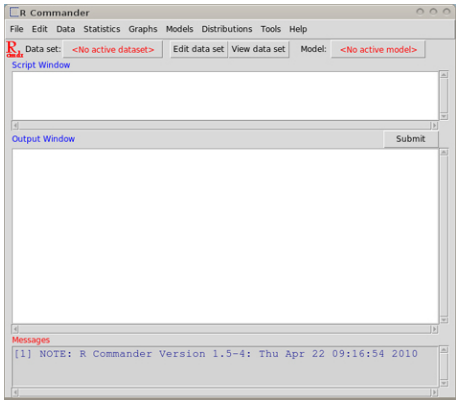
R Commander: Quattro domande di statistica (e quattro risposte)

Massimo Borelli
Dipartimento di Matematica e Geoscienze
Dipartimento di Scienze della Vita
Università degli Studi di Trieste
Via A. Valerio 12/1, 34127 Trieste, Italia
E-mail: `borelli@units.it`

Gennaio 2012

Sommario

In questo Quaderno Didattico vengono presentati alcuni esempi di analisi statistica inferenziale e di modellazione statistica per mezzo dell'interfaccia grafica denominata *R Commander*, implementata nel pacchetto statistico open source R.



Indice

1 Introduzione **3**
1.1 Convenzioni tipografiche 4

2 .. Vorrei capire se c'è *differenza tra i gruppi* .. **5**
2.1 Esercizio 11

3 .. Vorrei vedere se queste variabili *correlano tra loro* .. **11**
3.1 Esercizio 22

4 .. Dovrei *fare una multivariata* per capire quali fattori influenzano l'outcome .. **22**
4.1 Esercizio 26
4.2 Esercizio 27

5 .. Vorrei capire se *la sopravvivenza è diversa* .. **27**
5.1 Esercizio 32

Riferimenti bibliografici **32**

1 Introduzione

L'esperienza didattica e di ricerca di questi anni ci ha mostrato che nelle Scienze della Vita (biologiche o biotecnologiche, di medicina umana od animale) prima o poi capita di imbattersi in alcune 'domande statistiche':

1. .. vorrei capire se ci sono 'differenze tra i gruppi' ..
2. .. vorrei vedere se queste variabili 'correlano tra loro' ..
3. .. dovrei 'fare una multivariata' per capire quali fattori influenzano l'outcome ..
4. .. vorrei capire se 'la sopravvivenza è diversa' nei trattamenti ..

Per rispondere in maniera accettabile a queste domande lo studioso deve necessariamente affidarsi ad un software statistico; in tal senso, il software R (`http://www.r-project.org/`) è, a nostro modo di vedere, ottimo. Ma, se ci è concessa questa immagine, R è un po' come ricevere in regalo un jumbo jet: è vero che si vola veloci, ma non è banale mettersi ai comandi e pilotarlo se non si ha qualcuno che ti insegni a farlo. R infatti è un software creato con un'*interfaccia carattere*, il che si rivela poco attraente per chi abbia bisogno di fare, senza perder troppo tempo, qualche calcolo statistico o qualche grafico.

Fortunatamente negli anni sono state sviluppate alcune interfacce grafiche aggiuntive: qui vogliamo presentare quella che si chiama *R Commander* e che si utilizza installando il pacchetto denominato `Rcmdr` direttamente dal menu di R.

Per trattare le quattro istanze che abbiamo enunciato, utilizzeremo tre dataset che abbiamo pubblicato in rete. Nel primo dataset denominato *studentiannoscorso* ci sono semplici informazioni che abbiamo ottenuto chiedendo a 65 studenti di rivelarci il loro *anno* di nascita, il *genere* (*f* oppure *m*), la loro *statura* espressa in centimetri, il *peso* in chilogrammi, il numero di *scarpe* che indossano (spiegando ai burloni che non è ammessa la risposta '2'..), se avessero (*SI* o *NO*) il vizio del *fumo* e se praticassero in una qualche misura lo *sport* (stratificando le risposte di questa variabile in tre livelli possibili, *poco*, *saltuario* e *tanto*). Inoltre, durante il giorno dell'esame di Statistica, abbiamo misurato la loro *frequenza* cardiaca e la loro pressione arteriosa (*massima* e *minima*). Abbiamo infine preso nota della media aritmetica dei voti presenti sul loro libretto sino all'anno precedente (*mediaold*) e della nuova media aggiornata con gli esami dell'anno in corso (*medianew*).

Nel dataset *surgery* invece la variabile di interesse è la sopravvivenza (*outcome*) dei pazienti, individuati ciascuno in maniera univoca da un identificatore *id*. Le covariate che potenzialmente possono influire sul decorso

postoperatorio sono l'età del paziente (*age*), la stadiazione (*ajcc*) del tumore, la tecnica di intervento chirurgico adottata (*access*) e la regione intestinale (*region*) interessata dalla neoplasia.

Da ultimo, nel dataset *breastcancer* la variabile di interesse è la quantità dei giorni di degenza (*los*) postoperatoria, in due diversi interventi (*surgery*) effettuati con due tipi di bisturi (*treatment*) su pazienti di una determinata età (*age*), nei quali sono stati misurati i volumi di drenaggio toracico postoperatorio (*drainage*). La variabile *discharge* è un indicatore logico che vale 1 se la paziente è stata dimessa dal reparto, 0 altrimenti.

Il presente Quaderno Didattico riassume ed integra alcune lezioni che sono state proposte nel periodo novembre 2011 - gennaio 2012 agli studenti del corso di laurea magistrale in Biotecnologie Mediche della Facoltà di Medicina di Trieste, ai medici delle Scuole di Specialità della Facoltà di Medicina di Trieste, agli studenti del corso di laurea magistrale in Biotecnologie animali della Facoltà di Medicina Veterinaria dell'Università di Bologna.

Si parte dunque dal presupposto che chi legge questo fascicolo 'sappia già qualcosa di statistica', o almeno abbia dei vaghi ricordi sul significato delle parole *media*, *mediana*, *deviazione standard*, *retta di regressione*, *p-value*, *test di Student* ... Se questi ricordi sono troppo vaghi, per rinfrescare la memoria stando davanti al computer, il lettore potrebbe per esempio fare una ricerca su Google digitando le parole chiave `ccforum statistics` ed accedere ad un'ottima serie di reviews [1] pubblicate dalla rivista *Critical Care Medicine* e dedicate alla statistica medica.

Diamo infine per scontato che nel computer che vi ritrovate davanti a voi in questo momento qualcuno vi abbia già installato R ed *R Commander*. Se così non fosse, e non sapete come fare a farlo, andate su Google e digitate le parole chiave `massimo borelli medici`: troverete lì da qualche parte due *guide illustrate per l'installazione*.

1.1 Convenzioni tipografiche

In queste note abbiamo cercato di utilizzare alcune convenzioni tipografiche:

- le DICITURE IN MAIUSCOLETTO riguardano voci che sono presenti direttamente nei menu o nelle finestre di dialogo di *R Commander*;
- i comandi a `carattere isospaziato` riguardano per esempio l'output che viene fornito in colore rosso e blu dalla finestra dei risultati di *R Commander*, oppure i termini che si devono digitare da tastiera nelle caselle di dialogo e di inserimento dati.
- i nomi dei dataset e le variabili contenute in essi sono rappresentati in *carattere corsivo*.

