Analisi dei dati

Introduzione all'analisi esplorativa dei dati con `R`

Nicola Torelli

2025-05-05

Indice

Capitolo 1

Introduzione

Il materiale che segue è predisposto a beneficio degli studenti del I anno del corso di laurea in Intelligenza Artificiale e Data Analytics dell'Università di Trieste.

A parte le conoscenze di base di matematica e informatica non vi sono prerequisiti formali, al fine di apprezzare pienamente i contenuti è però preferibile che gli studenti abbiano acquisito le nozioni fondamentali del corso di Calcolo delle Probabilità.

La presente dispensa copre la seconda parte del corso, la prima parte è dedicata all'introduzione al software `R` che viene ampiamente utilizzato per illustrare i concetti e le tecniche introdotte e per fornire importanti esemplificazioni. Questo consente di ampliare in modo sostanziale anche la conoscenza degli strumenti di visualizzazione e di analisi dei dati specifici di `R`.

Molto rilevanti sono inoltre i materiali che verranno introdotti durante le sessioni di tutorato ove si riprenderanno e amplieranno molti degli esempi visti durante le lezioni e trattati nella presente dispensa. A tal fine, saranno resi disponibili attraverso moodle i dati da analizzare quando non si tratti di dati già disponibili in pacchetti `R`.

Il testo non sarà esente da errori. Cerco di migliorarlo ogni anno e sarò molto grato a coloro che vorranno segnalarmi sviste o imprecisioni.

Capitolo 2

I dati

Nel seguito faremo riferimento al termine **dati** intendendo un insieme di informazioni relative a un insieme di **unità**.

Ci concentreremo ora sull' "Analisi Esplorativa dei Dati" (EDA) o "analisi descrittiva". L'obiettivo è mettere in luce aspetti interessanti dei dati applicando tecniche di analisi, riassunti numerici e rappresentazioni grafiche.

L'aspetto essenziale è che in tale fase si vorrebbe che siano i dati stessi a parlare, senza ricorrere a assunzioni specifiche, e a rivelare caratteristiche salienti e interessanti.

2.1 Popolazione e unità statistiche

I dati possono essere raccolti per diversi motivi:

- a supporto della ricerca scientifica in diversi ambiti,
- raccolti dalla pubblica amministrazione nel gestire un servizio o a seguito dell'utilizzo di un software o nella gestione di un sito web.

In generale, va detto che l'obiettivo ultimo è quello di conoscere le caratteristiche di una **popolazione**.

Una popolazione è una collettività e gli elementi di tale collettività sono detti **unità statistiche**, sono esempi:

- la popolazione degli italiani di sesso maschile con oltre 18 anni al 01/01/2012;
- le famiglie italiane al 01/01/2012;
- i 218 comuni del FVG;
- i clienti di un negozio;
- coloro che accedono a un sito web.

La popolazione può essere finita (ad es. la popolazione italiana) o infinita (ad es. tutte le persone affette da una patologia, oggi o in futuro).

2.1.1 Dati e ricerca scientifica

Occorre che i dati vengano raccolti utilizzando protocolli che permettano di generalizzare quello che emergerà dalla loro analisi. La statistica, e in particolare quella inferenziale, stabilisce criteri e regole perchè si possa attribuire ai dati raccolti un valore scientifico. A tal fine si distingue fra:

1. dati ottenuti secondo disegni sperimentali controllati. Essi consentono di valutare correttamente l'esistenza di relazioni causali - come l'efficacia di un farmaco - o valutare un processo di produzione;
2. dati osservazionali. Questi sono disponibili in un numero di casi forse più ampio. Essi vengono spesso raccolti in indagini o rilevazioni statistiche che sono di tipo:

- totale (o censuario) se osservo tutte le unità della popolazione (è appunto il caso del censimento),
- o parziale (cosa inevitabile se la popolazione è infinita) quindi osservando solo alcuni elementi della popolazione.

Nel caso della rilevazione parziale è cruciale che la raccolta dei dati avvenga secondo schemi che li rendano rappresentativi dell'intera popolazione.

La migliore garanzia è offerta da una selezione (campionamento) degli elementi da osservare che segua criteri di scelta casuale. Solo se si può contare su un rigoroso schema di campionamento casuale è possibile utilizzare correttamente i metodi della statistica per formulare conclusioni riferite alla intera popolazione.

Si noti che nel caso i dati vengano rilevati in un contesto osservazionale, anche se si utilizzano schemi di campionamento rigorosi, non è tuttavia possibile, o agevole, poter trarre conclusioni sulla esistenza di una relazione fra le variabili osservate. Occorre quindi molta cautela nell'interpretare in senso causale le associazioni osservate e spesso occorre ricorrere ad assunzioni non verificabili empiricamente.

2.2 Analisi esplorativa dei dati

L'analisi esplorativa dei dati (EDA) o (in inglese *data analysis*) o **analisi statistica descrittiva** non si pone l'obiettivo di ricavare conclusioni su un aggregato diverso da quello osservato (cosa di estremo rilievo in contesti scientifici).

L'attenzione è invece sulle tecniche per fornire efficaci sintesi dei dati (anche con opportune tecniche grafiche di visualizzazione) così da mettere in evidenza alcune caratteristiche essenziali con l'obiettivo di **monitorare un fenomeno, effettuare confronti, elaborare congetture da sottoporre poi ad analisi più accurate**.

Le conclusioni che si traggono non voglione quindi avere carattere di generalità: non si vuole estendere quanto si osserva sull'insieme di dati disponibile a popolazioni più ampie utilizzando apparati formali (come quello della statistica inferenziale dove si riesce a misurare anche l'attendibilità delle conclusioni che si traggono).

Tuttavia i *pattern* osservati nei dati sono evidenze utili seppure riferibili esclusivamente all'insieme di dati osservato. Si noti che se i dati si riferiscono a un'intera popolazione (come per il censimento) ottenere una efficace sintesi degli stessi costituisce informazione valida per l'intera popolazione.

Non si fa quindi riferimento a priori a modelli stocastici che potrebbero aver generato i dati come nel caso dell'inferenza statistica o all'esistenza di relazioni specifiche fra le quantità osservate.

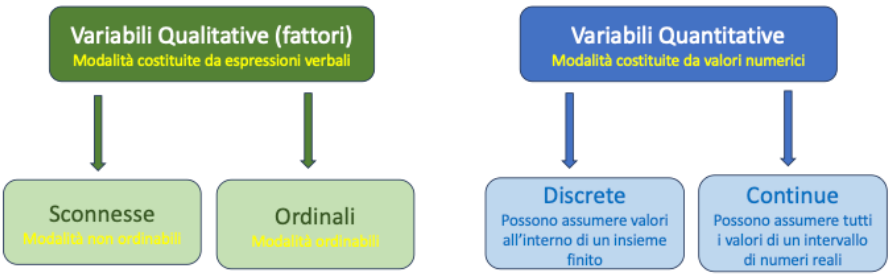
Si prescinde inoltre dall'idea che i dati siano "perfetti" e si ammette che essi possano essere sporchi, inaccurati, osservati in modo incompleto e con errori. Per cui l'analisi dei dati che introdurremo dovrà spesso includere una fase non banale di "pulizia" dei dati preliminare alla fase di analisi esplorativa.

2.2.1 Tipi di dati

2.2.1.1 Variabili statistiche

Un dato statistico è il risultato della rilevazione (misurazione/osservazione) di **variabili** o **caratteri** su un'**unità statistica** appartenente a una popolazione.

2.2.1.2 Tipi di variabili (o caratteri)



2.2.1.3 Variabili qualitative

- Una variabile è **qualitativa** se i valori che può assumere, detti *modalità*, si presentano espressi in forma verbale;
  - una variabile qualitativa è **sconnessa** se le sue modalità non implicano una graduazione;

  - una variabile qualitativa è **ordinale** se le sue modalità implicano una graduazione;
- le modalità possono essere predefinite a priori;
- a volte, in rilevazioni con questionari, le modalità vengono desunte a posteriori a partire dalla descrizione dettagliata dello stato della singola unità relativamente al carattere in questione.

Le variabili qualitative ai fine delle analisi dei dati che verranno condotte con R sarà opportuno definirle come **fattori**.

2.2.1.4 Variabili quantitative

- Una variabile è **quantitativa** se assume valori espressi in forma numerica che corrispondono a una misurazione o a un conteggio;
- rispetto ai valori che possono assumere
  - una variabile quantitativa è **discreta** se l'insieme dei valori numerici che può assumere è finito oppure numerabile;
  - una variabile quantitativa è **continua** se l'insieme dei valori numerici che assume è, almeno concettualmente, associabile con i valori di un intervallo reale, limitato o illimitato.

NB. Per la limitata precisione utilizzabile nel rilevare le misure, la distinzione tra variabile discreta e continua è di fatto convenzionale.

2.2.2 La matrice dei dati

La più semplice forma con cui rappresentare i dati relativi ad alcune variabili, diciamo p, su un collettivo di n unità è la **matrice dei dati**. Ovvero una matrice che ha n righe e p colonne. Così che una riga rappresenta i dati raccolti per una generica unità e una colonna contiene il vettore di valori osservati su ciascuna variabile per l'insieme delle unità.

Di solito n è molto maggiore di p e l'obiettivo dell'analisi dei dati è quello di analizzare le colonne della matrice:

- se si prende in esame una variabile (colonna) per volta si parla di analisi di una singola variabile o **analisi univariata**
- se si prendono in esame più variabili (più colonne) congiuntamente si parla di analisi **bivariata** nel casi di due variabili o **multivariata** se considero più di due colonne congiuntamente.

Capitolo 3

Indici e grafici

3.1 Visualizzazione dei dati

Gran parte dell'analisi (statistica) dei dati si basa sull'idea di fornire sintesi efficaci degli stessi così da estrarre al meglio le informazioni che essi contengono su un determinato collettivo.

Vi sono varie strategie per ottenere **riassunti** dei dati in ambito descrittivo/esplorativo. Nell'ambito dell'inferenza statistica il concetto di riassunto dei dati può essere impostato in modo più formale.

In estrema sintesi tuttavia gli strumenti che si utilizzano in ambito esplorativo sono di fatto due e sono complementari:

- utilizzo di **misure sintetiche** di specifiche caratteristiche dei dati (esempi banali sono la media e la varianza come misure, rispettivamente, di centralità e dispersione dei dati);
- utilizzo di **sintesi grafiche** e di tecniche di visualizzazione.

L'idea di fondo è che l'elaborazione dei dati grezzi (pensiamo alla matrice dei dati) sia necessaria così da comprimere e rendere comprensibile l'informazione in essi contenuta oltre a renderla più semplice da comunicare e conservare.

Uno dei principi fondamentali è che l'elaborazione dei dati e la loro sintesi (con indici o grafici) comporti inevitabilmente una perdita di informazione: è il prezzo da pagare per poter *leggere* i dati e per trarre da essi informazione utile. Ci si aspetta tuttavia che l'informazione che si perde sia non rilevante per interpretare il fenomeno e ottenere una visione di insieme delle caratteristiche salienti del collettivo che si sta esaminando.

3.1.1 Le tabelle di frequenza

La più semplice elaborazione è quella che conduce alle tabelle di frequenza: essa è un primo esempio di sintesi dei dati.

Per chiarire le idee, consideriamo il data frame del `package` "insuranceData" che contiene i dati relativi ai danni liquidati per sinistri dichiarati a un'assicurazione. Esso contiene sia variabili categoriali che quantitative continue (una descrizione dei dati ed è stata già introdotto nella prima parte del corso). Costruiamo la semplice tabella relativa alla variabile categoriale `AutoBi$Attorney`.

```
library(insuranceData)
data("AutoBi")
table(AutoBi$ATTORNEY)
```

```
##
##   1   2
## 685 655
```

Le informazioni sulla variabile riguardante il ricorso all'avvocato nei dati `AutoBi` sono in una colonna del data frame e l'obiettivo è sapere se in quel collettivo si tenda a ricorrere spesso all'avvocato. La tabella di frequenza riassume tali informazioni in modo efficace e ci permette di sapere che nell'insieme di dati si ricorre all'avvocato irca nella stessa misura con cui non si ricorre. Se si conservasse solo la tabella si perderebbe il dettaglio informativo sul singolo caso, sulla scelta o meno dell'avvocato che è contenuta nella specifica riga della matrice dei dati, tuttavia questa informazione è irrilevante per rispondere all'obiettivo conoscitivo. Il pezzo di informazione che ho perso è in questo caso trascurabile.

Si consideri, ad esempio, ora la variabile `LOSS` (sempre del data frame `AutoBi`). Anche per essa costruiamo una tabella di frequenza. Questa tabella contiene le frequenze assolute.

Tuttavia per fare questa semplice elaborazione e costruire la tabella della variabile quantitativa è necessario considerare le classi di valori e quindi trasformare la variabile in fattore: invece del singolo dato (che in questo caso sarà verosimilmente diverso per ogni unità) conservo solo la categoria ovvero la classe cui quel valore apparteneva.

```
AutoBi$LOSSclass<-cut(AutoBi$LOSS,breaks=c(0,0.5,2,4,8,1100))
table(AutoBi$LOSSclass) # tabella delle frequenze assolute
prop.table(table(AutoBi$LOSSclass)) # tabella delle frequenze relative
```

```
##
##     (0,0.5]     (0.5,2]       (2,4]       (4,8] (8,1.1e+03]
##         288         324         396         195         137
##
##     (0,0.5]     (0.5,2]       (2,4]       (4,8] (8,1.1e+03]
##   0.2149254   0.2417910   0.2955224   0.1455224   0.1022388
```

La tabella fornisce una sintesi del fenomeno in esame: ad esempio ora si vede che coloro che hanno subito un danno sotto i 2000 dollari rappresentano circa il 45% dei casi.

La sintesi è molto più leggibile rispetto alla lista di 1340 valori riportati nella relativa colonna del data frame. Però nel fare la tabella si sono perse informazioni. Se si disponesse della sola tabella, non sarebbe possibile sapere se vi sono e quanti sono i casi che hanno un danno fra 1500 e 2000 dollari. Nel fare la sintesi si sono persi quindi alcuni dettagli e occorre chiedersi quale sia il livello di sintesi ottimale che bilanci la necessità di chiarezza e la perdita di informazione.

Il compromesso fra l'esigenza di sintesi e quella di mantenere il dettaglio informativo è un motivo ricorrente dell'analisi dei dati e in generale dell'elaborazione statistica.

Nella scelta di una opportuna sintesi numerica o di una tecnica grafica questo sarà quindi un tema sempre presente. In molti casi le visualizzazioni grafiche risultano più flessibili e maggiormente in grado di conservare buona parte dell'informazione originale.

3.1.2 Primi semplici grafici con R

Si introdurranno ora prime semplici rappresentazioni grafiche, immediatamente comprensibili, e alcune ben note, che ci consentiranno anche di iniziare a esplorare le enormi potenzialità grafiche di R.

Le tabelle di frequenza rappresentano una delle più elementari forme di elaborazione e sintesi dei dati. Ad esse si possono associare alcune rappresentazioni grafiche che quindi utilizzeranno direttamente i dati riassunti nell'oggetto generato dal comando `table()`.

3.1.2.1 Il diagramma a torta

Si prenda in esame il data set `Cars93` (già visto in precedenza)

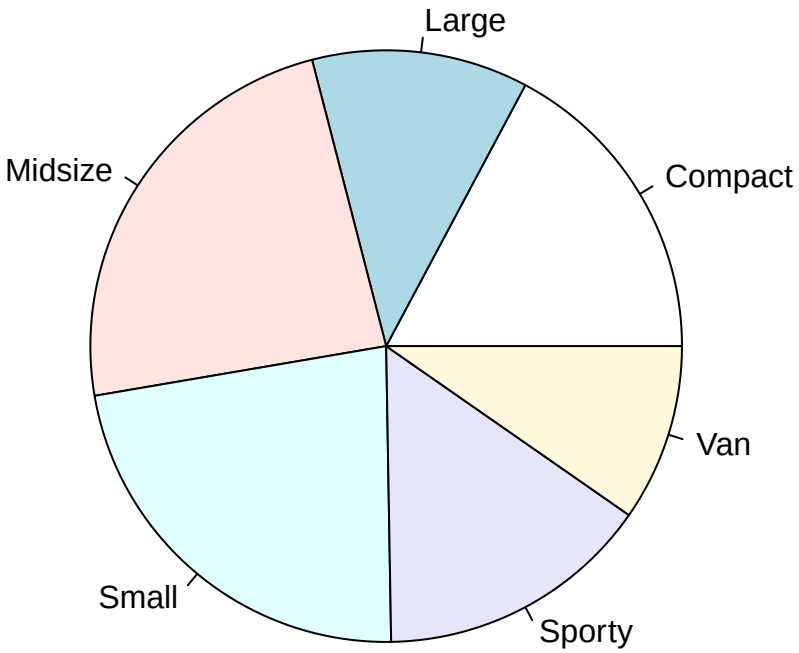
```
library(MASS)
data("Cars93")
```

La variabile `Type`, il tipo di auto, è qualitativa per cui possiamo ottenere la tabella di frequenza

```
tabtipo<-table(Cars93$Type)
```

Possiamo associare a tale tabella il cosidetto "pie chart" o, in italiano, **diagramma a torta** con la funzione `pie()`.

```
pie(tabtipo)
```



La fetta di torta corrispondente a ciascuna modalità ha un'area proporzionale alla frequenza della stessa.

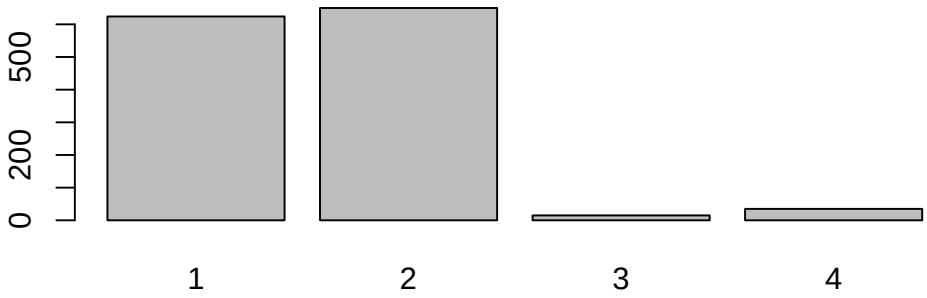
Il diagramma a torta è tanto noto quanto, spesso, da sconsigliare. Se si legge la nota in fondo all'help su `pie()` in `R` si trova: "Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data".

3.1.2.2 Il diagramma a barre: `barplot()`

Un diagramma adatto a rappresentare variabili qualitative, o meglio le tabelle di frequenza ottenute con fattori qualitativi è costituito dal **diagramma a barre** (*barplot* o *barchart* in inglese).

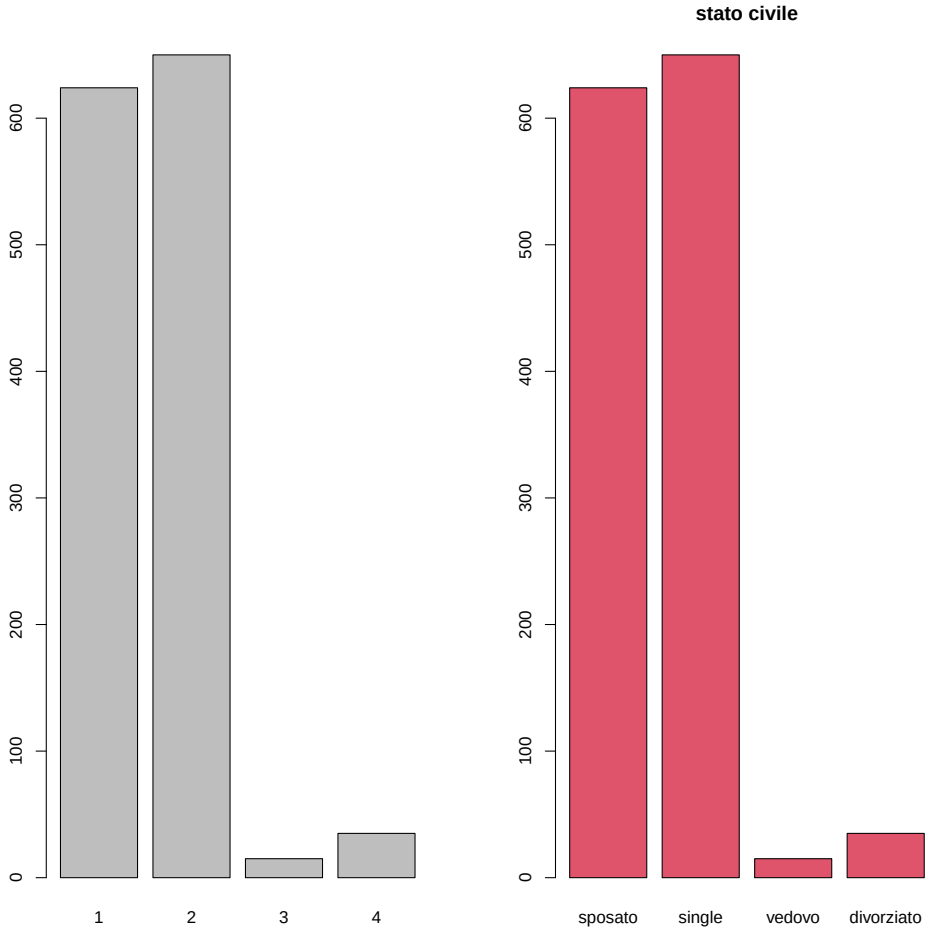
Anche in questo caso la funzione `barplot()` si applica non al vettore di dati originale ma alla tabella di frequenze ricavata da esso. Vediamo un esempio con la variabile `MARITAL` del data frame `AutoBi`:

```
maritab<-table(AutoBi$MARITAL)
barplot(maritab)
```

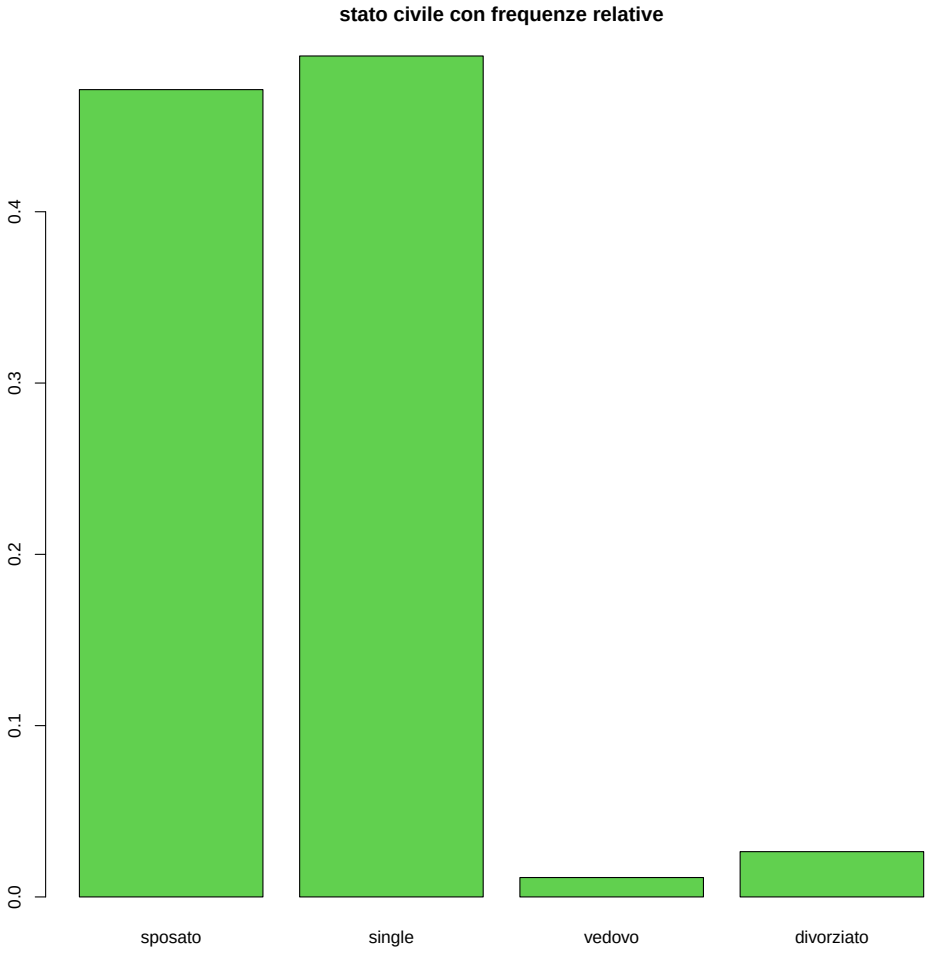


Stavolta le frequenze sono proporzionali alle altezze dei rettangoli ciascuno dei quali rappresenta una modalità. La funzione contiene alcuni parametri che consentono di personalizzare il grafico. ad esempio:

```
par(mfrow=c(1,2))
# questa funzione serve per fissare alcuni parametri della finestra grafica:
# in questo caso divide la finestra grafica in due colonne e una riga. Per cui
# affiancherà i due grafici successivi. Si noti che nel secondo grafico
# sono anche utilizzati alcuni parametri per mettere il titolo al grafico
# (main="titolo" o per cambiare il colore col="2")
barplot(maritab)
barplot(maritab, main="stato civile", names.arg=c("sposato", "single",
                                    "vedovo", "divorziato"), col=2)
```




```r
par(mfrow=c(1,1))  # così torniamo ad avere un solo grafico
                   # per ciascuna finestra grafica
maritab1<-table(AutoBi$MARITAL)/
  length(AutoBi$MARITAL[!is.na(AutoBi$MARITAL)])
# In questo caso consideriamo le frequenze relative
# si noti che abbimao diviso per il numero di casi escludendo gli NA
barplot(maritab1, main="stato civile con frequenze relative",
        names.arg=c("sposato", "single", "vedovo", "divorziato"), col=3)
```




```
# diagramma a barre con le frequenze relative
```

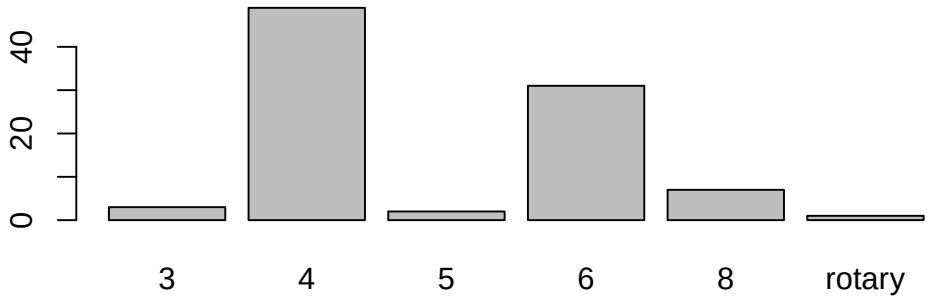
Si vede bene che ogni rettangolo è separato, la sua larghezza e la posizione sull'asse orizzontale è arbitraria. Questo deve far riflettere sul fatto che tale grafico possa essere poco appropriato se la tabella di frequenze riguarda una variabile quantitativa trasformata in fattore con il metodo delle classi di valori.

Infatti in quest'ultimo caso i valori numerici postrebbero essere rappresentati più apportunamente sull'asse orizzontale. In particolare, ma non solo, se si tratta di una variabile quantitativa continua.

Il messaggio è quindi: **no utilizzare il barplot per rappresentare una variabile quantitativa continua (in classi)**.

Talvolta esso si può tuttavia utilizzare con una variabile discreta con un numero di valori molto basso. ad esempio nel caso del numero di cilindri nel data set delle auto.

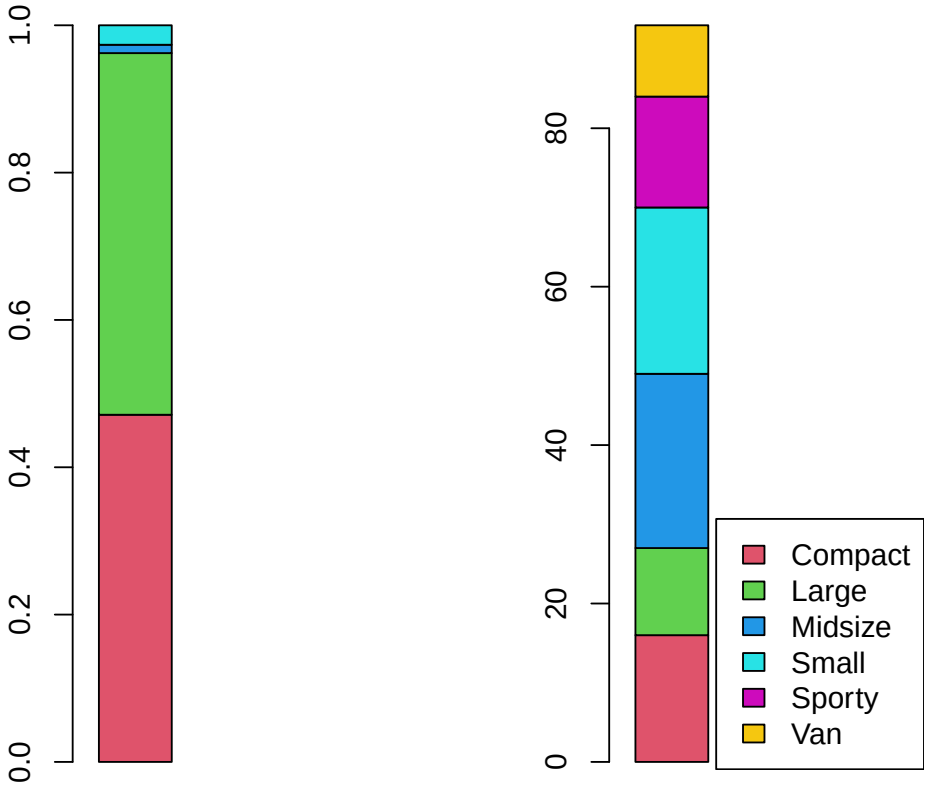
```r
barplot(table(Cars93$Cylinders))
```



3.1.2.3 Il diagramma a barre sovrapposte

Un altro metodo per rappresentare una distribuzione di frequenza è quello di ricorrere al diagramma a barre "impilato". Invece di avere una barra separata per ogni frequenza si rappresentano in un'unica barra le porzioni relative a ciascna modalità con le frequenze rappresentate dalla lunghezza di ciascuna porzione. Si suddividono le modalità con dei segmenti di un unico rettangolo la cui lunghezza è proporzionale alla frequenza (versione 'stacked'). Occorre che i dati della tabella di frequenza siano trasformati in un vettore colonna (una matrice con tante righe quante le modalità e una solo colonna).

```r
# esempio con stato civile
par(mfrow=c(1,2))
freq<-matrix(prop.table(table(AutoBi$MARITAL)),
             nrow=length(table(AutoBi$MARITAL)), ncol=1)
barplot(freq, xlim=c(0,4), col=(2:5), )
# esempio con auto
nrig=length(table(Cars93$Type))
autof<-matrix(table(Cars93$Type), nrig, ncol=1)
colr=(2:(nrig+1))
barplot(autof, xlim=c(0,4), col=(2:(nrig+1)))
legend(x="bottomright", legend=levels(Cars93$Type), fill=colr)
```




```
par(mfrow=c(1,1))
```

3.1.3 Rappresentare graficamente variabili quantitative: primi elementi

Nel caso dell'analisi di variabili quantitative abbiamo già osservato che occorre maggiore attenzione poichè l'eventuale riassunto grafico (o numerico) può comportare una più sensibile perdita di alcune informazioni. Occorre sempre chiedersi se la perdita di informazione è ben bilanciata dalla qualità informativa del riassunto stesso.

