienti ricoverati in UCIC per SCA risulta essere espressa come somma dei seguenti valori:

*3,24 + 0,031*età(in anni)+0,03*(Troponina I in ingresso)*

```
> summary(fit_gamma_id_19)
Call:
glm(formula = GGDEG ~ ETA + TnIng, family = Gamma(link =
identity),
    data = dati)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3717  -0.3975  -0.1295   0.1941   1.8924
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.238041   0.618632   5.234 2.35e-07 ***
ETA         0.030873   0.009303   3.319 0.000963 ***
TnIng       0.028451   0.010779   2.639 0.008534 **
---
    Null deviance: 157.02  on 563  degrees of freedom
Residual deviance: 150.08  on 561  degrees of freedom
AIC: 2649.1

Number of Fisher Scoring iterations: 6
```

3.5 Discussione

Il modello lineare generalizzato della famiglia gamma e con funzione di legame identità risulta dal punto di vista pratico per il computo della stima della degenza dei pazienti con SCA ricoverati in UCIC dell'Azienda Ospedaliera di Padova soddisfacente sia come valori predittivi che nella sua composizione analitica.
Risultano abbastanza intuitive soprattutto per i clinici cogliere le implicazioni tra i valori di citomiolisi cardiaca (troponina I) e l'età come fattori che possono influenzare positivamente (farla aumentare) la durata della degenza.
Utilità del modello è quella di poter analizzare la congruità delle durate della degenza dei pazienti in UCIC dove i costi per posto

letto giornalieri sono considerevoli, come i rischi di infezione e per una più mirata progettualità del percorso assistenziale del paziente in UCIC.

4 Analisi della sopravvivenza

4.1 Introduzione

L'analsi della sopravvivenza esamina e modella il tempo in cui occorre un evento. Generalmente l'evento in esame è la morte, da questo deriva il nome di "analisi della sopravvivenza" e di molta della terminologia derivata, anche se l'ambito di applicazione dell'analisi della sopravvivenza è molto più ampio. Sostanzialmente gli stessi metodi sono impiegati in una varietà di discipline come ad esempio la sociologia, la storia, la fisica, … per analisi degli eventi storici, dei divorzi, degli svezzamenti, del tempo di dimezzamento di un isotopo radioattivo, …
Il focus dell'analisi della sopravvivenza è sulla distribuzione dei tempi di sopravvivenza.
Anche se sono ben noti i metodi per la stima incondizionata delle distribuzioni della sopravvivenza, la maggior parte dei modelli di sopravvivenza si interessano del rapporto tra sopravvivenza e uno o più variabili chiamate predittori, di solito definite in ambito statistico come covariate.

4.2 Le rappresentazioni grafiche: il metodo di Kaplan-Meier

Nell'analisi dei tempi di sopravvivenza, si usano diverse terminologie e notazioni. In particolare il rischio di morte viene anche chiamato tasso di rischio, tasso di fallimento, forza di mortalità e tasso istantaneo di mortalità; il fallimento (in inglese: failure) implica qualcosa che si rompe o si deteriora, mentre la forza di mortalità e il tasso di mortalità implicano qualcuno che muore. La notazione usata varia da caso a caso e sono comuni le seguenti: h(x) o H(x) (dall'inglese hazard, "pericolo"). La notazione usata quindi sarà h(t). In molti studi di sopravvivenza, il numero di pazienti arruolati potrebbe essere troppo piccolo perché le curve di sopravvivenza possano essere costruite impiegando i metodi usati per calcolare S_t nelle tavole di mortalità convenzionali. Kaplan e Meier hanno suggerito un metodo (detto anche "metodo del prodotto-limite") che fa un uso efficiente dei dati limitati disponibili in tali studi, e che può anche tener conto dei tempi di sopravvivenza troncati o censurati (censored).
Tale metodo si basa sull'idea di considerare i dati raggruppati in un numero di intervalli temporali relativamente piccoli, sulla base degli eventi verificatisi. In questo modo l'intervallo di tempo non è fisso, ma dipende da quando si realizzano gli eventi di interesse, in genere la morte. Se per esempio, a partire dal tempo iniziale t_0 si verifica il primo decesso dopo 3 mesi e il secondo decesso dopo 8 mesi, gli intervalli considerati saranno t_1=3 e t_2=8.
Il metodo di Kaplan-Meier per la stima della probabilità di sopravvivenza, suddivide i tempi in completi quando si verifica l'evento morte e troncati o censurati quando il tempo di

sopravvivenza è incompleto. Per il calcolo si utilizzano i seguenti dati:

- t: i tempi completi e censurati in ordine crescente,
- n_t: i soggetti vivi e non censurati (a rischio di morte) al momento t
- d_t: il numero di decessi osservati al tempo t
- n_t-d_t: la differenza dei precedenti termini
- p_t: rappresenta la probabilità di condizionata di sopravvivenza al tempo t ed è $(n_t-d_t)/n_t$
- S(t): rappresenta la probabilità non condizionata di sopravvivenza al tempo t e viene calcolata moltiplicando fra loro tutti i valori nella colonna p_t includendo p_t.

In genere, le curve di sopravvivenza riportano i valori di S(t) come una curva a gradini, in cui a ogni decesso osservato (o più se contemporanei) corrisponde un gradino verticale.
La stima della varianza della proporzione che sopravvive al tempo t può essere calcolata usando la seguente formula:

$$Var(S(t)) = S^2(t)*\sum(d_i/n_i(n_i-d_i))$$

e l'errore standard la radice quadrata di questa quantità.
Due questioni dovrebbero essere considerate quando si interpretano le curve di sopravvivenza ottenute da un insieme di dati:
la precisione della curva peggiora all'aumentare di t (e dunque al diminuire di n) anche se il sottostante rischio appare costante;
la curva di sopravvivenza osservata non può dare alcuna informazione sull'andamento della sopravvivenza per tempi più lunghi di quelli considerati nello studio.

4.2.1 Il test del log-rank

Si supponga di voler confrontare gli effetti di due trattamenti, A e B, osservando le esperienze di mortalità di due gruppi di pazienti. L'ipotesi nulla è che i due trattamenti sono uguali. Ciò significa che si potranno calcolare gli aattesi per un gruppo (per esempio A) calcolati mettendo insieme i due gruppi, e confrontarli con gli osservati, rapportando la differenza all'errore standard degli eventi osservati. La logica del log-rank test è che, assunta per vera l'ipotesi nulla, a parità di tempo t di osservazione, la proporzione dei morti fra i pazienti che ricevono il trattamento A dovrebbe essere uguale a quella attesa, calcolata mettendo insieme i pazienti dei due gruppi.
È opportuno costruire una tabella per ciascun tempo t, in cui è stato osservato almento un decesso.
Il numero di morti attese nel gruppo con trattamento A è, uguale al prodotto dei marginali diviso il totale delle osservazioni:

$$E(d_t)=d*n_t/n$$

con varianza:

$$Var(dt)=[d(n-d)n_1n_2]/[(n-1)n^2]$$

Il test del log-rank implica il confronto fra il numero osservato di morti, nel gruppo con il trattamento A, e il numero atteso, assunta vera l'ipotesi nulla. Questo procedimento viene ripetuto per ogni tabella 2 per 2 corrispondente a ciascun valore di t, in cui vengono osservate una o più morti. Si procede a sommare il totale dei casi osservati e attesi e a confrontare i totali.
Talvolta questa espressione viene elvata al quadrato e e viene riferita a una distribuzione del chi-quadrato con 1 grado di libertà, ma tale modo di procedere ha lo svantaggio che senza l'ispezione dei dati, non è ovvio se più o meno morti sono osservate nel gruppo che riceve il trattamento A, rispetto a quanto atteso sotto l'ipotesi nulla.
Una formula approssimata per il chi-quadrato più semplice da calcolare, e che può essere estesa al caso di tre o più trattamenti, deriva dalla formulazione generale del test:

$$\chi^2=\sum(O-E)^2/E$$

Il log-rank test viene considerato un test non parametrico perché non assume alcuna distribuzione della funzione di sopravvivenza.

4.3 Alcune considerazioni sui modelli per l'analisi della sopravvivenza

Supponiamo T rappresenti il tempo di sopravvivenza. Consideriamo T come una variabile casuale con funzione di distribuzione cumulativa

$P(t) = Pr (T \leq t)$

e la funzione di densità di probabilità

$p(t) = dP(t) / dt.$

La funzione di sopravvivenza S(t) è il complemento della funzione di distribuzione P(t),

$S (t) = Pr (T> t) = 1 - P (t).$

Una quarta rappresentazione della distribuzione dei tempi di sopravvivenza è la funzione di rischio, che valuta il rischio di scomparsa istantanea (morte) al tempo t, a condizione che la sopravvivenza a quel momento:

$$h(t) = \lim_{\Delta t \to 0} Pr [(t \leq T < t+\Delta t)|T \geq t]$$

$$= f(t) / S(t)$$

Per realizzare un modello sui di sopravvivenza di solito si impiega la funzione di rischio o il logaritmo di tale funzione. Per esempio, assumendo un rischio costante, $h (t) = v$, si implica una distribuzione esponenziale dei tempi di sopravvivenza, con funzione di densità $P (t) = v \exp(- v t)$.
Altri comuni modelli di rischio includono :

$\text{Log } h (t) = v + \rho t$

che conduce alla distribuzione di Gompertz dei tempi di sopravvivenza, e
Log h (t) = v + ρ log (t)
che porta alla distribuzione di Weibull dei tempi di sopravvivenza (si veda, per esempio, Cox e Oakes, 1984: Sezioni 2,3)
In entrambe le distribuzioni di Gompertz e Weibull, il rischio può aumentare o diminuire con il passare del tempo, inoltre, in entrambi i casi, l'impostazione di ρ = 0 rende il modello esponenziale.
Una complicazione per l'analisi dei dati di sopravvivenza è la cosiddetta censura, di cui la forma più comune di censura è quella a destra:
- rappresenta la scadenza del periodo di osservazione o quando un singolo individuo viene rimosso dallo studio, prima che l'evento si verifica. Per esempio, alcuni individui possono essere ancora vivi alla fine di una sperimentazione clinica, o possono cadere fuori dello studio per vari motivi diversi da quelli di morte prima del termine del periodo di osservazione.
L'osservazione è censurata a sinistra, se non si conosce il suo tempo iniziale di comparsa del rischio. La stessa osservazione può essere censurata sia a destra che a sinistra, una circostanza denominata intervallo-censurato. Censurare complica la funzione di probabilità, e quindi la stima, dei modelli di sopravvivenza.
Inoltre, la censura deve essere indipendente dal valore futuro di rischio di morte. Censure che soddisfano questa esigenza si definiscono non informative. Se questa condizione non è soddisfatta, avremo che le stime della distribuzione di sopravvivenza possono essere seriamente distorte. Per esempio, se gli individui tendono ad uscire fuori da una sperimentazione clinica poco prima di morire, e quindi la loro morte passa inosservata, il tempo di sopravvivenza sarà ovviamente sovrastimato.
Un esempio corretto e comune di censure non informative si verifica quando uno studio termina in una data prestabilita e quindi i soggetti in vita vengono tutti censurati.