Inoltre, con la scrittura:

Comando1 > Comando2 > Comando3

intendiamo indicare il fatto che in *R Commander* per compiere l'azione di cui vogliamo parlarvi è sufficiente fare clic con il mouse sulla barra del menu in alto, dove è riportato il Comando1. Apparirà un menu a cascata in cui sarà presente la scelta Comando2, e cliccando su di essa apparirà una finestra di dialogo denominata Comando3.

2 .. Vorrei capire se c'è *differenza tra i gruppi* ..

In questa sezione iniziamo un discorso che si completerà appena nelle prossime due sezioni successive. Scopriremo infatti che determinare differenze tra gruppi in un certo senso equivale a valutare la correlazione lineare che intercorre tra due variabili: entrambe queste istanze infatti conducono allo studio dei **modelli lineari**, i quali si possono ulteriormente ampliare nei **modelli lineari generalizzati**, che sono gli strumenti pertinenti alla terza istanza. Precisiamo che in questa sezione la variabile risposta che ci interessa è il *peso* degli studenti: vogliamo infatti cercare di capire quali possano essere i fattori che in senso statistico hanno un effetto su di esso.

Dopo aver lanciato l'interfaccia grafica *R Commander* direttamente dalla console di R con il comando `library(Rcmdr)`, importiamo dal web il dataset `studentiannoscorso`. A tale proposito, è sufficiente scegliere il percorso:

Dati > Importa dati > da file di testo o blocco appunti o URL

e attivare il pulsante Internet URL, incollando nella successiva finestra di dialogo l'indirizzo:

`http://www.dmi.units.it/~borelli/dataset/studentiannoscorso.txt`

Possiamo provare a visualizzare con un boxplot la Variabile *peso* degli studenti, con il comando

Grafici > Grafico a scatola e baffi ..

ed usando il pulsante Graficare per gruppi impostando il fattore *sport*. Otteniamo i tre boxplot raffigurati in Figura 1. Quando si vuole 'vedere se ci siano differenze tra i gruppi' molti studiosi non esitano a ricorrere all'**Anova**. Proviamo farlo anche noi, con il comando:

Statistiche > Medie > Anova ad una via ..

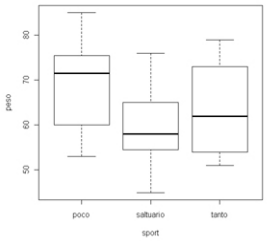


Figura 1: I tre boxplot del peso stratificato rispetto all'attività sportiva.

Figura 2: La finestra di dialogo per impostare un'Anova.

Come si vede nella Figura 2, si apre una finestra grafica che ci chiede di scegliere un GRUPPO (*sport*) ed una VARIABILE DI RISPOSTA (il *peso*). Il software fornisce un output molto esaustivo, ma che non è di immediata comprensione per tutti i lettori:

```
> anovamodel.2 <- aov(peso ~ sport, data=dataset)

> summary(anovamodel.2)
            Df sum sq mean sq f value   pr(>f)
port         2 1020.4  510.20  6.4876 0.002765 **
Residuals   62 4875.8   78.64
---
Signif. Codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> numsummary(dataset$peso , groups=dataset$sport,
statistics=c('mean', 'sd'))
              Mean       sd data:n
poco      69.10000 9.112744     20
saltuario 59.43478 8.462893     23
tanto     62.72727 9.056341     22
```

Per fare un po' di chiarezza, ci conviene ripartire dai ricordi del celebre test t di Student, il test statistico che va a vedere se vi sia differenza in senso statistico tra le medie di due campioni. Vogliamo infatti convincerci che l'Anova non è altro che una generalizzazione del test t qualora si sia in

presenza non di due, ma di tre (o più) gruppi. A tale proposito, usiamo il comando:

STATISTICHE > MEDIE > T- TEST PER CAMPIONI INDIPENDENTI..

ed eseguiamo il test t relativamente al *peso* suddiviso per il *genere* degli studenti:

```
> t.test(peso~genere, alternative='two.sided', conf.level=.95,
+   var.equal=false, data=dataset)

Welch two sample t-test

Data:  peso by genere
T = -8.0893, df = 62.656, p-value = 2.583e-11
Alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -16.94397 -10.23027
Sample estimates:
Mean in group f mean in group m
       56.62500        70.21212
```

Se invece del *t-test* avessimo fatto una 'one-way anova' *peso* vs. *genere* come avevamo poc'anzi fatto per *sport*, avremmo ottenuto il seguente output:

```
> anovamodel.3 <- aov(peso ~ genere, data=dataset)

> summary(anovamodel.3)
            Df sum sq mean sq f value    pr(>f)
Genere       1 2999.2 2999.20  65.222 2.627e-11 ***
Residuals   63 2897.0   45.98
---
Signif. Codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> numsummary(dataset$peso , groups=dataset$genere,
statistics=c('mean', 'sd'))
      Mean       sd data:n
F 56.62500 6.409418     32
M 70.21212 7.122840     33
```

Confrontando i due output ci accorgiamo che il messaggio è sempre lo stesso: le ragazze hanno un peso medio di circa 56.6 kg, i ragazzi di circa 70.2 kg, e queste due medie sono significativamente diverse tra loro; lo rivela il p-value, che è un numero praticamente uguale a zero (precisamente 2.6 per dieci alla meno undici). O, se si preferisce, lo rivelano le tre stelline, ovvero

l'intervallo di fiducia al 95 per cento, che è compreso tra -16.9 e -10.2 kg (e lo zero non appartiene all'intervallo).

Avremmo forse potuto eseguire il t-test *peso* vs. *sport*? No, il software non lo consente, proprio perché non vi sono due, ma tre **livelli** (*poco*, *saltuario* e *tanto*) nel **fattore** *sport*. Ma non è in effeti una carenza del software, è proprio la questione matematica che è diversa. Nel test t di Student si va a considerare la differenza tra due medie; nella Anova (***an**alysis **of** **va**riance*) l'approccio è ben diverso: si va a vedere se la dispersione attorno alla media generale sia o non sia 'differente' considerando la presenza o l'assenza del 'gruppo'. Se infatti calcoliamo le statistiche descrittive del peso:

Statistiche > Informazioni riassuntive > Indici riassuntivi numerici..

```
> numsummary(dataset[,'peso'], statistics=c('mean', 'sd', 'quantiles'),
+   quantiles=c(0,.25,.5,.75,1))
     Mean       sd 0% 25% 50% 75% 100%  n
 63.52308 9.598352 45  56  62  73   85 65
```

vediamo che la dispersione attorno alla media, quantificata dalla deviazione standard (ossia la radice quadrata della varianza), vale circa 9.6. Invece le statistiche descrittive del peso che si ottengono attivando il pulsante Riassumi per gruppo.. e scegliendo come variabile di raggruppamento il *genere*, mostrano che la dispersione attorno alla media è minore (6.4 per le ragazze, 7.1 per i ragazzi):

```
> numsummary(dataset[,'peso'], groups=dataset$genere,
statistics=c('mean', 'sd', 'quantiles'),
quantiles=c(0,.25,.5,.75,1))
      Mean       sd 0% 25% 50%   75% 100% data:n
F 56.62500 6.409418 45  53  56 58.25   75     32
M 70.21212 7.122840 54  65  72 75.00   85     33
```

Se invece, come per il fattore *fumo*, l'effetto non fosse stato significativo (si verifichi con il t-test per campioni indipendenti che il p-value vale circa 0.53) allora le deviazioni standard sarebbero variate di poco (9.5 per i non fumatori, 10.3 per i fumatori) rispetto al 9.6 generale.