Nel seguito si introdurranno elementi per analizzare variabili quantitative mediante rappresentazioni grafiche e attraverso riassunti numerici. Vale la pena di segnalare che gli strumenti che verranno introdotti sono riferiti all'analisi di variabili quantitative continue. Tuttavia essi sono in genere del tutto adeguati anche nel caso di variabili discrete e specialmente nel caso di variabili discrete che assumono tanti valori distinti: ad esempio, se le unità statistiche fossero i comuni italiani e la variabile la popolazione degli stessi (si tratta evidentemente di una variabile discreta), ci si aspetta che la variabile assuma valori dell'ordine delle centinaia e che verosimilmente osserverei quasi esclusivamente valori distinti.

Se si tratta di variabili discrete (spesso di conteggio) che assumono solo valori bassi, quindi con pochi valori distinti (esempio sono la variabile numero di fratelli, o il numero di sinistri con l'auto che l'assicurato denuncia in un anno) l'uso dei semplici grafici visti per le variabili categoriali si rivela spesso del tutto adeguato. In tal caso anche i riassunti numerici che saranno illustrati, come ad esempio media, varianza, etc., sono comunque appropriati e possono essere utilizzati.

Si ricorda, invece, che i riassunti numerici e le rappresentazioni grafiche che si introdurranno, **non** devono essere utilizzate **mai** per variabili qualitative/fattori/categoriali. Tale errore spesso si fa se le variabili sono rappresentate invece che con la modalità con un valore numerico (ad esempio dopo una operazione di ricodifica). È per tale motivo che risulta sempre opportuno, nell'analisi con `R`, la trasformazione di tali variabili in fattori perchè il software sarà spesso in grado di decidere correttamente quale strumento va utilizzato.

3.1.3.1 Il diagramma ramo e foglie (stem and leaf)

Un semplice grafico che ha il vantaggio peraltro di non perdere informazione relativamente alla variabile quantitativa è il **diagramma ramo e foglie**. In `R` viene evocato con la funzione `stem()`. Ad esempio, si consideri la lunghezza in pollici delle auto:

```
Cars93$Length
stem(Cars93$Length)
```

```
##  [1] 177 195 180 193 186 189 200 216 198 206 204 182 184 193 198 178 194 214 179
## [20] 203 183 203 174 172 181 175 192 180 174 202 141 171 177 180 179 176 192 212
## [39] 151 164 175 173 185 168 172 166 184 200 188 191 205 219 164 172 184 190 169
## [58] 175 187 166 199 172 190 170 181 190 188 188 190 194 201 173 177 181 196 195
## [77] 177 184 176 146 175 179 161 162 174 188 187 163 187 180 159 190 184
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   14 | 16
##   15 | 19
##   16 | 123446689
##   17 | 012222334445556677778999
##   18 | 000011123444445677788889
##   19 | 000001223344556889
##   20 | 001233456
##   21 | 2469
```

Come si vede il grafico è costituito da valori a destra di una linea verticale che corrispondono alla prime due cifre della variabile (il ramo). A sinistra della linea viene messa la terza cifra (ordinando dalla più bassa alla più alta). Tali cifre si impilano e la lunghezza della pila (la foglia) rappresenta direttamente

la frequenza assoluta. La scelta di cosa usare come ramo è arbitraria e va fatta considerando i valori assunti dalla variabile.

Il grafico viene utilizzato per piccoli insiemi di dati (di solito non superiori al centinaio di casi). Se i casi sono più numerosi o se vi sono variabili molto disperse (per cui risulta difficile trovare efficacemente i valori da usare per il ramo) la rappresentazione risulta meno efficace e sarà necessario perdere qualche dettaglio: ad esempio se provo a usarla con `AutoBi$LOSS`

```
stem(AutoBi$LOSS)
```

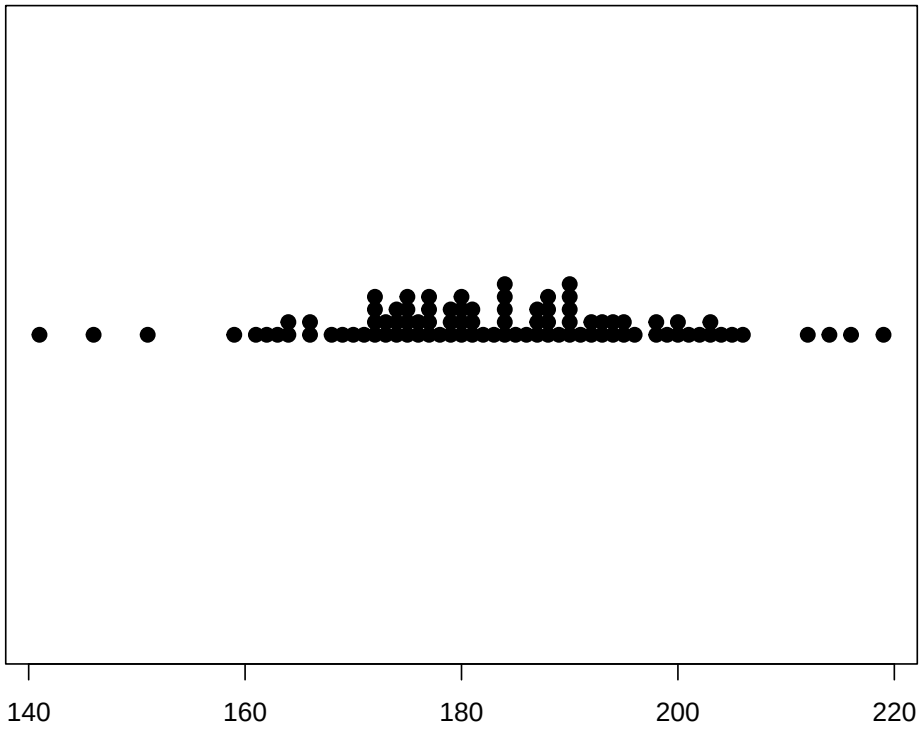
```
##
##   The decimal point is 2 digit(s) to the right of the |
##
##    0 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+1251
##    1 | 015699
##    2 | 27
##    3 |
##    4 |
##    5 |
##    6 |
##    7 |
##    8 |
##    9 |
##   10 | 7
```

è evidente che, in questo caso, lo stem & leaf non è una rappresentazione grafica efficace.

3.1.3.2 Il diagramma a punti

Il diagramma a punti, di fatto rappresenta su un grafico i singoli valori osservati. Per cui anche in questo caso non si perde alcun dettaglio. La funzione `stripchart()` del package `graphics` può essere usata a tale scopo.

```
stripchart(Cars93$Length, pch=19, method="stack", cex=1.2, ylim=c(0,2))
```




```
# Si noti come nella funzione vengano fissati alcuni parametri. Si tratta di
# parametri grafici che potranno essere utilizzati anche in altri casi Ad esempio
# pch permette di scegliere quale simbolo utilizzare per rappresentare un punto
# mentre cex permette di variare la grandezza del simbolo stesso
```

Il parametro `method=stack` consente di sovrapporre dati osservati che risultano spesso avere il medesimo valore. Il default è invece quello di mostrare il valore osservato senza dare conto della eventuale molteplicità.

Si considerino ad esempio i dati `Ozone` nel package `mlbench` e in particolare la variabile `V4` che contiene il livello di concentrazione dell'ozono osservato (che appare misurato come una variabile che assume valori interi discreti) in California in tutti i giorni dell'anno

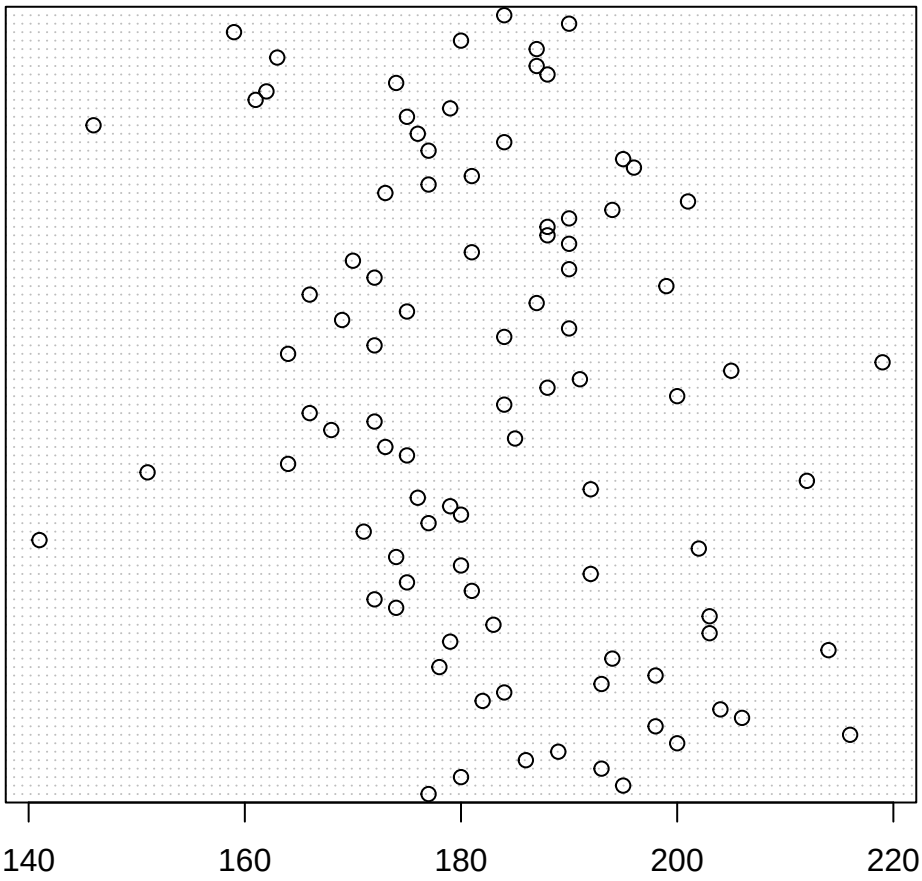
Anche in questo caso sono visibili i singoli valori.

3.1.3.3 Il diagramma a punti e la funzione `plot()`

Il diagramma a punti fornisce semplicemente la rappresentazione di ogni singolo punto osservato. Non vi è quindi alcuna sintesi dei dati. Tuttavia il suo utilizzo è certamente di interesse se si dispone di un numero limitato di casi da analizzare.

La funzione `dotchart` fa essenzialmente questo: rappresenta su un sistema di assi cartesiano i singoli punti.

```
dotchart(Cars93$Length)
```

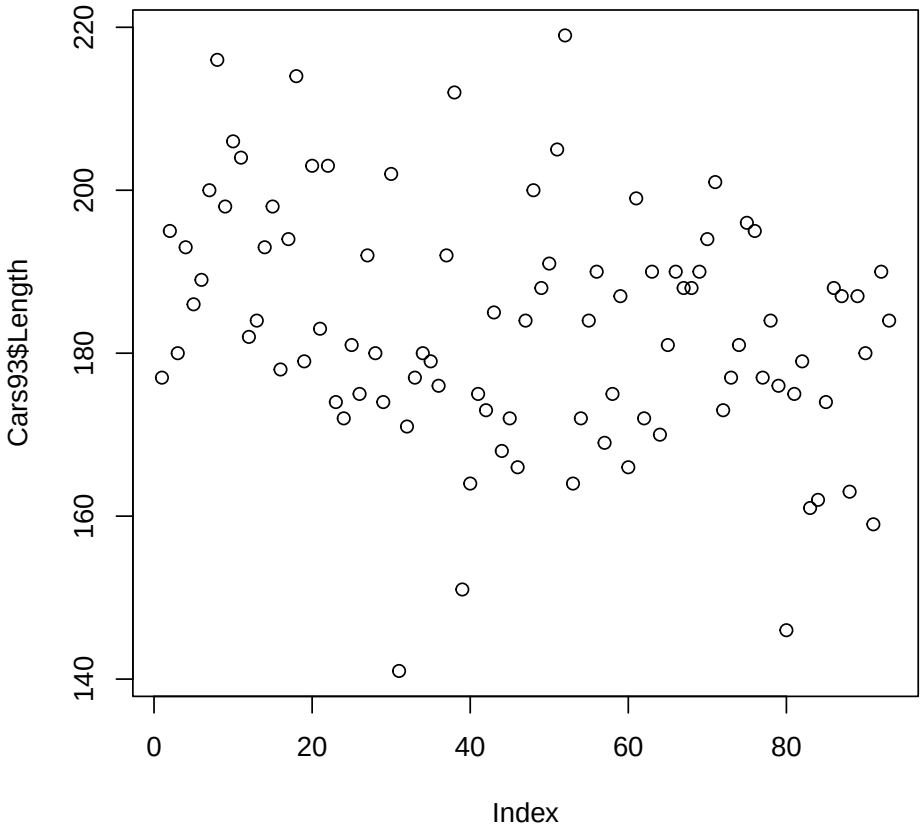


Le coordinate sull'asse verticale riflettono solo l'ordine con cui appaiono nella variabile le singole osservazioni.

E' importante osservare che tale grafico è del tutto equivalente (nel senso che si sono solo invertiti gli assi) alla funzione **`plot()`**, che si introduce ora ma che verrà trattata con maggior dettaglio in seguito. Essa consente di riprodurre su un sistema di assi un'insieme di punti fornendo le loro coordinate. Di base quindi si dovrebbe fornire una coppia di vettori che rappresentano le coordinate dei singoli punti. Tale funzione è estremamente generale e consentirà di ottenere facilmente molte rappresentazioni grafiche di insiemi di dati bivariati (oltre che essere utilizzato per rappresentare graficamente, ad esempio, curve o funzioni di una variabile).

In effetti si può usare semplicemente

```
plot(Cars93$Length)
```



Il sistema si aspetta una coppia di variabili ma se se ne indica solo una, la coordinata sull'asse delle ascisse riporta semplicemente l'ordine con cui il valore appare nel vettore (che come si vede è chiamato `Index` nel grafico).

3.2 Riassunti numerici per variabili quantitative

Prima di procedere a illustrare altri metodi grafici per presentare un insieme di dati relativi a una variabile quantitativa converrà richiamare alcuni concetti di base che riguardano le misure di sintesi per riassumere le caratteristiche salienti di una variabile quantitiva (discreta o continua che sia).

Come si è già anticipato, quando si dispone dei dati per una variabile quantitativa è necessario sintetizzare e elaborare i dati così da fare emergere i fatti e le informazioni salienti. Spesso è necessario rinunciare a dettagli minuti che sono relativi a singole unità per tentare di dare una lettura di insieme del fenomeno con riferimento al collettivo da cui i dati sono tratti.

È una operazione essenziale per trasformare i dati grezzi in informazione.

L'attenzione va quindi concentrata su quali caratteristiche del collettivo sono di interesse e decidere se e come queste possono essere riassunte attraverso opportune sintesi numeriche.

I due aspetti dell'insieme di dati sui quali si concentrano le maggiori attenzioni sono:

- **la tendenza centrale**: ovvero posso con un singolo valore riassumere l'ordine di grandezza del fenomeno di interesse?
- **la dispersione (o variabilità)**: cioè, un singolo valore può fornire un'idea di quanto i dati realtivi alla variabile siano diversi tra loro o, viceversa, quanto si somiglino?

Al fine di trattare tale argomento, e anche per sviluppi successivi, vale la pena di introdurre una notazione generale.

Rappresenteremo i valori assunti da una variabile quantitiva X per un insieme di n unità con la notazione $x_1, x_2, \dots, x_n$.

3.2.1 Indici di tendenza centrale (un solo valore al posto di tanti): le medie

Gli indici di tendenza centrale sono anche noti come **medie**. Si suole distingure far *medie analitiche* e *medie di posizione*.

3.2.1.1 Medie analitiche

1. **La media aritmetica**

La media aritmetica M è di gran lunga la media analitica più nota e si calcola come

$$M = \sum_{i=1}^{n} \frac{x_i}{n}$$

È già noto che in `R` esiste la funzione `mean()` che calcola la media aritmetica degli elementi di un vettore di dati quantitativi.

La media aritmetica ha alcune importanti proprietà:

a. assume un valore compreso fra $x_{(1)} = min(x_i)$ e $x_{(n)} = max(x_i)$
b. la somma delle differenze fra ciascun valore osservato e la media, in valore assoluto, è nulla: $\sum_{i=1}^{n}(x_i - M) = 0$
c. ha la seguente proprietà di minimo $M = \arg\min_{a \in R} \sum_{i=1}^{n}(x_i - a)^2$

La terza proprietà si può leggere come segue:

se si decide che invece di ogni singolo dato x_i si usa una sintesi che è a potrei avere una perdita di informazione che decido di misurare con la differenza (scarto)

dalla media al quadrato $(x_i - a)^2$: la perdita di informazione totale è quindi $\sum_{i=1}^{n}(x_i - a)^2$. La media M è quel valore che rende minima tale perdita.

Inoltre dalla seconda proprietà consegue che $\sum_{i=1}^{n} x_i = nM$. Per cui si può dire che la media M è quel valore che sostituito a ogni singolo dato lascia invariata la somma.

Vale la pena di sottolineare che in molte situazioni ad ogni valore x_i può essere associato un peso p_i che denota l'importanza di tale valore. Un esempio riguarda i voti conseguiti negli esami: com'è noto essi hanno pesi diversi in quanto possono avre diversi CFU. Ma vale anche per il caso in cui alcuni valori vengono osservati più volte, accade ad esempio per le variabili quantitative discrete, e i dati quantitativi vengono riassunti in una tabella di frequenza: in tal caso la frequenza denota l'importanza di tale valore, il suo peso. In tal caso nel calcolare la media aritmetica occorre tenere conto dei pesi e si definisce la **media aritmetica ponderata**

$$M = \sum_{i=1}^{n} \frac{x_i p_i}{\sum_i^n p_i}$$

2. **Altre medie analitiche**

Se si parte da questa ultima notazione, si potrebbe cercare un valore che analogamente alla media aritmetica, lascia invariata una certa funzione dei dati.

a. Ad esempio se invece della somma si considera il prodotto dei dati e si cerca la costante M_g che lascia invariata questa funzione, cioè: $\prod_{i=1}^{n} x_i = M_g^n$ si ottiene la **media geometrica**.

$$M_g = \left[\prod_{i=1}^{n} x_i\right]^{\frac{1}{n}}$$

b. Se il valore cercato lascia invariata la somma dei reciproci, si ottiene la **media armonica**

$$M_{ar} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

c. Analogamente, si ottengono le medie di potenze r-esima (con r intero)

$$M_r = \left[\sum_{i=1}^{n} \frac{x_i^r}{n}\right]^{\frac{1}{r}}$$

che hanno la proprietà di esser funziona monotona non decrescente di r. In virtù di ciò, e osservando che si potrebbe dimostrare che consoiderando il limite per r che tende a 0 si ottiene la media geometrica, si ha quindi il seguente ordinamento $M_{ar} \leq M_g \leq M \leq M_2$ (ove l'uguaglianza vale se tutti i dati hanno lo stesso valore. Da questo segue anche che $M_2^2 \geq M^2$.

d. Si noti che la potenza r-esima della media potenziata di ordine r definisce il **momento empirico r-esimo**.

3.2.1.2 I Quantili e le medie di posizione

Le **medie (e gli indici) di posizione** sono particolari valori che hanno una posizione definita rispetto la sequenza (ordinata) dei dati (si noti che non necessariamente essi appartengano alla sequenza stessa). A tal fine è importante introdurre il concetto di **quantile empirico**.

Fissata una proporzione p (con $0 \leq p \leq 1$) si definisce x_p **quantile empirico** di ordine p, con $p \in [0, 1]$, quel valore $x \in \mathbb{R}$ tale che

$$x_p : \frac{\#(x_i \leq x_p)}{n} = p$$

.

Si noti che dato un valore p:

1. non è detto che esista un unico valore x_p che realizzi la condizione data;
2. potrebbe non esistere alcun valore che realizzi la condizione data, ma si potranno ritrovare valori di quantili $x_{p'}$ con p' molto prossimo a p.

Se si considera la sequenza ordinata $x_{(1)}, x_{(2)} \ldots, x_{(n-1)}, x_{(n)}$ per cui

$x_{(1)} \leq x_{(2)} \leq \ldots, \leq x_{(n-1)} \leq x_{(n)}$

i valori definiscono i quantili di ordine $\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, \frac{n}{n}$. Pertanto, il generico valore della sequenza ordinata $x_{(i)}$ rappresenta il quantile di ordine p, x_p con $p = \frac{i}{n}$.

Si noti, inoltre, il rilievo che assume in questo contesto la definizione che specifica x_p come il valore per cui sia p la proporzione di valori **minori o uguali** a esso. Se disponiamo di un insieme di dati non particolarmente numeroso chiedere di contare i dati **strettamente minori** di x_p o quelli minori o uguali può fare differenza. In tal caso infatti la sequenza ordinata fornirebbe i quantili $\frac{0}{n}, \frac{1}{n}, \ldots, \frac{n-2}{n}, \frac{n-1}{n}$.

Per tale motivo a volte si preferisce adottare una diversa convenzione per cui i dati ordinati $x_{(1)} \leq x_{(2)} \leq \ldots, \leq x_{(n-1)} \leq x_{(n)}$ rappresentano i quantili di ordine $\frac{0.5}{n}, \frac{1.5}{n}, \ldots, \frac{n-1.5}{n}, \frac{n-0.5}{n}$. Cioè secondo tale convenzione $x_{(i)}$ rappresenta il quantile di ordine $x_{\frac{i-0.5}{n}}$ di ordine $p = \frac{i-0.5}{n}$.

Si noti che se n è grande ci si attende che le differenze derivanti dalle diverse convenzioni saranno di scarso rilievo.

3.2.1.3 La mediana e i quartili

Come detto. alcuni particolari quantili hanno uno speciale rilievo. In particolare, i quantili $x_{0.25}, x_{0.5}, x_{0.75}$ sono detti **quartili** e denominati, ripettivamente, primo, secondo e terzo quartile.

Particolare rilievo ha poi il secondo quartile che è anche detto **mediana**.

La mediana Me è quindi definita come quel valore per cui il 50% dei dati è inferiore (o uguale) a esso e una analoga percentuale è superiore.

Come si è osservato sopra, se cerchiamo il quantile $Me = x_{0.5}$ per un insieme di dati (con la condizione che siano il 50% quelli minori o uguali di Me) potremmo:

- trovare infiniti valori che realizzino esattamente la condizione data (se ad esempio n è pari)
- non trovare alcun valore che realizzi la condizione data (ad esempio se n è dispari).

La cosa si risolve convenzionalmente ponendo:

- $Me = x_{(\frac{n+1}{2})}$ se n è dispari
- $Me = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$ se n è pari.

Si noti tuttavia che nel caso di n pari, qualsiasi valore $x_{(\frac{n}{2})} \leq x \leq x_{(\frac{n}{2}+1)}$ avrebbe la proprietà richiesta alla mediana.

La **mediana** è usata spesso in coppia con la media aritmetica per caratterizzare il valore centrale di una distribuzione. Si noti che la mediana ha una proprietà analoga a quella della media aritmetica. In particolare

$$Me = \arg\min_{a \in \mathbb{R}} \sum_{i=1}^{n} |x_i - a|$$

È cioè quel valore per cui è minima la somma dei valori assoluti delle distanze dei singoli dati da esso.

Come vedremo meglio, con esempi su vari insiemi di dati, la mediana è molto meno sensibile alla presenza di valori estremi (e anomali rispetto alla massa dei dati). In tal senso essa è più **resistente**. La media arimetica invece tende a essere molto influenzata da pochi valori estremi molto diversi dalla massa dei dati.

È quindi una buona regola guardare a entrambi i valori nel riassumere un insieme di dati.

Come per la media, esiste una funzione in `R` per calcolare la mediana usando la convenzione esposta.

```
median(Cars93$Length)
median(AutoBi$LOSS)
```

```
## [1] 183
## [1] 2.331
```

Anche il primo e il terzo quartile sono molto informativi sulla tendenza centrale della distribuzione: si noti infatti che, per definizione, il 50% dei valori centrali sono compresi fra di essi.

Per quanto riguarda la determinazione di tali due quartili, in particolare se n non è molto elevato, vi sono difficoltà analoghe a quanto visto per la mediana. Per individuare i quantili occorre, come per la mediana, ricorrere a qualche convenzione. In R si usa l'interpolazione lineare fra i due quantili successivi osservati x_{p_1} e x_{p_2} tali che $p_1 \leq p \leq p_2$. (se si vuole approfondire si veda l'help della funzione `quantile` in cui vengono proposte anche altre strategie di approssimazione dei quantili)

Va infine ricordato che per caratterizzare la distribuzione dei dati oltre ai quartili (che dividono i dati ordinati in quattro porzioni ugualmente numerose) spesso si definiscono i **decili** (cioè i quantili $x_{0.1}, x_{0.2}, \ldots, x_{0.9}$) o i percentili (cioè i quantili $x_{0.01}, x_{0.02}, \ldots, x_{0.98}, x_{0.99}$).

In R la funzione `quantile()` consente di calcolare i quantili per diversi valori di p. La convenzione in questo caso è che la più piccola osservazione sia il quantile di ordine 0 e la più grande sia quello di ordine 1.

```
# se non si indica nulla vengono calcolati i quartili e i quantili di
# ordine 0 e 1, cioè il massimo e il minimo
quantile(AutoBi$LOSS)
```

```
##         0%        25%        50%        75%       100%
##    0.00500    0.64000    2.33100    3.99475 1067.69700
```

```
# si possono calcolare i decili per la variabile LOSS
quantile(AutoBi$LOSS, probs=seq(0,1,.1))
```

```
##        0%       10%       20%       30%       40%       50%       60%       70%
##    0.0050    0.2380    0.4842    0.9970    1.6746    2.3310    2.9700    3.6555
##       80%       90%      100%
##    4.5148    8.0765 1067.6970
```

```
quantile(AutoBi$LOSS)
```

```
##         0%        25%        50%        75%       100%
##    0.00500    0.64000    2.33100    3.99475 1067.69700
```

```
# Si verifichi infine che, banalmente, il quantile di ordine 0.5 è la mediana
median(AutoBi$LOSS)==quantile(AutoBi$LOSS, probs=0.5)
```

```
##  50%
## TRUE
```

3.2.1.4 Riassunto dei 5 valori

Si è visto che i 3 quartili danno un riassunto interessante sulla distribuzione.

La mediana è un indice di tendenza centrale e i due quartili esterni danno un'idea di dove sia collocato il 50% centrale della distribuzione. Aggiungendo a questi il valore massimo e il valore minimo osservati nella sequenza di dati abbiamo un'idea di come i dati siano distribuiti sulle code.

I 5 valori (minimo, I quartile, mediana, III quartile, massimo) sono detti riassunto dei 5 valori (in inglese **five numbers summary**) e esiste una funzione `fivenum()`che li restituisce se applicata a un vettore di dati (numerici), inoltre gli stessi vengono restituiti anche nella funzione `summary()`, che, quando applicata a un vettore numerico, fornisce inoltre anche la media aritmetica. In entrambi i casi i valori mancanti vengono rimossi automaticamente. Ad esempio:

```
#  per la variabile LOSS  si ottengono il riassunto dei 5 valori e il summary
fivenum(AutoBi$LOSS)
summary(AutoBi$LOSS)
```

```
## [1]    0.0050    0.6400    2.3310    3.9955 1067.6970
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.005    0.640    2.331    5.954    3.995 1067.697
```

3.2.1.5 Le medie sfrondate (trimmed)

Come si è detto, il principale indice di posizione, cioè la media aritmetica, rischia di essere molto influenzata dalla presenza di (pochi) valori difformi rispetto alla gran parte dei dati.

Un tema rilevante è se tali dati anomali (outliers) debbano essere esclusi dalle successive analisi o essere mantenuti. Come vedremo, su tale aspetto non esiste una regola generale perchè dipende dalla natura dei dati e in alcuni casi i fenomeni che osserviamo sono tali da poter ammettere che vi siano osservazioni estreme e quindi eliminare tali osservazioni potrebbe essere del tutto arbitrario e inappropriato.

Il tema che si pone è quindi quello comunque di trovare riassunti della tendenza centrale dei dati che non siano eccessivamente influenzati da tali dati anomali. A tal fine si possono calcolare delle medie dei dati disponibili *sfrondando* l'insieme dei dati dei valori troppo grandi e/o troppo piccoli.

Questo conduce a calcolare la **media sfrondata** fissando una proporzione $\alpha \in [0, 1]$ di valori da escludere tra quelli estremi.

La funzione `R` che calcola la media di un vettore ha fra i parametri la possibilità di indicare se si vuole una media sfrondata. Ad esempio:

```
# calcoliamo la media sfrondata al 5% per la variabile LOSS
mean(AutoBi$LOSS, na.rm=TRUE, trim=0.05)
mean(AutoBi$LOSS, na.rm=TRUE)
```

```
## [1] 2.858701
## [1] 5.953461
```

3.2.2 Indici di dispersione (o variabilità)

L'altro importante aspetto di una distribuzione di dati è costituito dalla loro **dispersione**, ovvero si tratta di valutare e sintetizzare quanto i dati siano diversi

l'uno dall'altro. Tale aspetto è anche detto **variabilità**.

Anche in questo caso si possono calcolare degli indici numerici che riassumano tale caratteristica e si possono proporre versioni più resistenti ai valori anomali.

3.2.2.1 La varianza, lo scarto quadratico medio, il coefficiente di variazione

La più nota misura di dispersione per un insieme di n dati è la varianza. Essa è basata sulla quantità

$$DEV = \sum_{i=1}^{n}(x_i - M)^2$$

ovvero la somma degli scarti dalla media aritmetica che è detta **devianza**.

La **varianza** V è la devianza media pertanto essa è definita come

$$V = \frac{\text{devianza}}{n} = \sum_{i=1}^{n}\frac{(x_i - M)^2}{n}$$

Quando tuttavia la varianza viene calcolata nell'ambito di un approccio inferenziale, ovvero essa non è semplicemente un indice di dispersione ma assume il ruolo di stimatore della varianza di una popolazione infinita da cui è tratto un campione casuale, si conviene usare la divisione per $n-1$. Poichè quest'ultimo contesto è piuttosto comune, non deve stupire che in `R` la varianza, calcolata con la funzione `var()`, è definita come $\frac{\text{devianza}}{n-1} = V\frac{n}{n-1}$.

È evidente che per n elevato i valori che si ottengono nei due casi non si discostano molto e l'interpretazione in termini descrittivi non cambia nella sostanza. Tuttavia è bene essere consapevoli di quale versione si sta usando.

La devianza, e quindi la varianza, è tanto più grande quanto più i dati sono distanti da un valore centrale, che è la media aritmetica, che è candidato a rappresentarli tutti.

La varianza è calcolata in termini di distanze al quadrato e quindi non è espressa nella stessa unità di misura della variabile. Per tale motivo è frequente ricorrere alla radice quadrata dell'indice, che è detto **scarto quadratico medio** (SQM) o **deviazione standard** (in inglese *standard deviation*)

$$\text{SQM} = sd = \sqrt{V}$$

che è espressa nella stessa unità di misura della variabile. In `R` la funzione che calcola lo scarto quadratico medio è `sd()` (basata sempre sulla varianza divisa per n-1).