4.4 Il modello di Cox a rischi proporzionali

Come accennato, l'analisi di sopravvivenza di solito esamina la relazione che lega la distribuzione della sopravvivenza alle covariate.
Più comunemente, questo esame comporta la specificazione di un modello lineare per il logaritmo dei rischi. Per esempio, un modello parametrico basato sulla distribuzione esponenziale può essere scritto come

$$\log h_i (t) = \alpha + \beta_1 x_{i1} + \beta_2 x_{ik} + \ldots + \beta_k x_{ik}$$

o equivalentemente,

$$h_i (t) = \exp (\alpha + \beta_1 x_{i1} + \beta_2 x_{ik} + \ldots + \beta_k x_{ik})$$

che è, come un modello lineare per il logaritmo del rischio o come un modello moltiplicativo del rischio. Qui, i, è un pedice per l'osservazione, e le x sono le covariate. La costante α in questo modello rappresenta una sorta di linea basale logaritmica del rischio, del momento che log hi (t) = α [o hi (t) = exp(α)] quando tutte le x sono pari a zero. Ci sono comunque altri modelli parametrici di regressione simili basati su altre distribuzioni di sopravvivenza.
Il modello di Cox, al contrario, lascia la funzione di rischio basale α (t) = log h0 (t), non specificata:
$\log h_i(t) = \alpha(t) + \beta_1 x_{i1} + \beta_2 x_{ik} + \ldots + \beta_k x_{ik}$
o, ancora equivalentemente,
$h_i(t) = h_0(t) \exp(\beta_1 x_{i1} + \beta_2 x_{ik} + \ldots + \beta_k x_{ik})$
Questo modello è semi-parametrico, perché il rischio di base può assumere qualsiasi forma, al covariare del modello lineare basato sui regressori. Si consideri, adesso, due osservazioni i e i' che differiscono per i loro valori delle x, con il corrispondente predittore lineare:

$\eta_i = \beta_1 x_{i1} + \beta_2 x_{ik} + \ldots + \beta_k x_{ik}$
e
$\eta_{i'} = \beta_1 x_{i'1} + \beta_2 x_{i'k} + \ldots + \beta_k x_{i'k}$
Il rapparto del rischio per queste due osservazioni,

$h_i(t)/h_{i'}(t) = h_0(t)\exp(\eta_i) / h_0(t)\exp(\eta_{i'}) = \exp(\eta_i) / \exp(\eta_{i'})$

che è indipendente dal tempo t. Di conseguenza, il modello di Cox è un modello a rischi proporzionali.
Sorprendentemente, anche se il rischio basale non è specificato, il modello di Cox può ancora essere stimato dal metodo della verosimiglianza parziale, sviluppato da Cox (1972) in concomitanza della presentazione del suo modello. Sebbene le stime risultanti non siano così efficienti come le stime di massima verosimiglianza per un correttamente specificato parametrico modello di regressione dei rischi, non appare arbitraria, e incorretto, assumere questa forma del rischio di base per compensare le virtù del modello stesso.

4.5 Analisi della sopravvivenza mediante le Random Survival Forest

Prima di analizzare la sopravvivenza utilizzando le Random Survival Forest (RSF) vediamo alcuni metodi di complemento alle stesse RSF di cui si rimanda per una trattazione più esauriente a Azzalini A. e Scarpa B. (2004), Analisi dei dati e data mining, Milano, Sprinter-Verlag Italia .

4.5.1 Metodo bootstrap

È una tecnica di campionamento per approssimare la distribuzione campionaria di una statistica. Permette perciò, di approssimare media e varianza di uno stimatore, costruire intervalli di confidenza e calcolare p-values di test quando, in particolare, non si conosce la distribuzione della statistica di interesse.
Nel caso di campionamento casuale semplice, il funzionamento è il seguente:

- consideriamo un campione effettivamente osservato di numerosità pari ad n, diciamo $x = (x_1,...,x_n)$.
- da x si ricampionano m altri campioni di numerosità costante pari ad n; diciamo $x_1^*,\ldots,x_m^*$, che definiamo campioni bootstrap,
- in ciascun campione bootstrap, i dati provenienti dal primo elemento del campione, cioè x_1, possono essere estratti più di una volta e ciascun dato ha probabilità pari a $1/n$ di essere estratto.

Sia T lo stimatore di θ che ci interessa studiare. Si calcola tale quantità per ogni campione bootstrap, In questo modo si hanno a disposizione m stime di θ, dalle quali è possibile calcolare la media bootstrap, la varianza bootstrap, i percentili bootstrap etc. che sono approssimazioni dei corrispondenti valori ignoti e portano informazioni sulla distribuzione di $T(x)$.
Tale metodo è utile quindi partendo da queste quantità stimate per calcolare intervalli di confidenza, saggiare ipotesi, etc.

4.5.2 Random forest

Le foreste casuali sono una tecnica di randomizzazione (Breiman, 2001) utilizzata alla base di un processo di analisi utile per migliorare le prestazioni del processo di analisi stesso.
Nello specifico una foresta casuale è un classificatore che consiste di molti alberi di decisione e di uscite che sono le modalità di classe di ogni singolo albero decisionale.
L'algoritmo per indurre una foresta casuale è stato sviluppato da Leo Breiman e Adele Cutler, e "Random Forest" è il loro marchio. Il termine proviene da foreste casuali decisionali (random decision forest) che è stata proposta da Tin Kam Ho dei Bell Labs nel 1995. Il metodo combina la tecnica di “bagging” di Breiman e l’idea di Ho "metodo del sottospazio casuale" per costruire una collezione di alberi di decisione con controllate variazioni.
Ogni albero è costruito utilizzando la seguente procedura:

4 Supponiamo che il numero di casi in oggetto sia N, e il numero di variabili da trattare nel classificatore sia M.
5 Con $m<<M$ (con m di molto inferiore a M) numero di variabili in ingresso da utilizzare per determinare la decisione ad un nodo di un albero.

6 Si sceglie un set di prova per questo albero prendendo N volte, con reinserimento, da tutti gli N casi di set di prova disponibili (vale a dire prendere un campione di bootstrap). Utilizza il resto dei casi per stimare l'errore dell'albero, di prevedere le loro classi.

7 Per ciascun nodo di un albero, scegliere casualmente m variabili su cui basare la decisione sua quel nodo. Calcolare la migliore divisione sulla base di tali m variabili nella formazione impostata.

8 Ogni albero è pienamente sviluppato e non potato (come un normale albero classificatore).

I vantaggi delle random forest sono:

- Per molti insiemi di dati, produce un albero classificatore molto accurato.
- Permette di gestire un gran numero di variabili d'ingresso.
- Stima l'importanza delle variabili nel determinare la classificazione.
- Genera una stima priva di bias della generalizzazione dell'errore con la crescita della foresta.
- Comprende un buon metodo per la stima dei dati mancanti e mantiene la precisione, quando una gran parte dei dati stessi sono mancanti.
- Fornisce un metodo sperimentale per individuare le interazioni tra variabili.
- Può bilanciare l'errore (campionario) negli insiemi di dati non bilanciati.
- Calcola la vicinanze tra i casi, utile per il clustering, l'individuazione delle rilevazioni outliers, e la visualizzazione dei dati.
- È di facile apprendimento.

4.5.3 Random Survival Forest (RSF)

È strettamente collegata alle Random forest e da quest'ultima metodologia di analisi dei dati ne eredita tutti suoi vantaggi.
Due sono le caratteristiche delle RSF che sono da sottolineare:

- facili da usare, metodica abbastanza robusta, preliminarmente devono essere settati i seguenti parametri (numero delle variabili predittrici che verranno casualmente scelte, numero degli alberi decisionali da elaborare e quale regola utilizzare per la creazione dei nodi dell'albero)
- metodo estremamente adattivo ai dati e virtualmente non vincolato alle usuali assunzioni che si applicano ai classici modelli statistici.

Quest'ultima proprietà è particolarmente utile per l'analisi della sopravvivenza. Le analisi statistiche standard della sopravvivenza si basano spesso su ipotesi restrittive, quali ad esempio i rischi proporzionali (vedi appunto il modello di Cox). Inoltre, con tali

metodi vi è sempre la preoccupazione se le associazioni tra le variabili predittrici del modello e i rischi sono stati opportunamente modellati, e se o meno effetti non lineari o di ordine superiore per interazioni dovrebbero essere inclusi. Al contrario, tali problemi sono trattati automaticamente e gestiti senza complicazioni utilizzando il metodo delle RSF.

5 Applicazione di varie metodiche per l'analisi della sopravvivenza in campo biomedico: analisi della durata della degenza in unità di cure intensive cardiologiche (UCIC) dei pazienti ricoverati per SCA in UCIC da ottobre 2006 a ottobre 2007

5.1 Analisi esemplificativa della sopravvivenza in UCIC per i pazienti affetti da SCA attraverso le curve di Kaplan-Meier

Riprendendo in considerazione il dataset utilizzato per l'analisi della durata della degenza dei pazienti ricoverati in UCIC nel perido ottobre 2006 – ottobre 2007 con la diagnosi d'ingresso di SCA, si è registrata una mortalità del 2,3% (vedi fig.28)

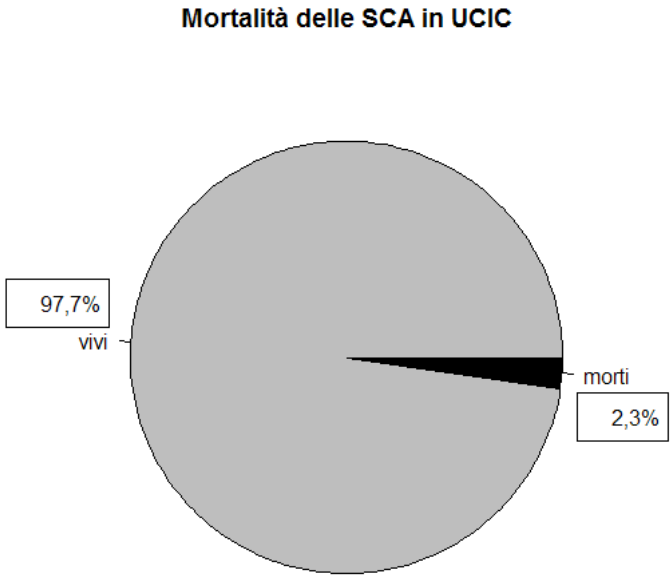


Fig.28

Con semplici comandi in R dopo aver caricato il pacchetto survival risulta:

```
> ucicfit <- survfit(Surv(GGDEG, DEC == 1), data = dati)
> ucicfit
Call: survfit(formula = Surv(GGDEG, DEC == 1), data = dati)
```

n	events	median	0.95LCL	0.95UCL
564	13	Inf	22	Inf

Il software R conferma il risultato lusinghiero della bassa mortalità intrareparto dando un valore mediano non calcolabile e un intervallo di confidenza (al 95%) per la mortalità compresa tra lo 0% e il 3,9%.
Graficamente tutto quanto esposto nella precedente proposizione si traduce nella seguente curva di Kaplan-Meier (fig.)

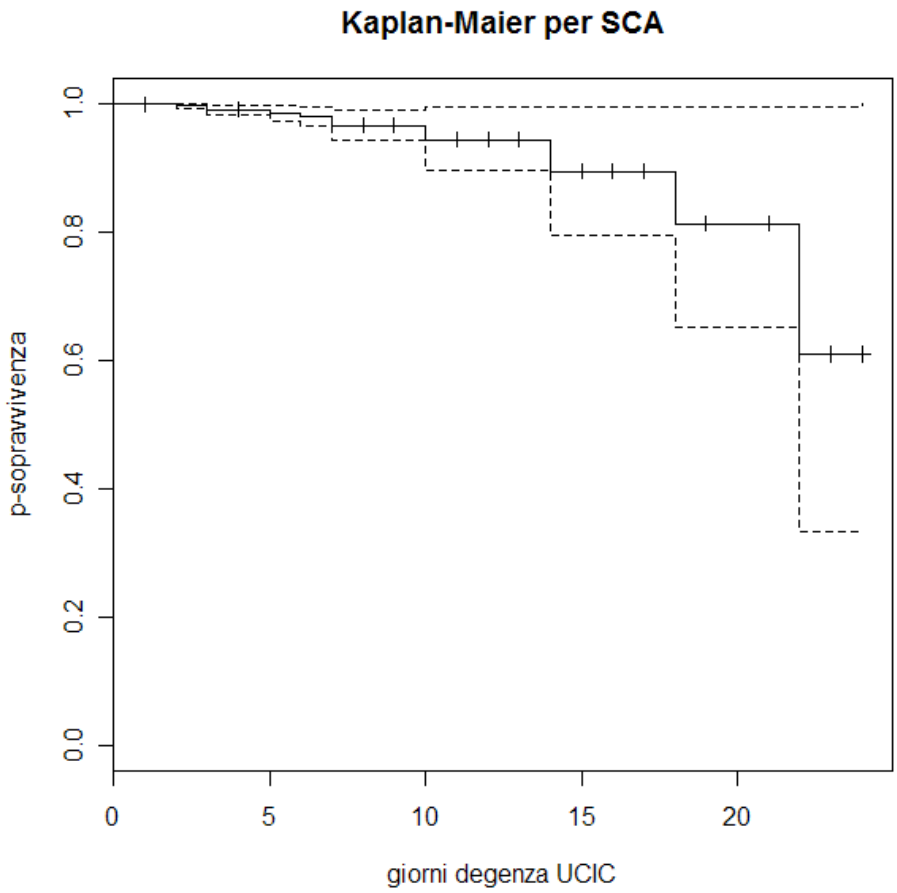
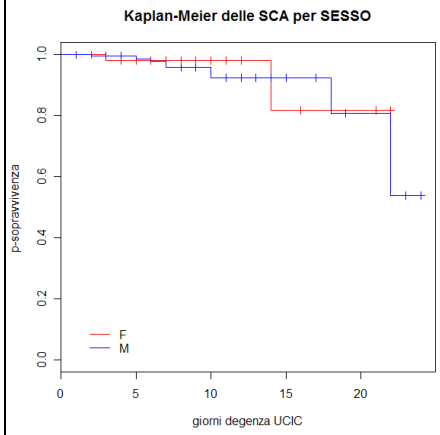
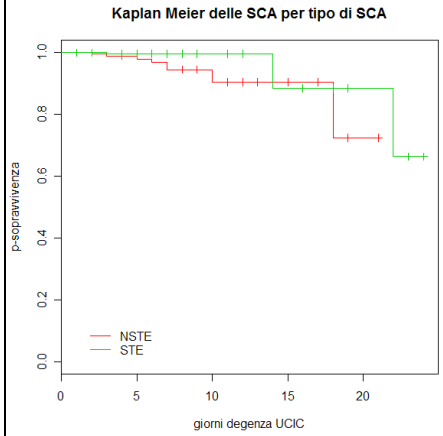
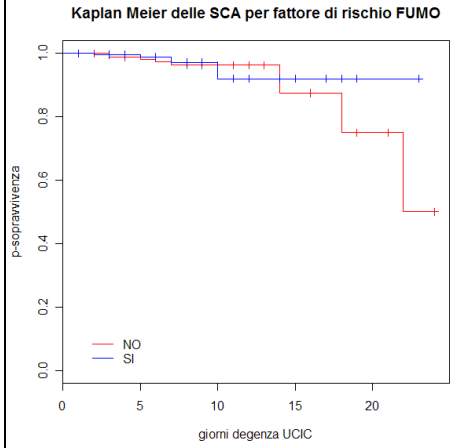
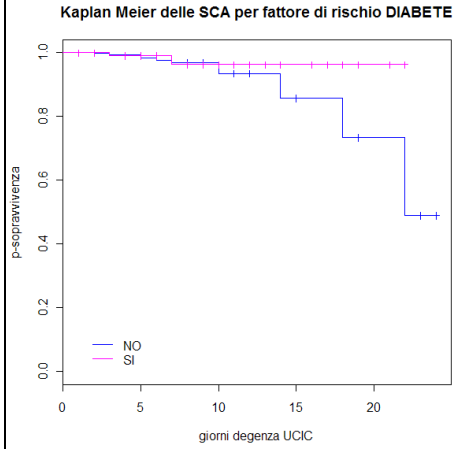