Sarebbe ora il caso di occuparci della *Anova a due vie* per convincersi che si tratta di una sorta di 'doppio test' di Student, ma per il momento non ci addentriamo nella questione. Dobbiamo però ricordare alcuni fatti molto importanti prima di passare alla prossima sezione.

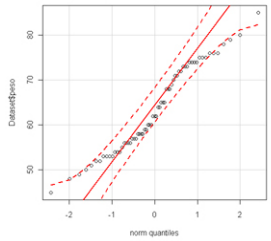


Figura 3: Il grafico quantile-quantile per stabilire la normalità del *peso*.

Innanzitutto, la teoria prescrive che è opportuno applicare il metodo parametrico dell'Anova esclusivamente se i campioni presi in esame provengono da una popolazione distribuita normalmente. Per decidere se i dati di cui ci occupiamo siano o non siano distribuiti normalmente, abbiamo a disposizione dei metodi grafici, come il **qq plot** raffigurato in Figura 3:

Grafici > grafico quantili-quantili..

Il fatto che i punti del grafico in Figura 3 non si adagino lungo la diagonale (*retta interquartile*) ma assumano una forma 'ad S' ci fa presagire che siamo in presenza di dati non distribuiti normalmente. La conferma formale si ha con il test di Shapiro e Wilk:

Statistiche > Informazioni riassuntive > Test di normalità di Shapiro-Wilk..

```
> shapiro.test(dataset$peso)

Shapiro-wilk normality test

Data:  dataset$peso
W = 0.9563, p-value = 0.02207
```

Il basso p-value ottenuto ci fa rifiutare l'ipotesi di normalità. Dunque, già questa violazione ci suggerisce di evitare di effettuare l'Anova, ma ricorrere ad un approccio non parametrico - ad esempio, possiamo considerare il test della somma dei ranghi di Kruskal e Wallis:

Statistiche > Tests non parametrici > Test di Kruskal-Wallis..

```
> kruskal.test(peso ~ sport, data=dataset)

Kruskal-wallis rank sum test

Data:  peso by sport
Kruskal-wallis chi-squared = 10.3988, df = 2, p-value = 0.00552
```

Ancora un'osservazione. Supponendo che invece il test di Shapiro-Wilk avesse confermato l'ipotesi di normalità fornendoci un p-value 'grande' (i.e. maggiore o uguale di 0.05), avremmo comunque potuto affidarci spensieratamente all'Anova per testare differenze in media tra i gruppi? No. La media è un inicie di centralità, ma noi abbiamo bisogno di essere sicuri che un'ulteriore ipotesi matematica sia verificata: ossia che i campioni tra loro siano **omoschedastici**, cioè abbiano - in senso statistico - la medesima dispersione. Questo si può fare in effetti utilizzando parecchi test, ma con *R Commander* di solito si preferisce usare il test di Levene sulle mediane:

Statistiche > Varianze > Test di Levene..

```
> leveneTest(dataset$peso, dataset$sport, center=median)
Levene's test for homogeneity of variance (center = median)
      Df f value pr(>f)
Group  2  0.4052 0.6686
      62
```

Perché è necessario verificare che i campioni non siano tra loro **eteroschedastici** (ossia, come dice Crawley [3], *when variances are different do not compare the means*)? Non riusciamo a mostrarlo direttamente da *R Commander* (e questo non fa onore all'applicazione), ma dobbiamo un momento ricorrere ad R, copiando ed incollando le seguenti istruzioni:

```
www <- 'http://www.dmi.units.it/~borelli/dataset/studentiannoscorso.txt'
dataset <- read.table( www , header = true )
attach(dataset)
shapiro.test(mediaold)$p.value
shapiro.test(medianew)$p.value
wilcox.test(mediaold, medianew, paired = true)$p.value
leveneTest(mediaold, medianew)
```

Le medie attuali dei voti riportati sui libretti, rispetto a quelle dell'anno precedente, non sono distribuite normalmente (in entrambi i casi p-value $<$ 0.01), ma il test di Levene (p-value 0.056) ci impone cautela nel considerare le due distribuzioni omoschedastiche. In questo caso particolare, per uno statistico, c'è imbarazzo a dire se a distanza di un anno gli studenti abbiano o no migliorato il loro rendimento universitario (anche se uno studio di Conover et al. [2] suggerisce di affidarsi al test di Fligner e Killeen per valutare l'omoschedasticità).

2.1 Esercizio

In R sono già presenti per scopo didattico, senza necessità di doverli importare dalla rete, alcuni dataset. Con il comando:

Dati > Dati presenti nei pacchetti > Leggi i dati da un pacchetto caricato..

compare la finestra Leggi dati da un pacchetto. Scegliete a sinistra il Pacchetto `datasets`, e scorrete in basso il menu di destra Set di dati fino ad imbattervi in `chickwts` (se non lo trovate, non disperate: state appena vedendo i dataset che iniziano con la lettera maiuscola, mentre a voi interessa quello scritto con la minuscola). Date Ok per caricare il dataset, che riporta il peso (*weight*) di alcuni polli nutriti con differenti alimenti (*feed*). C'è differenza in senso statistico tra i diversi gruppi? Quale tecnica avete utilizzato, parametrica o non parametrica?

3 .. Vorrei vedere se queste variabili *correlano tra loro* ..

Attenzione! Sgombriamo il campo da una certa confusione: '*correlazione*' ed '*associazione*' sono due nozioni statistiche diverse. Quando abbiamo due variabili di tipo categorico, possiamo andare a vedere se vi sia o meno una **associazione** tra queste, come nel caso del *genere* e del *fumo*. A tale scopo, la cosa migliore è creare una '**tavola due per due**':

Tabelle di contingenza > Tabella a due entrate..

Selezionando la variabile di riga (*genere*), la variabile di colonna (*fumo*) e spuntando, per semplicità, tutte e quattro le checkbox dei Test d'ipotesi, otteniamo un output piuttosto corposo:

```
> .table
      Fumo
Genere no si
```

```
     F 26  6
     M 25  8

> .test <- chisq.test(.table, correct=false)

> .test

Pearson's chi-squared test

Data:  .table
X-squared = 0.29, df = 1, p-value = 0.5902


> .test$expected # expected counts
      Fumo
Genere       no       si
     F 25.10769 6.892308
     M 25.89231 7.107692

> round(.test$residuals^2, 2) # chi-square components
      Fumo
Genere   no   si
     F 0.03 0.12
     M 0.03 0.11

> remove(.test)

> fisher.test(.table)

Fisher's exact test for count data

Data:  .table
P-value = 0.7642
Alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.3598084 5.5771220
Sample estimates:
Odds ratio
  1.379709
```

Il test del Chi Quadrato (p-value 0.59) ed il test esatto di Fisher (p-value 0.76) ci rivelano che non vi è associazione tra le due variabili: le ragazze ed

i ragazzi hanno il dannoso vizio del fumo all'incirca nella medesima proporzione. Per inciso, molti autori consigliano di preferire tra i due test quello esatto di Fisher, stante il fatto che il test del Chi Quadrato è per sua natura un test asintotico (i.e. un test che 'lavora bene' con tabelle 'di grandi numeri'). Dunque, in questo esempio, in gergo diremmo che '*l'odds ratio*[1] (1.38) *non è significativo*'.