Per definizione la varianza (e lo SQM) assumono solo valori positivi e sono pari a 0 solo nel caso che tutti i valori sono uguali tra loro (variabilità nulla).

L'interpretazione dell'indice non è sempre agevole non essendo definito un limite superiore. È spesso interessante utilizzarla per comparare, ad esempio, la variabilità in due insiemi di dati che hanno medie non troppo diverse tra loro.

Si riporta un esempio in cui con R si calcola la varianza in due insiemi di dati.

```
# si calcola la varianza  e la deviazione standard per la variabile LOSS
# per i maschi e per le femmine
"Maschi"
var(AutoBi$LOSS[AutoBi$CLMSEX==1], na.rm=TRUE)
sd(AutoBi$LOSS[AutoBi$CLMSEX==1], na.rm=TRUE)
"Femmine"
var(AutoBi$LOSS[AutoBi$CLMSEX==2], na.rm=TRUE)
sd(AutoBi$LOSS[AutoBi$CLMSEX==2], na.rm=TRUE)
##
```

```
## [1] "Maschi"
## [1] 301.0511
## [1] 17.35082
## [1] "Femmine"
## [1] 1745.729
## [1] 41.78192
```

Le medie nei due gruppi sono simili

```
# si ottiene la media per la variabile LOSS per i maschi e per le femmine
mean(AutoBi$LOSS[AutoBi$CLMSEX==1], na.rm=TRUE)
mean(AutoBi$LOSS[AutoBi$CLMSEX==2], na.rm=TRUE)
##
```

```
## [1] 5.652647
## [1] 6.213765
```

In questo caso, il confronto darebbe un'indicazione che vi è più variabilità nel caso delle femmine. Tuttavia è facile notare che tale maggiore variabilità è fortemente condizionata dalla presenza di valori anomali. A tal fine si consideri il summary dei due insiemi:

```
# varianza per la variabile LOSS per i maschi e per le femmine
summary(AutoBi$LOSS[AutoBi$CLMSEX==1])
summary(AutoBi$LOSS[AutoBi$CLMSEX==2])
##
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.030   0.628   2.372   5.653   3.901 222.405      12
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
##    0.0050    0.6905    2.2270    6.2138    4.0252 1067.6970        12
```

Si vede chiaramente come nel caso delle femmine si ottiene un valore estremamente elevato. Esso ha impatto sulla media (non sulla mediana) e di conseguen-

za anche sulla varianza. La varianza è quindi, come la media, un indice poco "resistente".

3.2.2.2 Altre misure di variabilità: coefficiente di variazione, scarto interquartile, MAD

Come per la tendenza centrale anche per la variabilità esistono altre misure che hanno caratteristiche che li rendono in alcuni casi preferibili alla varianza o quanto meno è spesso consigliabile in una fase esplorativa affiancarli ad essa.

1. **lo scarto interquartile: SI**: è semplicemente definito come la differenza fra il III quartile e il I quartile, ovvero $SI = x_{0.75} - x_{0.25}$. Ovviamente risente meno dei valori estremi, ma tiene conto solo della dispersione nella parte centrale della distribuzione. Ovviamente è anche calcolabile direttamente dai dati che fornisce `summary`
2. **il MAD**: Scarto assoluto mediano dalla mediana. esso è definito come

   $$MAD = \text{Mediana}(|x_i - Me|)$$

   È un indice resistente, cioè non risente di eventuali valori anomali. Si calcola con la funzione `mad()`.
3. il **coefficiente di variazione**: spesso il confronto fra la variabilità in due collettivi ha senso se nei due collettivi la variabile presenta valori che hanno circa lo stesso ordine di grandezza (i due gruppi cioè hanno media non eccessivamente diversa). Per ovviare a questo problema si introduce il coefficiente di variazione che è quindi un indice di variabilità relativo. Esso è semplicemente definito come il rapporto fra scarto quadratico medio e media aritmetica.

Vediamo ancora l'esempio precedente con i dati di `LOSS` separatamente per maschi e femmine.

```
# si ottiene la varianza e il MAD per la variabile LOSS sia per i maschi e
# per le femmine
"varianza"
var(AutoBi$LOSS[AutoBi$CLMSEX==1], na.rm=TRUE)
var(AutoBi$LOSS[AutoBi$CLMSEX==2], na.rm=TRUE)
"MAD"
mad(AutoBi$LOSS[AutoBi$CLMSEX==1], constant=1, na.rm=TRUE)
mad(AutoBi$LOSS[AutoBi$CLMSEX==2], constant=1, na.rm=TRUE)
##
```

```
## [1] "varianza"
## [1] 301.0511
## [1] 1745.729
## [1] "MAD"
## [1] 1.6635
## [1] 1.6755
```

Si noti come la variabilità misurata dal MAD è molto simile nei due gruppi a testimonianza dell'impatto che l'unico valore estremo nel gruppo delle femmine aveva sul calcolo della varianza.

3.2.3 Simmetria e curtosi

Sintetizzare solo la tendenza centrale o la dispersione di una variabile quantitativa lascia in ombra molte altre caratteristiche della distribuzione dei dati. È pertanto buona regola guardare a altre specificità della distribuzione dei valori e eventualmente cercare dei valori che li sintetizzino.

Una rilevante caratteristica di un insieme di dati è legata alla tendenza degli stessi a distribuirsi in modo molto diverso nella parte sinistra della distribuzione (i valori bassi quindi) rispetto alla parte destra.

Se si fa riferimento a un indice di tendenza centrale, ad esempio la mediana, potrebbe accadere che i dati a destra siano molto più dispersi rispetto a quelli a sinistra ovvero la distribuzione presenta una coda a destra più lunga. Si dirà che la distribuzione ha una asimmetria positiva. Viceversa, se vi è una coda a sinistra più lunga si parla di asimmetria negativa.

Se accade questo i dati evidenziano una **asimmetria** (*skewness* in inglese) nella distribuzione della variabile. Il concetto complementare è quindi quello di **simmetria** che riguarda la situazione in cui i dati a destra e a sinistra del centro della distribuzione esibiscono un comportamento analogo.

È facile apprezzare la simmetria (o l'asimmetria) guardando ad alcuni percentili della distribuzione. Vediamo due esempi in cui consideriamo i decili di due variabili nei dati di `AutoBI`:

```
# si otengono i decili per le variabile LOSS e CLMAGE
"LOSS"
quantile(AutoBi$LOSS, probs = seq(0,1,0.1), na.rm=T, digits=2)

"CLMAGE"
quantile(AutoBi$CLMAGE, probs = seq(0,1,0.1), na.rm=T)
```

```
## [1] "LOSS"
##       0%      10%      20%      30%      40%      50%      60%      70%
##   0.0050   0.2380   0.4842   0.9970   1.6746   2.3310   2.9700   3.6555
##      80%      90%      100%
##   4.5148   8.0765 1067.6970
## [1] "CLMAGE"
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    0   13   18   21   26   31   36   41   47   55   95
```

Confrontando i quantili a destra e a sinistra della mediana è evidente che la distribuzione della variabile età è meno asimmetrica della variabile `LOSS`.

Si tratta quindi di proporre degli indici sintetici che riassumano tale caratteristica (ovvero l'asimmetria).

3.2.3.1 Indici di simmetria basati su particolari quantili

Come si è visto un modo per valutare la simmetria è quello di guardare la posizione di alcuni quantili rispetto alla mediana. Avendo già discusso del rilievo che si dà ai tre quartili, un indice può esser costruito guardando alla posizione del secondo quartile $Q2 = x_{.5}$, ovvero la mediana, rispetto agli altri quartili $Q1 = x_{0.25}$ e $Q3 = x_{0.75}$.

- in assenza di asimmetria sarà $(Q3 - Q2) = (Q2 - Q1)$
- se vi è asimmetria positiva (coda a destra lunga) $(Q3 - Q2) > (Q2 - Q1)$
- se vi è asimmetria negativa (coda a destra lunga) $(Q3 - Q2) > (Q2 - Q1)$

Allora un indice (di Galton) basato sui quartili è il seguente:

$$G = \frac{Q3 + Q1 - 2Q2}{Q3 - Q1}$$

La stessa idea potrebbe applicarsi a quantili più estremi, ad esempio, il primo e il nono decile $x_{0.1}$ e $x_{0.9}$ confrontati con la mediana.

$$K = \frac{x_{0.9} + x_{0.1} - 2x_{0.5}}{x_{0.9} - x_{0.1}}$$

3.2.3.2 Indici basati sul confronto fra media e mediana

Abbiamo già notato che il diverso comportamento di media e mediana è legato alle quantità che minimizzano e alla tendenza della media a essere influenzata da valori sulla coda della distribuzione. Quindi:

- se la media è circa uguale alla mediana è un sintomo di assenza di asimmetria
- se la media è superiore (inferiore) alla mediana è un sintomo di asimmetria positiva (negativa) in quanto avrò prevalenza di valori (alcuni molto distanti dalla media)sulla coda destra (sinistra)

Pertanto si può costruire un indice (coefficiente di asimmetria) come segue

$$\frac{3(M - Me)}{sd}$$

3.2.3.3 L'indice di asimmetria

L'altra possibilità è quella di costruire un indice basato sugli scarti dalla media elevati al cubo $(x_i - M)^3$. Ovviamente è rilevante in questo caso se uno scarto è positivo o negativo.

L'**indice di asimmetria** è definito come:

$$\gamma = \frac{\sum_{i=1}^{n} \frac{(x_i - M)^3}{n}}{sd^3}$$

Tale indice è nullo per una distribuzione simmetrica (a ogni valore superiore alla media ne corrsponderebbe un altro equidistante e inferiroe alla media) mentre assume valori positivi nel caso di asimmetria positiva (coda destra lunga) e negativi nel caso di asimmetria negativa (coda sinistra lunga).

In `R` esso può essere calcolato con la funzione `skewness()` che si trova nel pacchetto `moments`

```
library(moments)
# si ottengono varie misure di asimmetria per le variabili LOSS e CLMAGE
"LOSS"
"Indice di Galton"
Qloss<-fivenum(AutoBi$LOSS, na.rm=T)
g<-(Qloss[4]+Qloss[2]-2*Qloss[3])/(Qloss[4]-Qloss[2])
"Indice di asimmetria"
skewness(AutoBi$LOSS)
"CLMAGE"
```

```
## [1] "LOSS"
## [1] "Indice di Galton"
## [1] "Indice di asimmetria"
## [1] 25.68795
## [1] "CLMAGE"
```

3.2.3.4 Indice di curtosi

La curtosi è un'ulteriore caratteristica della distribuzione di una variabile quantitativa. Essa misura la tendenza, in distribuzioni simmetriche, di mostrare code corte che scendono velocemente allontanandosi dal centro della distribuzione oppure code più 'pesanti' Per cui non è difficile osservare valori distanti dal centro anche allontanandosi da esso.

Il punto di riferimento è il modello della gaussiana e in genere si valuta la curtosi proprio andando a comparare il comportamento delle code con quello della gaussiana.

L'indice di **curtosi** δ è definito come segue

$$\delta = \frac{\sum_{i=1}^{n} \frac{(x_i - M)^4}{n}}{sd^4}$$

Anch'esso è disponibile nel pacchetto sopracitato (`moments`) è sempre maggiore di 0 ed è pari a 3 se la curtosi è pari a quella della gaussiana. Se è minore di 3 allora la distribuzione ha code piu corte (rispetto alla gaussiana, leptocurtosi) nel caso contrario più lunghe (platicurtosi).

```
# si calcola l'indice di curtosi per la variabile CLMAGE
"CLMAGE"

kurtosis(AutoBi$CLMAGE, na.rm=TRUE)

# si osservi che la curtosi non è molto diversa da quella della gaussiana.
```

```
## [1] "CLMAGE"
## [1] 3.042604
```

3.2.4 Altri grafici per variabili quantitative

3.2.4.1 Il diagramma a scatola con baffi (boxplot)

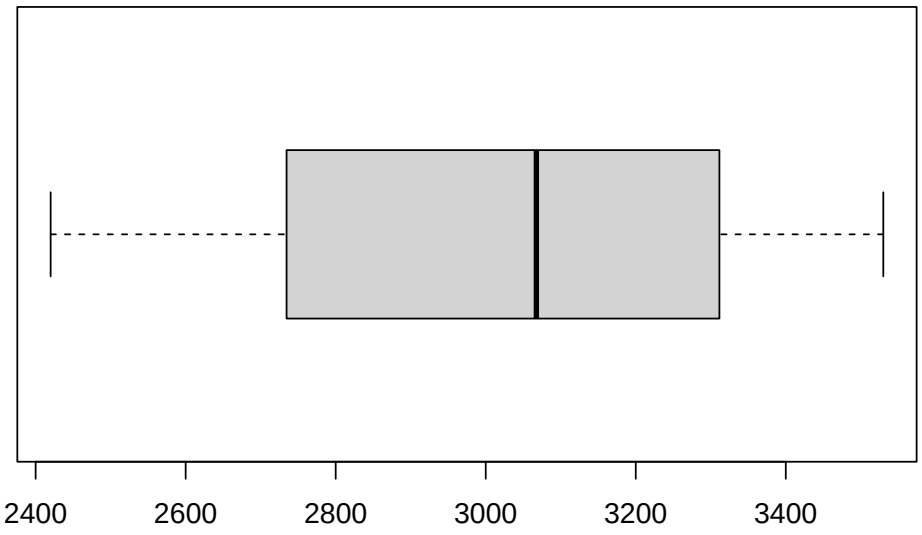
Come si è visto il riassunto dei 5 numeri fornisce i 3 quartili che danno idea di diversi aspetti della distribuzione: tendenza centrale (la mediana), dispersione (scarto interquartile), simmetria (indice di Galton). Inoltre il minimo e il massimo danno indicazione di cosa accade sulla coda destra e sinistra.

Un'idea molto semplice ed efficace per riassumere graficamente le caratteristiche essenziali di una variabili quantitativa è quella che conduce al cosiddetto **diagramma a scatola con baffi** (*box and whiskers plot*). Conviene illustrarlo con un esempio utilizzando la funzione di `R` che permette di ottenerlo cioè `boxplot()`.

```
fivenum(neo$Peso)
```

```
## [1] 2420.0 2734.5 3067.5 3311.5 3530.0
```

```
boxplot(neo$Peso, horizontal=TRUE)
```



```
# Si noti il parametro horizontal che ci consente di mettere la
# scatola verticalmente o orizzontalmente
```

Come si vede si sono usati i dati forniti dal riassunto dei 5 numeri per disegnare la scatola. La scatola è posizionata con riferimento a un asse lungo cui sono i valori della variabile così che i bordi corrispondano al I e al III quartile (Q1 e Q3) mentre al suo centro la linea è in corrispondenza della mediana (Q2):

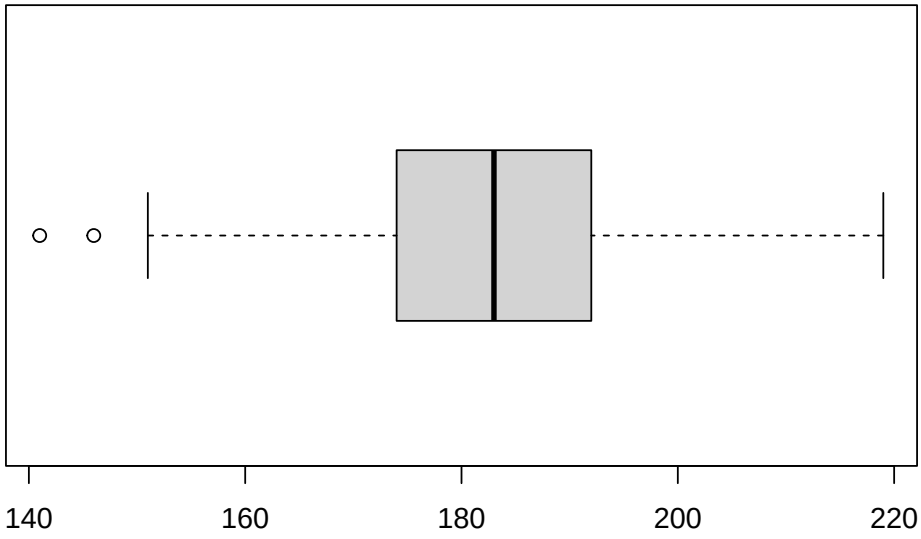
1. la larghezza della scatola ci dà quindi un'indicazione della dispersione;
2. la sua posizione ci da una indicazione di dove si trova il 50% dei valori centrali della variabile e assieme alla linea in corrispondenza della mediana ci dà indicazione dei valori centrali;
3. la posizione della mediana nella scatola ci informa sulla simmetria;
4. i baffi ci danno indicazione del comportamento sulle code.

Se tuttavia proviamo lo stesso grafico con una diversa variabile

```
fivenum(Cars93$Length)
```

```
## [1] 141 174 183 192 219
```

```
boxplot(Cars93$Length, horizontal=TRUE)
```



In questo caso il baffo sinistro non si estende fino al minimo. Inoltre vediamo sul grafico due punti rappresentati separatamente.

Questo perchè in realtà nel diagramma a scatola si evidenziano quei punti che sono ritenuti distanti dal resto dei dati (a destra o a sinistra). Tali valori sono detti *outliers* e vengono rappresentati isolatamente sul grafico.

La definizione di *outlier* nel diagramma a scatola si deriva dalla seguente regola:

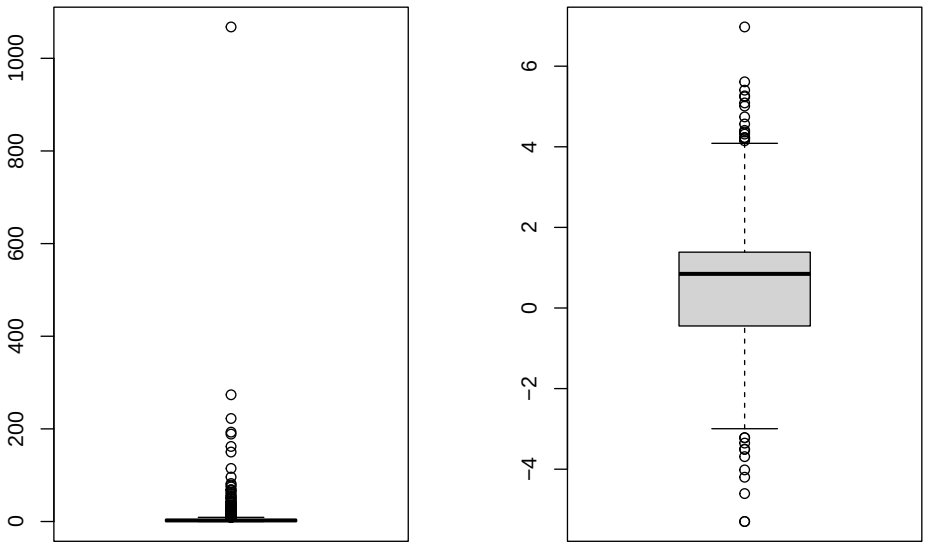
- sono *oulier* (valori anomali) quei valori che sono più distanti dai bordi della scatola (cioè dai quartili) più di una volta e mezza la differenza interquartile $SI = Q3 - Q1$. Pertanto tutti i punti che sono superiori a $Q3 + 1.5SI$ o inferiori a $Q1 - 1.5SI$ verranno annotati separatamente sul grafico;
- se essi esistono (a destra o a sinistra) di conseguenza il baffo non va esteso fino al massimo o al minimo valore osservato, va invece esteso:
  - a destra, fino al più grande valore che non sia segnalato come outlier (cioè il massimo valore osservato che risulti inferiore a $Q3 + 1.5SI$);
  - a sinistra, fino al più piccolo valore che non sia segnalato come outlier (cioè il minimo valore osservato che risulti superiore a $Q1 - 1.5SI$).

Il diagramma a scatola è una rappresentazione grafica usata molto frequentemente anche perchè è adeguata anche nel caso di un insieme molto limitato di casi (n=20 o 30).

Attenzione però che nel caso i valori siano pochi la regola con cui si determinano i quartili (che ha margini di arbitrarietà come abbiamo già discusso) può cambiare l'apparenza del grafico.

Anche nel caso del diagramma a scatola la presenza di code molto lunghe e/o dati molto estremi esso potrebbe risultare inefficace. Vediamo cosa accade ad esempio con la variabile *AutoBi*LOSS$ che abbiamo visto esser caratterizzata da una coda a destra lunghissima (forte asimmetria positiva) asimmetrica.

```
par(mfrow=c(1,2))
boxplot(AutoBi$LOSS)

# i valori elevatissimi presenti nella variabile vengono rappresentati
# separatamente, ma questo fa sì che la scatola diventi difficile da
# apprezzare sul medesimo grafico.

# Proviamo a usare la trasformata logaritmica
boxplot(log(AutoBi$LOSS))
```



3.2.4.2 La funzione di ripartizione empirica

Una semplice idea per rappresentare graficamente un insieme di dati è quello di ottenere una versione empirica della funzione di ripartizione (o della funzione cumulata) corrispondente quindi alla analoga funzione $F(x)$ definita per una variabile aleatoria X.

Com'è noto per una variabile aleatoria continua X la funzione di ripartizione (anche detta distribuzione cumulata di probabilità) è definita come $F(x) = Pr(X \leq x)$, ovvero fornisce la probabilità che si ottenga un valore aleatorio della X inferiore o uguale a x.

Tale funzione è definita per $x \in \mathbb{R}$

- monotona non decrescente;
- pari a 0 per valori inferiori o uguali al limite inferiore del supporto della variabile X;
- pari a 1 per valori superiori al limite superiore del supporto della variabile X.

Inoltre l'inversa della funzione di ripartizione fornisce i quantili per cui il quantile $x_p = F^{-1}(p)$. Ovviamente il quantile per ogni $0 \leq p \leq 1$ esiste se la funzione di ripartizione non ha discontinuità.

L'equivalente empirico di tale funzione, denotato con $\hat{F}(x)$ avendo osservato l'insieme di dati $x_1, x_2, \ldots, x_n$ per una variabile quantitativa, si ottiene cercando il valore che $\forall x \in \mathbb{R}$ fornisce la proporzione di unità inferiori o pari a x, ovvero

$$\hat{F}(x) = \text{proporzione}(x_i \leq x) = \frac{\text{numero di valori} \leq x}{n}$$

Riprendendo la notazione introdotta per i dati ordinati $x_{(1)} \leq x_{(2)} \leq \ldots, \leq x_{(n-1)} \leq x_{(n)}$ la funzione di ripartizione empirica (cumulata empirica) sarà pari a:

- $\hat{F}(x) = 0$ se $x \leq x_{(1)}$;
- in $x = x_{(1)}$ la funzione fa un salto ed è pari a $\frac{1}{n}$ e rimane costante fino al prossimo valore $x_{(2)}$;
- in $x = x_{(2)}$ la funzione fa un ulteriore salto ancora di altezza $\frac{1}{n}$;
- e così via fino a $x_{(n)}$ partire dal quale la funzione assume il valore 1;
- se vi sono valori ripetuti il salto sarà pari a $\frac{m}{n}$ ove m è il numero di valori ripetuti.

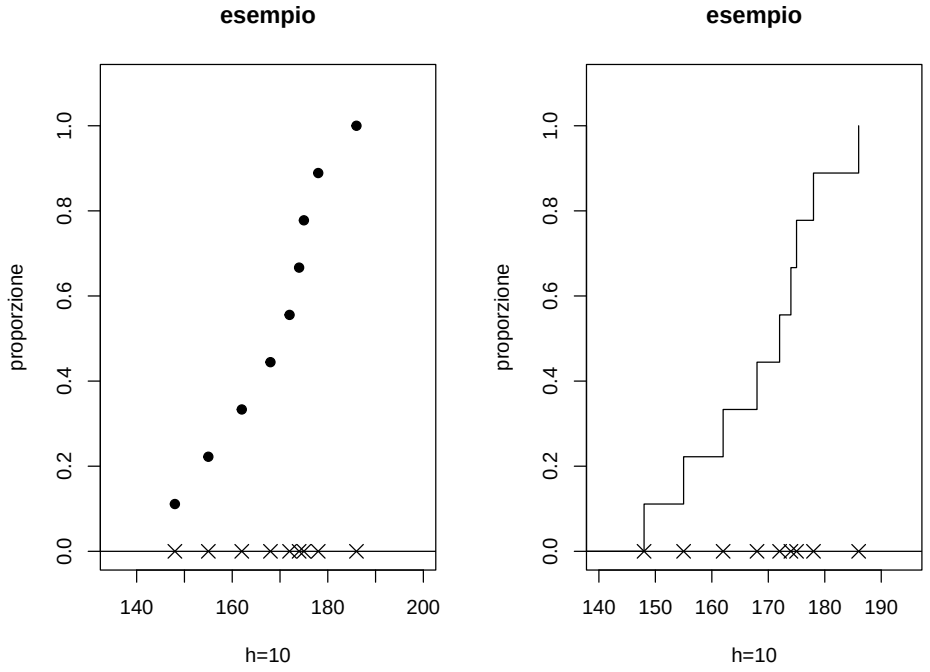
Si tratta quindi di una funzione a scalini definita come:

$$\hat{F}(x) = \begin{cases} 0 & \text{se} \quad x < x_{(1)} \\ \frac{i}{n} & \text{se} \quad x_{(i)} \leq x < x_{(i+1)} \\ 1 & \text{se} \quad x \geq x_{(n)} \end{cases} \qquad \text{per} \quad i = 1, 2, \ldots, n-1$$

.

Si può facilmente tracciare tale grafico in R come si vede in questo semplice esempio:

```
par(mfrow=c(1,2))
# consideriamo l'insieme di 9 dati nel vettore xx
xx<-c(148, 155, 162, 168, 172, 174, 175, 178, 186)

# rappresentiamo i punti su un grafico
# i parametri ylim e xlim permettono di definire
# i limiti per le coordinate x e y, pch permette di scegliere il simbolo
# (4 è il simbolo X), cex le dimensioni del simbolo.
nn<-length(xx)
plot(xx,rep(0,nn), ylim=c(0,1.1), xlim=c(135,200), pch=4, cex=1.5,
     main="esempio", xlab="h=10", ylab="proporzione")
abline(0,0)
# abline(a,b) traccia una retta con intercetta e coefficiente angolare
# con a e b assegnati
# points() permette di aggiungere al grafico esistente nuovi punti
points(c(0,xx),c(0,1:9)/nn, pch=19)
# e il parametro type="s" consente di ottenere una funzione a gradini
plot(xx,rep(0,nn), ylim=c(0,1.1), xlim=c(140,195), pch=4, cex=1.5,
     main="esempio", xlab="h=10", ylab="proporzione")
abline(0,0)
points(c(0,xx),c(0,1:9)/nn, type="s")
```

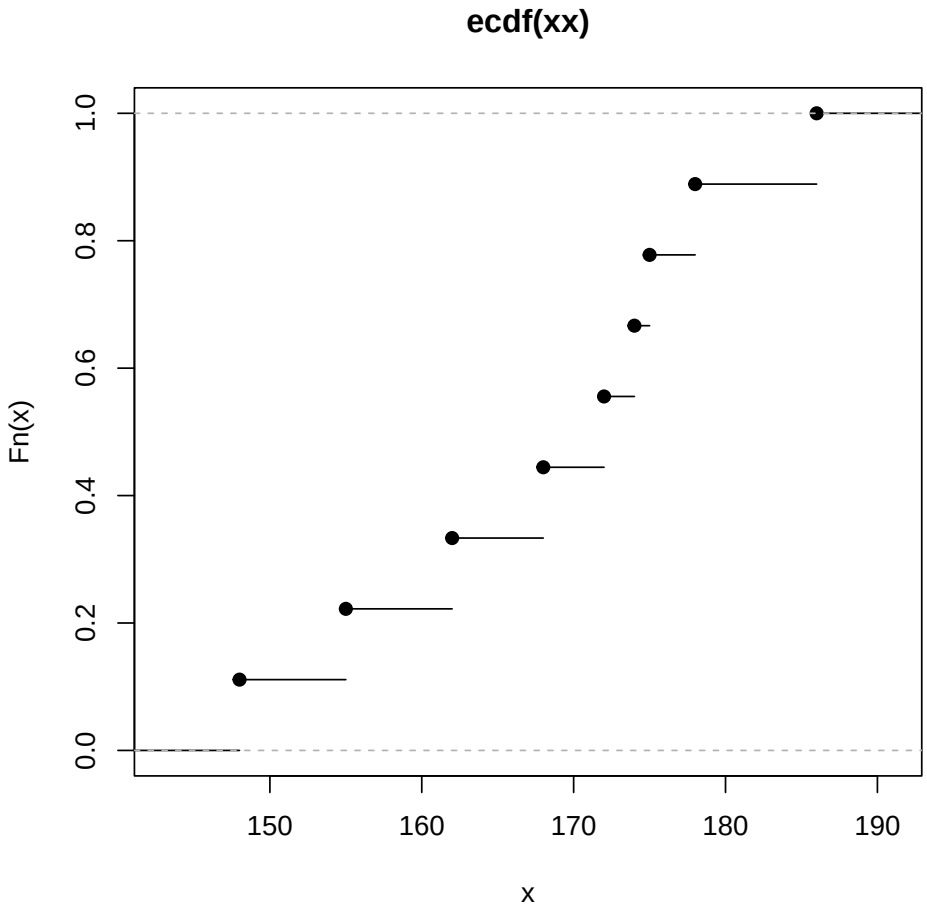



```
par(mfrow=c(1,1))
```

Come si vede la funzione ha incrementi pari a $1/n$ ogni volta che si è in corrispondenza di un nuovo dato osservato. La linea verticale nel gradino è graficamente utile ad apprezzare l'altezza dello stesso ma dal punto di vista formale la funzione dovrebbe presentarsi come una funzione costante a tratti.