Fig.29

Ora analizzeremo se vi sono delle differenze significative della sopravvivenza mediante le curve di Kaplan-Meier e il successivo calcolo del log-rank test mettendo in relazione proprio la sopravvivenza e alcuni dei fattori considerati nella precedente analisi della durata della degenza (fig.30).
L'analisi non è esaustiva per tutti i fattori considerati ma solo esemplificativa dell'uso di tali curve. Successivamente le variabili verranno utilizzate complessivamente per la realizzazione del modello di Cox a rischi proporzionali e le RSF.

			
survdiff(formula = Surv(GGDEG, DEC == 1) ~ SESSO, data = dati) N Observed Expected (O-E)^2/E (O-E)^2/V SESSO=F 174 4 4.23 0.01209 0.0181 SESSO=M 390 9 8.77 0.00582 0.0181 Chisq= 0 on 1 degrees of freedom, p= 0.893	> survdiff(Surv(GGDEG, DEC == 1) ~ WHY, data = dati) Call: survdiff(formula = Surv(GGDEG, DEC == 1) ~ WHY, data = dati) N Observed Expected (O-E)^2/E (O-E)^2/V WHY=NSTE 321 10 6.52 1.85 4.09 WHY=STE 243 3 6.48 1.87 4.09 Chisq= 4.1 on 1 degrees of freedom, p= 0.0432	> survdiff(Surv(GGDEG, DEC == 1) ~ FUMO, data = dati) Call: survdiff(formula = Surv(GGDEG, DEC == 1) ~ FUMO, data = dati) N Observed Expected (O-E)^2/E (O-E)^2/V FUMO=No 291 9 7.44 0.327 0.78 FUMO=Si 273 4 5.56 0.437 0.78 Chisq= 0.8 on 1 degrees of freedom, p= 0.377	> survdiff(Surv(GGDEG, DEC == 1) ~ DIAB, data = dati) Call: survdiff(formula = Surv(GGDEG, DEC == 1) ~ DIAB, data = dati) N Observed Expected (O-E)^2/E (O-E)^2/V DIAB=No 438 11 9.59 0.207 0.802 DIAB=Si 126 2 3.41 0.582 0.802 Chisq= 0.8 on 1 degrees of freedom, p= 0.37
NON SIGNIFICATIVO	SIGNIFICATIVO	NON SIGINIFICATIVO	NON SIGNIFICATIVO

5.2 Analisi esemplificativa della sopravvivenza in UCIC per i pazienti affetti da SCA attraverso il modello di regressione di Cox a rischi proporzionali

Ora verrà utilizzato il modello di Cox per stimare la sopravvivenza del campione di pazienti ricoverati in UCIC precedentemente utilizzato per l'analisi della stima della durata della degenza.

Dal dataset (dati) introduciamo una nuova variabile (DEC) la quale indica lo stato di censurato (trasferito/dimesso dal reparto = 0) o deceduto (1)) per ogni singolo paziente (unità statistica).

L'istruzione corrispondente in R è la seguente:

```
>fit_cox1<-
coxph(Surv(GGDEG,DEC)~ETA+SESSO+BMI+OBE+FAM+IXT+FUMO+DIS+DI
AB+IRC+BPCO+WHY+PreIMA+PreAng+PrePTCA+PreCABG+ICD+PM+AngPre
IMA+PreSTROKE+TnIng)
Warning message:
In coxph(Surv(GGDEG, DEC) ~ ETA + SESSO + BMI + OBE + FAM +
IXT +  :
  X matrix deemed to be singular; variable 4 16
```

Purtroppo incontriamo il primo problema: la matrice X risulta singolare e quindi il calcolo viene arrestato.
Proviamo ad eliminare la quarta covariata (OBE) come consigliato anche dalla diagnostica di R ottendo il seguente comando e il corrispettivo output:

```
>fit_cox2<-
coxph(Surv(GGDEG,DEC)~ETA+SESSO+BMI+FAM+IXT+FUMO+DIS+DIAB+I
RC+BPCO+WHY+PreIMA+PreAng+PrePTCA+PreCABG+ICD+PM+AngPreIMA+
PreSTROKE+TnIng)
Warning message:
In coxph(Surv(GGDEG, DEC) ~ ETA + SESSO + BMI + FAM + IXT +
FUMO +  :
  X matrix deemed to be singular; variable 15
```

Si ripete il problema di matrice singolare e quindi R sospende il computo del modello.
Reiteriamo questa "potatura" di covariate escludendo la 15° (PreCABG).

```
>  fit_cox2 <- coxph(Surv(GGDEG, DEC) ~
ETA+SESSO+BMI+FAM+IXT+FUMO+DIS+DIAB+IRC+BPCO+WHY+PreIMA+Pre
Ang+PrePTCA+ICD+PM+AngPreIMA+PreSTROKE+TnIng)
```

Il software è ora riuscito a computare il modello e qui sotto è in dettaglio riportato:

```
> summary(fit_cox2)
Call:
coxph(formula = Surv(GGDEG, DEC) ~ ETA + SESSO + BMI + FAM
+ IXT + FUMO + DIS + DIAB + IRC + BPCO + WHY + PreIMA +
PreAng + PrePTCA + ICD + PM + AngPreIMA + PreSTROKE +
TnIng)

  n= 564
```

```
                  coef exp(coef) se(coef)       z     p
ETA           0.03609    1.0367   0.0341  1.0584 0.290
SESSOM       -0.36272    0.6958   0.7310 -0.4962 0.620
BMI          -0.15718    0.8546   0.0957 -1.6430 0.100
FAMSi        -1.38718    0.2498   0.9018 -1.5382 0.120
IXTSi         0.19176    1.2114   0.8953  0.2142 0.830
FUMOSi       -0.18132    0.8342   0.7493 -0.2420 0.810
DISSi         0.16899    1.1841   0.7010  0.2411 0.810
DIABSi       -0.81977    0.4405   0.8691 -0.9433 0.350
IRCSi         1.15038    3.1594   0.7665  1.5008 0.130
BPCOSi        1.48103    4.3975   0.7556  1.9602 0.050
WHYSTE       -1.21224    0.2975   0.9281 -1.3062 0.190
PreIMASi      1.27006    3.5611   0.7136  1.7798 0.075
PreAngSi      0.24008    1.2713   0.7762  0.3093 0.760
PrePTCASi    -1.03171    0.3564   1.1935 -0.8644 0.390
ICDSi         3.41786   30.5041   2.9503  1.1585 0.250
PMSi         -2.78493    0.0617   2.9550 -0.9425 0.350
AngPreIMASi  -0.04788    0.9532   0.9154 -0.0523 0.960
PreSTROKESi   0.01402    1.0141   1.1618  0.0121 0.990
TnIng         0.00887    1.0089   0.0109  0.8162 0.410

             exp(coef) exp(-coef) lower .95 upper .95
ETA             1.0367     0.9646  0.969723      1.11
SESSOM          0.6958     1.4372  0.166044      2.92
BMI             0.8546     1.1702  0.708448      1.03
FAMSi           0.2498     4.0035  0.042650      1.46
IXTSi           1.2114     0.8255  0.209502      7.00
FUMOSi          0.8342     1.1988  0.192067      3.62
DISSi           1.1841     0.8445  0.299724      4.68
DIABSi          0.4405     2.2700  0.080209      2.42
IRCSi           3.1594     0.3165  0.703310     14.19
BPCOSi          4.3975     0.2274  1.000171     19.33
WHYSTE          0.2975     3.3610  0.048256      1.83
PreIMASi        3.5611     0.2808  0.879324     14.42
PreAngSi        1.2713     0.7866  0.277669      5.82
PrePTCASi       0.3564     2.8058  0.034357      3.70
ICDSi          30.5041     0.0328  0.093981   9900.98
PMSi            0.0617    16.1987  0.000188     20.22
AngPreIMASi     0.9532     1.0490  0.158495      5.73
PreSTROKESi     1.0141     0.9861  0.104021      9.89
TnIng           1.0089     0.9912  0.987654      1.03

Rsquare= 0.05   (max possible= 0.198 )
Likelihood ratio test= 28.8  on 19 df,   p=0.07
Wald test            = 25.1  on 19 df,   p=0.158
Score (logrank) test = 38  on 19 df,   p=0.00593
```

Relativamente all'output da notare che nella colonna **z** sono riportati i valori della statistica Wald la quale è asintoticamente una normale standard nell'ipotesi che il corrispondente repressore **β** sia 0 e nella colonna **p** i corrispondenti p-value.
La seconda colonna del primo pannello e la prima del secondo sono gli esponenziali dei coefficienti delle covariate e sono interpretabili cone effetti moltiplicativi sul rischio.
Il rapporto di verosimiglianza (likelihood-ratio), Wald e Score (logrank) test sono tutti test asintoticamente equivalenti sotto l'ipotesi nulla che tutti i coefficienti **β** siano 0.