Se invece le due variabili non sono dei fattori ma sono delle quantità numeriche, come ad esempio accade nella misura di grandezze fisiche, allora è più corretto parlare di **correlazione**. Anche qui sono possibili due approcci, quello parametrico (raccomandato per dati distribuiti normalmente) e quello non parametrico (adatto per distribuzioni non normali, ovvero per variabili di tipo **ordered factor** - come avviene per esempio nelle scale likert usate nei questionari, o nei codici di gravità attribuiti al triage del pronto soccorso).

Teniamo però a mente che il *peso* dei nostri studenti non è distribuito normalmente. Quindi se si va a cercare la correlazione tra il *peso* e la *statura*, sarebbe forse più indicato ricorrere all'approccio non parametrico, calcolando il coefficiente di correlazione di Spearman (oppure quello di Kendall, che nella pratica vengono percepiti come equivalenti):

STATISTICHE > INFORMAZIONI RIASSUNTIVE > TEST DI CORRELAZIONE ..

Per scegliere le due variabili (*peso* e *statura*) si deve tenere cliccato sulla tastiera del computer il tasto CTRL. Ecco l'output che si ottiene:

```
> cor.test(dataset$peso, dataset$statura, alternative='two.sided',
+   method='spearman')

Spearman's rank correlation rho

Data:  dataset$peso and dataset$statura
S = 12566.91, p-value = 8.248e-12
Alternative hypothesis: true rho is not equal to 0
Sample estimates:
      Rho
0.7253735
```

C'è una correlazione del 73 per cento circa, ed è significativa (p-value $<$ 0.001), nel senso che essa è diversa da zero in senso statistico. Se invece avessimo voluto considerare la più consueta correlazione ρ di Pearson, avremmo ottenuto una stima puntale del 74 per cento (stima ad intervallo da 61 per cento a 84 per cento, con fiducia del 95 per cento):

[1] Per rivedere il significato di odds ratio: `http://ccforum.com/content/8/1/46`

Figura 4: La finestra di dialogo per impostare una retta di regressione. Si badi che la cosiddetta VARIABILE DI RISPOSTA è quella che verrà rappresentata sull'asse verticale delle ordinate.

```
> cor.test(dataset$peso, dataset$statura, alternative='two.sided',
+   method='pearson')

Pearson's product-moment correlation

Data:  dataset$peso and dataset$statura
T = 8.8495, df = 63, p-value = 1.18e-12
Alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6115788 0.8364607
Sample estimates:
      Cor
0.7444353
```

E' interessante notare che le medesime informazioni sarebbero state disponibili se avessimo considerato una **retta di regressione**, come in Figura 4 :

STATISTICHE > FITTA I MODELLI > REGRESSIONE LINEARE..

```
> regmodel.1 <- lm(peso~statura, data=dataset)

> summary(regmodel.1)

Call:
Lm(formula = peso ~ statura, data = dataset)
```

Figura 5: La finestra di dialogo Modelli lineari

```
Residuals:
    Min      1q  median      3q     max
-13.546  -4.209  -1.569   4.431  17.139

Coefficients:
             Estimate std. Error t value Pr(>|t|)
(intercept) -83.89056   16.67708   -5.03 4.34e-06 ***
Statura       0.85392    0.09649    8.85 1.18e-12 ***
---
Signif. Codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.459 on 63 degrees of freedom
Multiple r-squared: 0.5542,adjusted r-squared: 0.5471
F-statistic: 78.31 on 1 and 63 df,  p-value: 1.18e-12
```

Nel complesso output che ci viene fornito, oltre alla pendenza m della retta di regressione $y = mx + q$ (*statura* 0.85) ed alla quota q (*intercept* -83.89) , viene riportato il valore *Multiple R-squared* 0.55, il quale prende il nome di **coefficiente di determinazione** ed altri non è che il coefficiente di correlazione ρ di Pearson (0.744) elevato al quadrato. Non a caso, il p-value del modello di regressione, nonché della pendenza della retta (1.18e-12), coincide con quello ottenuto poco fa dal test di correlazione.

Dunque, ricapitolando: in presenza di due variabili categoriche usiamo le *misure di associazione*; ma con due variabili numeriche ricerchiamo una *correlazione*, la quale appare intimamente legata al calcolo della *retta di regressione*. Cosa dobbiamo fare nella situazione intermedia, quella in cui la variabile risposta è numerica (come ad esempio il *peso*) e la variabile esplicativa è categorica (come nel *genere*, nel *fumo* o nello *sport*)? Proviamo a vedere cosa capita se impostiamo un **modello lineare**:

Statistiche > Fitta i modelli > Modelli lineari..

```
> linearmodel.2 <- lm(peso ~ genere, data=dataset)

> summary(linearmodel.2)

Call:
Lm(formula = peso ~ genere, data = dataset)

Residuals:
    Min      1q  median      3q     max
-16.212  -4.625   0.375   3.788  18.375

Coefficients:
            Estimate std. Error t value Pr(>|t|)
(Intercept)   56.625      1.199  47.237  < 2e-16 ***
genere[t.m]   13.587      1.682   8.076 2.63e-11 ***
---
Signif. Codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.781 on 63 degrees of freedom
Multiple r-squared: 0.5087,adjusted r-squared: 0.5009
F-statistic: 65.22 on 1 and 63 df,  p-value: 2.627e-11
```

Come direbbe Super Pippo: *ta-dah!* Nella sezione precedente, avevamo eseguito il test t relativamente al *peso* versus il *genere* degli studenti ed avevamo trovato proprio lo stesso p-value. Inoltre, il peso medio delle ragazze (56.6) è proprio l'intercetta `(Intercept)` del modello lineare, mentre il peso medio dei ragazzi (70.2) è esattamente uguale alla somma dell'intercetta e della pendenza: $56.6 + 13.6 = 70.2$. Questo accade perché possiamo pensare alla retta $y = mx+q$, dove qui x è una variabile 'muta', di tipo booleano, che vale 0 oppure 1 a seconda del fatto che si consideri il primo o il secondo dei due livelli del fattore *genere*, che sono (in ordine alfabetico) f ed m. Perciò, per una ragazza, $13.6 \cdot 0 + 56.6 = 56.6$, mentre per un ragazzo $13.6 \cdot 1 + 56.6 = 70.2$.