Esiste la funzione `ecdf` predefinita in `R` per ottenere la funzione di ripartizione empirica. Essa consente numerose varianti e genera un oggetto che può essere direttamente fornito alla funzione `plot`

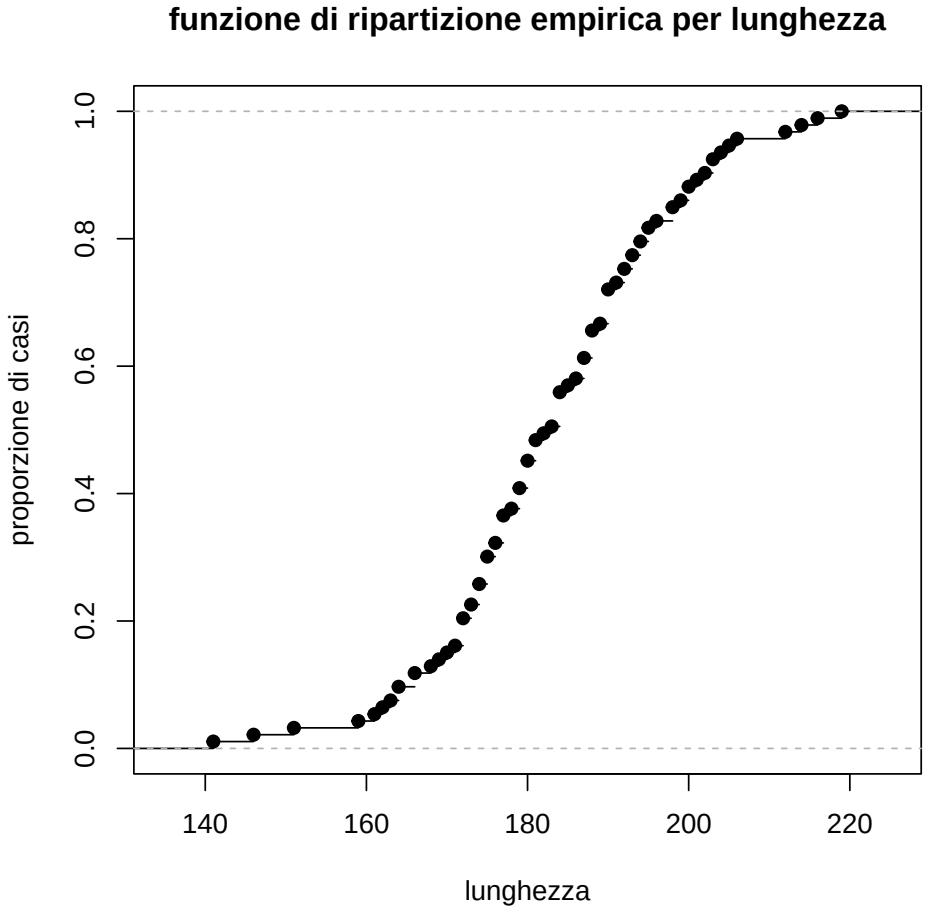
```
mio<-ecdf(xx)
plot(mio)
```



La funzione `ecdf()` è basata sulla funzione `stepfun()` che è appositamente costruita per trattare funzioni a gradini cui si rinvia per dettagli.

Proviamo a ottenere ora la funzione di ripartizione empirica per i dati sulla lunghezza delle auto.

```
plot(ecdf(Cars93$Length), main="funzione di ripartizione empirica per lunghezza",
   xlab="lunghezza", ylab="proporzione di casi")
```



Tale rappresentazione grafica è di interesse per vari motivi:

- conserva l'informazione su ogni singolo valore;
- è del tutto adeguata anche con variabili quantitative discrete;
- permette di visualizzare dove si trovi la mediana o alcuni percentili (la funzione dei quantili è in effetti l'inversa della funzione di ripartizione). Da questo si intende chiaramente come solo alcuni quantili sono definiti, mentre per altri occorre ricorrere ad approssimazioni (ad esempio considerando interpolazioni lineari fra i due quantili con l'ordine più prossimo);
- la sua interpretazione non è agevole e immediata, ma è evidente che i punti in cui la curva cresce più veleocemente sono quelli in cui si sono osservati molti casi;
- si potrebbe ricorrere a definizioni alternative (ad esempio considerando numero di $x_i < x$). Ottenendo una curva che sarà leggermente spostata rispetto a quella già definita ma con la stessa forma;
- a partire da essa si potrebbe costruire una versione empirica della funzione di sopravvivenza $S(x) = 1 - F(x)$. A questa funzione occorrerebbe dedica-

re un capitolo a parte vista la sua rilevanza per l'analisi di dati di durata soprattutto in ambito medico (relativi, ad esempio, all'analisi dei tempi impiegati per guarire da una malattia). In effetti esistono in `R` pacchetti dedicati a questo;
- è estremamente vantaggiosa se si vuole confrontare la distribuzione empirica con una distribuzione teorica (aspetto che verrà trattato più avanti).

Resta evidente che una più accurata rappresentazione della distribuzione dei dati si avrebbe considerando piuttosto che la versione empirica della loro funzione di ripartizione la versione empirica delle funzione di densità (che, ricordiamo, in distribuzioni di probabilità per variabili aleatorie continue sarebbe la derivata della funzione di ripartizione).

I modi per rappresentare empiricamente la funzione di densità empirica per un insieme di dati saranno oggetto dei prossimi paragrafi.

3.2.5 L'istogramma

L'istogramma è senz'altro una delle più diffuse e conosciute tecniche di rappresentazione grafica di un carattere quantitativo, i principi che ne ispirano la costruzione e l'interpretazione sono semplici e sono descritti in un qualsiasi testo di statistica di base.

L'istogramma in realtà costituisce uan forma semplice per ottenere una approssimazione empirica della funzione di densità ed è dall'estensione di tale idea che si possono poi considerare tecniche più complesse per la determinazione della curva di densità e per ottenere una versione "liscia" dell'istogramma.

Si suppone di disporre di un insieme di n dati $x_1, x_2, \dots, x_n$ relativi ad una variabile quantitativa X.

Abbiamo già visto come sia possibile sintetizzare i dati di una variabile quantitativa in una tabella di frequenza: la variabile X è opportunamente categorizzata e trasformata in un fattore mediante un operazione di raggruppamento in classi; si determina cioè una sequenza di valori $z_0 < z_1 < \dots < z_{I-1} < z_I$ che definiscono una successione di intervalli disgiunti (classi) $(z_{i-1}, z_i]$ (con $i = 1, 2, \dots, I$) e si determinano le frequenze relative

$$f_i = \frac{\text{numero di valori nell'intervallo}(z_{i-1}, z_i]}{n}.$$

La rappresentazione grafica mediante istogramma della distribuzione di frequenze relative riassunta dalle I coppie $((z_{i-1}, z_i], ni_i)$ si effettua costruendo dei rettangoli la cui base coincida con gli intervalli $(z_{i-1}, z_i]$ e la cui altezza è tale da rendere le aree dei rettangoli proporzionali alle frequenze n_i. Di particolare interesse è il caso in cui le aree dei rettangoli sono pari alle frequenze relative f_i.

Tale risultato si ottiene introducendo il concetto di **densità di frequenza relativa**. L'altezza di ogni rettangolo viene quindi determinata in modo tale da

consentire una corrispondenza fra l'area del rettangolo A_i che insiste su $(z_{i-1}, z_i]$ e la frequenza relativa f_i: ovvero $f_i \propto A_i$. Quindi l'altezza del rettangolo che insiste sull'intervallo $(z_{i-1}, z_i]$ deve essere proporzionale a

$$\frac{f_i}{z_i - z_{i-1}}. \tag{3.1}$$

Di solito si calcola l'altezza così che l'area di ogni rettangolo sia pari esattamente alla frequenza relativa, così che $A_i = f_i$, e in tal caso l'area complessiva all'interno dei rettangoli che compongono il grafico risulterà pari ad 1.

Le altezze dei rettangoli sono quindi poste pari alla frequenza relativa nell'intervallo diviso per l'ampiezza dell'intervallo stesso. Esse **non** rappresentano quindi le frequenze relative per gli intervalli in questione bensì la **densità di frequenze relative** (salvo nel caso in cui le basi dei i rettangoli abbiano tutti ampiezza pari a 1), ovvero la quota di frequenze relative che insiste su un generico intervallo unitario in $(z_{i-1}, z_i]$.

L'istogramma quindi rappresenta un primo, semplice, tentativo di ottenere una versione empirica della funzione di densità sottostante ai dati. Quindi

$$\hat{f}(x) \geq 0$$

$$\text{prop}(a \leq \text{units} \leq b) = \int_a^b f(x)dx$$

In linea di principio possiamo calcolare il valore approssimativo della proporzione di casi fra due numeri reali a e b sommando le aree dei rettangoli rappresentate dall'istogramma e compresi nell'intervallo (considerando porzioni dei rettangoli se a o b non coincidono con gli estremi delle classi).

Come detto, attuando una suddivisione in classi si sono perse informazioni e quindi la vera proporzione di casi (che potrei ottenere solo se tornassi alle informazioni originali rinunciando al riassunto grafico) nei dati sarà verosimilmente diversa.

L'assunzione sottostante alla rappresentazione grafica è la costanza della densità nell'intervallo, ovvero su ogni sottointervallo Δx interno a una classe di valori la frequenza relativa dipende esclusivamente dall'ampiezza Δx (uniforme distribuzione nelle classi).

L'istogramma presenta caratteristiche che lo rendono particolarmente utile in molte situazioni applicative, ma ha anche alcuni inconvenienti.

I principali vantaggi sono:

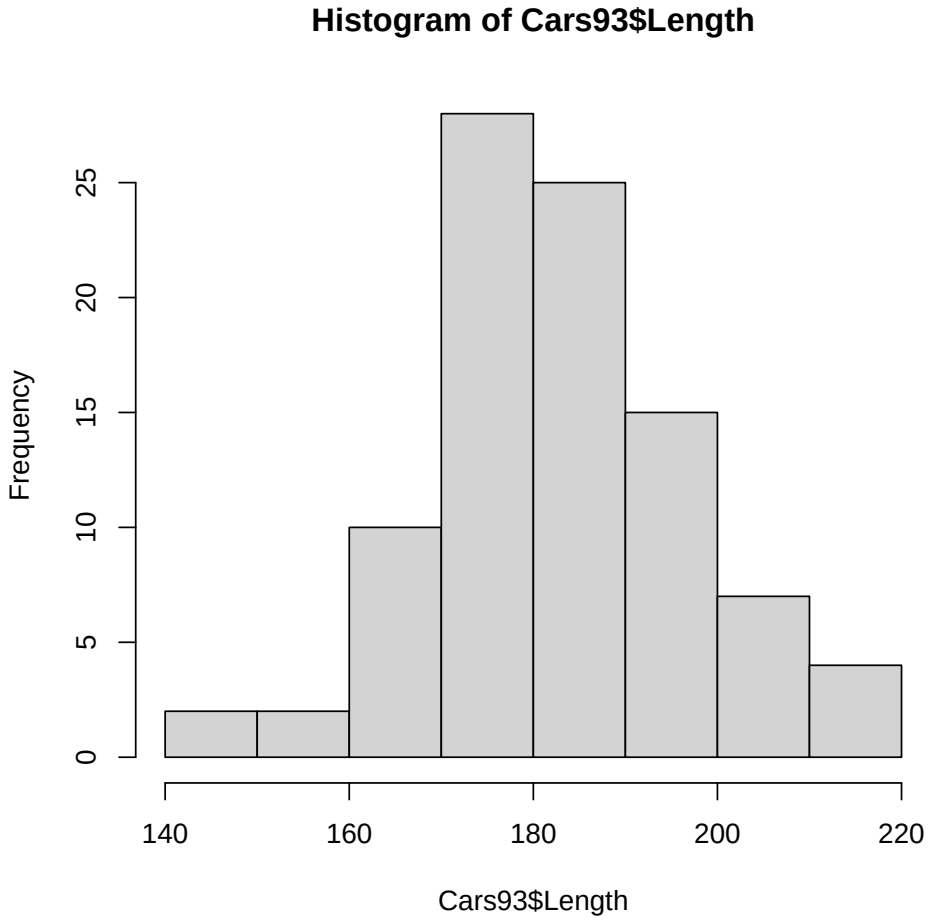
1. è semplice da interpretare;
2. è immediato da costruire;
3. è facile da utilizzare.

I suoi limiti sono invece i seguenti:

1. il grafico presenta forti discontinuità. In particolare si potrebbe dedurre che vi è un salto fra la densità in due valori di x molto prossimi ma che appartengono a due diverse classi (è ovvio che tale conclusione potrebbe essere diversa per una scelta alternativa delle classi) e talvolta (con pochi dati) il grafico potrebbe anche essere sensibilmente diverso per scelte alternative delle classi: in genere la scelta di classi più o meno ampie equivale alla scelta di un diverso grado di lisciamento e di regolarità della rappresentazione grafica;
2. la densità è assunta costante in ogni intervallo e tale assunto è spesso discutibile.

Ottenere un istogramma in `R` è semplice in quanto la funzione `hist()` automaticamente, determina le classi, calcola le frequenze nelle classi, e produce il grafico. ad esempio:

```
hist(Cars93$Length)
```



Come si vede il grafico presenta classi tutte della medesima ampiezza e presenta

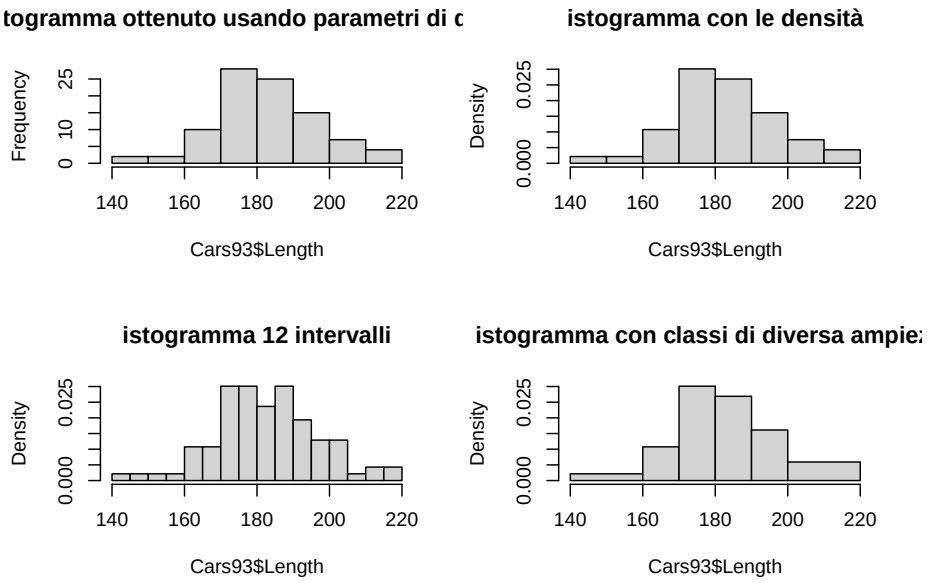
sull'asse i valori relativi alle frequenze assolute. Il grafico è in questo caso non rappresenta le densità, tuttavia la forma resta invariata in quanto le ampiezze delle classi sono uguali e le altezze dei rattangoli sono proporzionali alle densità.

Se si vuole però un grafico di densità, occorre specificare il parametro `prob=TRUE`

Il numero delle classi M che vengono usate in `R` è determinato con una regola detta regola di Sturges per cui $M = 1 + \frac{log(n)}{log(2)}$. Quindi cresce (lentamente) con la dimensione del vettore di dati. E' possibile usare regole diverse (si veda l'help della funzione `hist`).

Ovviamente i parametri della funzione ci consentono di variare tale scelta e anche di usare classi di ampiezza diversa. Il parametro `breaks=` consente di fare entrambe le cose. Gli esempi sotto che illustrano ancora quanto detto sopra: si noti come aumentando il numero delle classi si rischi di avere un grafico meno regolare troppo dipendente dalla specificita dell'insieme di dati.

```
par(mfrow=c(2,2))
hist(Cars93$Length, main="istogramma ottenuto usando parametri di default")
hist(Cars93$Length, prob=TRUE, main="istogramma con le densità")
hist(Cars93$Length, prob=TRUE, breaks=12, main="istogramma 12 intervalli")
hist(Cars93$Length, prob=TRUE, breaks=c(140,160,170,180,190,200,220),
     main="istogramma con classi di diversa ampiezza")
```

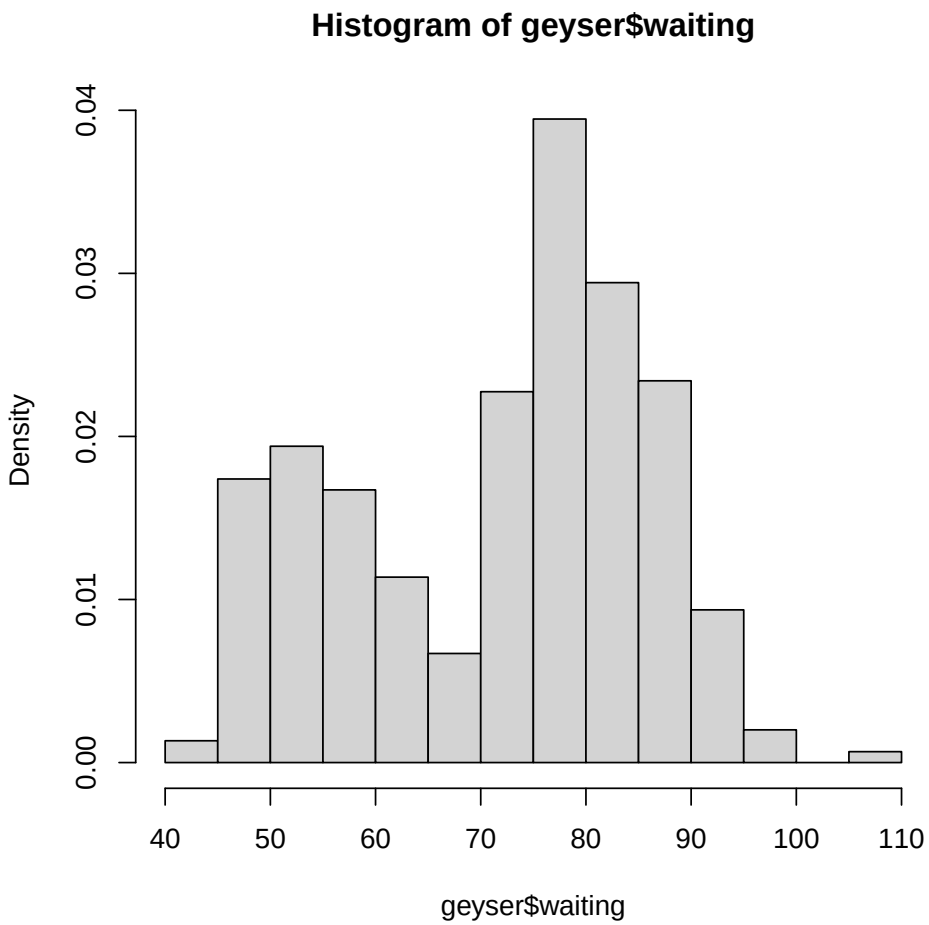



```
par(mfrow=c(1,1))
```

Per illustrare ancora quanto la rappresentazione con istogramma possa fornire grafici diversi variando alcuni criteri di costruzione, si considerino i dati `geyser` presenti nel package `MASS`. Essi si riferiscono ai tempi di attesa fra due eru-

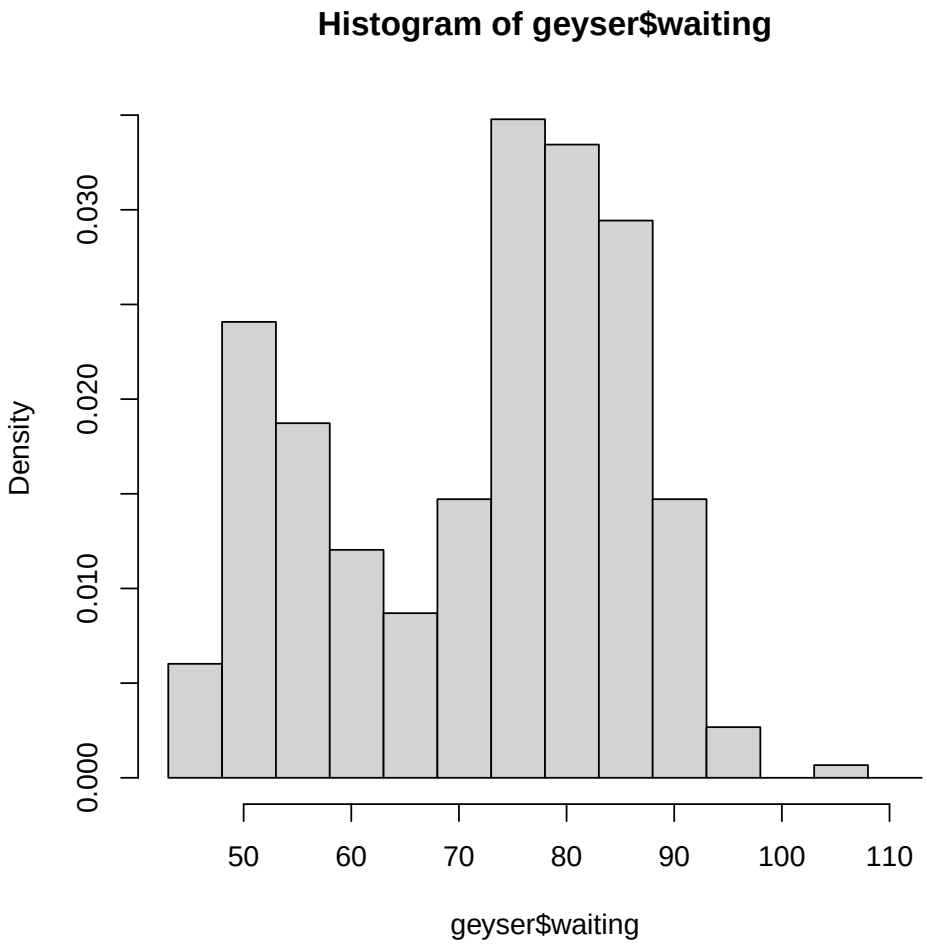
zioni del geyser *Old faithful* nel parco di Yellowstone negli USA. Otteniamo l'istogramma

```
library(MASS)
data(geyser)
hist(geyser$waiting, prob=TRUE)
```



Si consideri ora una diversa suddivisione in classi mantenendo la stessa ampiezza per la variabile cambiando semplicemente l'origine della prima classe.

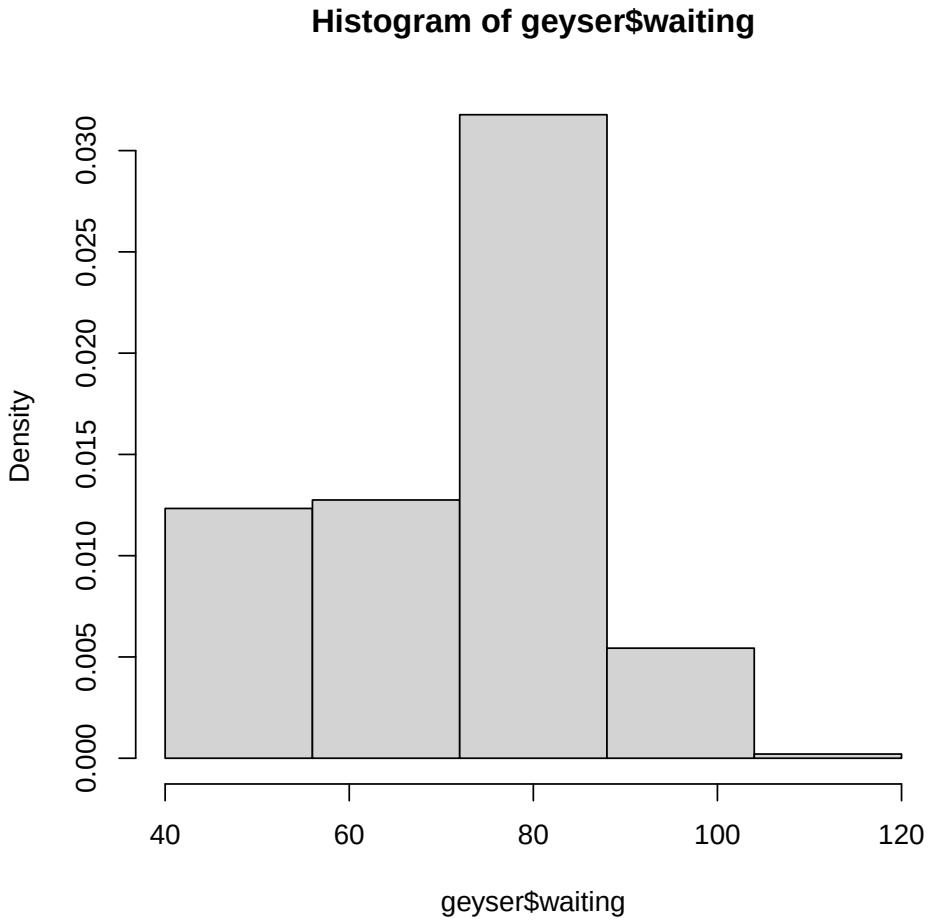
```
library(MASS)
data(geyser)
hist(geyser$waiting, breaks=c(43,(5*(1:14)+43)), prob=TRUE)
```



L'ampiezza delle classi è la stessa nei due casi, il punto di origine delle classi è invece spostato. Si noti che pur essendo le caratteristiche del grafico simili (entrambe le rappresentazioni grafiche consentono di cogliere le due gobbe della distribuzione), la collocazione dei due massimi nelle due gobbe e la distanza che li separa appare diversa. Se, inoltre si guarda al primo grafico si deduce che la densità di frequenza nel punto $x = 72$ è la stessa che nel punto $x = 74$, mentre si arriverebbe a ben altra conclusione riguardo la densità locale nei due punti se si guarda al secondo grafico.

Si noti che se utilizzassi solo 4 classi perderei la caratteristica saliente di questi dati che è data dalla presenza delle due gobbe

```
library(MASS)
data(geyser)
hist(geyser$waiting, breaks=c(40,56,72,88,104,120), prob=TRUE)
```



3.2.5.1 La moda

Nel caso di un modello probabilistico per una variabile aleatoria continua X con densità $f(x)$ è possibile definire la **moda Mo** come

$$Mo = \arg\max_{x \in R} f(x)$$

ovvero il valore cui corrisponde la densità più alta.

La moda è un indice di tendenza centrale e, ovviamente, essa risulta definita anche per una variabile aleatoria discreta: in questo caso si tratta di determinare il valore del supporto della variabile aleatoria cui corrisponde la probabilità più elevata.

La versione empirica della moda, calcolata per un insieme di dati, è quindi definita nei diversi casi come segue:

- nel caso di una variabile categoriale (un fattore) essa è definita come la modalità cui corrisponde la frequenza relativa più elevata nel collettivo esaminato. La moda rappresenta **l'unico indice di tendenza centrale**, che è possibile determinare **per una variabile categoriale**.
- nel caso di una variabile quantitativa discreta essa è definita come il valore cui corrisponde la frequenza più elevata;
- nel caso di una variabile quantitativa continua, essa si può definire dopo la trasformazione della variabile in fattore utilizzando il raggruppamento in classi. In tal caso si determina la ***classe modale*** che è quella classe cui corrisponde la densità di frequenza più elevata (si noti che non si guarda in questo caso alla semplice frequenza nella classe se le classi hanno ampiezza diversa).

È interessante notare che se si vuole determinare la moda di una variabile categoriale (o anche di una variabile quantitativa discreta) non esiste una funzione predefinita in `R`, come ad esempio`mean()`o `median()`, rispettivamente per media e mediana, occorre definire una funzione apposita. Sotto si fornisce un esempio:

```
moda1<-function(x){
t<-table(x)
return(labels(t[t==max(t)]))
}
# se ora si considera un vettore categoriale
pp<-c("A","A","A","A","B","B","A","C","A","C","B","B")
table(pp)
```

```
## pp
## A B C
## 6 4 2
```

```
# si vede che la moda è "A" e, in effetti,
moda1(pp)
```

```
## [1] "A"
```

```
# proviamo con una variabile quantitativa discreta
po<-rpois(100,3)
table(po)
```

```
## po
##  0  1  2  3  4  5  6  7
##  8 24 22 21 13  9  2  1
```

```
# si vede che la moda è pari a 3
moda1(po)
```

```
## [1] "1"
```

Nel caso di variabili quantitative (continue o discrete), essendo la moda il valore cui corrisponde il punto di massimo assoluto per la funzione che rappresenta la

densità (o la frequenza), si possono talvolta reperire altri valori però corrispondenti a punti di massimo relativo (mode secondarie). Se una distribuzione oltre alla moda principale (quella per cui si osserva il massimo assoluto) mostra una o più mode secondarie si parla di distribuzione **multimodale** (**bimodale**, ad esempio, se i punti di massimo sono due).

Si noti che in distribuzioni unimodali la posizione della moda rispetto alla mediana e alla media fornisce indicazioni sulla asimmetria della distribuzione. In distribuzioni simmetriche i tre indici sono circa nella stessa posizione, mentre se c'è asimmetria positiva (coda destra lunga) si ha $Mo < Me < M$. L'ordinamento si inverte nel caso di asimmetria negativa.

Nel caso dei dati del geyser visti sopra la distribuzione presenta una evidente bimodalità.

3.2.6 Il metodo del nucleo per il "lisciamento" di una curva di densità

A partire dalla definizione di densità di frequenza introdotta per l'istogramma è possibile ottenere la densità di frequenza relativa locale in un qualsiasi punto x come

$$d(x) = \frac{numero\ di\ valori\ in\ (x - \frac{h}{2}; x + \frac{h}{2}]}{hn} \tag{3.2}$$

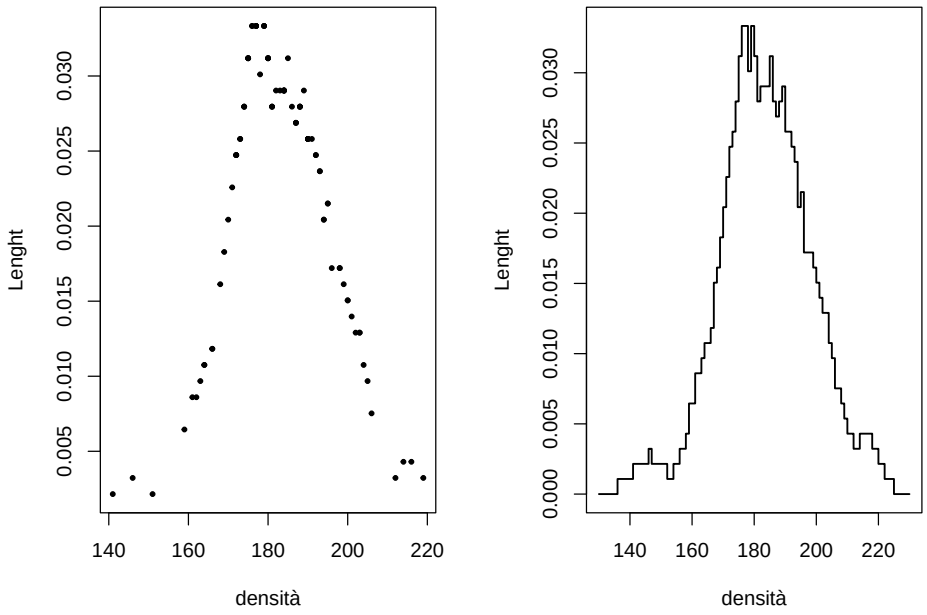
Equivale a considerare una classe di ampiezza h e a muoverla posizionandola in corrispondenza di ciascun valore osservato

Si potrebbe quindi tracciare il grafico della curva $d(x)$, osservando che essa risulterà avere delle discontinuità nei punti $x_i \pm \frac{h}{2}$ e sarà costante nei tratti intermedi. L'inconveniente legato alle discontinuità presenti nel grafico, peraltro meno rilevanti se n è grande, può essere tuttavia superato come si vedrà nel successivo paragrafo.

La funzione $d(x)$ quindi associa ad ogni valore x la densità di frequenze relative misurata con riferimento ad un intervallo di ampiezza h centrato su x. La curva di densità che descrive tale funzione è uno strumento utile per l'analisi di un insieme di dati quantitativi: come per l'istogramma, esso fornisce infatti una rappresentazione grafica di immediata comprensione. Dove la densità di frequenza è elevata il numero di unità che insiste sull'intervallo di ampiezza h centrato su x è maggiore che in punti ove la densità risulta minore. Si ha in definitiva uno sguardo sull'intera distribuzione di valori senza tuttavia dover imporre come per l'istogramma una particolare scelta di classi di valori.