Togliendo una ad una le covariate non significative, si ottiene il seguente modello:

```
>  fit_cox19 <- coxph(Surv(GGDEG, DEC) ~ BMI+IRC)
>  fit_cox19
Call:
coxph(formula = Surv(GGDEG, DEC) ~ BMI + IRC)

        coef exp(coef) se(coef)     z      p
BMI   -0.175      0.84   0.0692 -2.52 0.0120
IRCSi  2.064      7.88   0.6942  2.97 0.0029

Likelihood ratio test=10.4  on 2 df, p=0.00548  n= 564
```

Il corrispondente grafico di stima della funzione di sopravvivenza S(t) mediante la regressione di Cox in funzione del tempo misurato in giorni è qui sotto riportato in fig.31

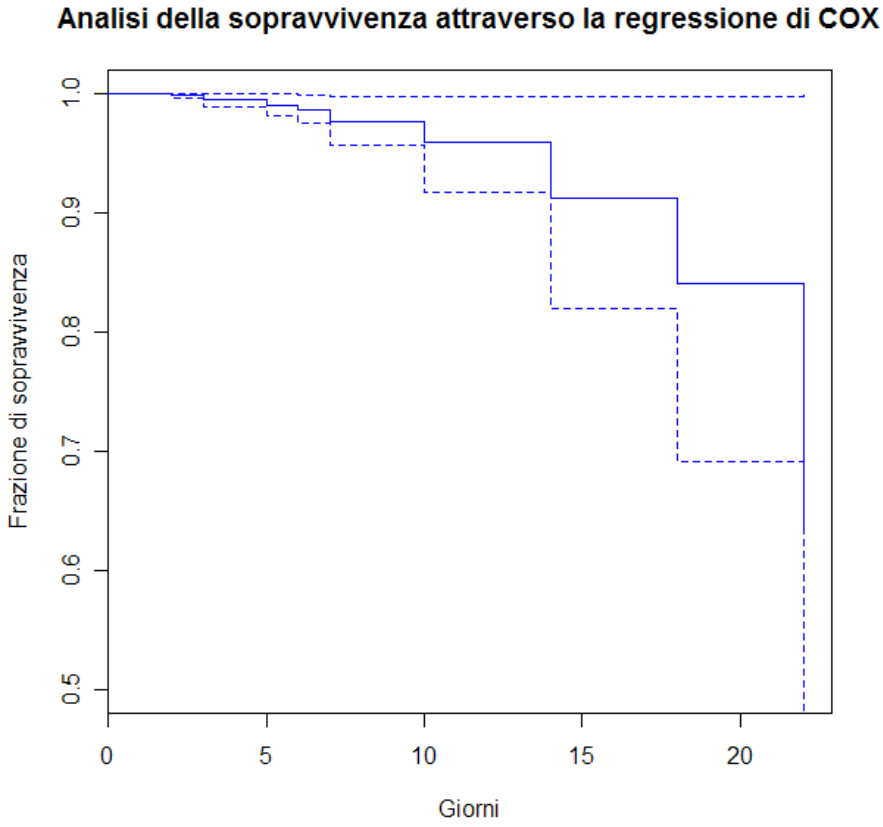


Fig.31

5.2.1 Diagnostiche sul modello di Cox considerato

Come nel caso dei modelli lineari generalizzati visti in precedenza, è desiderabile chiedersi quanto il modello di regressione di Cox descriva i dati. Essenzialmente esistono tre tipi di problematiche da verificare:

- Violazione di assunzione di proporzionalità dei rischi;
- Osservazioni influenti;

- La non linearità nella relazione tra il logaritmo del rischio e le covariate.

Tutte queste diagnostiche si operano analizzando i residui (Therneau T.M. e Grambsch P.M (2000), Modelling Survival Data, Springer).

5.2.1.1 Controllo dell'assunto di proporzionalità dei rischi

I test sia analitici che grafici si basano sui residui di Schoenfeld scalati. Per verificare l'assunto di proporzionalità per ogni covariata si correla il corrispondente set dei residui di Schoenfeld scalati (colonne della matrice dei residui che si ottengono con il comando residuals(model, "scaledsch")) con una trasformata del tempo (il default in R è basato sulla stima della funzione di sopravvivenza di Kaplan-Meier, K(t)).

Con i dati a disposizione si ottengono i seguenti valori:

```
> cox.zph(fit_cox19)
            rho    chisq     p
BMI    -0.06398 0.027163 0.869
IRCSi   0.00625 0.000639 0.980
GLOBAL       NA 0.027354 0.986
```

La funzione offre un test dell'assunzione di proporzionalità del rischio per ogni covariata e per l'intero modello. Valori del p-value del test ($p<0.05$) indicano una violazione dell'assunto sopra espresso.
Una rappresentazione grafica della diagnostica è qui sotto riportata in fig.32

RESIDUI DI SCHOENFELD SU TEMPO

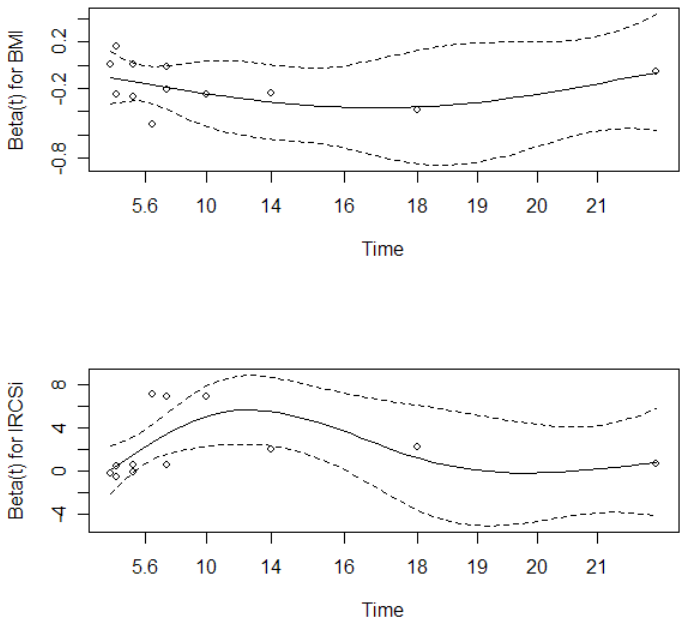


Fig.32

La linea continua dei grafici indica una funzione di lisciamento per il modello e le linee tratteggiate che rappresentano un ±2 deviazione standard attorno al modello.
L'interpretazione dei grafici è facilitata proprio dal lisciamento prodotto dalle linee dei grafici: perché l'assunto di proporzionalità del rischio sia soddisfatto non vi devono essere sistematiche distanze dei punti dall'area compresa tra le linee tratteggiate.

Con i dati in nostro possesso sia il test analitico che i grafici sono suggestivi per l'assunto di non proporzionalità dei rischi.

5.2.1.2 Osservazioni influenti

Vengono analizzate mediante l'osservazione della magnitudine dei residui beta (procedura analoga viene adottata anche nelle diagnostiche dei modelli lineari generalizzati). Più i punti si discostano dalla linea centrale più le osservazioni che li hanno generati sono influenti.
Con il dataset dei pazienti dell'UCIC e con i seguenti comandi otteniamo i grafici riportati in fig.33

```
> par(mfrow=c(2,1))
> for (j in 1:2){
+ plot(dfbeta[,j], ylab=names(coef(fit_cox19))[j])
+ abline(h=0,lty=2) }
```

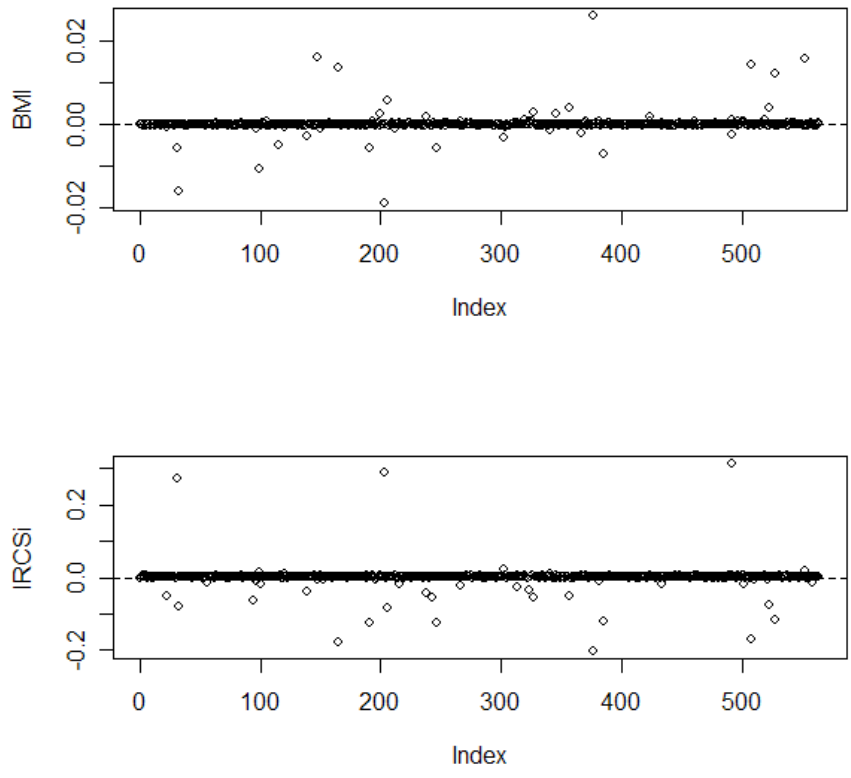


Fig.33

Risultano evidenti solo 4 casi (uno rispetto al repressore di BMI e 3 di IRC) corrispondenti rispettivamente a 376° 29° 203° 491° osservazione. Si ritiene opportuno non apportare modifiche ulteriori al dataset dei dati.

5.2.1.3 La non linearità nella relazione tra il logaritmo del rischio e le covariate

La non linearità è un sinonimo di una incorretta specificazione della forma della funzione nella parte parametrica del modello. Utile per una analisi della non linearità sono i residui di Martingala.
I residui di Martingala possono essere rappresentati graficamente contrapposti alle singole covariate.
Per interpretare il grafico bisogna analizzare la linearità dei residui evidenziata dalla regressione lineare locale generata usando la funzione (lowess).

Con i dati provenienti dal dataset dell'UCIC, non si evincono particolari problematiche rispetto al requisito della linearità (vedi fig.34). Da notare che non si possono realizzare tali grafici di diagnostica per le variabili dicotomiche.

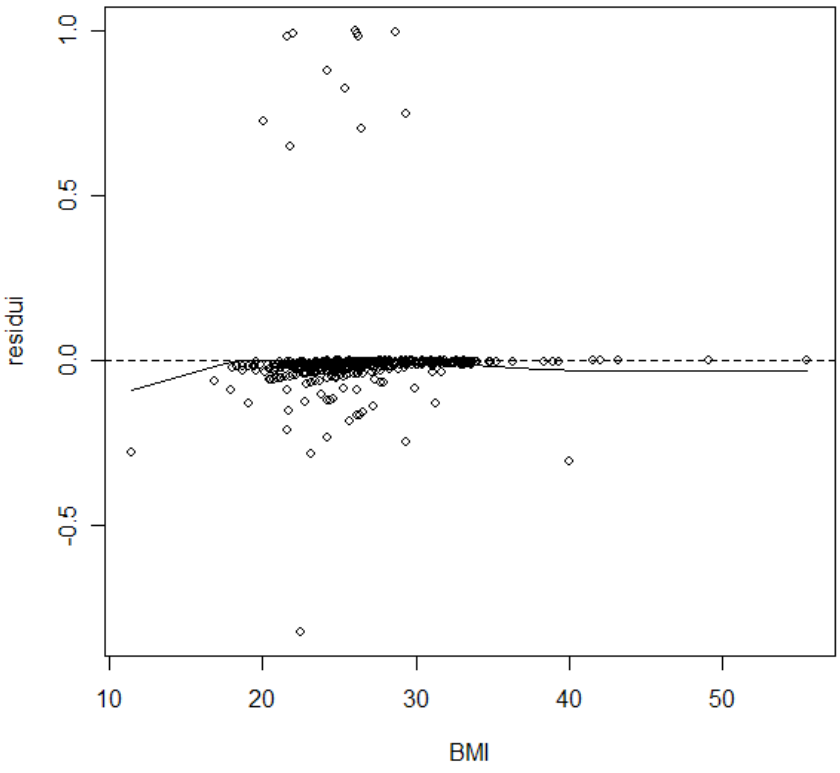


Fig.34

5.3 Analisi della sopravvivenza in UCIC per i pazienti affetti da SCA attraverso il modello delle Random Survival Forest

Scopo di questa sezione dell'elaborato è quella di ricondurre un analisi inferenziale sulla sopravvivenza dei pazienti ricoverati presso l'UCIC dell'Azienda Ospedaliera di Padova utilizzando un approccio non parametrico basato sulle Random Survival Forest e specificatamente sul software sviluppato da Ishwaran ed altri[].
Successivamente si procederà ad un confrontarlo con il precedente modello di Cox così massicciamente utilizzato nella letteratura biomedica[1].

Qui di seguito sono riportati i comandi di R per richiamare le RSF per l'analisi della sopravvivenza dei pazienti ricoverati in UCIC.

```
>fit_RSF<-rsf(Survrsf(GGDEG,DEC)~
ETA+SESSO+BMI+OBE+FAM+IXT+FUMO+DIS+DIAB+IRC+BPCO+WHY+PreIMA
+PreAng+PrePTCA+PreCABG+ICD+PM+AngPreIMA+PreSTROKE+TnIng,da
ta=dati,ntree=10000)
```

Viene riportata in fig. la frazione di errore (*error rate*) per il modello Random Survival Forest come funzione del numero di alberi (parte sinistra del grafico) assieme all' *out-of-bag importance values* dei predittori (parte destra del grafico).