Ora però si deve prestare molta attenzione al corretto significato delle stelline che appaiono accanto alle stime del modello lineare. Calcoliamo il modello lineare *peso* vs. *sport*:

```
> linearmodel.3 <- lm(peso ~ sport, data=dataset)

> summary(linearmodel.3)
```

```
Call:
Lm(formula = peso ~ sport, data = dataset)

Residuals:
     Min       1q   median       3q      max
-16.1000  -7.4348   0.5652   6.9000  16.5652

Coefficients:
                   Estimate std. Error t value Pr(>|t|)
(intercept)          69.100      1.983  34.847  < 2e-16 ***
Sport[t.saltuario]   -9.665      2.711  -3.565 0.000708 ***
Sport[t.tanto]       -6.373      2.740  -2.326 0.023307 *
---
Signif. Codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.868 on 62 degrees of freedom
Multiple r-squared: 0.1731,adjusted r-squared: 0.1464
F-statistic: 6.488 on 2 and 62 df,  p-value: 0.002765
```

L'Anova effettuata nella sezione precedente ci assicurava che c'era differenza tra i gruppi (p-value 0.003), ma non era chiaro se chi facesse *poco* allenamento fosse in media più pesante di chi invece ne fa *tanto*, o di chi lo fa in modo *saltuario*. Uno statistico, con 'un po' d'occhio', considerando le stime dei parametri ed i loro errori standard è in grado di dire chi sia diverso da chi. Ma per far bene le cose, a questo punto, è meglio ritornare alla ANALISI DELLA VARIANZA AD UNA VIA della sezione precedente ed attivare la checkbox CONFRONTO DI MEDIE A COPPIE per ottenere questo lunghissimo output e la Figura 6:

```
> library(multcomp, pos=4)

> anovamodel.4 <- aov(peso ~ sport, data=dataset)

> summary(anovamodel.4)
            Df sum sq mean sq f value   pr(>f)
Sport        2 1020.4  510.20  6.4876 0.002765 **
Residuals   62 4875.8   78.64
---
Signif. Codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> numsummary(dataset$peso , groups=dataset$sport,
statistics=c('mean', 'sd'))
```

```
              Mean       sd data:n
Poco      69.10000 9.112744     20
Saltuario 59.43478 8.462893     23
Tanto     62.72727 9.056341     22

> .pairs <- glht(anovamodel.4, linfct = mcp(sport = 'tukey'))

> summary(.pairs) # pairwise tests

 Simultaneous tests for general linear hypotheses

Multiple comparisons of means: tukey contrasts


Fit: aov(formula = peso ~ sport, data = dataset)

Linear hypotheses:
                       Estimate std. Error t value Pr(>|t|)
Saltuario - poco == 0    -9.665      2.711  -3.565  0.00203 **
Tanto - poco == 0        -6.373      2.740  -2.326  0.05963 .
Tanto - saltuario == 0    3.292      2.645   1.245  0.43154
---
Signif. Codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(adjusted p values reported -- single-step method)


> confint(.pairs) # confidence intervals

 Simultaneous confidence intervals

Multiple comparisons of means: tukey contrasts


Fit: aov(formula = peso ~ sport, data = dataset)

Quantile = 2.4011
95% family-wise confidence level


Linear hypotheses:
                       Estimate lwr      upr
Saltuario - poco == 0   -9.6652 -16.1754  -3.1551
Tanto - poco == 0       -6.3727 -12.9513   0.2059
Tanto - saltuario == 0   3.2925  -3.0574   9.6424
```

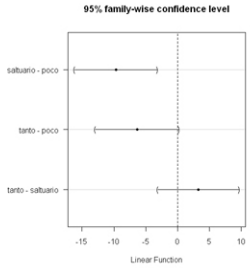


Figura 6:

Il cosiddetto 'post-hoc onesto' di Tukey ci rivela che nel nostro dataset non c'è differenza di *peso* in senso statistico (p-value 0.43) tra chi fa *tanto sport* e chi comunque lo fa in modo *saltuario*. In altri termini, non vi sono tre gruppi distinti di attività sportive, ma due, chi fa poco, e gli altri due gruppi accomunati assieme.

Una cosa che è poco nota tra i non-statistici che fanno ricerca nelle Scienze della Vita, è che di ogni modello lineare è opportuno giudicarne la adeguatezza effettuando la cosiddetta **diagnostica del modello**, che si ottiene con il comando:

Modelli > Grafici > Grafici diagnostici di base

Non abbiamo qui tempo di dilungarci sul significato di ciascuno dei grafici. Ma come sapevamo già in partenza, ci stiamo trascinando dietro il problema della normalità dei residui, che deriva dal fatto che *peso* è un insieme di dati non distribuiti normalmente. Questa è una situazione molto comune, che talvolta può venir modificata con una opportuna trasformazione della variabile; comunemente si va a cercare una funzione di tipo logaritmo, di tipo quadratico, o di tipo esponenziale, che 'migliorino' la asimmetria o quanto meno riescano ad 'attenuare' la possibile plurimodalità dei dati. Direttamente da *R Commander* però non è possibile fare questo tipo di analisi e quindi impartiamo i seguenti comandi in R:

```
library(MASS)
boxcox(peso ~ statura)
```

Siamo alla ricerca di un esponente opportuno cui elevare il *peso* per cercare di renderlo normale. La curva concava visualizzata ci fa intuire che si potrebbe provare con un esponente λ positivo prossimo allo zero. Vediamo in maggior dettaglio:

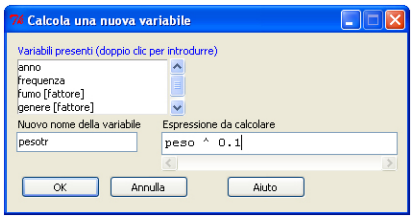


Figura 7:

```
boxcox(peso ~ statura, lambda = seq(0, 0.5, length = 100))
```

Bene: proviamo a scegliere $\lambda = 0.1$ come esponente per trasformare la variabile *peso*:

```
pesotr <- peso^0.1
shapiro.test(pesotr)
```

Il p-value del test di Shapiro-Wilk migliora un pochino (0.04), anche se dobbiamo decidere se vogliamo accontentarci perché non siamo ancora in presenza di una distribuzione di dati che potrebbe essere considerata normale, oppure buttare tutto al mare e usare i metodi non parametrici. Torniamo dunque ad *R Commander* per creare da lì la nuova variabile trasformata *pesotr*:

DATI > GESTIONE VARIABILI DEL SET DI DATI ATTIVO > CALCOLA UNA NUOVA VARIABILE..

Come nuovo nome della variabile digitiamo PESOTR e come espressione da calcolare impostiamo *peso* elevato alla 0.1 , come si vede nella Figura 7. Potete ora controllare con il pulsante VISUALIZZA I DATI di *R Commander* che effettivamente è stata aggiunta al dataset un'ulteriore colonna con la nuova variabile *pesotr*. A questo punto è possibile provare a ripetere le analisi condotte per ottenere dei p-value maggiormente attendibili.