Per esempio, usando tale criterio (con h=10) per i dati sulla lunghezza delle auto si otterrebbe:

```r
par(mfrow=c(1,2))
# versione della densità ottenuta applicando la definizione in (2)

nucl<-function(x,xx,h=1) {
  y=0
  n=length(xx)
  nucl=0
  for (i in 1:n){
    nucl=nucl+(abs(x-xx[i]) <= h/2)*1/(n*h)}
  nucl
}
plot(sort(Cars93$Length),nucl(sort(Cars93$Length),Cars93$Length,10),
     cex=.5, pch=19, ylab="Lenght", xlab="densità")

# versione del grafico in cui il calcolo viene fatto per tutti
# i punti di x e non solo in quelli osservati

plot(seq(130,230),nucl(seq(130,230),Cars93$Length,10), lwd=1.5,
     type="s", ylab="Lenght", xlab="densità")
```




```r
par(mfrow=c(1,1))
```

I valori di densità che descrive tale funzione forniscono uno strumento utile per l'analisi di un insieme di dati quantitativi: come per l'istogramma, esso fornisce infatti una rappresentazione grafica di immediata comprensione. Dove la densità di frequenza è elevata il numero di unità che insiste sull'intervallo di ampiezza

h centrato su x è maggiore che in punti ove la densità risulta minore. Si ha in definitiva uno sguardo sull'intera distribuzione di valori senza tuttavia dover imporre come per l'istogramma una particolare scelta di classi di valori.

Si noti che tracciare un istogramma equivale alla determinazione della densità secondo la definizione introdotta nella (3.2) sopra solo per i punti x al centro degli intervalli $x = \frac{z_{z-i}+z_i}{2}$. La particolarità è che tale valore della densità viene estesa a tutti i valori $x \in (z_{i-1}; z_i]$. L'istogramma è quindi una soluzione molto particolare e non del tutto efficiente del problema di determinare la funzione di densità di frequenza $d(x)$.

3.2.6.1 Determinazione di una funzione di densità empirica con il metodo del nucleo

Al fine di superare alcuni degli inconvenienti già citati nel caso dell'istogramma è quindi possibile approssimare la curva di densità semplicemente applicando la definizione (3.2).

La (3.2)} può essere opportunamente riscritta come segue. Si definisca

$$W(x) = \begin{cases} 1 & \text{se} \quad |x| < \frac{1}{2} \\ 0 & \text{altrimenti} \end{cases}$$

allora

$$d(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} W\left(\frac{x - x_i}{h}\right) \tag{3.3}$$

La (3.3) è equivalente alla (3.2) ma, come vedremo, rende naturale l'estensione e la generalizzazione di quest'ultima.

Come già notato la curva che si ottiene con la (3.2) è discontinua nei punti $x_i \pm \frac{h}{2}$, ove x_i è un generico valore corrispondente ad uno dei dati osservati, anche se l'entità delle discontinuità è ridotta rispetto a quello che accade con l'istogramma. Ciò è dovuto alla discontinuità della funzione W.

In realtà la formulazione in (3.3) equivale ad assumere che l'impatto sulla curva di densità di un qualsiasi valore x_i sia distribuito in ugual misura con peso $\frac{1}{h}$ su tutto l'intervallo $(x - \frac{h}{2}; x + \frac{h}{2}]$, cosicchè comunque il peso complessivo di ogni singola osservazione x_i sia pari a $\frac{1}{n}$.

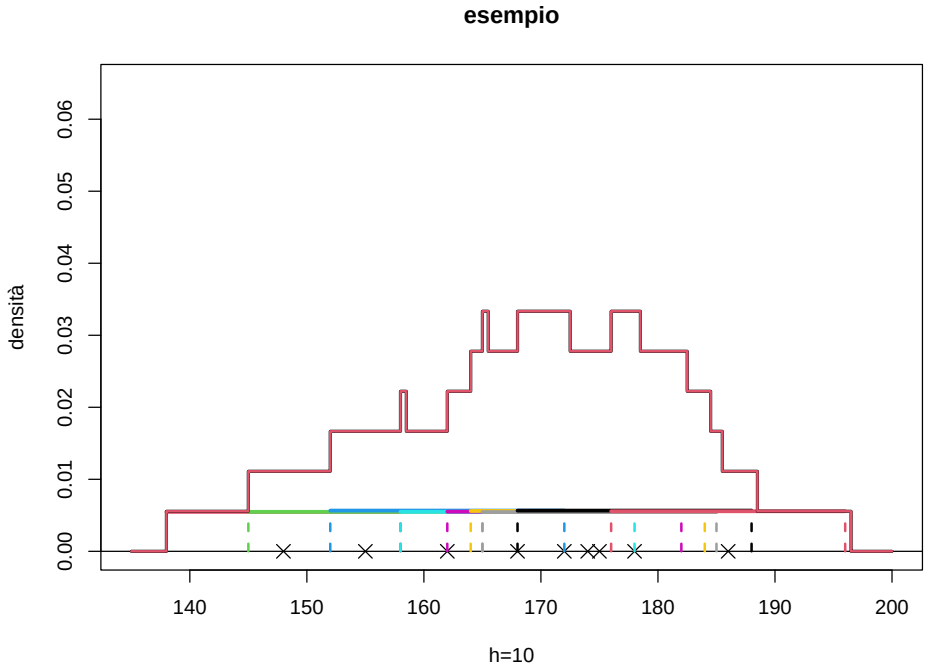
La funzione $W(u)$ è un rettangolo di ampiezza h ed il calcolo nella (3.3)} equivale a posizionare un rettangolo di ampiezza h in corrispondenza di ogni valore osservato x, a valutare la frequenza relativa dei casi che sono all'interno dell'intervallo e a determinare la densità di frequenze relative in x come il rapporto fra la frequenza relativa ottenuta e l'ampiezza dell'intervallo stesso.

Una interpretazione equivalente si ha immaginando di posizionare una scatola di ampiezza h e altezza $\frac{1}{n}$ in corrispondenza di ogni dato osservato x_i, in modo

che il centro della scatola coincida con x_i. Il valore di $d(x)$ è pari alla somma delle altezze di tutte le scatole che comprendono il valore x diviso per n.

Illustriamo quanto detto nel seguente esempio

```
par(mfrow=c(1,1))
# consideriamo ancora l'insieme di 9 dati nel vettore xx
xx<-c(148, 155, 162, 168, 172, 174, 175, 178, 186)

plot(xx,rep(0,9), ylim=c(0,0.065), xlim=c(135,200), pch=4, cex=1.5,
     main="esempio", xlab="h=10", ylab="densità")
abline(0,0)

# infine rappresentiamo le funzioni W, ovvero dei rettangoli centrati su ciascun dato
yc=0
for (i in 1:9){
yc[i]<- jitter(1/180) #la funzione jitter() consente di spostare un dato
                      #di una piccola quantità, è utile per rappresentare
                     # dati che apparirebbero altrimenti sovrapposti
segments(xx[i]-10,yc[i],xx[i]+10,yc[i], col=(i+1), lwd=3) # la funzione segments
                                  # aggiunge una linea su un grafico fra
segments(xx[i]-10,0,xx[i]-10,yc[i], col=(i+1),lwd=2, lty=2)
segments(xx[i]+10,0,xx[i]+10,yc[i], col=(i+1),lwd=2, lty=2)}

# Successivamente costruiamo una funzione che calcoli per ogni x
# la somma delle ordinate delle funzioni W(x) (la somma delle altezze
# dei rettangoli di ampiezza h)

rett<-function(x,xx,h=1) {
  n=length(xx)
  rett=0
  for (i in 1:n){
    rett=rett+(abs(x-xx[i]) <= h/2)*1/(n*h)}
  rett
}

# ora sovrapponiamo al grafico: la funzione points() permette di aggiungere
# punti su un grafico esistente
xr<-seq(135,200,.5)
points(xr,rett(xr,xx,20), lwd=2.5, type="s")

# aggiungiamo un altro esempio con rettangoli più ampi
points(xr,rett(xr,xx,20), lwd=2, type="s", col=2,
  main="esempio con rettangoli più ampi", ylab="h=20", xlab="densità")
```



Una naturale estensione della (3.3) si ha se si assume che ogni singolo dato osservato x_i abbia un impatto sulla densità in x decrescente con la distanza $x - x_i$. Ciò implica che ad esempio si possa usare una funzione per $W(u)$ diversa dal rettangolo.

In generale possiamo introdurre una funzione $K(u)$, detta **nucleo**, che abbia le seguenti caratteristiche

- $\int_{\infty}^{-\infty} K(u)du = 1$,
- $K(u)$ è una funzione simmetrica rispetto a 0,
- $K(u)$ assume solo valori positivi.

E la funzione $d(x)$ avendo osservato un insieme di dati viene determinata come

$$d(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right). \tag{3.4}$$

Si noti che la funzione $W(u)$ definita precedentemente ha le caratteristiche di una funzione nucleo. Altre scelte ragionevoli per $K(u)$ sono, ad esempio le seguenti:

$$K(u) = \begin{cases} \frac{3}{4}(1 - \frac{1}{5}u^2)5^{-\frac{1}{2}} & \text{se} |u| < 5 \\ 0 & \text{altrimenti} \end{cases}$$

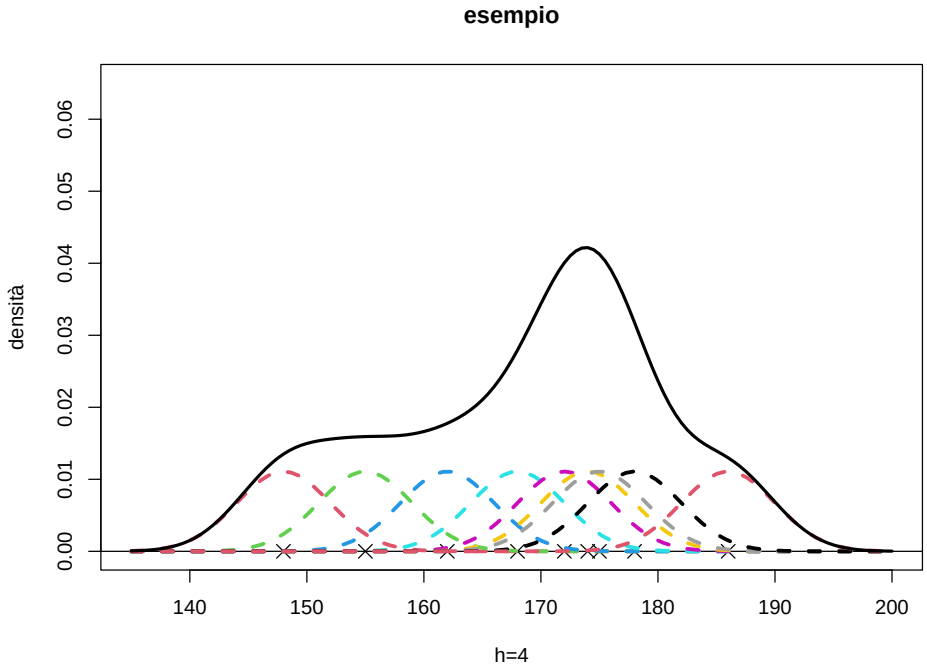
$$K(u) = \begin{cases} 1 + cos2\pi & \text{se} |u| < 0.5 \\ 0 & \text{altrimenti} \end{cases}$$

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}.$$

L'ultima funzione è facilmente identificabile come una funzione di densità di una normale standard (nucleo gaussiano). L'interpretazione della (3.4) data per la funzione $W(u)$ (detta nucleo rettangolare) è valida anche in questo caso, basti pensare che in corrispondenza di ogni osservazione x_i stavolta si posiziona invece che una scatola un cumulo centrato su x_i la cui forma è data da una delle funzioni nucleo descritte precedentemente divise per hn. Il valore di $d(x)$ è pari alla somma del valore dell'ordinata di ciascuna degli n cumuli.

Riconsideriamo i dati dell'esempio precedente e utilizziamo il nucleo gaussiano. Nel grafico rappresentiamo delle piccole gaussiane (che hanno area sottostante pari a $1/n$) centrate su ciascun dato. La curva che si ottiene è pari, per ogni valore x, alla somma delle ordinate in x di ciascun nucleo.

```
par(mfrow=c(1,1))

xx<-c(148, 155, 162, 168, 172, 174, 175, 178, 186)
plot(xx,rep(0,9), ylim=c(0,0.065), xlim=c(135,200), pch=4, cex=1.5,
     main="esempio", xlab="h=4", ylab="densità")
abline(0,0)

# rappresentiamo le funzioni K, come delle guassiane divise per n
n=length(xx)
for (i in 1:n){
  curve(dnorm(x,xx[i],4)/n, col=(i+1), lwd=3, lty=2, add=TRUE)}

# Successivamente costruiamo la funzione che calcoli per
# ogni x la somma delle ordinate delle funzioni K (la somma
# delle altezze delle gaussiane)

nuclg<-function(x,xx,h=1) {
  nucl=0
  nn<-length(x)
  n=length(xx)
  for (i in 1:nn){
  nucl[i]= sum(dnorm(x[i],xx,h)/n)}
  nucl
}
lines(xr,nuclg(xr,xx,4),lwd=2.5)
```



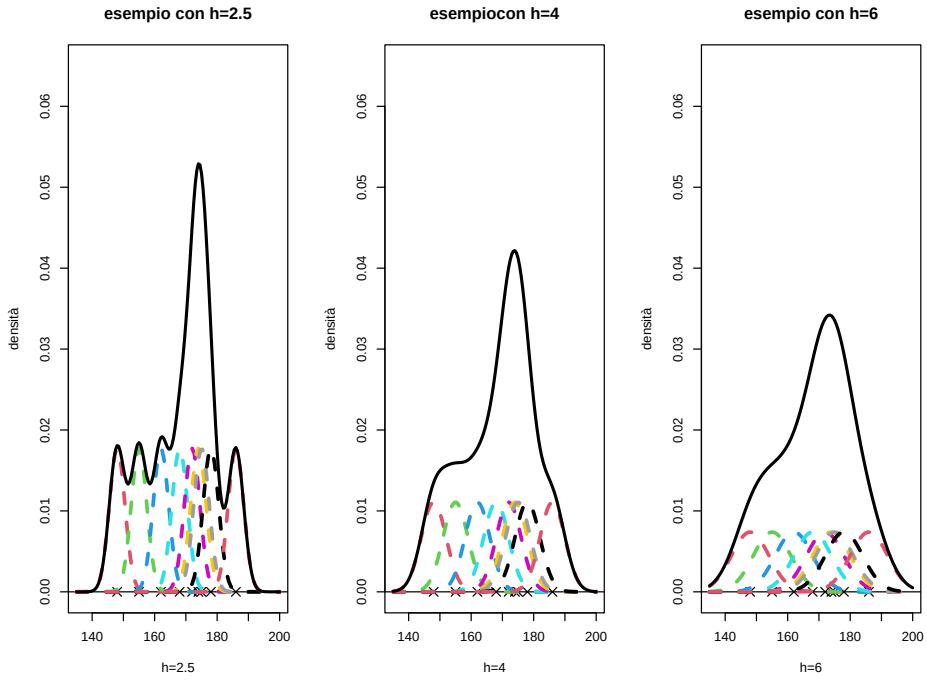
3.2.6.2 Il parametro di lisciamento

Il comportamento della funzione $d(x)$ ottenuta impiegando la (3.4) eredita alcune delle proprietà del nucleo utilizzato. In particolare, se $K(u)$ è essa stessa una funzione di densità come nel caso visto del nucleo gaussiano, allora:

- $d(x)$ ha le stesse proprietà di una funzione di densità;
- se la funzione nucleo è derivabile allora lo è anche $d(x)$.

Il grado di lisciamento della curva $d(x)$ dipende tuttavia dal valore di h che è appunto detto parametro di lisciamento. Infatti come si deduce dalla formula (3.4), h è un fattore di scala che compare fra gli argomenti della funzione K. Valori di h grandi implicano un impatto su $d(x)$ di un generico dato x_i anche per valori molto distanti da esso mentre valori di h piccoli fanno sì che il peso di x_i abbia un ruolo nel determinare il valore di $d(x)$ solo quando x non è molto distante da x_i stesso.

Riprendiamo l'esempio visto e otteniamo le densità ottenute con il metodo del nucleo con tre diversi valori di lisciamento diversi.



La figura a destra con h è più piccolo e con nuclei più appuntiti dà come risultato risultato una curva meno liscia. L'opposto vale per la figura a destra ove i nuclei troppo ampi nascondono le variazioni della densità in alcune aree.

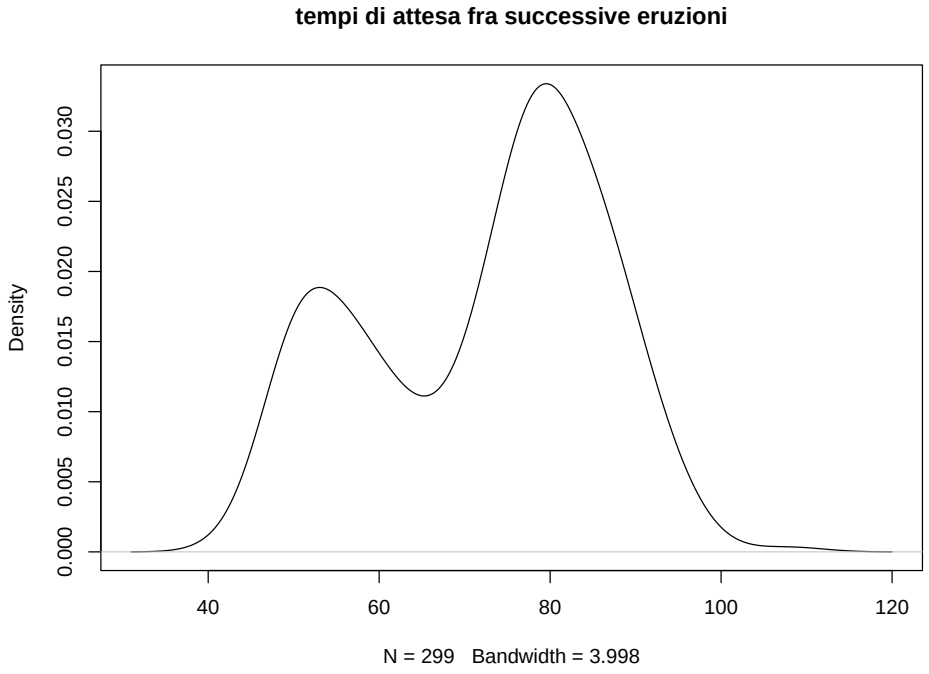
La scelta di h dipende da considerazioni analoghe a quelle fatte per la scelta dell'ampiezza delle classi nel caso dell'istogramma. Un valore del parametro di lisciamento piccolo rischia di introdurre un alto numero di brusche variazioni nella curva di densità in particolar modo nelle regioni dove si osservano pochi dati mentre un valore alto di h permette di determinare una funzione liscia al prezzo di oscurare caratteristiche locali della curva di densità (ad esempio nascondendo effettive caratteristiche bimodali della distribuzione dei dati).

`R` contiene alcune funzioni che consentono il calcolo della curva di densità, e a partire da questo produca un grafico della curva di densità scegliendo un appropriato nucleo e il valore di h opportuno (in `R` il parametro di lisciamento è denotato con `bw` riferito al termine *bandwidth*) .

In particolare, se si vuole un grafico della densità con il metodo del nucleo si può utilizzare la funzione `density()`. In essa il grado di lisciamento viene controllato mediante il parametro `bw=` ed è anche possibile scegliere funzioni nucleo alternative (il default è il nucleo gaussiano anche se va detto che, in genere, la scelta del nucleo è meno cruciale rispetto a quella del parametro di lisciamento).

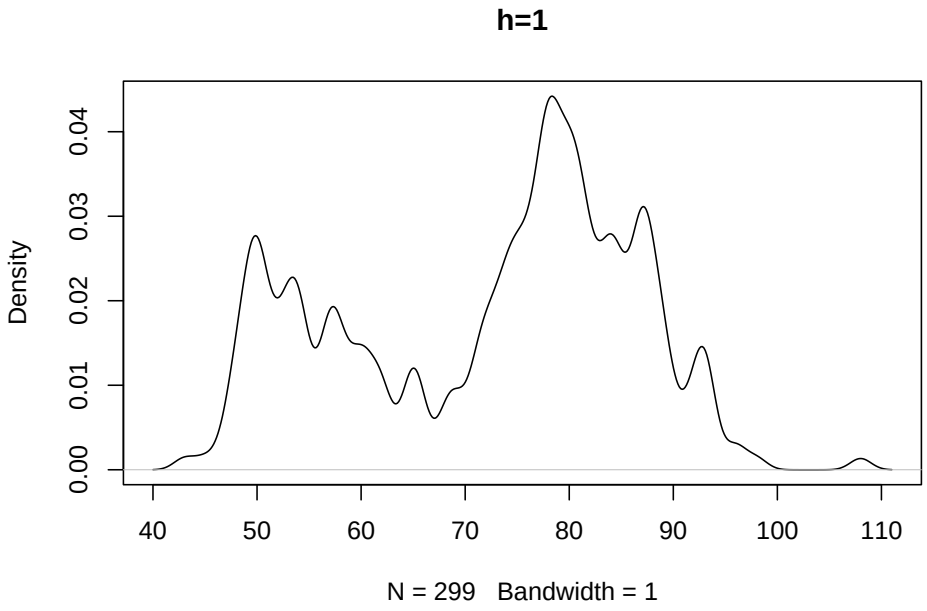
Vediamo cosa accade usando tale funzione con i dati geyser.

```
par(mfrow=c(1,1))

den<-density(geyser$waiting)
plot(den, main="tempi di attesa fra successive eruzioni")
```

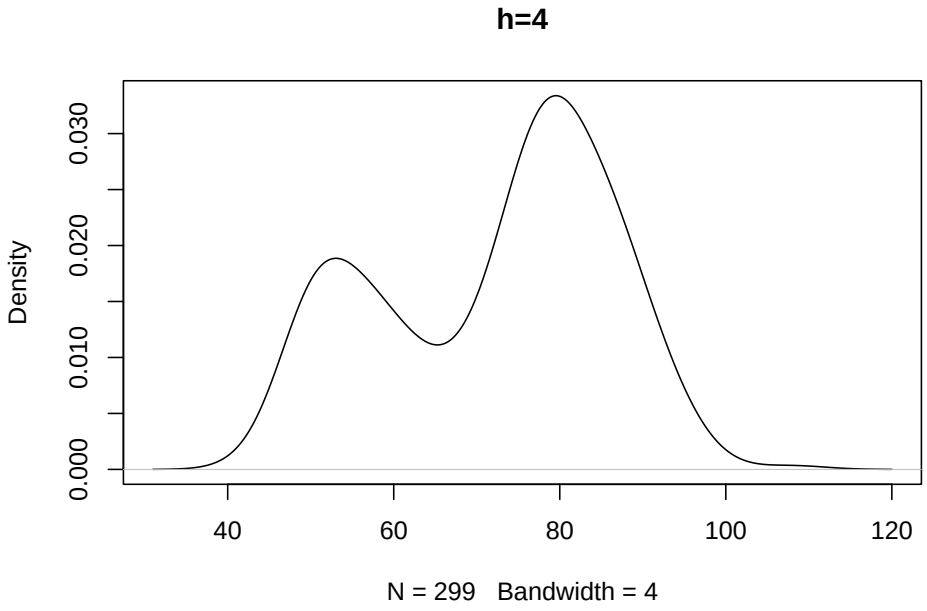



```
# Nel grafico è riportato il valore del parametro di lisciamento.
# La funzione density() potrebbe quindi essere chiamata direttamente
# all'interno di altre funzione come la plot() per ottenere il grafico
# della curva di densità.
# Si guardi cosa accade se si varia il parametro di lisciamento
```

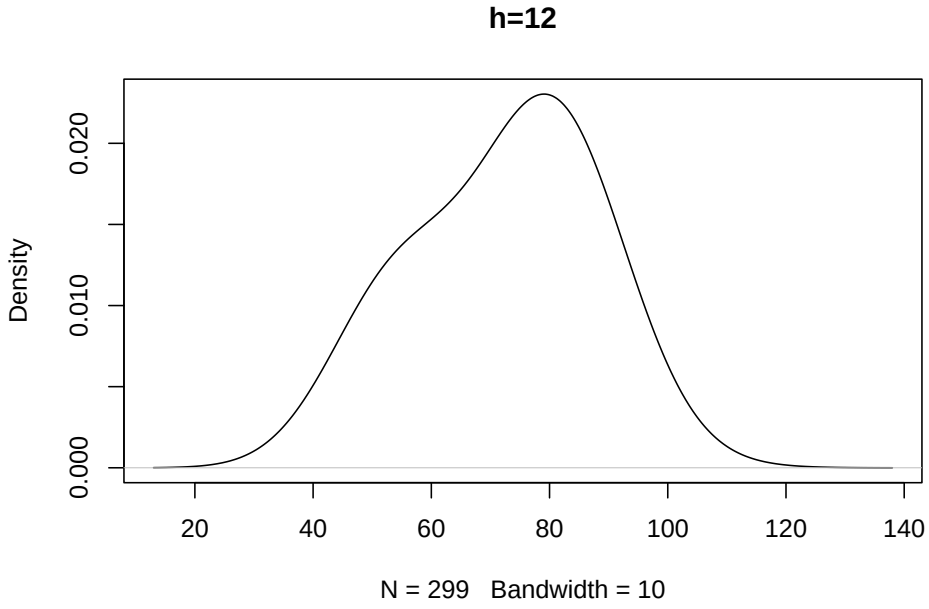
```
plot(density(geyser$waiting, bw=1), main="h=1")
```




```
plot(density(geyser$waiting, bw=4), main="h=4")
```




```
plot(density(geyser$waiting, bw=10), main="h=12")
```




```
# con un parametro di lisciamento troppo alto si perde la bimodalità
# con uno troppo piccolo appaiono delle  variazioni brusche nella
# densità che riflettono densità locali (è come fare un istogramma
# con troppe classi rispetto alla numerosità dei dati)
```

Con insiemi di dati che presentano forte asimmetria il valore di h scelto potrebbe essere adeguato per le regioni in cui la densità è alta e essere invece troppo piccolo per ottenere un buon grado di lisciamento lungo la coda. è possibile in tal caso introdurre criteri flessibili che permettano di utilizzare valori di h diversi in relazione alla diversa densità locale dei dati.

In generale, è consigliabile ottenere differenti curve di densità per diversi valori di h e giudicare a posteriori quale valore di h mostri di cogliere adeguatamente le caratteristiche salienti dell'insieme di dati.

Tuttavia è possibile tenere presenti alcuni semplici criteri pratici. è ad esempio ragionevole attendersi che il valore di h ideale sia inversamente proporzionale ad una funzione di n. In genere si consiglia di scegliere h proporzionale a $n^{-\frac{1}{5}}$. Se si usa una funzione nucleo gaussiana il valore di h di default è $h = 0,9An^{-\frac{1}{5}}$ ove

$$A = \frac{min(\text{scarto quadratico medio}, \text{scarto interquartile})}{1.34}.$$

Per le altre opzioni della funzione `density` si rinvia a quanto contenuto nella documentazione in linea di `R`. Esiste nella `Base` di `R` la funzione `ksmooth()` che funziona in modo analogo alla `density()`. Una seconda funzione, anche più generale, per determinare la funzione di densità con il metodo del nucleo è la `bkde()` nel package `KernSmooth`.

Capitolo 4

Il trattamento preliminare e la fase di pulizia dei dati

I dati che sono oggetto delle analisi esplorative spesso vengono acquisiti in una forma grezza e quindi possono presentare difetti, inaccuratezze o imperfezioni. Inoltre, la fase di acquisizione e raccolta dei dati e quella della loro organizzazione informatica in una base di dati hanno un carattere generale e non sempre si ha in mente la successiva analisi dei dati che può avre molteplici finalità. Pertanto i dati vanno preliminarmente trattati per renderli idonei e funzionali agli specifici obiettivi conoscitivi.

La fase di pulizia e trasformazione dei dati è di grande importanza e spesso la qualità delle successive analisi dipende in maniera cruciale da essa. Tale fase, per la quale spesso si usa il termine *data wrangling* o *data munging*, è quindi il processo di trasformazione e strutturazione dei dati da una forma grezza a un formato "pulito" con l'intento di migliorare la qualità dei dati e renderli più fruibili e utili.

Alcune delle tecniche di pre-trattamento dei dati richiedono conoscenze che vanno oltre gli aspetti elementari tipici di un corso come il presente che ha un carattere introduttivo, ciò nonostante possono essere formulati i concetti di base e compresi i temi e i problemi fondamentali.

4.1 Dati mancanti e valori anomali

4.1.1 Dati mancanti

Si è già avuto modo di osservare che alcune matrici dei dati, che poi in `R` sono strutturati in `dataframe`, hanno dei *buchi*, ovvero per alcune unità statistiche, in corrispondenza di una o più variabili, manca l'informazione. Si tratta di situazioni in cui non è stato possibile ottenere il valore della variabile per i

motivi più svariati (se i dati provengono da indagini sulla popolazione i soggetti potrebbero avere rifiutato di fornire la risposta ad alcuni quesiti, lo strumento che doveva rilevare alcune variabili non ha funzionato, il dato è stato magari rilevato ma è palesemente errato per cui viene omesso).

Tale mancanza di informazione, come si è già visto, nel data frame in `R` viene poi identificata con il codice speciale `NA` (*Not Available*). Occorre prestare attenzione al fatto che, in alcuni casi, si decide di assegnare al valore mancante un codice numerico speciale (che non può corrispondere a nessuno dei valori previsti per la variabile, ad esempio il valore "999") o un altra sigla. In tal caso, se si usa `R` è consigliabile trasformare questi in `NA` .