L'*importance value* per ogni singolo predittore è la differenza nel out-of-bag error rate quando il predittore è permutato casualmente compare sull'out-of-bag error rate senza nessuna permutazione. Valori positivi indicano variabili informative, piccolissimi valori o valori negativi indicano variabili non informative.

Si considera la scarsa l'inadeguatezza di ogni predittore mediante la regola di suddivisione del log-rank anche se sono utilizzabili nel pacchetto randomSurvivalForest di R altri 3 criteri (conserve, logrankscore, logrankapprox).

Per una rapida pamonaramica di quanto prodotto vediamo il seguente comando che ci fornisce informazioni sulla numerosità campionaria, numero dei decessi, numero degli alberi creati, grandezza del noto terminale minimo, percentuale del numero di nodi terminali, numero delle variabili provate ad ogni suddivisione dell'albero, totale del numero delle variabili, regole di suddivisione dell'albero, stima della frazione di errore:

```
> print(fit_RSF)

Call:
 rsf.default(formula = Survrsf(GGDEG, DEC) ~ ETA + SESSO +
BMI + OBE + FAM + IXT + FUMO + DIS + DIAB + IRC + BPCO +
```

[1] 12266 abstract individuati nella banca dati PubMed (www.pubmed.gov) con la chiave di ricerca "Cox model" in data 29/01/2008

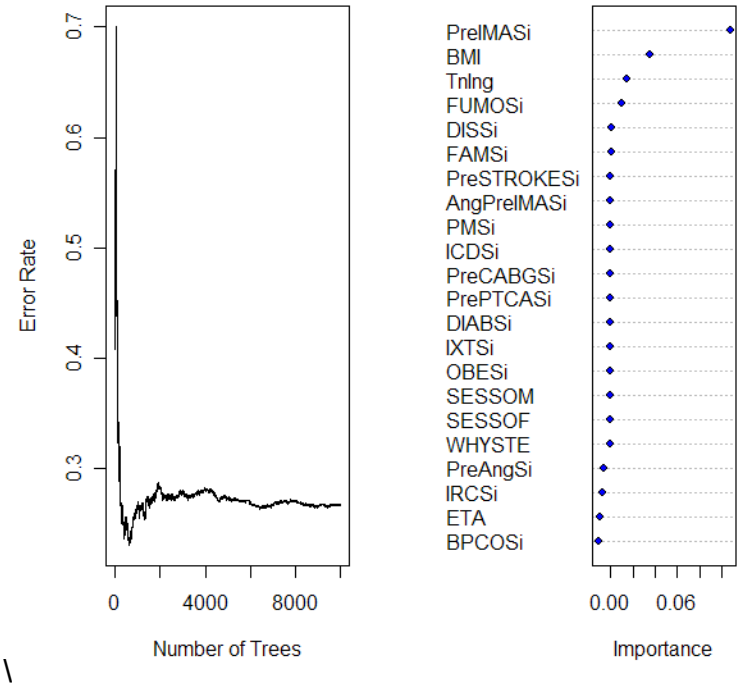
```
WHY + PreIMA + PreAng + PrePTCA + PreCABG + ICD + PM +
AngPreIMA + PreSTROKE + TnIng, data = dati, ntree = 10000,
proximity = TRUE,      variable = TRUE)

                         Sample size: 564
                  Number of deaths: 13
                   Number of trees: 10000
         Minimum terminal node size: 3
       Average no. of terminal nodes: 1.8478
No. of variables tried at each split: 4
              Total no. of variables: 22
                      Splitting rule: logrank
              Estimate of error rate: 25.2%
```

La fig.35 riporta il valore dell'importanza di tutti e 21 i predittori che si era tentato precedentemente di analizzare con il modello di Cox. Di seguito sono anche riportati il comando e l'analitico dei risultati.

```
> plot.error(fit_RSF)

               Importance   Relative Imp   predictorWt
PreIMASi           0.1073         1.0000             1
BMI                0.0360         0.3352             1
TnIng              0.0148         0.1375             1
FUMOSi             0.0098         0.0917             1
DISSi              0.0015         0.0143             1
FAMSi              0.0009         0.0086             1
PreSTROKESi        0.0000         0.0000             1
AngPreIMASi        0.0000         0.0000             1
PMSi               0.0000         0.0000             1
ICDSi              0.0000         0.0000             1
PreCABGSi          0.0000         0.0000             1
PrePTCASi          0.0000         0.0000             1
DIABSi             0.0000         0.0000             1
IXTSi              0.0000         0.0000             1
OBESi              0.0000         0.0000             1
SESSOM             0.0000         0.0000             1
SESSOF             0.0000         0.0000             1
WHYSTE            -0.0006        -0.0057             1
PreAngSi          -0.0055        -0.0516             1
IRCSi             -0.0074        -0.0688             1
ETA               -0.0098        -0.0917             1
BPCOSi            -0.0111        -0.1032             1
```



\

Fig.35

```
> plot.ensemble(fit_RSF)
```

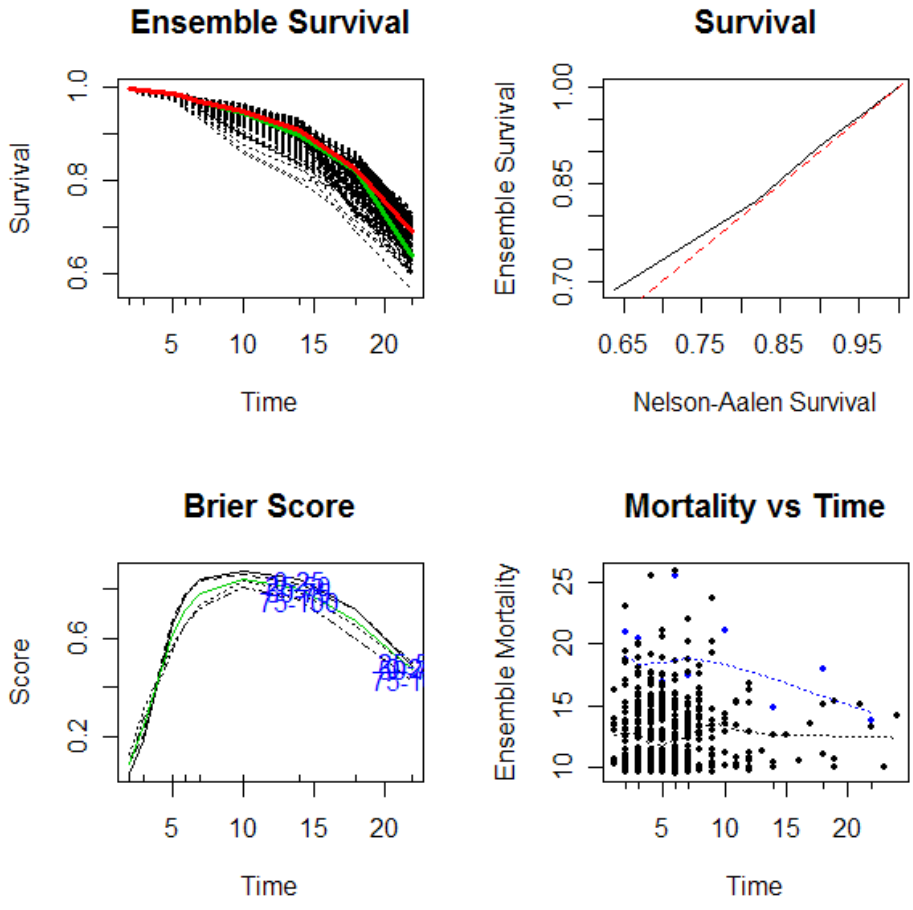


Fig.36

Il comando produce quattro grafici (fig.36). Andando dall'alto verso il basso, da sinistra a destra:

- funzione di sopravvivenza: la spessa linea rossa rappresenta la curva di sopravvivenza globale, mentre quella verde è la stima della sopravvivenza secondo Nelson-Aalen,

- si tratta di un confronto tra le precedenti curve (la curva stimata della sopravvivenza con la RSF e la stima di Nelson-Aalen).

- il punteggio di Brier, è un valore compreso tra 0 e 1 per valutare la bontà della stima prodotta (dove 0 = perfetta stima, 1 = pessima stima, e stima buona = 0,25), Sono riportate con le linee nere i valori corrispondenti ai 4 gruppi corrispondenti a 0-25, 25-50, 50-75 e 75-100 percentile della mortalità mentre la linea verde rappresenta il punteggio non stratificato

- grafico della stima della mortalità osservata rispetto al tempo. I punti blu corrispondono ai decessi, mentre i punti neri sono le osservazioni censurate ovvero i pazienti trasferiti o dimessi.

Si può ora verificare graficamente il contributo di ogni singolo predittore sulla mortalità disposta sull'asse verticale ed interpretabile come numero totale dei decessi. Ovvero se l'individuo i-esimo del nostro dataset ha una mortalità pari a 100 allora se tutti gli altri individui del dataset sono simili all'i-esimo considerato allora in percentuale ci sarà 100 morti nel valore dell'asse delle y nel grafico. Qui di seguito nella fig. è evidenziato il contributo delle variabili più predittive dell'evento morte in UCIC (figg.37-38).