Abbiamo visto dunque che con un modello lineare possiamo eseguire *regressioni* ed *anova*. La **Ancova** (ossia la analisi della *covarianza*), intesa come 'combinazione' di una regressione ed un'Anova, è anche immediata: ad esempio, vogliamo vedere se la *statura* e il *genere* siano predittori significativi del *pesotr*. Possiamo provare a considerare un modello puramente additivo, *statura + genere* (ossia, le rette di regressione per le ragazze e per i ragazzi vengono assunte essere parallele tra loro), oppure un modello con interazione *statura * genere* (ossia, è ammesso che le rette di regressione per le ragazze e per i ragazzi abbiano due pendenze significativamente diverse tra loro).

```
> linearmodel.8 <- lm(pesotr ~ genere*statura, data=dataset)

> summary(linearmodel.8)

Call:
Lm(formula = pesotr ~ genere * statura, data = dataset)

Residuals:
      Min        1q    median        3q       max
-0.034159 -0.010246 -0.001167  0.009049  0.044365

Coefficients:
                      Estimate std. Error t value Pr(>|t|)
(intercept)          1.2971231  0.0719476  18.029  < 2e-16 ***
genere[T.m]         -0.0231210  0.1025779  -0.225  0.82242
statura              0.0011955  0.0004312   2.773  0.00737 **
genere[T.m]:statura  0.0002350  0.0005948   0.395  0.69422
---
Signif. Codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01425 on 61 degrees of freedom
Multiple r-squared: 0.6327,adjusted r-squared: 0.6146
F-statistic: 35.02 on 3 and 61 df,  p-value: 2.746e-13
```

Dal fatto che il termine di interazione GENERE[T.M]:STATURA non è significativo (p-value 0.69), allora è meglio preferire il modello più parsimonioso, quello additivo:

```
> linearmodel.10 <- lm(pesotr ~ genere + statura, data=dataset)

> summary(linearmodel.10)

Call:
Lm(formula = pesotr ~ genere + statura, data = dataset)

Residuals:
      Min        1q    median        3q       max
-0.034530 -0.009282 -0.000792  0.008962  0.044458
```

```
Coefficients:
            Estimate std. Error t value Pr(>|t|)
(intercept) 1.276534   0.049256  25.916  < 2e-16 ***
genere[T.m] 0.017351   0.004900   3.541 0.000762 ***
statura     0.001319   0.000295   4.471 3.38e-05 ***
---
Signif. Codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01415 on 62 degrees of freedom
Multiple r-squared: 0.6317,Adjusted r-squared: 0.6199
F-statistic: 53.18 on 2 and 62 df,  p-value: 3.558e-14
```

Le due rette di regressione hanno la medesima pendenza m (0.0013) e per le ragazze la quota q è 1.276, mentre per i ragazzi la retta parte un po' più sopra, $q = 1.276 + 0.017 = 1.294$. Si noti che le rette si riferiscono alla variabile *pesotr* e per tornare al peso occorre elevare le formule per 1/0.10, ossia alla decima, ottenendo dunque non una *retta di regressione* ma una *regressione polinomiale*.

3.1 Esercizio

Il dataset *iris* di Edgar Anderson è veramente molto celebre tra gli statistici. In esso vengono riportate le misure di lunghezza e di larghezza dei petali e dei sepali di centocinquanta piante, suddivise nei tre gruppi (specie): *setosa*, *versicolor* e *virginica*. Esplorate il dataset per benino e, riutilizzando tutti i comandi che abbiamo sin qui incontrato, proponete alcuni modelli lineari.

4 .. Dovrei *fare una multivariata* per capire quali fattori influenzano l'outcome ..

Nelle due sezioni precedenti ci siamo occupati del *peso* degli studenti, ossia di una variabile risposta di tipo numerico. Quando invece la risposta di nostro interesse è una variabile binaria del tipo successo/insuccesso, come accade per la variabile *outcome* del dataset *surgery*, non è affatto opportuno ricorrere ad un modello lineare per l'analisi dei dati. La **regressione logistica** invece è la soluzione più adeguata: per un rapido richiamo, rimanendo comodamente seduto davanti al vostro calcolatore, consultate `http://ccforum.com/content/9/1/112`. A noi basta qui ricordare che se $0 < p < 1$ è la probabilità che un certo evento accada, allora il rapporto $p/(1-p)$ tra la probabilità dell'evento e la probabilità del suo complemen-

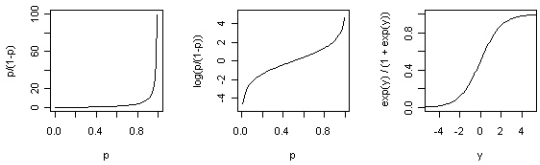


Figura 8:

tare viene denominato *odds*[2]. L'odds è un numero prossimo a zero se p è molto bassa; al contrario, l'odds è molto grande se p è vicina ad 1: si veda la Figura 8 nel pannello a sinistra.

Il logaritmo di un *odds* viene detto in gergo *logit* ed, al variare della probabilità p, il grafico del *logit* è illimitato, come si vede nel pannello centrale della Figura 8.

L'idea vincente dunque è quella di 'ribaltare' il grafico (i.e. come dicono i matematici, invertire la funzione), esplicitando la probabilità p in funzione del *logit* ed osservare che la funzione ottenuta ha valori compresi tra zero ed uno - che sono le 'etichette' tipiche attribuite agli eventi binari. Con un po' di algebra è facile scoprire che questa curva, rappresentata a destra nella Figura 8 ha equazione $p = \frac{exp(y)}{1+exp(y)}$.

Detto questo, vediamo come effettuare una regressione logistica. Importiamo da `http://www.dmi.units.it/~borelli/dataset/surgery.txt` il dataset in *R Commander* come abbiamo visto in precedenza, e proviamo ad impostare quello che gli statistici in gergo chiamano un '*glim*', ossia modello lineare generalizzato:

Statistiche > Fitta i modelli > Modelli lineari generalizzati..

Si apre una finestra di dialogo moto simile a quella del modello lineare. In più però viene richiesto di saper scegliere una Famiglia (e si sottointende che si tratta di una famiglia di variabili aleatorie adatte a modellare la risposta della regressione; nel caso della logistica è corretto scegliere la variabile aleatoria binomiale, `binomial`) ed una Funzione di link (che è il 'trucco matematico' che consente di trattare in maniera lineare ciò che lineare non è, come abbiamo visto nei grafici precedenti; la funzione `logit` è quella consigliata in questo caso). Come formula del modello ipotizziamo che la sopravvivenza possa essere messa in relazione additiva con tutte e quattro le covariate del dataset, come si vede nella Figura 9

Otteniamo il seguente output:

[2]Nella sezione precedente avevamo parlato di *odds ratio*, ossia di rapporto (frazione) tra due *odds*.

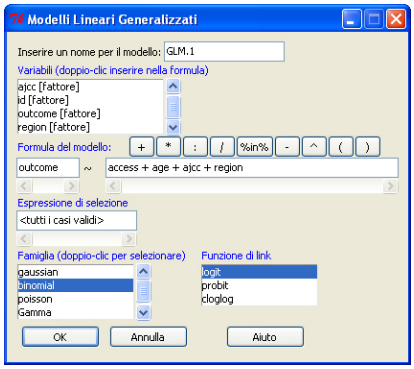


Figura 9: La finestra di dialogo Modelli Lineari Generalizzati.

```
Call:
Glm(formula = outcome ~ access + age + ajcc + region,
family = binomial(logit), Data = dataset)

Deviance residuals:
    Min       1q   median       3q      max
-1.9811  -0.5411  -0.4088  -0.3187   2.4405

Coefficients:
                  Estimate std. Error z value pr(>>z>)
(intercept)       0.738033   1.006996   0.733   0.4636
Access[t.open]    0.553434   0.291086   1.901   0.0573 .
Age               0.007163   0.013445   0.533   0.5942
Ajcc[t.low]      -4.150978   0.503382  -8.246  < 2e-16 ***
Ajcc[t.medium]   -2.759667   0.495458  -5.570 2.55e-08 ***
Region[t.rectum]  0.514300   0.290934   1.768   0.0771 .
---
Signif. Codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(dispersion parameter for binomial family taken to be 1)

    Null deviance: 478.30  on 494  degrees of freedom
Residual deviance: 362.57  on 489  degrees of freedom
  (4 observations deleted due to missingness)
Aic: 374.57

Number of fisher scoring iterations: 5
```

Sembrerebbe che molta importanza sia da attribuire alla stadiazione del tumore (*ajcc*), mentre l'età dei pazienti non sia rilevante in senso statistico. A questo punto, si dovrebbe riprovare ad impostare un nuovo 'glim', sbarazzandosi via della covariata *age*. In realtà, quello che si può fare per economizzare tempo è chiedere ad R Commander di semplificare il modello in maniera automatica:

MODELLI > SELEZIONE DEL MODELLO MEDIANTE 'STEPWISE'..