Le diverse funzioni di `R` prevedono alcune soluzioni standard riguardo a come trattare la mancata risposta: - per alcune funzioni, come ad esempio `media()`, `var()`, `median`, se la variabile da analizzare contiene `NA`, la funzione potrebbe non funzionare e occorre specificare esplicitamente come trattare il dato mancante attraverso il parametro `na.rm=TRUE`, che è un parametro logico che indica se rimuovere o meno i valori mancanti, così che la funzione svolge i calcoli solo sui dati disponibili.

```
# Riprendiamo ancora i dati di AutoBi.
# L'ultima riga del summary del data frame contiene per ogni variabile
# il conteggio degli `NA`
summary(AutoBi)[7,]
```

```
##         CASENUM        ATTORNEY          CLMSEX          MARITAL        CLMINSUR
##              NA              NA  "NA's   :12  "  "NA's   :16  "  "NA's   :41  "
##        SEATBELT          CLMAGE            LOSS       LOSSclass
##  "NA's   :48  " "NA's   :189  "              NA              NA
```

```
# Ora proviamo a calcolare la media di
mean(AutoBi$CLMAGE)
```

```
## [1] NA
```

```
# essa non è calcolata per la presenza di valori mancanti, il default del parametro
# `na.rm` è FALSE
mean(AutoBi$CLMAGE, na.rm=T)
```

```
## [1] 32.53084
```

- Per altre funzioni, è presente una soluzione di default (ad esempio, vengono eliminati tutti i casi in cui il dato è mancante). Ad esempio se facciamo una tabella di una variabile categoriale i dati mancanti vengono ignorati semplicemente

```
table(AutoBi$MARITAL)
```

```
##
##   1   2   3   4
```

```
## 624 650  15  35
```

```
# se si volessero evidenziare i dati mancanti allora si può utilizzare il parametro `useNA`
table(AutoBi$MARITAL, useNA="ifany")
```

```
##
##    1    2    3    4 <NA>
##  624  650   15   35   16
```

- La gestione dei dati mancanti diviene più complessa se si analizzano più variabili congiuntamente. Se, ad esempio, si devono analizzare congiuntamente una coppia di variabili (o anche più di due variabili, tema questo che sarà oggetto dei prossimo capitoli), potrebbe diminuire notevolmente il numero di casi disponibili in quanto viene esclusa ogni unità per cui in almeno una della variabili oggetto di analisi vi sia un `NA`.

```
auto<-na.omit(AutoBi) # tale funzione conserva solo i casi completi
anyNA(auto) # con questo comando si controlla che nel data set non ci siano dati mancanti
```

```
## [1] FALSE
```

```
dim(auto)   # si noti che vi è stata una significativa riduzione del numero di casi disponibili
```

```
## [1] 1091    9
```

Si vedrà più avanti che, in casi simili, nelle funzioni adatte ad analisi di più variabili, si possono specificare altre strategie riguardo l'esclusione dei dati mancanti.

Vi sono molte soluzioni che sono state proposte per il trattamento e la compensazione dei dati mancanti, tuttavia esse coinvolgono spesso metodi e concetti che vanno oltre quello che è il livello di un corso introduttivo all'analisi dei dati. Si fornisce tuttavia una breve introduzione intuitiva a alcune idee e tecniche per compensare i dati mancanti.

4.1.1.1 Le caratteristiche dei dati mancanti

Come già ricordato, è sempre indispensabile avere consapevolezza di quanti dati validi sono disponibili per l'analisi di ciascuna variabile. Occorre sempre tenere a mente che quando per una variabile non vengono rilevati dati per tutti le unità del collettivo esaminato, questo può dare origine a problemi:

1. a seguito della eliminazione dei dati mancanti, i casi disponibili potrebbero ridursi in misura così marcata da rendere l'analisi priva di significato;
2. vi è la possibilità che i dati mancanti, anche quando essi riguardano una piccola porzione dell'intero collettivo oggetto di analisi, possano essere relativi a unità con caratteristiche diverse da quelle per le quali l'informazione è disponibile. I dati mancanti in tal caso sono **selettivi** e limitare l'analisi ai soli dati disponibili porterebbe a una raffigurazione dei fenomeni parziale, distorta e non veritiera. Si pensi, ad esempio, al caso in

cui si sono raccolte con un questionario informazioni sul reddito da lavoro e si voglia svolgere una analisi descrittiva su tale variabile. Se i dati mancanti riguardassero in larga prevalenza coloro che hanno i redditi più elevati (spesso i più restii a fornire il dato), l'analisi dei dati disponibili fornirebbero una raffigurazione distorta del fenomeno.

La soluzione di ignorare semplicemente i dati mancanti e utilizzare solo i casi disponibili può quindi essere ritenuta ragionevole se:

(a) la proporzione di casi mancanti è una quota molto limitata, una porzione molto piccola del totale dei casi presenti nel data frame e
(b) vi sono buone ragioni per escludere che questa (possibilmente piccola) porzione di dati mancanti non sia del tutto diversa dai dati disponibili. In particolare, se vale questa seconda proprietà, può essere accettabile anche una proporzione maggiore di dati mancanti.

Al fine poi di fornire supporto all'idea che i dati mancanti siano o meno selettivi, può essere talvolta utile vedere se per le altre variabili del data frame, per le quali si dispone dell'osservazione completa, vi siano o meno differenze distributive per il gruppo dei dati osservati e dei dati mancanti. Si badi bene però che, anche se questo accadesse, esso sarebbe solo un indizio che non vi sia mancata risposta selettiva, ma non una prova. Può infatti accadere che le due variabili abbiano uguale distribuzione riguardo ad altre variabili ma continui ad essere presente selettività.

4.1.1.2 Tipi di dati mancanti

Le caratteristiche fondamentali del fenomeno dei dati mancanti fanno riferimento alla formalizzazione probabilistica di alcuni concetti esposti sopra. Nella letteratura sui dati mancanti si parla di

1. "dati mancanti totalmente a caso" (MCAR), in cui occorre immaginare che i dati che non si osservano abbiano caratteristiche del tutto simili a quelli osservati;
2. "dati mancanti a caso" (MAR), in cui i dati mancanti hanno caratteristiche simili a quelli osservati all'interno di gruppi specifici. Ad esempio, si immagini di studiare la statura e che per tale variabile sia assente l'informazione per un buona parte dei casi. Si immagini poi che per l'intero collettivo si osservi però il genere. Se nei dati osservati si osserva una quota simile di maschi e femmine mentre nei dati per cui manca la statura sono quasi tutti maschi: i dati sulla statura disponibili mostreranno una tendenza ad avere valori più bassi perchè vi sono più donne che sono tendenzialmente di statura minore. In questo caso, i valori mancanti sono relativi più spesso a casi con statura più elevata ma solo perchè vi è una selettività che riguarda un'altra variabile (che però si osservo completamente) e non perchè vi è minore propensione ad osservare le stature alte. I dati quindi sono "mancanti completamente a caso" all'interno dei due gruppi "maschio" e "femmina". Possiamo eventualmente correggere

la distorsione tenendo conto della maggiore propensione a non osservare la statura nei due gruppi di diverso genere.

3. "dati mancanti non casualmente" (MNAR). È il caso peggiore perchè i dati mancanti sono diversi da quelli disponibili e i possibili correttivi potrebbero non eliminare la distorsione dovuta alla osservazione parziale del fenomeno.

Esistono ragionevoli soluzioni proposte per risolvere o alleviare il problemi legati alla presenza dei dati mancanti nei primi due casi.

4.1.1.3 Illustrazione dei meccanismi di dati mancanti

La possibilità di potere compensare per i dati mancanti dipende in modo cruciale dai meccanismi citati sopra. Per comprenederli meglio può servire un esempio in cui si simula l'esistenza di dati mancanti.

Consideriamo un data frame relativo ai dati sul diabete di un gruppo di indiani americani. Per illustrare il problema, l'insieme di dati è completo, ovvero senza dati mancanti. Ci concenteremo sulla variabile `glucose` e sul fattore `diabetes` che indica se il soggetto ha o meno il diabete (se `diabetes=="pos"` ha il diabete), e si considera il caso semplificato in cui l'obiettivo sia esclusivamente quello di ottenere la media del glucosio per il collettivo esaminato.

```
    # carico il data frame con i dati completi.
load("pimanmd.RData")
attach(pimanmd)
```

```
## The following object is masked from package:datasets:
##
##     pressure
```

```
summary(glucose) # questa è la varibile in cui simuleremo i dati mancanti.
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    56.0    99.0   119.0   122.6   143.0   198.0
```

```
                # La quantità che ci interessa è la
                # media di tale variabile
dia<-table(diabetes) # la variabile diabetes è un fattore con due modalità
                     # ed è completamente osservata.
```

Ora verrà simulata la situazione in cui vi sono dati mancanti, circa il 40%) nella variabile `glucose` secondo i tre meccanismi e si osserverà se in questo insieme di dati ridotto comunque si riesce ad avere una idea corretta del valore medio del glucosio.

Dapprima i dati mancanti completamente a caso (MCAR).

```
# creo dati mancanti completamente a caso MCAR
set.seed(1111)
```

```
n<-length(glucose)
MCAR<-sample(1:n, 0.4*n) # si seleziona casualmente un gruppo di casi (il 40%)
glucoseMCAR<-glucose
glucoseMCAR[MCAR]<-NA # si creano i dati mancanti
mean(glucoseMCAR)
```

```
## [1] NA
```

Il summary della variabile con il 40% di casi in meno mostra che, anche se sono tanti i casi mancanti, la media non si discosta di molto dal valore osservato per i dati completi.

Si crea ora un nuovo insieme di dati in cui ancora complessivamente ci sarà circa il 40% di casi mancanti. Tuttavia questi sono mancanti più spesso nel caso di coloro che hanno il diabete che però è una variabile completamente osservata. All'interno dei due gruppi tuttavia i dati mancano secondo un processo casuale. Si tratta di dati MAR

```
# si creano dati mancanti a caso MAR

MARneg<-sample(which(diabetes=="neg"),0.25*dia[1])
      # si seleziona casualmente un gruppo di casi (il 25%) fra i non diabetici
MARpos<-sample(which(diabetes=="pos"),0.55*dia[2])
      # si selezioa casualmente un gruppo di casi (il 55%) fra i diabetici
glucoseMAR<-glucose
      # assegniamo ai valori selezionati il dato mancante
glucoseMAR[MARneg]<-NA
glucoseMAR[MARpos]<-NA

mean(glucoseMAR)
```

```
## [1] NA
```

Si noti che stavolta la media si discosta da quella dei dati completi (si ottiene una media del glucosio più bassa perchè mancano più spesso i diabetici).

Infine, si genera un insieme di dati in cui il 40% dei casi mancanti dipende dalla variabile oggetto di studio. Questi saranno NON mancanti a caso (MNAR).

```
# esempio creazione dati NON mancanti a caso NMAR
probMD=glucose/sum(glucose)
MNAR<-sample(1:n, 0.4*n, prob=probMD)
  # si selezionano i dati mancati con probabilità proporzionale
  # al valore del glucosio
glucoseMNAR<-glucose
glucoseMNAR[MNAR]<-NA
mean(glucoseMNAR)
```

```
## [1] NA
```

Come si vede anche in questo caso il `summary`mostra che se analizzo i dati disponibili otterrei una visione distorta del fenomeno (un valore medio del glucosio più basso).

Si illustra ora una delle soluzioni più popolari per recuperare i dati mancanti: l'**imputazione**. Con essa si ricostruisce un data set completo da analizzare imputando un dato plausibile ove esso manca.

4.1.1.4 L'imputazione dei dati mancanti

Nel caso in cui i dati siano "mancanti completamente a caso" o "mancanti a caso", l'uso di tecniche che prevedono la sostituzione del dato mancante con un valore presunto (si parla di **imputazione del dato mancante**) permettono di ricostruire artificialmente un data set completo senza comprometter la valutazione delle quantità di interesse.

La più semplice tecnica di imputazione del dato mancante è quella per cui si sostituisce ad esso un valore che è pari a una sintesi dei dati disponibili: un valore centrale ad esempio. Al posto del valore mancante, si può ad esempio sostituire la media dei dati disponibili (o la loro mediana).

```
# imputazione con la media: creo nuove variabili cui aggiungere i valori imputati
glucoseMCARi<-glucoseMCAR
glucoseMARi<-glucoseMAR
glucoseMNARi<-glucoseMNAR

# Imputo con la media nel caso di dati mancanti completamente a caso

glucoseMCARi[is.na(glucoseMCAR)] <- mean(glucose, na.rm = TRUE)
mean(glucoseMCARi)
```

```
## [1] 123.1477
```

```
# Imputo con la media nel caso di dati mancanti a caso
glucoseMARi[is.na(glucoseMAR)] <- mean(glucose, na.rm = TRUE)
mean(glucoseMARi)
```

```
## [1] 120.7636
```

```
# Imputo con la media nel caso di dati NON mancanti a caso
glucoseMNARi[is.na(glucoseMNAR)] <- mean(glucose, na.rm = TRUE)
mean(glucoseMNARi)
```

```
## [1] 121.6324
```

Come si vede, l'imputazione con la media dei casi disponibili, nel caso MAR e MNAR, non elimina la sottostima che avevamo osservato.

Tuttavia, nel caso di dati "mancanti a caso" si potrebbe, ad esempio, imputare non la media di tutti i casi disponibili, ma la media di specifici gruppi. Ri-

prendendo il caso, riportato per illustrare i dati "mancanti a caso", si dovrebbe imputare la media delle stature dei rispondenti donna, se il caso mancante riguardava una unità di sesso femminile, o la media degli uomini, nei restanti casi.

```
# preparo le variabili per le imputazioni con le medie per sesso
glucoseMARim<-glucoseMAR
glucoseMNARim<-glucoseMNAR

# Imputo con la media per sesso nel caso MAR
glucoseMARim[is.na(glucoseMAR)&diabetes=="neg"] <-
  mean(glucose[diabetes=="neg"], na.rm = TRUE)
glucoseMARim[is.na(glucoseMAR)&diabetes=="pos"] <-
  mean(glucose[diabetes=="pos"], na.rm = TRUE)
mean(glucoseMARim)
```

```
## [1] 122.9941
```

```
# Imputo con la media per sesso nel caso di dati NON MAR
glucoseMNARim[is.na(glucoseMAR)&diabetes=="neg"] <-
  mean(glucose[diabetes=="neg"], na.rm = TRUE)
glucoseMNARim[is.na(glucoseMAR)&diabetes=="pos"] <-
  mean(glucose[diabetes=="pos"], na.rm = TRUE)
mean(glucoseMNARim)
```

```
## [1] NA
```

```
detach(pimanmd)
```

Si noti che l'uso di imputazione con la media per i due gruppi ha eliminato la distorsione nel caso dei dati mancanti a caso (MAR) e la riduce (ma non la elimina) nel caso MNAR.

L'idea di utilizzare un metodo di imputazione che tenga conto di altre variabili, osservate completamente o in misura molto ampia, è peraltro alla base delle tecniche di imputazione più sofisticate che non saranno trattati con completezza qui (anche se richiameremo brevemente il tema più avanti quando parleremo di analisi di regressione). Si possono usare tecniche di imputazione in cui si tiene conto contemporaneamente di più varaibili, completamente o quasi completamente osservate, nel fare l'imputazione.

Va tenuto comunque a mente che anche se si usano tali tecniche, non sarà mai garantito che la presenza di dati mancanti non distorca le analisi perchè non si può escludere a priori il caso dei dati non mancanti a caso. È sempre consigliabile quindi riportare all'inizio delle analisi le statistiche sui dati mancanti per ciascuna delle variabili e segnalare l'eventuale uso di tecniche di imputazione.

Ovviamente esistono in `R` numerosi pacchetti che consentono l'imputazione di dati mancanti, anche utilizzando tecniche molto sofisticate che qui non sono

trattate poiché coinvolgono modelli di analisi e algoritmi predittivi più avanzati.

4.1.2 Gli outliers (valori anomali)

Un valore anomalo, o, in inglese **outlier** (letteralmente "che giace al di fuori") è un dato che si discosta in modo "significativo" dalla maggioranza dei dati. È quindi un valore "strano" cui viene naturale dedicare una speciale attenzione. Si noti che si parla di valori anomali in particolare quando si è interessati all'analisi di variabili quantitative. Non viene in genere definito un concetto analogo nel caso di variabili categoriali: per esse, al più, si potrà notare che alcune modalità sono molto rare.

L'identificazione di un outlier deriva dal confronto di ciascun dato con le altre osservazioni del collettivo che si vuole analizzare, non viene quindi definito in termini assoluti: il valore della statura di 2.10 cm. può essere un valore anomalo se riferito alla popolazione generale, ma non lo sarebbe se considero la popolazione dei giocatori di basket maschi. La sua definizione è poi strettamente legata alle caratteristiche del fenomeno che si sta analizzando. Nei dati sulla variabile `LOSS` del dataframe `AutoBi` si sono trovati valori elevatissimi ma che potrebbero essere del tutto coerenti con la natura di una variabile caratterizzata da forte asimmetria positiva con lunga coda a destra.

Esistono metodi, alcuni molto complessi, computazionalmente onerosi e/o che coinvolgono strumenti di statistica inferenziale, per identificare valori anomali. I vari metodi si propongono di assegnare a ciascuna valore un punteggio che ne denoti il grado di anomalia (rispetto alla massa dei dati): si tratta cioè di segnalare i valori sospetti. In questa sede ci limiteremo a elencare solo i più semplici strumenti analitici e grafici.

Tuttavia una volta chi si decide di usare un metodo che identifichi un valore come anomalo, nella fase di preparazione dei dati, occorre decidere come agire: si tiene il dato, lo si elimina dall'insieme dei dati da analizzare, lo si trasforma o lo si sostituisce?

4.1.2.1 Come individuare i valori anomali

Nella precedente sezione è stata introdotta il diagramma a scatola con baffi, **boxplot**. Si è visto che nel grafico vengono riportati i singoli valori che risultano lontani dal centro della distribuzione. I centro della distribuzione è rappresentato dalla scatola che, ricordiamo, ha come limiti il I e il III quartile. Nel boxplot vengono rappresentati individualmente quei valori che sono distanti dagli estremi della scatola più di una volta e mezza la lunghezza della scatola stessa (che è pari allo scarto interquartile SI).

Dati i valori osservati della variabile X, sono quindi da guardare con attenzione quei valori che sono al di fuori dell'intervallo

$$I = [x_{0.25} - 1.5SI \;\; ; \;\; x0.75 + 1.5SI]$$

Tali valori devono essere esaminate con attenzione in quanto sono potenziali valori anomali. Inoltre, i valori sono distanti dalla scatola più di 3 volte la lunghezza della scatola stessasono ancora più sospetti e vengono definiti come potenziali "ouliers distanti" (*far outliers*).

Tale criterio deriva dalla aspettativa che la variabile osservata presenti una distribuzione simmetrica e con code regolari (in particolare, con curtosi bassa). In effetti non appena la distribuzione della variabile esibisce una marcata simmetria il numero di valori che vengono segnalati come outliers da un boxplot diventa molto elevato: in tal caso è difficile pensare che si tratti di valori anomali.

Sotto è riportato il summary e il boxplot relativo alla variabile età per i dati di `AutoBi`.

```
summary(AutoBi$CLMAGE)
boxplot(AutoBi$CLMAGE)
```

Come si vede vengono evidenziati 7 valori che sono superiori a 80. Trattandosi di età è difficile ritenere un'anomalia che su oltre un migliaio di casi vi siano 7 casi di persone con oltre 80 anno (si noti che il valore più eevato è 95). Diverso sarebbe stato se avessimo osservato un valore pari, ad esempio, a 132: in tal caso si sarebbe trattato di un valore molto anomal, "sospetto" e certamente frutto di un errore.

La logica generale che è dietro alla segnalazione dei valori sospetti nel boxplot è quella di prendere come riferimento un valore centrale (ad esempio, la media, la mediana o la scatola stessa) e considerare valori anomali quelli che si discostano in valore assoluto dal valore centrale più k moltiplicato per una opportuna misura di variabilità. Il valore di k è in larga misura arbitrario salvo per il caso in cui ci aspettiamo che i dati abbiano una distribuzione ben precisa.

Se, ad esempio, ci aspettassimo che i dati siano ben rappresentati da un comportamento simile a quello di una gaussiana allora diventerebbe efficace un critero che segnali come outlier valori esterni all'intervallo

$$[media - k * sqm \;\; ; \;\; media + k * sqm]$$

ove sqm rappresenta lo scarto quadratico medio delle osservazioni. In effetti se si trattasse di dati dal comportamento comparabile con quello di una gaussiana allora ci aspettiamo che, ponendo ad esempio k=3.3, si osservino valori esterni a quell'intervallo in meno di 1 caso su 1000.

Più generale, è un criterio simile in cui consideriamo un indice di tendenza centrale e un indice di variabilità più resistente (così che non risulti influenzato dalla presenza di valori anomali). Questo condurrebbe al seguente intervallo (criterio di Hampel) per la definizione di valori anomali:

$$[mediana - k * MAD \;\; ; \;\; mediana + k * MAD]$$

k viene spesso posto pari a 1.5 o a 3.

Come detto, esistono metodi più sofisticati per definire i valori sospetti che non tratteremo in questa sede e va inoltre ricordato che l'individuazione o la definizione di un valore anomale quando si analizzano più variabili congiuntamente può rivelarsi un'operazione estremamente più complicata.

Tuttavia, una volta identificato un valore che sembra molto diverso da tutti gli altri si presenta il problema di decidere come agire.

4.1.2.2 Cosa fare con i valori anomali?

Possiamo identificare 3 principali azioni conseguenti all'identificazione dei valori anomali: 1. rimozione dei valori anomali; 3. valutazione dell'impatto dei valori anomali sulle analisi e uso di metodi resistenti; 2. trasformazione delle variabili.

4.1.2.2.1 Rimozione dei valori anomali La rimozione dei valori anomali è da evitare.

Vi sono almeno due casi in cui è però possibile procedere all'eliminazione del dato:

1. quando dall'esame del valore sospetto emerge con certezza che il dato sia conseguente ad un errore di misurazione, di trascrizione, di riporto. Se misuro stature di maschi adulti (in cm.) e osservo misure anomale come 475cm. o 1.82cm, sono certo che si tratta di misure errate perchè tali valori sono impossibili. Non posso fare altro che eliminare tale dato. In questo caso, si può attribuire al valore anomalo il valore `NA` e quindi trattarlo come un dato mancante. Ovviamente se è possibile correggere il dato avendo identificato l'errore questo porta a sostituire il valore errato con quello corretto. Nell'esempio sopra potrei avere verificato che la misura dell'altezza 1.82 era sbagliata perchè riportata in metri e non in cm.. Oppure, se posso tornare sui dati originali, potrei verificare che è stata male riportata la prima cifra e che invece di 475 il valore della statura era 175.
2. Se si verifica che il valore (o i valori anomali) sono relativi a unità che appartengono a una popolazione diversa da quella che si intende sottoporre ad analisi. Se si sta conducendo un'analisi sui consumi di autovetture con motore termico, potrebbe accadere che nella fase di rilevazione dati si inseriscano le misurazioni anche per alcune vetture ibride. Se osservassimo dei valori sospetti per queste vetture (consumi troppo bassi rispetto alle altre) si sarebbe autorizzati ad escluderle dall'analisi perchè non dovrebbero far parte del collettivo oggetto della analisi. In questo caso, la presenza di valori anomali è indicativa di errori nella fase di preparazione dei dati che possono eventualmente essere corretti tornando a effettuare la misura o ricontrollando il dato.
3. La trasformazione del valore anomalo, può essere conseguente a una trasformazione della variabile per tutti i suoi valori (aspetto trattato nella successiva sezione), alla sostituzione del singola valore sospetto magari con una tecnica di imputazione (come fosse un dato mancante).

4.1.2.2.2 Valutazione dell'influenza dei valori anomali sull'analisi Essendo in una fase esplorativa, non si hanno precise assunzioni sulla forma delle distribuzioni della variabili. Di solito non sappiamo, a priori, se la variabile oggetto di analisi ha una forma simmetrica o asimmetrica, se ha code pesanti o se ha una distribuzione irregolare (ad esempio multimodale). È noto però che alcune delle misure di sintesi o anche di visualizzazione grafica sono influenzate dalla presenza di valori eccezionalmente diversi da tutti gli altri:

1. una primo passo è quindi quello di valutare che impatto ha la presenza di valori anomali sulle analisi. Questo implica che si svolga un'analisi dei dati includendo tutti i dati e una seconda analisi escludendo i valori anomali. Dal confronto dei due risultati possiamo giudicare quanto "influenti" sono i valori anomali presenti nel data set. Esistono, nel caso di analisi più complesse che coinvolgono più variabili, specifici metodi per valutare l'influenza di ciascuana osservazione. Se si osserva che alcuni dati sono "influenti" e quindi modificano sensibilmente l'analisi,
2. un secondo aspetto porta a introdurre metodi di analisi che sono meno sensibili alla presenza di valori anomali. Abbiamo già definito alcune semplici tecniche di sintesi delle variabili come più "resistenti". Esiste un filone di analisi statistiche che si adattano al caso in cui sono presenti dati anomali: si tratta di tecniche dette "robuste" per le quali l'idea è di pesare opportunamente i dati così che venga limitato l'impatto delle osservazioni aberranti.

4.1.2.2.3 Trasformazioni delle variabili Le tecniche di identificazione dei valori anomali illustrate sopra non tengono in considerazione il fatto che alcuni fenomeni sono tali da generare raramente delle osservazioni molto diverse dalle altre. Questo riguarda, ad esempio, fenomeni estremi: se, ad esempio, si raccolgono dati sulla piovosità media giornaliera si troveranno misure che mostreranno

4.2 Trasformazione delle variabili e ricodifiche

4.2.1 (Ri)codifiche di variabili categoriali

Molto si è insistito sul distinguere con chiarezza, in tutte le fasi di analisi dei dati, se una variabile è quantitativa o categoriale perchè alcune tecniche di analisi o di rappresentazione grafica possono essere utilizzate in modo appropriato per ciascuna categori di variabili. Non posso calcolare media o varianza per una variabile categoriale e non è di solito una buona idea usare un barplot per variabili quantitative continue.

Tuttavia, risulterà utile poter dare una rappresentazione numerica adeguata anche per le variabili categoriali visto che in alcuni casi si dovranno analizzare congiuntamente variabili di diversa natura.

4.2.1.1 La codifica numerica di variabili qualitative (fattori) ordinali

Conviene considerare dapprima il caso di variabili qualitative ordinali: per esse la sostituzione con valori numerici è spesso considerata accettabile.

Ad esempio, si consideri la modalità con cui si rilevano le opinioni di un utente o di un consumatore su un prodotto o un servizio.

Si chiede di esprimere un giudizio scegliendo fra alcune modalità che sono ordinate in relazione al gradimento di un servizio indicando, ad esempio, in relazione a una frase che descrive la qualità del servizio, se si è "pienamente d'accordo", "abbastanza d'accordo", "indifferenti", "poco d'accordo","per niente d'accordo". Tale modalità di rilevazione è detta scala di Likert.

Si tratta, com'è evidente, di un fattore qualitativo ordinale e spesso ai fini di analisi successive si ricorre alla ricodifica numerica facendo corrispondere i valori da 1 a 5 ai diversi livelli.

A volta si usano scale simili su 7 livelli e si fanno corrispondere valori da 1 a 7.

Tale ricodifica numerica è considerata accettabile anche se va tenuto presente che i valori numerici implicano una precisa distanza fra le modalità che potrebbero costituire una forzatura se riferite alla variabile categoriale: cioè non è detto che fra essere "indifferenti" e essere "poco d'accordo" ci sia la medsima distanza che fra essere "poco d'accordo" e "completamente d'accordo".

4.2.1.2 La codifica numerica di variabili categoriali (fattori)

Nel caso di variabili categoriali non ordinali la sostituzione delle modalità con valori numerici è di solito più arbitraria. Tuttavia esistono alcune ricodifiche rispettose dell'informazione contenuta ma che consentono di ottenere variabili numeriche.

Il caso più semplice è quello relativo a variabili categoriali dicotomiche (con due sole modalità). In tal caso, si è già visto come sia accettabile far corrispondere i valori 1 e 0 alle due modalità essendo la variabile di fatto di tipo booleano. Talvolta, per lo sviluppo di alcuni algoritmi, si usa anche la codifica -1 e 1.

Più complessa è la ricodifica numerica nel caso di fattori con più di due modalità.

4.2.1.2.1 One hot-encoding (disgiuntiva completa) La tecnica più corretta è quella di creare tante variabili dicotomiche quante sono le modalità. Ogni variabile registra se per una data unità si osserva o meno la specifica modalità. Conviene vedere un esempio, utilizzando R.

```
varcat<-factor(c("A","B","A","C","B","A","A","C","B","C"))
# consideriamo il fattore varcat
dataf<-data.frame(varcat)
hotenc<-model.matrix(~dataf$varcat-1,)
# questo crea le tre variabili
```

```
dataf<-data.frame(varcat,hotenc)
dataf
```

```
##    varcat dataf.varcatA dataf.varcatB dataf.varcatC
## 1       A             1             0             0
## 2       B             0             1             0
## 3       A             1             0             0
## 4       C             0             0             1
## 5       B             0             1             0
## 6       A             1             0             0
## 7       A             1             0             0
## 8       C             0             0             1
## 9       B             0             1             0
## 10      C             0             0             1
```

```
# sono state create tre nuove variabili
```

Questo tipo di codifica è detto *"one hot encoding"*. Si noti che a volte è preferibile utilizzare la convenzione per cui si costituiscono tante variabili quante sono le modalità meno 1. La restante modalità può essere ottenuta per differenza da un vettore fatto tutto di "1". Il difetto principale di tale ricodifica è che se la variabile ha molte modalità allora verranno create tante variabili nuove e crescerà di conseguenza la dimensione del data set.

```
# Differente, ma equivalente, forma di encoding
hotenc1<-model.matrix(~dataf$varcat)
# questo crea le tre variabili ma la prima è detta "intercetta",
# chiariremo più avanti perchè, ed è identicamente pario a 1
dataf1<-data.frame(varcat,hotenc1)
dataf1
```

```
##    varcat X.Intercept. dataf.varcatB dataf.varcatC
## 1       A            1             0             0
## 2       B            1             1             0
## 3       A            1             0             0
## 4       C            1             0             1
## 5       B            1             1             0
## 6       A            1             0             0
## 7       A            1             0             0
## 8       C            1             0             1
## 9       B            1             1             0
## 10      C            1             0             1
```

4.2.1.2.2 Altre forme di ricodifica (sconsigliate) Esistono altre forme di ricodifica che sono talvolta utilizzate, ad esempio nell'ambito di procedure di machine learning, ma che non sono sostenute da ragionamenti rigorosi e NON

sono assolutamente da consigliare per analisi esplorative. Tuttavia le citiamo per completezza:

1. Assegnare un valore numerico a ciascuna categoria, ad esempio basandosi sull'ordine alfabetico (di fatto anche `R` fa una cosa simile). Se questo può essere a volte comodo perchè poterbbe consentire di utilizzare minore spazio di memoria, **non va assolutamente poi condotta un'analisi della variabile numerica ottenuta** perchè risulterebbe priva di senso.

```
# si noti che `R` fa esattamente questo con i fattori ma conservando l'informazione
# sulla corrispondenza fra valori numerici e categorie.
# Non sarà mai possibile su un fattore compiere analisi tipici di variabili numeriche
str(varcat)
```

```
##  Factor w/ 3 levels "A","B","C": 1 2 1 3 2 1 1 3 2 3
```

```
# ad esempio
mean(varcat)
```

```
## Warning in mean.default(varcat): argument is not numeric or logical: returning
## NA
```

```
## [1] NA
```

Tale procedura è detta "label encoding".

Una altro tipo di ricodifica numerica, si utilizza quando l'interesse è sull'analisi di una altra variabile, diciamo Y (quantitativa o qualitativa) detta target, e si assegnano alla variabile categoriale X, i valori medi di Y, o la frequenza assoluta o relativa di Y, in corrispondenza di ciascuna modalità di X. Questa procedura, detta di "target encoding", è utilizzata talvolta all'interno di alcuni algoritmi di machine learning.

Tuttavia anche tale procedura è in generale priva di solide giustificazioni teoriche e NON va mai usata per analisi esplorative.

4.2.2 Trasformazioni di variabili quantitative

Le variabili quantitative osservate sono di solito il risultato di misurazioni e valori ottenuti in relazione a un preciso sistema di riferimento. Questo implica che si possano effettuare trasformazioni delle variabili per riportarle a un'unità di misura convenzionale diversa. L'esempio che viene subito in mente è la temperatura per cui potrei avere ottenuto le misure in gradi centigradi e volere poi trasformare gli stessi in gradi fahrenait. In questo caso si tratta di una trasformazione lineare. O ancora si può pensare a esprimere misure monetarie utilizzando una diversa unità (ad es., conversione da dollari a euro).