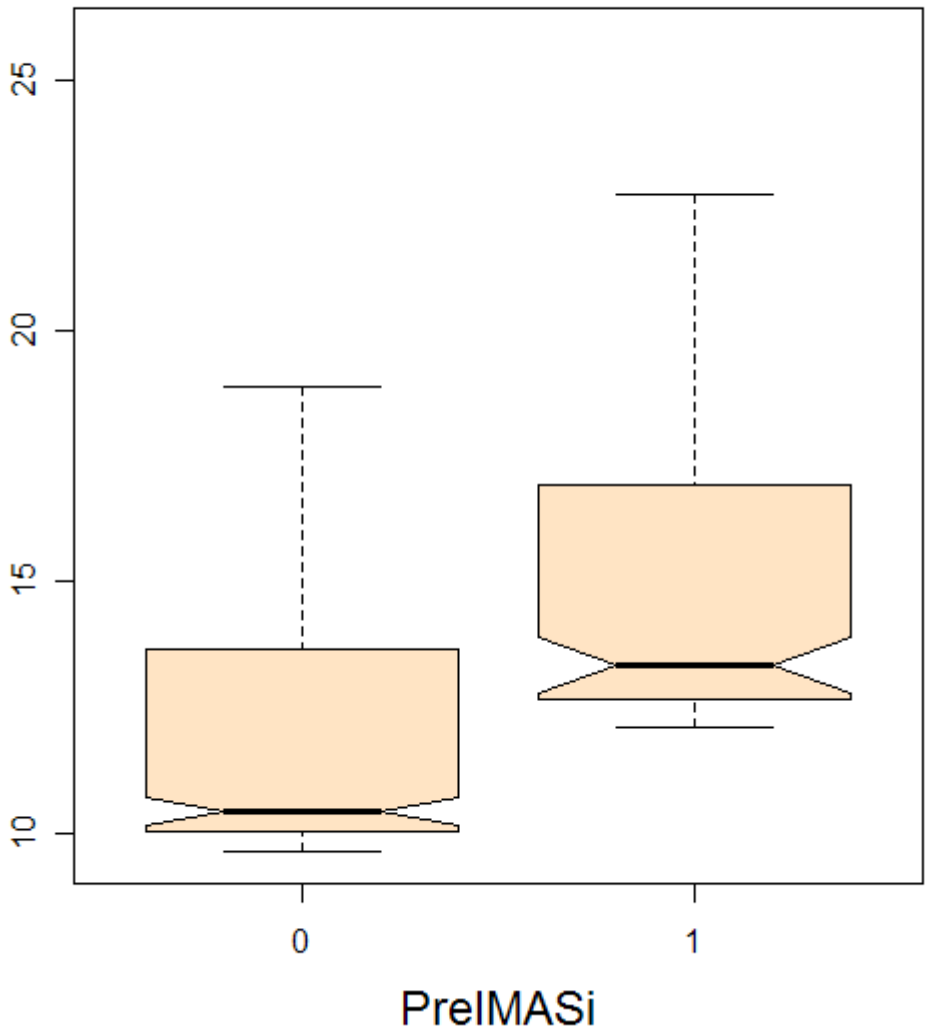


Fig.37

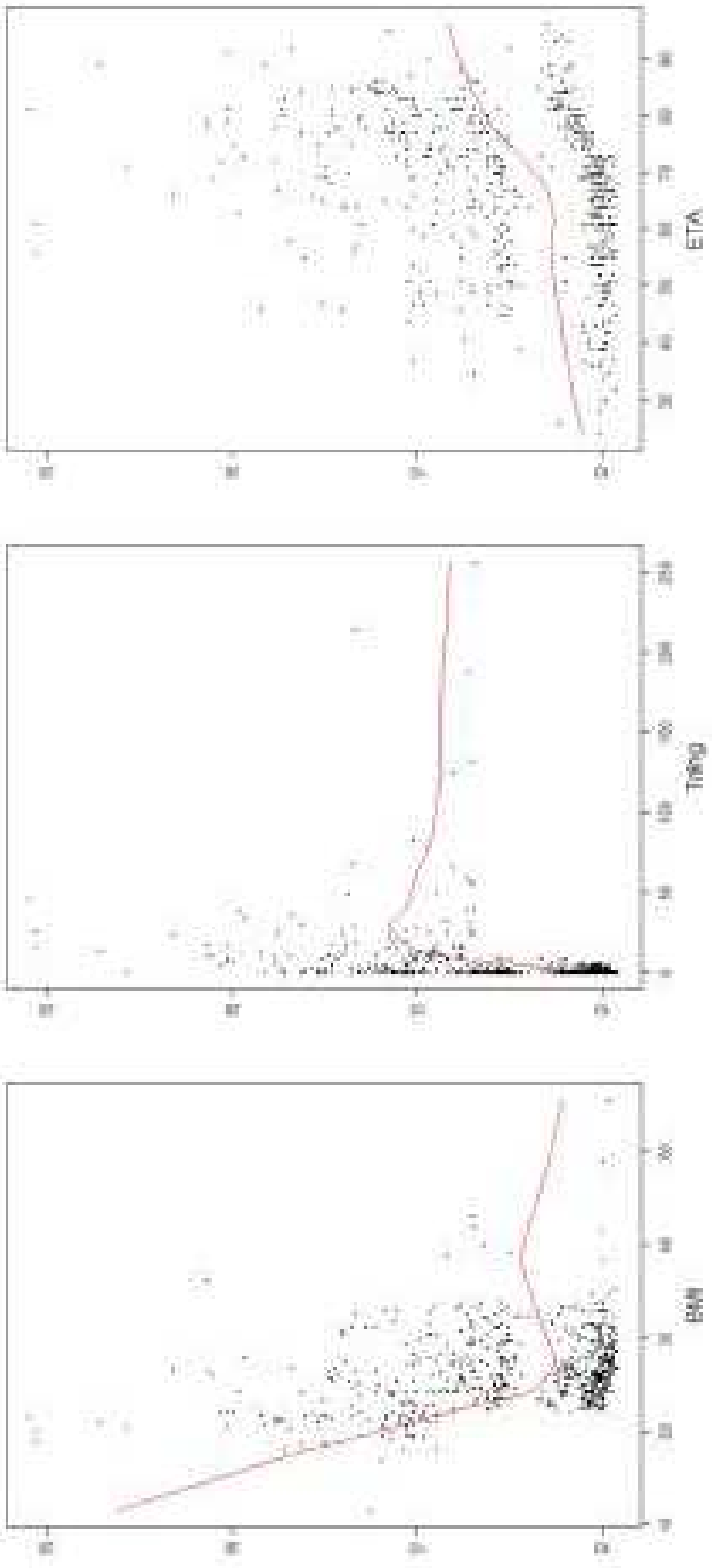


Fig.38

Fig.

5.4 Confronto tra le RSF e il modello di Cox

Risulta difficile dopo queste analisi confrontare i due modelli per i seguenti motivi:

- la scarsa bibliografia inerente confronti tra modelli parametrici o semiparetrici e le random survival forest
- la diversa natura (uno semiparametrico, l'altro non parametrico),
- il diverso dataset di origine a causa di un avviso di singolarità della matrice per il computo del modello di Cox,
- le diverse conclusioni tratte dai due metodi probabilmente dovute al problema espresso nel punto precedente.

Viste le criticità sopraccitate è intenzione dell'autore approfondire la ricerca di uno strumento di confronto congruo tra le RSF e il modello di regressione di Cox per l'analisi della sopravvivenza che per ragioni di tempo non potrà essere parte integrante di questo elaborato. Si segnala a riguardo il lavoro condotto da Figini Silvia (2006): Random Survival Forest models for SME Credit Risk Measurement nel quale vengono confrontate le performance delle RSF con un modello logistico attraverso le curve ROC.

5.5 Discussione

Per quanto riguarda le analisi sulla sopravvivenza dei pazienti affetti da sindrome coronarica acuta ricoverati in unità di cure intensive cardiologiche dell'Azienda Ospedaliera di Padova mediante il modello di Cox e le random survival forest (RSF) emergono le seguenti differenze sostanziali:

- le RSF non richiedono assunzioni da verificare come quella di proporzionalità dei rischi del modello di Cox,
- Maggiore lavoro computazionale nell'applicazione delle RSF rispetto all'elaborazione di un modello di Cox (riscontri ottenuti con il software *R* e con i pacchetti *survival* e *randomSurvivalForest* istallati),
- Complicazioni computazionali insorte nella realizzazione del modello di Cox con il dataset iniziale dei pazienti affetti da SCA dovuto al riscontro di matrice singolare,
- bagaglio di utilità grafiche contenute nella libreria delle RSF che sono estremamente chiarificatrici delle conclusioni sui predittori.

Nonostante il numero ancora ridotto delle applicazioni delle RSF rispetto all'utilizzo del modello di Cox in ambito di analisi sulla sopravvivenza su dati biomedici, le qualità evidenziate in precedenza pongono dei validi presupposti per un loro più consistente impiego in tale ambito.

Allegati

1. Elenco analitico delle variabili
2. Codice realizzato con il software R ver.2.6.1

Elenco analitico delle variabili considerate per l'analisi della durata della degenza e della sopravvivenza

ETA: variabile numerica per codificare l'età (in anni) del paziente al momento del ricovero;
GGDEG: variabile numerica per codificare la durata della degenza (in giorni);
SESSO: variabile dicotomica per coricare il sesso del paziente ("M"=maschio, "F"=femmina),
BMI (body mass index): variabile numerica per coricare l'indice di massa corporea del ricoverato definito come peso(in Kg)/altezza(in metri)^2,
DEC: variabile dicotomica per codificare lo stato del paziente (0=dimesso/trasferito, 1=deceduto),
OBE: variabile dicotomica per la codifica del fattore di rischio obesità ("No"=assenza del fattore, "Si"=presenza del fattore)
FAM: variabile dicotomica per la codifica del fattore di rischio persona con famigliarità per malattie cardiovascolari ("No"=assenza del fattore, "Si"=presenza del fattore)
IXT: variabile dicotomica per la codifica del fattore di rischio persona ipertesa ("No"=assenza del fattore, "Si"=presenza del fattore)
FUMO: variabile dicotomica per la codifica del fattore di rischio fumatore ("No"=assenza del fattore, "Si"=presenza del fattore)
DIS: variabile dicotomica per la codifica del fattore di rischio persona dislipidemica ("No"=assenza del fattore, "Si"=presenza del fattore)
DIAB: variabile dicotomica per la codifica del fattore di rischio persona diabetica ("No"=assenza del fattore, "Si"=presenza del fattore)
IRC: variabile dicotomica per la codifica di persona affetta da insufficienza renale cronica ("No"=assenza del fattore, "Si"=presenza del fattore)
BPCO: variabile dicotomica per la codifica di persona con broco-pneumopatia-cronico-ostruttiva ("No"=assenza del fattore, "Si"=presenza del fattore)
WHY: variabile qualitativa a 2 modalità (NSTE, STE) per la classificazione della tipologia di insulto ischemico subito dal miocardio all'ingresso;
PreIMA: variabile dicotomica per la codifica di persona con pregresso infarto del miocardio ("No"=assenza del fattore, "Si"=presenza del fattore);
PreANG: variabile dicotomica per la codifica di persona con pregressa angina ("No"=assenza del fattore, "Si"=presenza del fattore);
PrePTCA: variabile dicotomica per la codifica di persona con precedente intervento di angioplastica coronaria percutanea ("No"=assenza del fattore, "Si"=presenza del fattore);
PreCABG: variabile dicotomica per la codifica di persona con precedente intervento di bypass aortocoronarico ("No"=assenza del fattore, "Si"=presenza del fattore);

ICD: variabile dicotomica per la codifica di persona portatrice di defibrillatore cardiaco impiantabile (“No”=assenza del fattore, “Si”=presenza del fattore);
PM: variabile dicotomica per la codifica di persona portatrice di pace maker (“No”=assenza del fattore, “Si”=presenza del fattore);
AngPreIMA: variabile dicotomica per la codifica di persona con pregressi attacchi anginosi antecedenti l’attuale sindrome coronarica acuta (SCA) (“No”=assenza del fattore, “Si”=presenza del fattore);
PreSTROKE; : variabile dicotomica per la codifica di persona con pregresso ictus cerebrale (“No”=assenza del fattore, “Si”=presenza del fattore)
FE: variabile numerica per la codifica della funzione eiettiva cardiaca (in percentuale, da coronarografia(CNG));
VTD: variabile numerica per la codifica del volume tele diastolico (VTD)
NrVasiMal: variabile numerica intera per la codifica del numero dei vasi coronarici malati (da CNG);
IABP: variabile dicotomica indicante la presenza o no del contropulsatore aortico (“No”=assenza del fattore, “Si”=presenza del fattore;
PAIN: variabile dicotomica per la codifica della presenza di dolore al momento del ricovero (“No”=assenza del fattore, “Si”=presenza del fattore)
NrDerPatol: variabile numerica intera positiva per la codifica del numero delle derivazioni elettrocardiografiche patologiche all’ingresso(alterazioni del tratto ST);
Tning: variabile numerica indicante i livelli della troponina I all’arrivo in reparto;
TnPicco: variabile numerica indicante i livelli massimi rilevati della troponina I.

```
#Caricare il dataset UCIC e prime analisi
 dati=read.csv2(file.choose(), header = TRUE, sep
= ";", quote="\"", dec=",", fill = TRUE)
 attach(dati)
 str(dati)
summary(dati)
 names(dati)

#istogramma della distribuzione dei giorni di
degenza
hist(GGDEG,prob=T,breaks=200,col=3,main="Distribu
zione dei gg di degenza in UCIC per SCA")


#Analisi bivariata (grafici)
 boxplot(GGDEG ~ SESSO)
 boxplot(GGDEG ~ WHY)
 boxplot(GGDEG ~ OBE)
 boxplot(GGDEG ~ FAM)
 boxplot(GGDEG ~ IXT)
 boxplot(GGDEG ~ FUMO)
 boxplot(GGDEG ~ DIS)
 boxplot(GGDEG ~ DIAB)
 boxplot(GGDEG ~ PreIMA)
 boxplot(GGDEG ~ PreCABG)
 boxplot(GGDEG~NrVasiMal)
 plot(GGDEG ~ ETA)
 plot(GGDEG ~ BMI)
 plot(GGDEG ~ TnIng)
 plot(GGDEG ~ TnPicco)
#…test
 wilcox.test(GGDEG ~ SESSO)
 wilcox.test(GGDEG ~ WHY)
 wilcox.test(GGDEG ~ OBE)
 wilcox.test(GGDEG ~ FAM)
 wilcox.test(GGDEG ~ IXT)
 wilcox.test(GGDEG ~ FUMO)
 wilcox.test(GGDEG ~ DIS)
 wilcox.test(GGDEG ~ DIAB)
 wilcox.test(GGDEG ~ PreIMA)
 wilcox.test(GGDEG ~ PreCABG)
 kruskal.test(GGDEG~NrVasiMal)
#…varianza e covarianza
 list(var(ETA,GGDEG),cor(ETA,GGDEG))
 list(var(BMI,GGDEG),cor(BMI,GGDEG))
 cor(TnIng,GGDEG)
 length(TnIng[TnIng<10])
 list(var(TnIng,GGDEG),cor(TnIng,GGDEG))
 cor(TnPicco,GGDEG))
```

```
#GLM
 fit_gamma_log_1<-
glm(GGDEG~ETA+SESSO+BMI+OBE+FAM+IXT+FUMO+DIS+DIAB
+IRC+BPCO+WHY+PreIMA+PreAng+PrePTCA+PreCABG+ICD+A
ngPreIMA+PreSTROKE+NrVasiMal+TnIng,
family=gamma(link=log),data=dati)
 summary(fit_gamma_log_1)
…
 fit_gamma_log_20<-glm(GGDEG~ETA+TnIng,
family=gamma(link=log),data=dati)
 summary(fit_gamma_log_20)
Call:
glm(formula = GGDEG ~ ETA + TnIng, family =
Gamma(link = log),
    data = dati)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3745  -0.4078  -0.1320   0.1979   1.9057
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.284596   0.124617  10.308  < 2e-16
***
ETA         0.005771   0.001814   3.181  0.00155
**
TnIng       0.003778   0.001190   3.175  0.00158
**
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for Gamma family taken to
be 0.3229007)
    Null deviance: 157.02  on 563  degrees of
freedom
Residual deviance: 150.50  on 561  degrees of
freedom
AIC: 2650.8
Number of Fisher Scoring iterations: 6

 fit_gamma_id_1<-
glm(GGDEG~ETA+SESSO+BMI+OBE+FAM+IXT+FUMO+DIS+DIAB
+IRC+BPCO+WHY+PreIMA+PreAng+PrePTCA+PreCABG+ICD+A
ngPreIMA+PreSTROKE+NrVasiMal+TnIng,
family=gamma(link=identity),data=dati)
 summary(fit_gamma_log_1)
…
> summary(fit_gamma_id_19)
Call:
glm(formula = GGDEG ~ ETA + NrVasiMal + TnIng,
family = Gamma(link = identity),
    data = dati)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