Accettando le impostazioni di *default* (DIREZIONE DI RICERCA `backward / forward` e CRITERIO `BIC` (vi rimandiamo a Wikipedia, o meglio ancora, ai testi di Statistica per approfondimenti su questi dettagli) si giunge al **modello minimale adeguato**:

```
Call:  glm(formula = outcome ~ ajcc, family = binomial(logit),
data = dataset)

Coefficients:
   (intercept)     ajcc[t.low]  ajcc[t.medium]
         1.642          -4.134          -2.668

Degrees of freedom: 494 total (i.e. Null);  492 residual
  (4 observations deleted due to missingness)
Null deviance:     478.3
Residual deviance: 368.5  aic: 374.5
```

Nel campione in nostro esame l'unico fattore che importa per la sopravvivenza dal punto di vista statistico sarebbe il grado di stadiazione della neoplasia. I coefficienti della regressione si interpretano in questo modo:

1. caso $ajcc = high$: $log(p/(1-p)) = 1.642$; ossia $p = exp(1.642)/(1 + exp(1.642)) = 0.84$, ossia 84 per cento di probabilità di soccombere;

2. caso $ajcc = low$: $log(p/(1-p)) = 1.642 - 4.134 = -2.492$; dunque $p = exp(-2.492)/(1 + exp(-2.492)) = 0.08$, 8 per cento di probabilità di decesso;

3. caso $ajcc = medium$: $log(p/(1-p)) = 1.642 - 2.668 = -1.026$; dunque $p = exp(-1.026)/(1 + exp(-1.026)) = 0.26$, circa una probabilità su quattro di morte.

Questi valori si possono facilmente visualizzare con il GRAFICO DEGLI EFFETTI di Figura 10 :

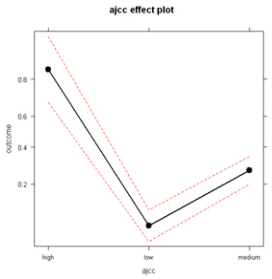


Figura 10: Il grafico degli effetti del modello lineare generalizzato `outcome` ∼ `ajcc`

Modelli > Grafici > Grafico degli effetti..

Si osservi che i 'pallini neri' rappresentano la probabilità di *outcome* infausto nei tre diversi gruppi, come abbiamo appena calcolato esplicitamente. L'occhio esperto riesce anche a fare un *multiple comparison* 'al volo', cogliendo il fatto che l'inaffidabilità delle stime è 'piccola' rispetto alla distanza che intercorre tra i pallini neri. Se non ve la sentite di affidarvi al colpo d'occhio, potete comunque basarvi sui numeri leggendo gli intervalli di fiducia (che tutti si ostinano a tradurre erroneamente con il termine di confidenza, quasi che gli intervalli non volessero veramente dire a tutti quanto valgono in realtà..):

Modelli > intervalli di confidenza..

```
> confint(glm.2, level=.95, type='lr')
                Estimate      2.5 %    97.5 %
(intercept)     1.642228  0.8396497  2.620170
Ajcc[t.low]    -4.134055 -5.1921456 -3.228968
Ajcc[t.medium] -2.668081 -3.7083833 -1.780801

               Exp(estimate)       2.5 %      97.5 %
(intercept)       5.16666667 2.315555721 13.73805228
Ajcc[t.low]       0.01601780 0.005560064  0.03959834
Ajcc[t.medium]    0.06938527 0.024517127  0.16850316
```

4.1 Esercizio

Reimportate in *R Commander* il dataset *studentiannoscorso* delle due sezioni precedenti e provate ad impostare un modello lineare generalizzato, che

spieghi il *peso* per mezzo dello sport, scegliendo la FAMIGLIA `gaussian` e la FUNZIONE DI LINK `identity`. Cosa ottenete?

4.2 Esercizio

Alla fine del 2011 apparve sul quotidiano *Libero* un articolo intitolato *Togliete i libri alle donne: torneranno a far figli*. Con il comando che abbiamo già visto:

DATI > DATI PRESENTI NEI PACCHETTI > LEGGI I DATI DA UN PACCHETTO CARICATO..

provate ad importare il dataset *swiss* della biblioteca *datasets*. Verificate innanzitutto che più di cento anni fa la media dell'indice di fertilità (*fertility*) era del 70.14, e che la media delle percentuale di persone che avevano istruzione superiore a quella obbligatoria elementare (*education*) era del 10.98 per cento. Eseguite la regressione binomiale descritta dalla formula: `fertility > 70.14` $\sim$ `education`. Secondo voi, il fatto che il coefficiente della regressione per *education* valga significativamente circa -0.14, conferma forse ciò che il giornalista Camillo Langone afferma in chiusura del suo pezzo: '*Così dicono i numeri: non prendetevela con me*'?

5 .. Vorrei capire se *la sopravvivenza è diversa* ..

R Commander nella sua 'versione base' non consente di effettuare l'analisi di sopravvivenza. E' necessario scaricare in R un *plug-in* dal CRAN, denominato `rcmdrplugin.survival` e, dopo averlo scaricato, lanciare *R Commander* ed applicarlo con il comando:

STRUMENTI > CARICA IL PLUG-IN..

R Commander a questo punto viene fatto riavviare, ed ora possiamo iniziare l'analisi importando il dataset:

`http://www.dmi.units.it/~borelli/dataset/breastcancer.txt`

Dobbiamo inizialmente specificare chi siano le variabili su cui valutare la sopravvivenza (nel nostro caso la length of stay, *los*) e chi sia l'indicatore dei possibili eventi censurati (nel nostro caso *discharge*). Lo facciamo come rappresentato in Figura 11 con il comando:

DATI > SURVIVAL DATA > SURVIVAL DATA DEFINITION..