In altri casi si ricorre a trasformazioni delle variabili per agevolare la rappresentazione grafica della variabile stessa o o per rendere più agevole il confronto di una variabile che in due collettivi ha ordini di grandezza molto diversi. Citiamo alcu-

ne trasformazioni il cui uso è consueto proprio per agevolare la rappresentazione grafica o l'analisi di taluni aspetti della variabile.

In generale, è spesso indispensabile operare delle trasformazioni quando si analizzano tante variabili congiuntamente ed è opportuno che esse siano tutte riportate a scale di misura comparabili.

Si ricorda infine che è una trasformazione di variabile quantitativa anche il raggruppamento in classi già introdotto per costruire tabelle per variabili quantitative o per rappresentarle graficamente con l'istogramma. In quel caso, si era già notato come il ridurre la variabile quantiativa alla stregua di un fattore ordinato comporta una perdità di dettaglio informativo.

4.2.2.1 La standardizzazione

Spesso si procede a operazioni molto semplici come aggiungere o togliere una costante o dividere/moltiplicare i dati per un valore. Si noti che in generale tali operazioni hanno un impatto noto sulle misure di centralità e dispersione principali.

In particolare se otteniamo una variabile Z a partire dai valori osservati su Y, ad esempio $Z_i = bX_i + a$ dove a e b sono costanti reali note, si noti che:

- $M_Z = bM_X + a$ dove M_Z e M_X sono rispettivamente la media della variabile trasformata e di quella originale
- $V_Z = b^2 V_X$ dove V_Z e V_X sono rispettivamente la varianza della variabile trasformata e di quella originale.

Molto rilevante è la seguente trasformazione: $Z_i = \frac{Y_i - M_Y}{\sqrt{V_Y}} = \frac{Y_i}{\sqrt{V_Y}} - \frac{M_Y}{\sqrt{V_Y}}$.

Essa è detta **standardizzazione**. Si verifica immediatamente che è

- $M_Z = 0$
- $V_Z = 1$.

Questo premette di confrontare variabili in relazione a altri aspetti (l'asimmetria ad esempio) eliminando l'impatto di media e varianza. Inoltre se analizzo molte variabili, tutte con scala diversa, la standardizzazione le riporta tutte a unità di misure confrontabili.

L'aspetto più rilevante è che in questo caso, la forma della distribuzione è invariata.

4.2.2.2 La riduzione a un intervallo unitario

Se l'obiettivo della trasformazione è quello di riportare tante variabili su scale confrontabili, si può ricorrere a un semplice espediente così che ogni variabile assume valori compresi nel medesimo intervallo. Ad esempio, si considera di solito l'intervallo $[0, 1]$ (o suoi multipli, $[0, 100]$). Tale operazione è spesso denotata come normalizzazione min-max (anche se non ha niente a che vedere con la distribuzione normale).

A tal fine, per i valori $x_i, x_2, \ldots, x_n$ di una generica variabile X basta porre

$$y_i = \frac{x_i - x_{(1)}}{x_{(n)} - x_{(1)}}$$

Si ricorda che la notazione $x_{(1)}$ e $x_{(n)}$ individua, rispettivamente, il più piccolo e il più grande fra i valori di X osservati.

La variabile Y assumerà valori compresi nell'intervallo $[0, 1]$. Le variabili così normalizzate non avranno però tutte uguale media e varianza come nel caso della standardizzazione.

4.2.2.3 Trasformazioni non lineari

L'uso di trasformazioni non lineari è piuttosto consueto quando una variabile assume valori estremi e con code lunghissime. Questo accade spesso, ad esempio, con variabili che assumono valori solo positivi e rappresentano grandezze monetarie (si pensi alla variabile `LOSS`già incontrata analizzando i dati di `AutoBI`). In tali casi è utile ricorrere a trasformazioni come $Z_i = Y_i^{1/k}$ con $k > 1$ (spesso si usa $k = 2$, la radice quadrata, o pari a 3, cubica).

In alcuni casi può essere opportuno considerare la trasfomazione reciproco $Z_i = Y_i^{-1}$.

Una delle trasformazioni più utilizzate nel caso di variabile con coda destra lunga (asimmetria positiva) è la trasformazione logaritmica $Z_i = log(Y_i)$ (si ricordi che era stata già utilizzata in occasione della illustrazione del boxplot) .

Ovviamente, si può ricorrere a trasformazioni come il reciproco o il logaritmo solo se la variabile originaria non contiene valori pari a 0 o negativi. Se si osservano valori pari a 0, talvolta si conviene di aggiungere ad essi un valore $c > 0$ molto piccolo (ad esempio 0.01) così da evitare il problema.

Si noti che in tutti questi casi, se pure si fosse già ottenuta la media della variabile originale, non è possibile ottenere media della variabile originale semplicemente applicando ad esse la medesima trasformazione (una conseguenza della disuguaglianza di Jensen).

Capitolo 5

Analisi statistica di due variabili (bivariata)

Prima di procedere all'analisi di due variabili, considerando i diversi casi possibili, 2 variabili entrambe qualitative, 1 variabile quantitativa e 1 qualitativa e, infine, 2 variabili quantitative, conviene introdurre alcuni strumenti che possono esser utili per confrontare distribuzioni di dati. Si considerano, in particolare, strumenti grafici che permettono di confrontare dati empirici con dati conformi a un modello teorico oppure di tecniche per confrontare due distribuzioni empiriche in due o più insiemi di dati osservati. Alcuni di tali strumenti saranno poi utili in seguito anche per le tecniche di analisi bivariata.

5.1 Confronto fra distribuzioni empiriche e teoriche

In particolare per l'analisi di variabili quantitative, risulta spesso utile poter disporre di strumenti per verificare se i dati osservati si conformano a qualche modello teorico. Ci si potrebbe chiedere se i dati abbiano una distribuzione che si discosta poco da un modello distributivo noto e del quale sono note le proprietà come potrebbe essere, ad esempio, la gaussiana.

Il problema può essere affrontato con un approccio formale e con tecniche di inferenza statistica e/o in modo più informale con strumenti grafici. Spesso in realtà i due approcci sono usati in forma complementare. Si rinvia a testi di statistica inferenziale per un approccio più formale al problema, gli strumenti grafici adatti a tale scopo sfrutteranno essenzialmente alcune delle tecniche di visulaizzazione già introdotte per l'analisi di variabili quantitative.

5.1.1 Confronto fra funzione di ripartizione empirica e teorica

La **funzione di ripartizione empirica** è stata già introdotta nel precedente capitolo. Uno strumento immediato per confrontare la distribuzione dei dati osservata si ottiene sovrapponendo alla funzione di ripartizione empirica dei dati la funzione di ripartizione della variabile che si vuole confrontare. Tale funzione di ripartizione deve quindi essere completamente nota. Per illustrare tale grafico in `R` si mettono a confronto i dati empirici con quelli costituiti da valori pseudocasuali generati da una variabile aleatoria nota. Ricordiamo che uno strumento importante di `R` sono le funzioni speciali per trattare modelli aleatori per i quali sono stati già mostrati alcuni esempi.

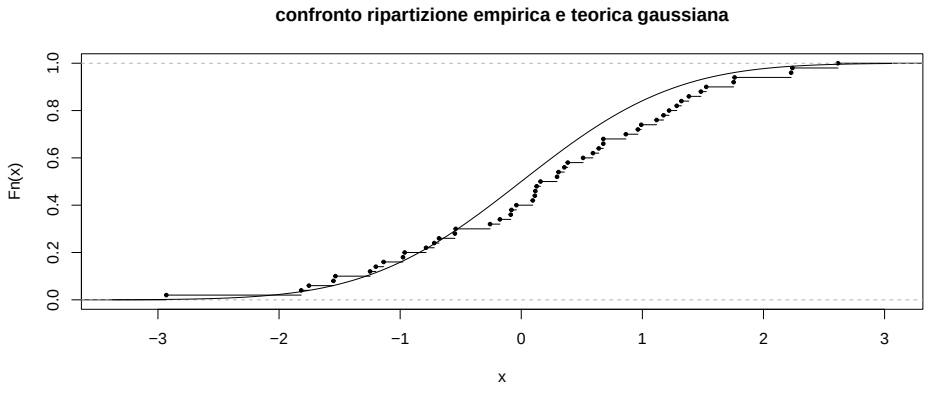
Si tratta di funzioni che consentono di valutare alcune quantità rilevanti per specifiche variabili casuali appartenenti alle principali famiglie parametriche. Si può ottenere, ad esempio, molto facilmente la funzione di probabilità (o di densità), la funzione di ripartizione, la funzione quantile e si possono generare numeri casuali per distribuzioni di probabilità, come la Binomiale, la Poisson, la Normale, la Gamma e molte altre. Le funzioni contengono un nome che identifica la famiglia di variabili casuali e se tale nome è preceduto da `d` viene calcolata la probabilità (o la densità), se preceduto da `p` si ottiene la funzione di ripartizione, con `q` si ottiene la funzione dei quantili e se preceduta da `r` si generano valori casuali. Utilizzando l'help in linea di R (`?Distributions`) si ottiene l'elenco tutte le distribuzioni di probabilità implementate nell abase di `R`. Ad esempio, la funzione `pnorm(x,m,s)` valuta la funzione di ripartizione in x per un modello gaussiano con media m e deviazione standard s.

Di seguito sono elencate alcune delle distribuzioni di probabilità attualmente disponibili in R:

Distribuzione	nome R	parametri aggiuntivi
beta	beta	shape1, shape2
binomial	binom	size, prob
Cauchy	cauchy	location, scale
chi-squared	chisq	df
exponential	exp	rate
F	f	df1, df2
gamma	gamma	shape, scale
geometric	geom	prob
hypergeometric	hyper	m, n, k
log-normal	lnorm	meanlog, sdlog
negative binomial	nbinom	size, prob
normal	norm	mean, sd
Poisson	pois	lambda
Student's t	t	df,
uniform	unif	min, max
Weibull	weibull	shape, scale

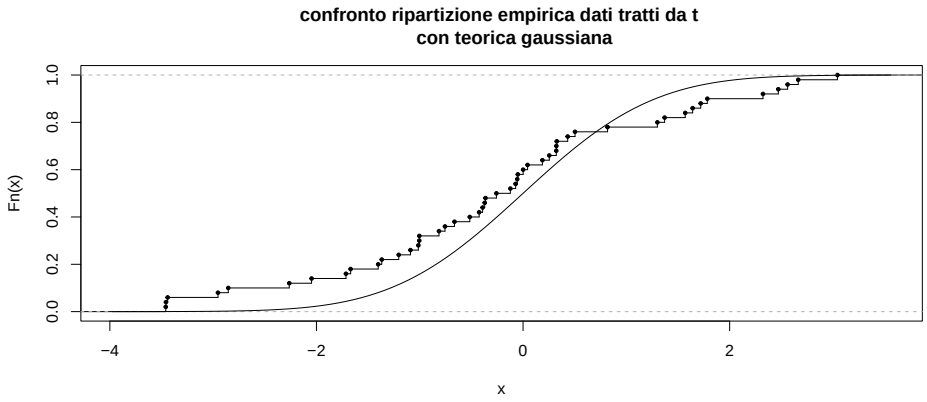
Molte altre famiglie di variabili aleatorie sono disponibili in pacchetti dedicati all'analisi di dati per specifici ambiti applicativi. Si simulano ora dei dati da una gaussiana e si ottiene il grafico della funzione di ripartizione empirica. Ci si propone di confrontare la funzione di ripartizione empirica dei dati osservati con una Gaussiana di media e varianza nota.

```
# si generano i dati pseudocasuali da una gaussiana standard

set.seed(1111)
dat1<-rnorm(50)
plot(ecdf(dat1),cex=.5,
     main="confronto ripartizione empirica e teorica gaussiana")
curve(pnorm(x), add=TRUE)
```




```
# si prova ora a generare da una distribuzione t di student con 1
# grado di libertà (simile alla gaussiana ma con code più pesanti)
```

```
# e si confrontno i due insiemi di dati con una gaussiana
# di media e varianza nulla
dat2<-rt(50,2)
plot(ecdf(dat2), cex=.5, verticals=TRUE,
     main="confronto ripartizione empirica dati tratti da t
     con teorica gaussiana")
curve(pnorm(x), add=TRUE)
```




```
par(mfrow=c(1,1))
```

Come si vede, nel secondo caso c'è uno scostamento sensibile fra la funzione di ripartizione empirica per i dati generati da una distribuzione con code pesanti e una ditribuzione gaussiana. In effetti, buona parte delle procedure statistiche per verificare se i dati osservati sono conformi a una distribuzione teorica si basano sul confronto fra tali curve.

In particolare la distanza fra le due curve viene spesso riassunta dalla statistica di Kolmogorov-Smirnov D definita come

$$D = \sup_x |\hat{F}(x) - F(x)|$$

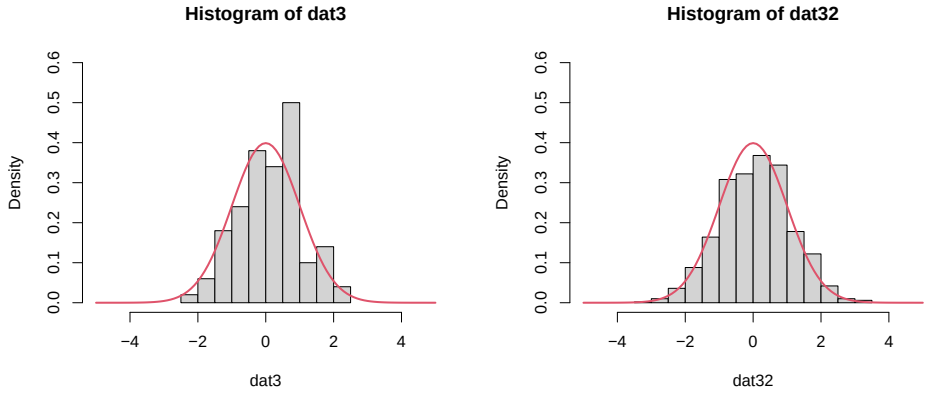
cioè la massima distanza verticale fra le due curve.

Si noti che il confronto può essere fatto anche con una distribuzione teorica per la quale si ipotizza la conoscenza della famiglia ma non la conoscenza dei parametri. Ad esempio, si potrebbe ipotizzare i dati provengano da una gaussiana per la quale media e varianza sono ignote. In tal caso si possono utilizzare lea media e la varianza calcolate sui dai dati raccolti e usarle come fossero i valori veri. Il confronto grafico fra le curve resta uno strumento utile anche in questo caso ma le proprietà statistiche, ad esempio quelle della statistica D non sono le stesse e questo ha conseguenze nell'ambito dei metodi statistici inferenziali.

5.1.2 Confronto delle funzioni di densità (empirica e teorica)

Un altro modo per confrontare una variabile osservata con con un modello teorico è quello di confrontare la funzione di densità del modello teorico (che si suppone sempre completamente noto). Si può quindi sovrapporre la funzione di densità teorica alle sue versioni empiriche tratte dai dati. Ad esempio, sovrapporre la funzione di densità teorica all'istogramma (che ricordiamo essere une versione grezza ed emprica della funzione di densità)

```
par(mfrow=c(1,2))
dat3<-rnorm(100) # si generano dati da una gaussiana standard
hist(dat3,prob=T, xlim=c(-5,5), ylim=c(0,0.6))
# e si sovrappone la funzione di densità della gaussiana standard
curve(dnorm(x,0,1),  col=2, lwd=2, add=TRUE)

# Ora si può provare con una numerosità dei dati più elevata
dat32<-rnorm(1000) # generiamo dati da una gaussiana standard
hist(dat32,prob=T, xlim=c(-5,5), ylim=c(0,0.6))

# sovrapponiamo la funzione di densità della gaussiana standard
curve(dnorm(x,0,1), col=2, lwd=2, add=TRUE)
```

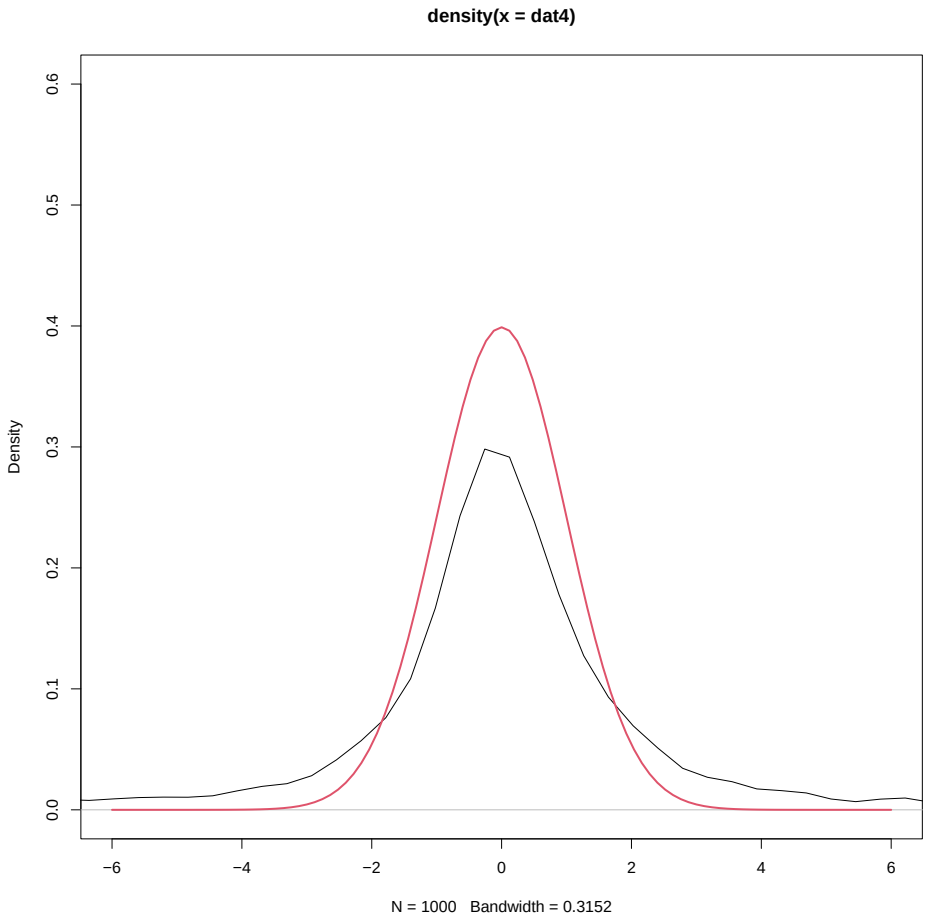



```
par(mfrow=c(1,1))
```

Con 100 dati, come si vede, l'istogramma risulta instabile e la somiglianza è vaga; come regola generale non vanno sopravvalutate le differenze nei grafici se si dispone di pochi dati. Con numerosità maggiori (e quindi con più classi) la somiglianza fra istogramma e modello teorico (nel caso in cui i dati peraltro provengono da quel modello teorico) aumenta.

In alternativa, si potrebbe usare la curva di densità empirica ottenuta con il metodo del nucleo. Si può provare a vedere se i dati empirici, generati da una distribuzione con code pesanti, come la t di student con 1 grado di libertà, differiscono dal modello teorico gaussiano.

```
dat4<-rt(1000,1) # si generano i dati da una t di studente con 1 gdl
plot(density(dat4), xlim=c(-6,6), ylim=c(0,0.6))
# se si sovrappone la funzione di densità di una gaussiana,
# dovremmo notare differenze sulle code
curve(dnorm(x,0,1),  col=2, lwd=2, add=TRUE)
```

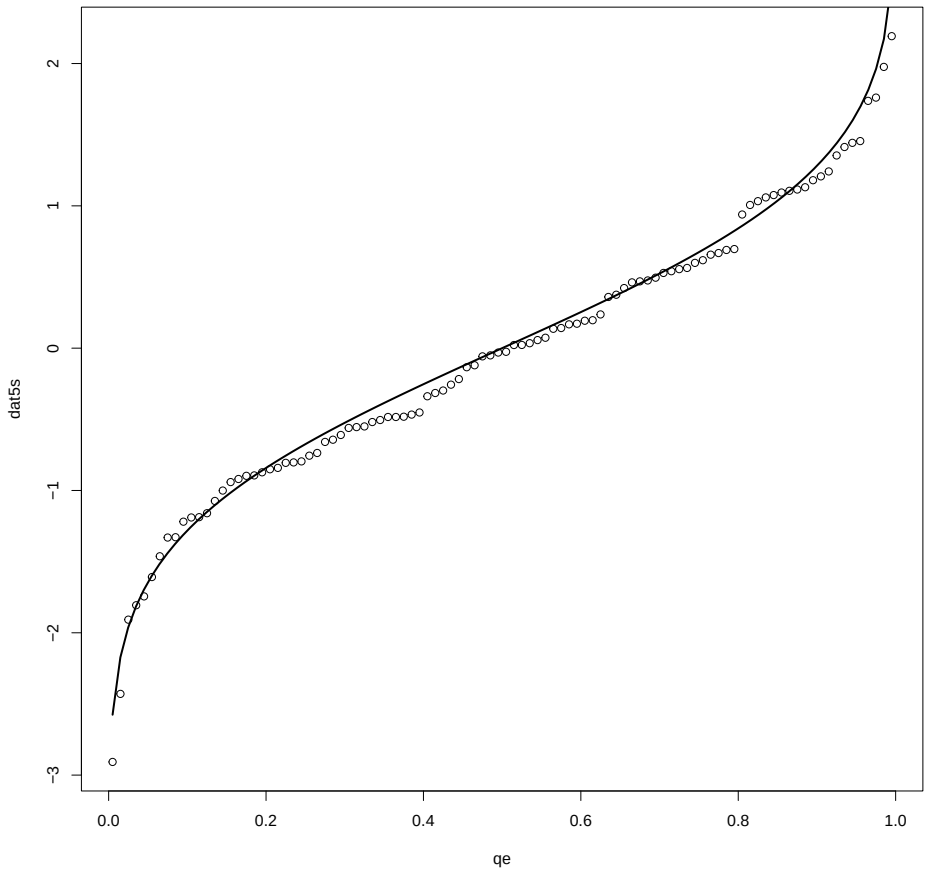


5.1.3 Confronto dei quantili: il grafico quantile-quantile

Per una variabile aleatoria continua X è definita la funzione dei quantili $F^{-1}(p) = x_p$ che è la funzione inversa della funzione di ripartizione. Essa associa ad ogni valore $p \in [0,1]$ il corrispondente quantile x_p. Si noti che in `R` per i diversi modelli teorici si ottiene il quantile di una variabile aleatoria appartenente a una famiglia parametrica premettendo `q` al nome che identifica la famiglia. Quindi `qnorm(p,m,s)` fornisce per un dato p il quantile della famiglia gaussiana di parametri dati.

Per un insieme di dati quantitativi osservati è possibile ottenere il grafico dei quantili empirici. Si ricorda che il valore $x_{(i)}$ rappresenta il quantile empirico $x_{\frac{i}{n}}$. Nel seguito usiamo la convenzione per cui il valore $x_{(i)}$ rappresenta il quantile $x_{\frac{i-0.5}{n}}$

```
set.seed(2222)
dat5<-rnorm(100)  # si generano i dati da una gaussiana standard e si
                  # ottengono i quantili empirici
dat5s<-sort(dat5)
qe<-((1:100)-0.5)/100
# e poi si traccia il grafico dei quantili
plot(qe, dat5s)
# cui si sovrappone il grafico dei quantili di una gaussiana standard
curve(qnorm(x,0,1),  lwd=2, add=TRUE)
```



Come si vede le due curve sono molto vicine a indicare che i dati provengono da quel modello teorico.

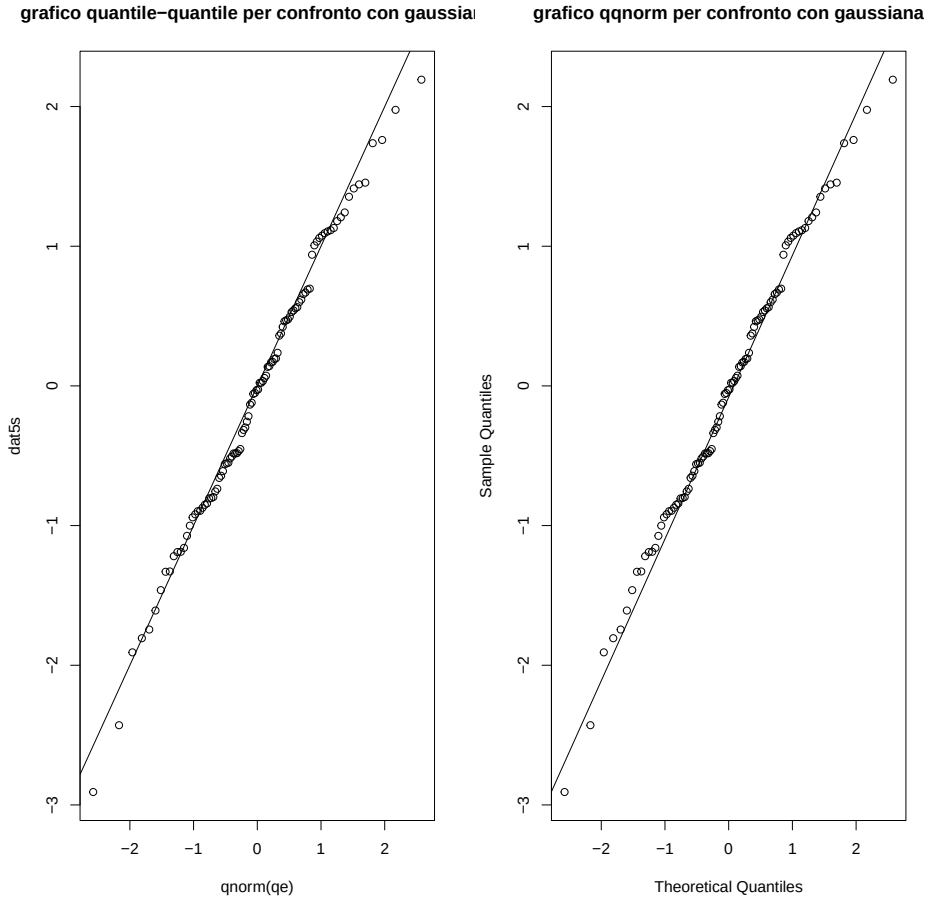
Si può tuttavia pensare a una strategia di confronto dei quantili diversa. Ci

aspettiamo che un quantile empirico sia molto vicino a quello teorico se le distribuzioni si somigliano. Se quindi si confrontano i quantili teorici x_p con i quantili empirici ottenuti per ogni valore $p = \frac{i}{n}$ in corrisponedenza di ogni dato osservato, essi dovrebbero disporsi lungo la retta bisettrice del I e IV quadrante se i dati provengono dal medesimo modello. Posso quindi considerare tutti i valori $p = \frac{i-0.5}{n}$ e a ciascun valore $x_{(i)}$ associare il quantile teorico.

Questa ultima curva rappresenta un importantissimo strumento per valutare l'aderenza fra dati empirici e modello teorico noto col nome di **grafico quantile-quantile** abbreviato spesso in **qqplot**.

Non è quindi sorprendente che esista una funzione per tracciare questo plot. In particolare per il confronto con una gaussiana si può usare la funzione `qqnorm()` (questo è un caso particolare della funzione `qqplot()` che illustreremo più avanti).

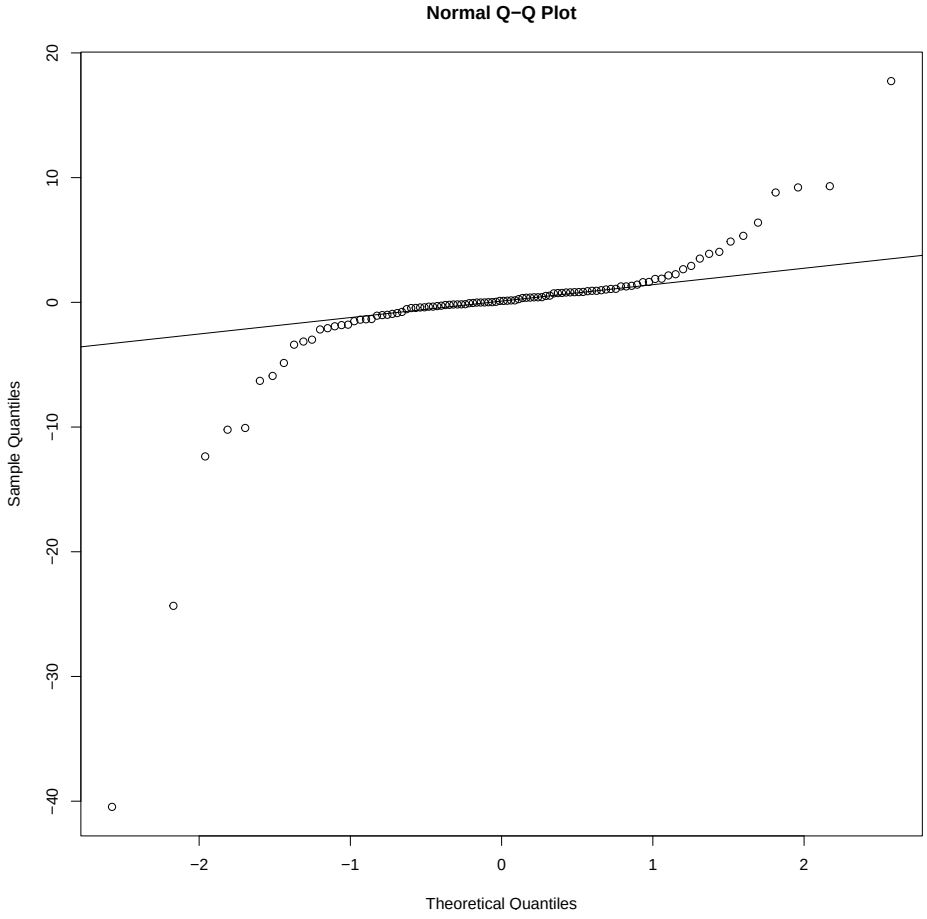
```
par(mfrow=c(1,2))
# prima si ottiene il grafico usando la definizione e la funzione plot()
plot(qnorm(qe), dat5s, main="grafico quantile-quantile per confronto con gaussiana")
abline(0,1)

# Ora si utilizza direttamente la funzione qqnorm()
qqnorm(dat5s, main="grafico qqnorm per confronto con gaussiana")
qqline(dat5s)      # aggiunge la linea per verificare il confronto
```




```
par(mfrow=c(1,1))
```

Proviamo ora a vedere cosa accade se confrontiamo i quantili calcolati per i dati tratti da un'altra distribuzione (la solita t di student con 1 gdl) con la gaussiana standard.

```
set.seed(1111)
dat6<-rt(100,1)
qqnorm(dat6)
qqline(dat6)
```



5.2 Strumenti grafici per il confronto fra due insiemi di dati osservati

I grafici introdotti per le variabili statistiche (in particolare quelle quantitative) e alcuni strumenti illustrati per il confronto tra distribuzione empirica e teorica possono essere utilizzati per confrontare due (o anche più di due) insiemi di dati osservati.