```
-1.4446  -0.3512  -0.1348   0.1820   1.8708
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.39931    0.65626   3.656 0.000287
***
ETA          0.04077    0.01043   3.909 0.000107
***
NrVasiMal    0.25645    0.14993   1.710 0.087894
.
TnIng        0.02968    0.01161   2.557 0.010885
*
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for Gamma family taken to
be 0.3033121)
    Null deviance: 118.72  on 444  degrees of
freedom
Residual deviance: 109.38  on 441  degrees of
freedom
  (119 observations deleted due to missingness)
AIC: 2095.7

Number of Fisher Scoring iterations: 6


fit_gauss_log_1<-
glm(GGDEG~ETA+SESSO+BMI+OBE+FAM+IXT+FUMO+DIS+DIAB
+IRC+BPCO+WHY+PreIMA+PreAng+PrePTCA+PreCABG+ICD+A
ngPreIMA+PreSTROKE+NrVasiMal+TnIng,
family=gaussian(link=log),data=dati)
 summary(fit_gamma_log_1)
…
> fit_gauss_log_12<-
glm(GGDEG~ETA+SESSO+OBE+DIS+DIAB+IRC+PreCABG+AngP
reIMA+NrVasiMal+TnIng,
family=gaussian(link=log),data=dati)
> summary(fit_gauss_log_12)
> fit_gauss_log_13<-
glm(GGDEG~ETA+SESSO+DIS+DIAB+IRC+PreCABG+AngPreIM
A+NrVasiMal+TnIng,
family=gaussian(link=log),data=dati)
> summary(fit_gauss_log_13)
> fit_gauss_log_14<-
glm(GGDEG~ETA+SESSO+DIS+DIAB+IRC+PreCABG+NrVasiMa
l+TnIng, family=gaussian(link=log),data=dati)
> summary(fit_gauss_log_14)
> fit_gauss_log_15<-
glm(GGDEG~ETA+SESSO+DIS+DIAB+IRC+NrVasiMal+TnIng,
family=gaussian(link=log),data=dati)
> summary(fit_gauss_log_15)
```

```
> fit_gauss_log_16<-
glm(GGDEG~ETA+SESSO+DIS+IRC+NrVasiMal+TnIng,
family=gaussian(link=log),data=dati)
> summary(fit_gauss_log_16)
> fit_gauss_log_17<-
glm(GGDEG~ETA+SESSO+IRC+NrVasiMal+TnIng,
family=gaussian(link=log),data=dati)
> summary(fit_gauss_log_17)
> fit_gauss_log_18<-
glm(GGDEG~ETA+SESSO+IRC+TnIng,
family=gaussian(link=log),data=dati)
> summary(fit_gauss_log_18)
> fit_gauss_log_19<-glm(GGDEG~ETA+SESSO+TnIng,
family=gaussian(link=log),data=dati)
> summary(fit_gauss_log_19)
> fit_gauss_log_20<-glm(GGDEG~ETA+TnIng,
family=gaussian(link=log),data=dati)
> summary(fit_gauss_log_20)


> fitnormlog<-fit_gauss_log_20
> fit_bin_log_1<-
glm(GGDEG~ETA+SESSO+BMI+OBE+FAM+IXT+FUMO+DIS+DIAB
+IRC+BPCO+WHY+PreIMA+PreAng+PrePTCA+PreCABG+ICD+A
ngPreIMA+PreSTROKE+NrVasiMal+TnIng,
family=binomial(link=logit),data=dati)
> fit_pois_log_1<-
glm(GGDEG~ETA+SESSO+BMI+OBE+FAM+IXT+FUMO+DIS+DIAB
+IRC+BPCO+WHY+PreIMA+PreAng+PrePTCA+PreCABG+ICD+A
ngPreIMA+PreSTROKE+NrVasiMal+TnIng,
family=poisson(link=log),data=dati)
> summary(fit_pois_log_1)

> fit_pois_log_2<-
glm(GGDEG~ETA+SESSO+BMI+OBE+FAM+IXT+FUMO+DIS+DIAB
+IRC+BPCO+WHY+PreAng+PrePTCA+PreCABG+ICD+AngPreIM
A+PreSTROKE+NrVasiMal+TnIng,
family=poisson(link=log),data=dati)
> summary(fit_pois_log_2)

> fit_pois_log_3<-
glm(GGDEG~ETA+SESSO+BMI+OBE+FAM+IXT+FUMO+DIS+DIAB
+IRC+BPCO+WHY+PreAng+PreCABG+ICD+AngPreIMA+PreSTR
OKE+NrVasiMal+TnIng,
family=poisson(link=log),data=dati)
> summary(fit_pois_log_3)

> fit_pois_log_4<-
glm(GGDEG~ETA+SESSO+BMI+OBE+IXT+FUMO+DIS+DIAB+IRC
+BPCO+WHY+PreAng+PreCABG+ICD+AngPreIMA+PreSTROKE+
NrVasiMal+TnIng,
family=poisson(link=log),data=dati)
```

```
> summary(fit_pois_log_4)


> fit_pois_log_5<-
glm(GGDEG~ETA+SESSO+OBE+IXT+FUMO+DIS+DIAB+IRC+BPC
O+WHY+PreAng+PreCABG+ICD+AngPreIMA+PreSTROKE+NrVa
siMal+TnIng, family=poisson(link=log),data=dati)
> summary(fit_pois_log_5)


> fit_pois_log_6<-
glm(GGDEG~ETA+SESSO+OBE+IXT+FUMO+DIS+DIAB+IRC+WHY
+PreAng+PreCABG+ICD+AngPreIMA+PreSTROKE+NrVasiMal
+TnIng, family=poisson(link=log),data=dati)
> summary(fit_pois_log_6)


> fit_pois_log_7<-
glm(GGDEG~ETA+SESSO+OBE+IXT+FUMO+DIS+DIAB+IRC+WHY
+PreAng+PreCABG+ICD+AngPreIMA+NrVasiMal+TnIng,
family=poisson(link=log),data=dati)
> summary(fit_pois_log_7)

> fit_pois_log_8<-
glm(GGDEG~ETA+SESSO+OBE+IXT+FUMO+DIS+DIAB+IRC+WHY
+PreCABG+ICD+AngPreIMA+NrVasiMal+TnIng,
family=poisson(link=log),data=dati)
> summary(fit_pois_log_8)


> fit_pois_log_9<-
glm(GGDEG~ETA+SESSO+OBE+FUMO+DIS+DIAB+IRC+WHY+Pre
CABG+ICD+AngPreIMA+NrVasiMal+TnIng,
family=poisson(link=log),data=dati)
> summary(fit_pois_log_9)

> fit_pois_log_10<-
glm(GGDEG~ETA+SESSO+OBE+FUMO+DIS+DIAB+IRC+PreCABG
+ICD+AngPreIMA+NrVasiMal+TnIng,
family=poisson(link=log),data=dati)
> summary(fit_pois_log_10)


> fit_pois_log_11<-
glm(GGDEG~ETA+SESSO+OBE+DIS+DIAB+IRC+PreCABG+ICD+
AngPreIMA+NrVasiMal+TnIng,
family=poisson(link=log),data=dati)
> summary(fit_pois_log_11)

> fit_pois_log_12<-
glm(GGDEG~ETA+SESSO+OBE+DIS+DIAB+IRC+PreCABG+AngP
```

```
reIMA+NrVasiMal+TnIng,
family=poisson(link=log),data=dati)
> summary(fit_pois_log_12)


> fit_pois_log_13<-
glm(GGDEG~ETA+SESSO+DIS+DIAB+IRC+PreCABG+AngPreIM
A+NrVasiMal+TnIng,
family=poisson(link=log),data=dati)
> summary(fit_pois_log_13)



> fit_pois_log_14<-
glm(GGDEG~ETA+SESSO+DIS+DIAB+IRC+PreCABG+NrVasiMa
l+TnIng, family=poisson(link=log),data=dati)
> summary(fit_pois_log_14)


> fit_pois_log_15<-
glm(GGDEG~ETA+SESSO+DIS+DIAB+IRC+NrVasiMal+TnIng,
family=poisson(link=log),data=dati)
> summary(fit_pois_log_15)


> fit_pois_log_16<-
glm(GGDEG~ETA+SESSO+DIS+DIAB+NrVasiMal+TnIng,
family=poisson(link=log),data=dati)
> summary(fit_pois_log_16)


> fit_pois_log_17<-
glm(GGDEG~ETA+SESSO+DIS+NrVasiMal+TnIng,
family=poisson(link=log),data=dati)
> summary(fit_pois_log_17)


> fit_pois_log_18<-
glm(GGDEG~ETA+SESSO+NrVasiMal+TnIng,
family=poisson(link=log),data=dati)
> summary(fit_pois_log_18)

> fitpoislog<-fit_pois_log_18
> fit_pois_sqrt_1<-
glm(GGDEG~ETA+SESSO+BMI+OBE+FAM+IXT+FUMO+DIS+DIAB
+IRC+BPCO+WHY+PreIMA+PreAng+PrePTCA+PreCABG+ICD+A
ngPreIMA+PreSTROKE+NrVasiMal+TnIng,
family=poisson(link=sqrt),data=dati)
> summary(fit_pois_sqrt_1)
> fit_pois_id_1<-
glm(GGDEG~ETA+SESSO+BMI+OBE+FAM+IXT+FUMO+DIS+DIAB
+IRC+BPCO+WHY+PreIMA+PreAng+PrePTCA+PreCABG+ICD+A
```

```
ngPreIMA+PreSTROKE+NrVasiMal+TnIng,
family=poisson(link=identity),data=dati)
> summary(fit_pois_id_1)

> summary(fitgammalog)

> summary(fitgammaid)

> names(fitgammaid)

> resgammalog <- resid(fitgammalog)
> resstgammalog <- rstandard(fitgammalog)
> resgammaid <- resid(fitgammaid)
> resstgammaid <- rstandard(fitgammaid)
> resgaussid <- resid(fitnormid)
> resstgaussid <- rstandard(fitnormid)
> resgausslog <- resid(fitnormlog)
> resstgausslog <- rstandard(fitnormlog)
> respoislog <- resid(fitpoislog)
> resstpoislog <- rstandard(fitpoislog)
> fitgammalog.val<-fitted(fitgammalog)
> fitgammaid.val<-fitted(fitgammaid)
> fitgaussid.val<-fitted(fitnormid)
> fitgausslog.val<-fitted(fitnormlog)
> fitpoislog.val<-fitted(fitpoislog)
> plot(resgammalog)
> plot(resgammaid)
> plot(resgaussid)
> plot(resgausslog)
> plot(resstgausslog)
> plot(respoislog)
> plot(resgaussid)
> plot(resgaussid,main="Residui dal modello
gaussiano")
> pairs(dati)
> hatmu=predict(fitgammalog,type="response")
> hatmu1=predict(fitgammalog,type="response")
> names(dati)