A questo punto, in STATISTICHE compare un ulteriore menu, SURVIVAL ANALYSIS. Da questo menu dovrebbe essere possibile determinare le curve

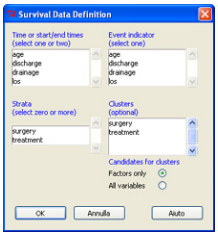


Figura 11: La finestra di dialogo del *plug-in* `survival`.

di sopravvivenza e testare i modelli di regressione; tuttavia, per ragioni che al momento non ci sono chiare, sul nostro computer appaiono degli errori che non ci consentono di proseguire. Naturalmente, non disperiamo, e abbandoniamo per il momento *R Commander* per fare ricorso ad R, caricando il pacchetto *survival* e importando il dataset *breastcancer* dalla rete:

```
library(survival)
www <- 'http://www.dmi.units.it/~borelli/dataset/breastcancer.txt'
breastcancer <- read.table(www, header = true)
attach(breastcancer)
```

Con i comandi `str`, `head` e `summary` visualizziamo rispettivamente la struttura delle variabili, le prime sei righe del dataset e le sue statistiche descrittive:

```
str(breastcancer)
head(breastcancer)
summary(breastcancer)
```

Ora, con il comando `survfit` possiamo calcolare gli elementi delle curve di sopravvivenza nei gruppi trattati con i bisturi dei tipi '*diart*' ed '*ultra*' mediante lo stimatore di Kaplan e Meier:

```
kaplanmeier <- survfit(surv(los, discharge) ~ treatment)
summary(kaplanmeier)
```

Le curve di sopravvivenza illustrate in Figura 12 si possono rappresentare con questi comandi:

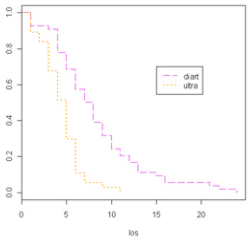


Figura 12: Curve di sopravvivenza del dataset `breastcancer`.

```
plot(kaplanmeier, lty=c(5,3), col = c('violet', 'orange'),
xlab='los', lwd = 2)
legend( 15, 0.7, levels(treatment), lty=c(5,3),
col = c('violet', 'orange') , lwd = 2)
```

In assenza di covariate, possiamo stabilire se vi sia differenza tra le due curve, in senso statistico, mediante il test nonparametrico *log-rank* (detto anche test di Mantel-Haenszel):

```
logrank <- survdiff(surv(los, discharge) ~ treatment)
logrank
```

Otteniamo il seguente output, dal quale risulta evidente che nei tempi di dimissione dei due gruppi la differenza è altamente significativa:

```
Call:
Survdiff(formula = surv(los, discharge) ~ treatment)

                 N observed expected (o-e)^2/e (o-e)^2/v
treatment=diart 54       54     70.9      4.02      24.5
treatment=ultra 37       37     20.1     14.17      24.5

 Chisq= 24.5  on 1 degrees of freedom, p= 7.61e-07
```

Siccome nel nostro dataset sono anche presenti le covariate *age* e *drainage*, il log-rank test non è adatto ai nostri scopi; proviamo dunque ad utilizzare l'approccio parametrico di Cox, dapprima introducendo l'età del paziente:

```
modello1 <- survreg(surv(los, discharge) ~ age + treatment )
summary(modello1)
```

A quanto sembra l'età ha un ruolo nel modello statistico (p-value = 0.0216, ossia 2.16e-02):

```
Call:
Survreg(formula = surv(los, discharge) ~ age + treatment)
                     Value std. Error     z        p
(intercept)         1.5632    0.30570  5.11 3.16e-07
age                 0.0105    0.00459  2.30 2.16e-02
treatment[T.ultra] -0.6486    0.10981 -5.91 3.49e-09
Log(scale)         -0.6669    0.08111 -8.22 2.00e-16

Scale= 0.513

Weibull distribution
Loglik(model)= -235.8   loglik(intercept only)= -250.6
        Chisq= 29.68 on 2 degrees of freedom, p= 3.6e-07
Number of newton-raphson iterations: 6
N= 91
```

Come interpretiamo i coefficienti della regressione? Per un paziente di età `age` operato con il bisturi `diart`, il tempo atteso t di degenza postoperatoria è dato da $t = exp(1.5362 + 0.0105 \cdot age)$, mentre per un paziente operato con il bisturi `ultra`, la degenza attesa si riduce: $t = exp(1.5362 - 0.6486 + 0.0105 \cdot age)$, ossia $t = exp(0.8876 + 0.0105 \cdot age)$. Vediamo però cosa succede se aggiungiamo l'effetto relativo ai volumi di drenaggio:

```
modello2 <- survreg(surv(los, discharge) ~ age + drainage + treatment )
summary(modello2)
```

```
Call:
Survreg(formula = surv(los, discharge) ~ age + drainage + treatment)
                       Value std. Error      z        p
(intercept)         1.464111   2.25e-01   6.50 7.99e-11
age                 0.006061   3.32e-03   1.83 6.76e-02
drainage            0.000565   9.44e-05   5.99 2.10e-09
```

```
treatment[T.ultra] -0.403277   8.75e-02  -4.61 4.00e-06
Log(scale)         -0.927736   8.20e-02 -11.31 1.19e-29

Scale= 0.395

Weibull distribution
Loglik(model)= -213.1   loglik(intercept only)= -250.6
        Chisq= 75.11 on 3 degrees of freedom, p= 3.3e-16
Number of newton-raphson iterations: 6
N= 91
```

Un altro effetto altamente significativo: maggiori volumi di drenaggio implicano un aumento dei tempi di degenza. La covariata *age* ora però ha perso significatività (p-value 0.0676) e quindi per parsimonia la togliamo dall'analisi. Sarebbe interessante scoprire se i volumi di drenaggio differiscano nei due trattamenti, ossia se il termine di interazione '`drainage:treatment`' sia significativo:

```
modello3 <- survreg(surv(los, discharge) ~ drainage * treatment )
summary(modello3)
```

Così non è, stante il p-value = 0.0676. Dunque, il nostro modello finale è puramente additivo:

```
modello4 <- survreg(surv(los, discharge) ~ drainage + treatment )
summary(modello4)
```

```
Call:
Survreg(formula = surv(los, discharge) ~ drainage + treatment)
                      Value std. Error      z         p
(intercept)         1.855291   7.60e-02  24.43 9.22e-132
drainage            0.000572   9.52e-05   6.00  1.92e-09
treatment[T.ultra] -0.408542   9.01e-02  -4.53  5.83e-06
Log(scale)         -0.896651   8.00e-02 -11.20  3.98e-29

Scale= 0.408

Weibull distribution
```

```
Loglik(model)= -214.6   loglik(intercept only)= -250.6
        Chisq= 71.97 on 2 degrees of freedom, p= 2.2e-16
Number of newton-raphson iterations: 6
N= 91
```

Concludiamo la nostra sessione rimuovendo il dataset:

```
detach(breastcancer)
rm(breastcancer)
```

5.1 Esercizio

Nella `library(MASS)` viene riportato il dataset *Melanoma*, relativo a 205 pazienti danesi. Importatelo con il comando `attach` e verificate se il tempo di sopravvivenza in giorni, *time*, possibilmente censurato con la variabile indicatrice `status == 1`, cambia con il genere *sex* dei pazienti e/o la loro età *age*.

Riferimenti bibliografici

[1] Jonathan Ball, Viv Bewick and Liz Cheek (2001 - 2005). *Medical statistics*, Critical Care.

`http://ccforum.com/series/CC_Medical`

[2] William J. Conover, Mark E. Johnson and Myrle M. Johnson (1981). *A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data.* Technometrics 23, 351:361.

[3] Michael J.Crawley (2005). *Statistics: An Introduction using R*, Wiley.