> par(mfrow=c(2,2))
> plot(fitgammalog)
> par(mfrow=c(2,2))
> plot(fitgammaid)
> par(mfrow=c(2,2))
> plot(fitnormid)
> par(mfrow=c(2,2))
> plot(fitnormlog)

> par(mfrow=c(2,2))
> plot(resgaussid,main="GLM: Normale")
```

```
> plot(resgausslog,main="GLM: Normale e legame
log")
> plot(resgammalog,main="GLM: Gamma e legame
log")
> plot(resgammaid,main="GLM: Gamma")


#Kaplan-Meier
 library(survival)
 ucicfit <- survfit(Surv(GGDEG, DEC == 1), data =
dati)
 plot(ucicfit)

 ucicfit <- survfit(Surv(GGDEG, DEC == 1), data =
dati)
 options(survfit.print.mean = TRUE)
 summary(ucicfit)

 plot(ucicfit,xlab="giorni degenza UCIC per
SCA",ylab="% ",main="Kaplan-Meier" )
 fit.bysex <- survfit(Surv(GGDEG, DEC == 1) ~
SESSO, data = dati)

 plot(fit.bysex, conf.int = TRUE, col = c("pink",
"blue"), lty = 1:2, legend.text = c("F",  "M"),
xlab="giorni degenza UCIC",ylab="%
",main="Kaplan-Meier")
 survdiff(Surv(GGDEG, DEC == 1) ~ SESSO, data =
dati)

 fit.bydia <- survfit(Surv(GGDEG, DEC) ~ WHY,
data = dati)
 survdiff(Surv(GGDEG, DEC == 1) ~ WHY, data =
dati)


Caricamento del pacchetto survival:
 library(survival)

 summary(dati)

#Tasso di mortalità UCIC con diagnosi di SCA:
 13/585

 fit_cox1 <- coxph(Surv(GGDEG, DEC) ~
ETA+SESSO+BMI+OBE+FAM+IXT+FUMO+DIS+DIAB+IRC+BPCO+
WHY+PreIMA+PreAng+PrePTCA+PreCABG+ICD+PM+AngPreIM
A+PreSTROKE+TnIng)


 ucic.out  <-  rsf(Survrsf(GGDEG,  DEC)~.,  data  =
dati, …..)
```

```
 print(ucic.out)


#Grafici quantile-quantile

 sigma2<-sum(GGDEG^(2))/length(GGDEG)-
(mean(GGDEG))^(2)
 s.hat<-sigma2/mean(GGDEG)
 a.hat<-(mean(GGDEG))^(2)/sigma2
 x<-qgamma(ppoints(GGDEG),
shape=a.hat,scale=s.hat)
 qqplot(x,GGDEG,xlab="quantili gamma",
ylab="quantili empirici")
 title(main="grafico quantilico")
 yy <- quantile(GGDEG, c(0.25, 0.75))
 xx<-
qgamma(c(0.25,0.75),shape=a.hat,scale=s.hat)
 slope<-diff(yy)/diff(xx)
 int<-yy[1]-slope*xx[1]
 abline(int,slope)

qqnorm(GGDEG)
qqline(GGDEG)


#Analisi della sopravvivenza con il modello di
#COX
 dati<-read.csv2(file.choose(), header = TRUE,
sep = ";", quote="\"", dec=",",fill=TRUE)
 attach(dati)
 library(survival)
 library(randomSurvivalForest)

 attach(dati)
 library(survival)
Carico il pacchetto richiesto: splines
 fit_cox1 <- coxph(Surv(GGDEG, DEC) ~
ETA+SESSO+BMI+OBE+FAM+IXT+FUMO+DIS+DIAB+IRC+BPCO+
WHY+PreIMA+PreAng+PrePTCA+PreCABG+ICD+PM+AngPreIM
A+PreSTROKE+TnIng)
Warning message:
In coxph(Surv(GGDEG, DEC) ~ ETA + SESSO + BMI +
OBE + FAM + IXT +  :
  X matrix deemed to be singular; variable 4 16
 fit_cox1 <- coxph(Surv(GGDEG, DEC) ~
ETA+SESSO+BMI+OBE+FAM+IXT+FUMO+DIS+DIAB+IRC+BPCO+
WHY+PreIMA+PreAng+PreCABG+ICD+PM+AngPreIMA+PreSTR
OKE+TnIng)
Warning message:
In coxph(Surv(GGDEG, DEC) ~ ETA + SESSO + BMI +
OBE + FAM + IXT +  :
  X matrix deemed to be singular; variable 4 15
```

```
 fit_cox1 <- coxph(Surv(GGDEG, DEC) ~
ETA+SESSO+BMI+OBE+FAM+IXT+FUMO+DIS+DIAB+IRC+BPCO+
WHY+PreIMA+PreAng+ICD+PM+AngPreIMA+PreSTROKE+TnIn
g)

 fit_cox1 <- coxph(Surv(GGDEG, DEC) ~
ETA+SESSO+BMI+FAM+IXT+FUMO+DIS+DIAB+IRC+BPCO+WHY+
PreIMA+PreAng+ICD+PM+AngPreIMA+PrePTCA+PreSTROKE+
TnIng)

 fit_cox2 <- coxph(Surv(GGDEG, DEC) ~
ETA+SESSO+BMI+FAM+IXT+FUMO+DIS+DIAB+IRC+BPCO+WHY+
PreIMA+PreAng+ICD+PM+AngPreIMA+PrePTCA+TnIng)
 fit_cox2

fit_cox2 <- coxph(Surv(GGDEG, DEC) ~
ETA+SESSO+BMI+FAM+IXT+FUMO+DIS+DIAB+IRC+BPCO+WHY+
PreIMA+PreAng+ICD+PM+AngPreIMA+PrePTCA+TnIng)

fit_cox3 <- coxph(Surv(GGDEG, DEC) ~
ETA+SESSO+BMI+FAM+IXT+FUMO+DIS+DIAB+IRC+BPCO+WHY+
PreIMA+PreAng+ICD+PM+PrePTCA+TnIng)
 fit_cox3

 fit_cox4 <- coxph(Surv(GGDEG, DEC) ~
ETA+SESSO+BMI+FAM+FUMO+DIS+DIAB+IRC+BPCO+WHY+PreI
MA+PreAng+ICD+PM+PrePTCA+TnIng)
 fit_cox4

 fit_cox5 <- coxph(Surv(GGDEG, DEC) ~
ETA+SESSO+BMI+FAM+FUMO+DIS+DIAB+IRC+BPCO+WHY+PreI
MA+ICD+PM+PrePTCA+TnIng)
 fit_cox5

 fit_cox6 <- coxph(Surv(GGDEG, DEC) ~
ETA+SESSO+BMI+FAM+FUMO+DIAB+IRC+BPCO+WHY+PreIMA+I
CD+PM+PrePTCA+TnIng)
 fit_cox6

 fit_cox7 <- coxph(Surv(GGDEG, DEC) ~
ETA+SESSO+BMI+FAM+DIAB+IRC+BPCO+WHY+PreIMA+ICD+PM
+PrePTCA+TnIng)
 fit_cox7

 fit_cox8 <- coxph(Surv(GGDEG, DEC) ~
ETA+BMI+FAM+DIAB+IRC+BPCO+WHY+PreIMA+ICD+PM+PrePT
CA+TnIng)
 fit_cox8
```

```
 fit_cox9 <- coxph(Surv(GGDEG, DEC) ~
ETA+BMI+FAM+DIAB+IRC+BPCO+WHY+PreIMA+ICD+PM+TnIng
)
 fit_cox9

 fit_cox10 <- coxph(Surv(GGDEG, DEC) ~
ETA+BMI+FAM+DIAB+IRC+BPCO+WHY+PreIMA+ICD+PM)
 fit_cox10

 fit_cox11 <- coxph(Surv(GGDEG, DEC) ~
ETA+BMI+FAM+IRC+BPCO+WHY+PreIMA+ICD+PM)
 fit_cox11

 fit_cox12 <- coxph(Surv(GGDEG, DEC) ~
+BMI+FAM+IRC+BPCO+WHY+PreIMA+ICD+PM)
 fit_cox12

 fit_cox13 <- coxph(Surv(GGDEG, DEC) ~
+BMI+FAM+IRC+BPCO+WHY+PreIMA+ICD)
 fit_cox13

 fit_cox14 <- coxph(Surv(GGDEG, DEC) ~
+BMI+FAM+IRC+WHY+BPCO+PreIMA)
 fit_cox14

 fit_cox15 <- coxph(Surv(GGDEG, DEC) ~
+BMI+FAM+IRC+BPCO+PreIMA)
 fit_cox15

 fit_cox16 <- coxph(Surv(GGDEG, DEC) ~
+BMI+IRC+BPCO+PreIMA)
 fit_cox16

 fit_cox17 <- coxph(Surv(GGDEG, DEC) ~
+BMI+IRC+PreIMA)
 fit_cox17

 fit_cox18 <- coxph(Surv(GGDEG, DEC) ~ +BMI+IRC)
 fit_cox18


plot(survfit(fit_cox18),xlab='Giorni',ylab='Frazi
one di sopravvivenza',main='Analisi della
sopravvivenza attraverso la regressione di COX')
#grafici di diagnostica per il mod. di Cox
 cox.zph(fit_cox19)
 par(mfrow=c(2,1))
 for (j in 1:2){
+ plot(dfbeta[,j],
ylab=names(coef(fit_cox19))[j])
+ abline(h=0,lty=2) }
```

```
#Random Survival Forest
fit_RSF<-rsf(Survrsf(GGDEG,DEC)~
ETA+SESSO+BMI+OBE+FAM+IXT+FUMO+DIS+DIAB+IRC+BPCO+
WHY+PreIMA+PreAng+PrePTCA+PreCABG+ICD+PM+AngPreIM
A+PreSTROKE+TnIng, data=dati, ntree=10000,
proximity=TRUE, forest=TRUE)
#RSF grafici
 plot.error(fit_RSF)
 plot.ensemble(fit_RSF)
 plot.error(fit_RSF)
plot.variable(fit_RSF,    predictorNames    =    c(
"PreIMA"))
 plot.variable(fit_RSF,      plots.per.page      =
3,,predictorNames = c( "TnIng", "FumoSi","BMI"))
```

Bibliografia

1. AAVV (2007), Linee guida per la diagnosi e il trattamento delle sindromi coronariche acute senza sopraslivellamento del tratto ST, Giornale italiano di cardiologia 8 (10): 599-675
2. Azienda Ospedaliera di Padova (2007), Il Bilancio Sociale di Mandato 2003-2007, Padova
3. Azzalini,A.(2001), Inferenza Statistica: Una presentazione basata sul concetto di verosimiglianza (2° edizione), Milano, Sprinter-Verlag Italia
4. Azzalini A. e Scarpa B. (2004), Analisi dei dati e data mining, Milano, Sprinter-Verlag Italia
5. Breiman L., Random Forest, http://oz.berkeley.edu/users/breiman/randomforest2001.pdf, Gennaio 2001
6. Carla I. (2005); Compendio di Statistica; Napoli, Esselibri
7. Ceri S., Mandrioli D. e Sbattella L.(2004), Informatica: arte e mestiere, Milano, McGraw-Hill
8. Cox, D.R. e Oakes D. (1996), Analysis of Survival Data, Cambridge, Chapman & Hall
9. Crivellari F. (2006), Analisi statistica dei dati con R, Milano, Apogeo
10. Fantazzini D. Figini S.(2007), Random Survival Forest models for SME Credit Risk Measurement, http://www.unipv.it/dipstea/workingpapers/47.pdf
11. Fox J., Cox (2002), Proportional-Hazards Regression for Survival Data, http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-cox-regression.pdf
12. Ishwaran H. e Kogalur U.B., “Random Survival Forest for R”, R News, Vol.7/2, pp. 25-31
13. Ishwaran H. e Kogalur U.B.(2007), The randomSurvivalForest Package., http://cran.r-project.org/doc/packages/randomSurvivalForest.pdf, 7 Settembre 2007
14. Marino P. (2001), Terapia intensiva, Milano, Masson
15. McCullagh e Nelder J.A. (1995), Generalized Linear Models second editin, Cambridge, Chapman & Hall
16. Thereau T.M. e Grambsch P.M. (2000), Modelling Survival Data: Extending the Cox Model, USA, Sprinter-Verlag