Si tratta in questo caso di affiancare o sovrapporre le rappresentazioni grafiche già introdotte.

5.2.1 Box-plot affiancati

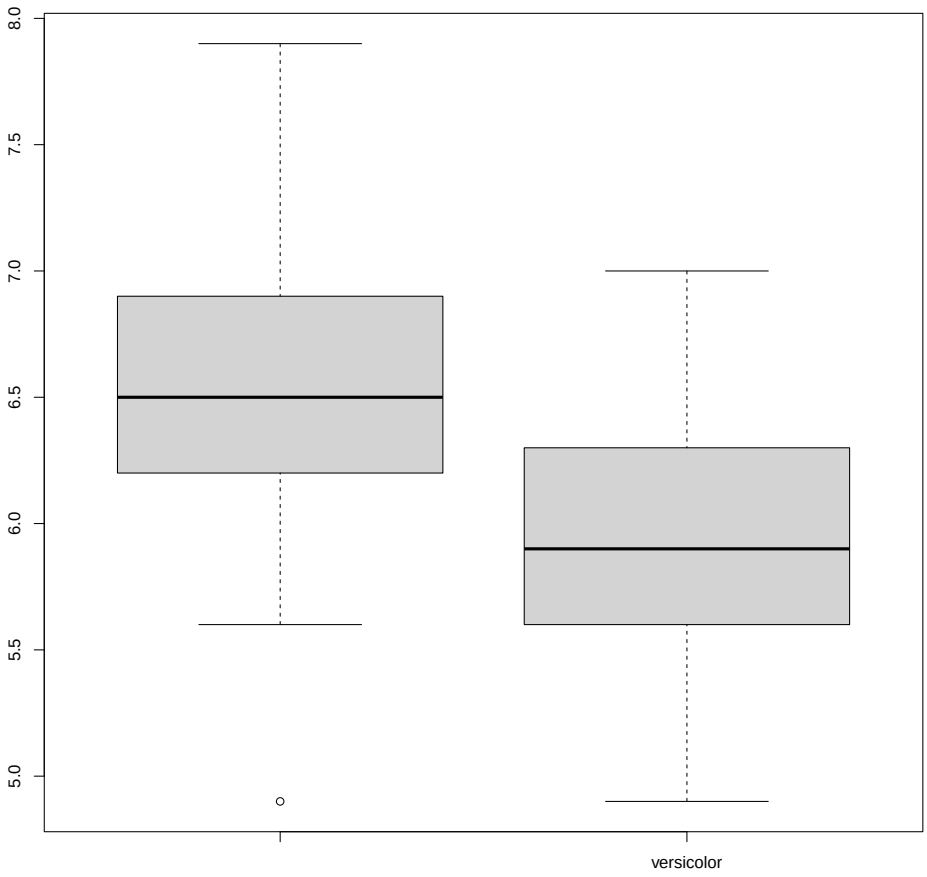
Come già discusso, il diagramma a scatola con baffi è una rappresentazione molto usata per rappresentare sinteticamente la distribuzione di un insieme di dati quantitativi. E si è anche osservato come essa possa essere efficace per

rappresentare anche piccoli insiemi di dati. In questo ultimo caso, ad esempio, usare un istogramma (con il conteggio dei casi entro classi di valori) non è sempre una buona idea.

Se si dispone di due insiemi di dati per i quali si osserva la medesima variabile si possono affiancare (o sovrapporre) i relativi boxplot.

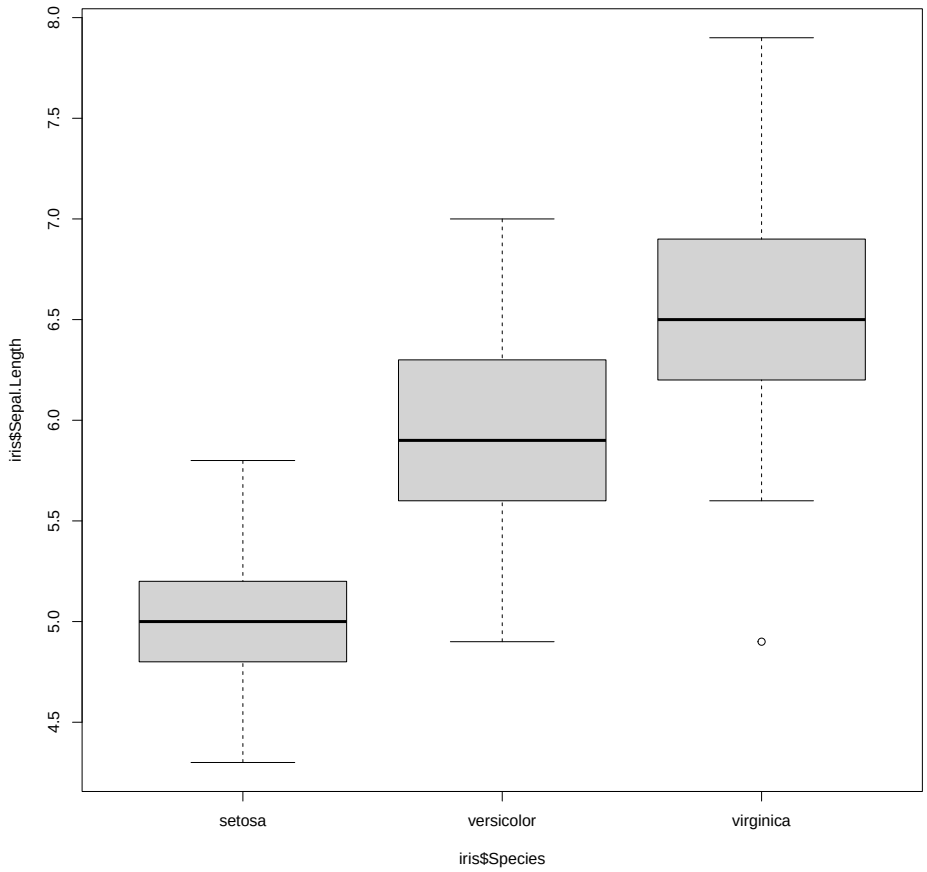
I dati che seguono (piuttosto famosi perchè introdotti da Ronald Fisher nel 1936 e spesso usati per illustrare alcune tecniche di analisi dei dati) riportano insiemi di misure relative a diverse specie di fiori di iris.

```
data(iris)
boxplot(iris$Sepal.Length[iris$Species=="virginica"],
        iris$Sepal.Length[iris$Species=="versicolor"],
        names=(c("","versicolor")))
```



In realtà, per i boxplot è molto semplice ottenere una versione in cui si affiancano le rappresentazioni delle distribuzioni di una variabile quantitativa anche più di due gruppi.

```
# proviamo ora con un boxplot per ogni tipo di iris
#

boxplot(iris$Sepal.Length~iris$Species)
```

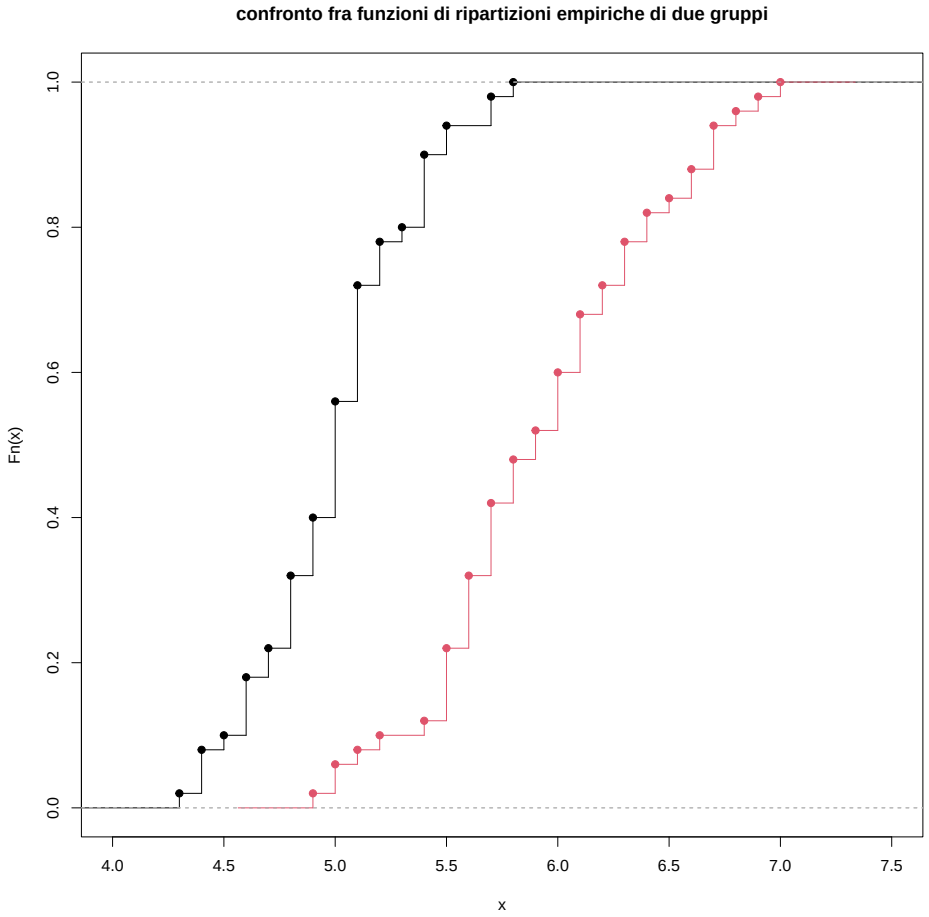



```
# si vedrà che il simbolo "~" è importante per definire quella che chiameremo
# in R "formula", in cui una variabile viene studiata condizionatamente
# a un'altra variabile (in questo caso a un fattore).
# Si noti che si potrebbe pure scrivere

# boxplot(Sepal.Length~Species, data=iris)

# ottenendo il medesimo risultato senza premettere a ciascuna
# variabile il nome del dataframe che le contiene.
```

5.2.2 Confrontare le funzioni di ripartizione empiriche

È facile ottenere per i diversi gruppi da confrontare le rispettive funzioni di ripartizione empiriche e sovrapporle nel grafico. Si vedrà ancora un esempio con i dati `iris`. Anche in questo caso possiamo confrontare le funzioni di ripartizione empirica dei due gruppi.

```
plot(ecdf(iris$Sepal.Length[iris$Species=="setosa"]), verticals=TRUE, xlim=c(4,7.5),
     main="confronto fra funzioni di ripartizioni empiriche di due gruppi")
plot(ecdf(iris$Sepal.Length[iris$Species=="versicolor"]),
     verticals=TRUE, col=2, add=TRUE)
```



Come si vede la funzione di ripartizione empirica ricavata dai dati per la specie versicolor X_v è spostata a destra rispetto a quella della specie setosa X_s (cioè $\hat{F}_v(x)) \leq \hat{F}_s(x_s)$).

Ovviamente si potrebbe aggiungere anche quella di un altro gruppo al grafico con il rischio che però risulti meno chiaro.

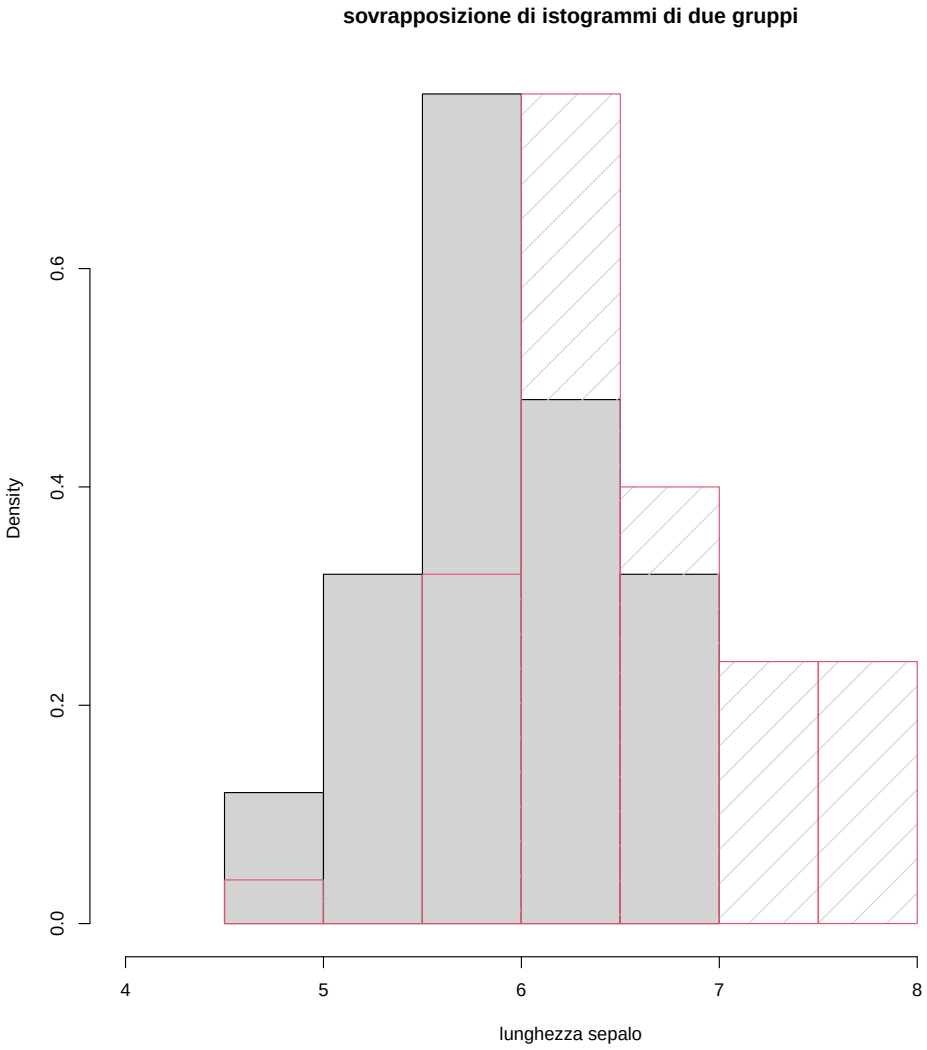
Anche in questo caso si potrebbe usare la statistica di Kolmogorof-Smirnof per misurare la distanza fra le due curve definita, ancora una volta, come il massimo scostamento fra le due curve.

$$D = \underset{x}{sup}|\hat{F}_s(x) - \hat{F}_v(x)|$$

5.2.3 Confronto tra le funzioni di densità empiriche

Si potrebbero anche confrontare le funzioni di densità empirica dei due gruppi. In particolare, si noti notare come sovrapporre istogrammi non conduca a una rappresentazione grafica efficace per mettere in luce le differenze fra i due gruppi.

```
# per brevità di notazione
hist(iris$Sepal.Length[iris$Species=="versicolor"], xlim=c(4,8.5), freq=FALSE,
     , xlab="lunghezza sepalo", main="sovrapposizione di istogrammi di due gruppi")
hist(iris$Sepal.Length[iris$Species=="virginica"], border=2, freq=FALSE,
     density=4.5, add=TRUE)
```

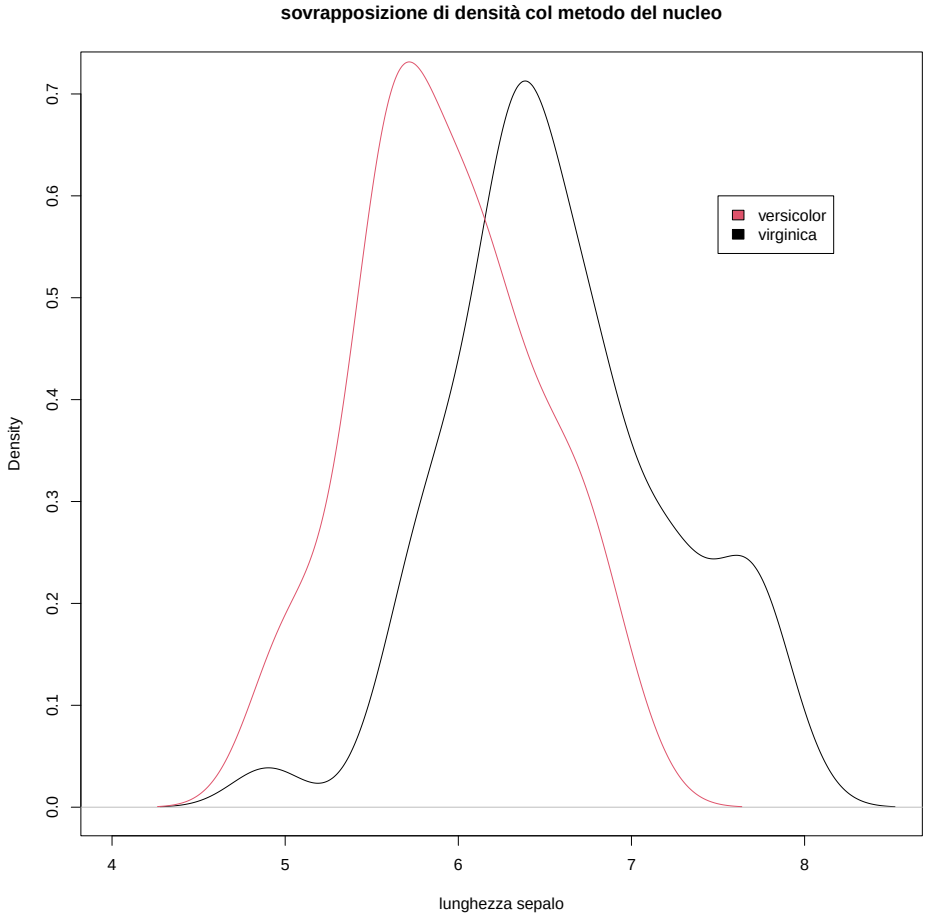


L'istogramma ricavato dai dati per la specie virginica X_v è spostata a destra rispetto a quella della specie versicolor come si vedeva benissimo con i box plot. Non è però molto efficace la rappresentazione grafica

Più chiaro risulta il grafico che sovrappone le curve di densità con il metodo del nucleo.

```
# per brevità di notazione si ottengano dei vettori con i
# dati sulle lunghezze dei sepali per i tre gruppi
virgi<-iris$Sepal.Length[iris$Species=="virginica"]
versi<-iris$Sepal.Length[iris$Species=="versicolor"]
setosa<-iris$Sepal.Length[iris$Species=="setosa"]
plot(density(virgi), xlim=c(4,8.5), xlab="lunghezza sepalo",
```

```
     main="sovrapposizione di densità col metodo del nucleo")
lines(density(versi), col=2)
legend(x=7.5,y=0.6,legend=c("versicolor","virginica"), fill=c(2,1))
```

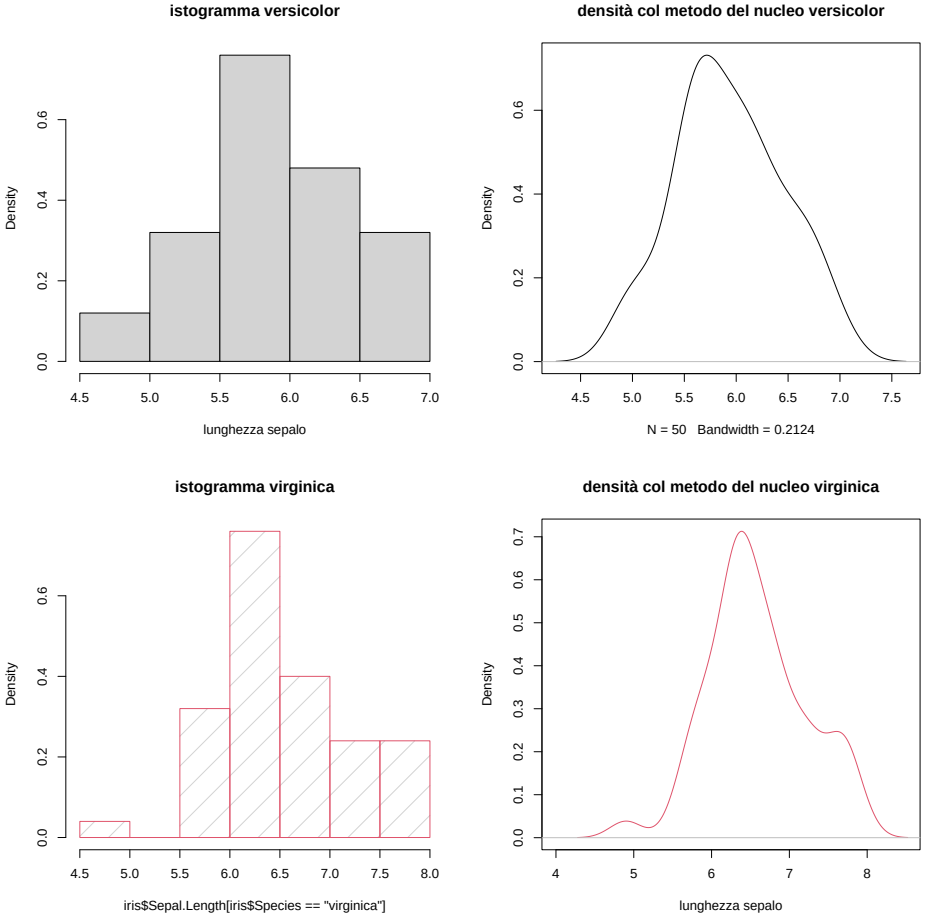


Come si vede, si tratta di un grafico più efficace dell'istogramma per fare confronti. Si ricordi che tali rappresentazioni potrebbero rivelarsi più affidabili con gruppi che contengono un numero di dati sufficientemente elevato (diciamo>100).

Potrebbe esser più opportuno affiancare o sovrapporre le figure. Tuttavia in tal caso gestire grafici multipli con la semplice suddivisione della finestra non è la cosa più opportuna. Vediamo un esempio:

```
par(mfrow=c(2,2))
hist(iris$Sepal.Length[iris$Species=="versicolor"], freq=FALSE,
      xlab="lunghezza sepalo", main="istogramma versicolor")
plot(density(versi),  main="densità col metodo del nucleo versicolor")
```

```
hist(iris$Sepal.Length[iris$Species=="virginica"], main="istogramma virginica",
     border=2, freq=FALSE, density=4.5)
plot(density(virgi), xlim=c(4,8.5), col=2,
      xlab="lunghezza sepalo", main="densità col metodo del nucleo virginica")
```

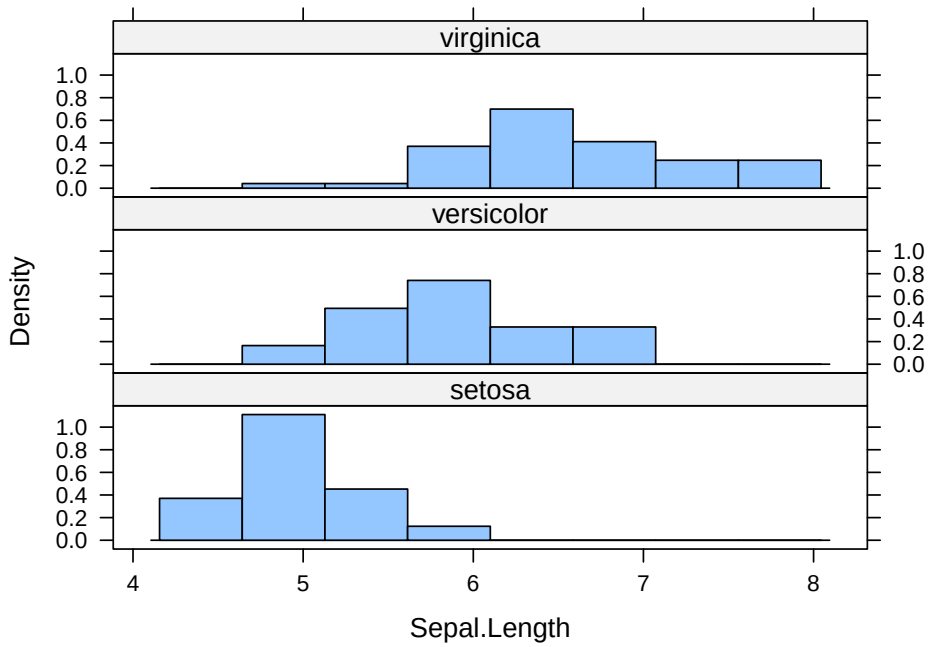



```
par(mfrow=c(1,1))
```

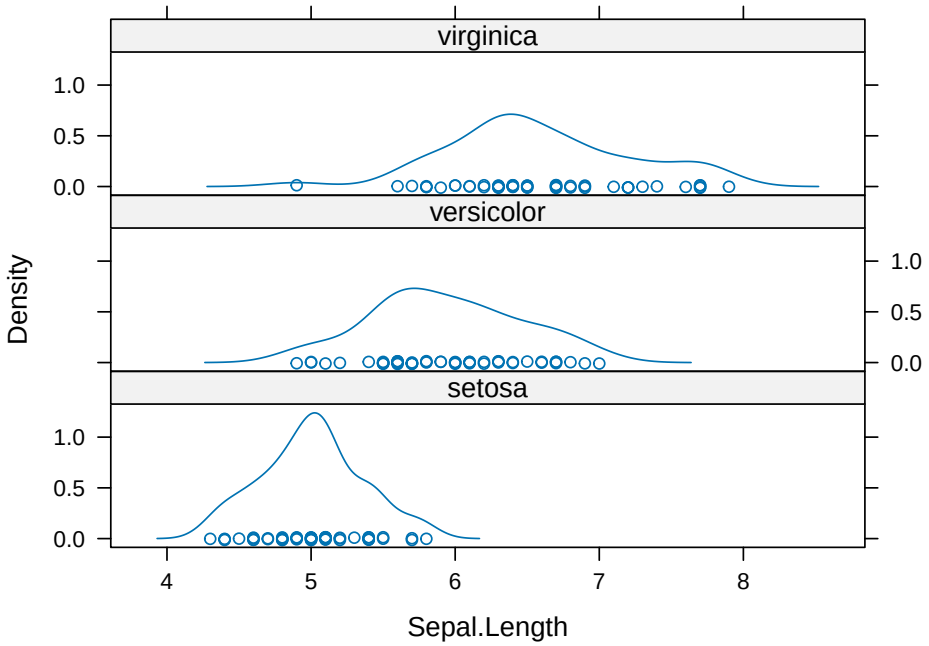
Come si vede il confronto è difficile perchè occorrerebbe adeguare le scale sull'asse orizzontale. Per fare grafici a più pannelli esistono pacchetti `R` appositi. Più aventi sarà introdotto, con qualche dettaglio, il più famoso e ampio pacchetto, `ggplot2`, che fornisce strumenti per ottenere visualizzazioni dei dati molto efficaci.

Ora si introduce, un altro il pacchetto, `lattice` che contiene varie funzioni che servono a realizzare facilmente grafici multipli (oggetti di tipo `trellis`). Di seguito un l'esempio con l'istogramma e la densità empirica.

```r
# ci si assicuri di avere scaricato il package lattice
library(lattice)

histogram(~Sepal.Length | Species, data=iris, type="density", layout=c(1,3))
```




```r
densityplot(~Sepal.Length | Species, data=iris, layout=c(1,3))
```



il package `lattice` contiene molti altri tipi di grafici che possono essere poi organizzati e gestiti in modo opportuno. Si rinvia all'help del pacchetto per dettagli.

5.2.4 Il Grafico quantile-quantile

Il confronto dei quantili empirici per corrispondenti valori di p è uno strumento utile a confrontare due distribuzioni. Se le due distribuzioni sono simili allora i corrispondenti quantili di ordine p dovrebbero essere gli stessi.

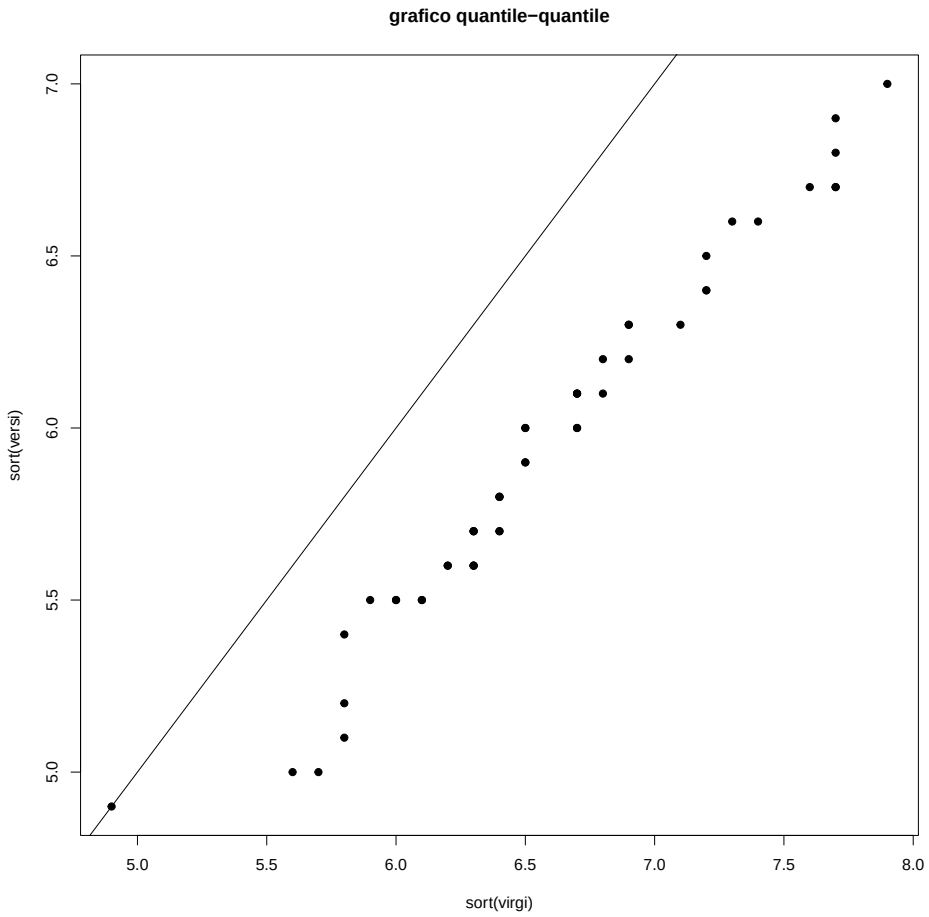
Si può quindi usare il medesimo grafico visto per confrontare quantili empirici e teorici, con la differenza che in questo caso avremo quantili empirici ottenuti per due diversi insiemi di dati.

Si noti che nel caso i due insiemi di dati che si confrontano sono della stessa numerosità n si tratta semplicemete di un grafico in cui si rappresentano i punti di coordinate corrispondenti ai due vettori ordinati.

Per cui il dato più piccolo in entrambi gli insiemi rappresenta un quantile di pari ordine cioè pari a $1/n$, il secondo valore ordinato saraà il quantile di ordine $2/n$ per i due gruppi, e così via.

```
# Si considerino i dati sulla lunghezza dei sepali per i tipo virginica e versicolor

length(virgi)
length(versi)
```

```
# hanno la stessa lunghezza. Il dato più piccolo in entrambi gli insiemi
# rappresenta un quantile di pari ordine cioè pari a 1/50, il secondo valore
# ordinato 2/50 etc.

plot(sort(virgi),sort(versi), pch=19,
     main="grafico quantile-quantile")

abline(0,1)
```




```
## [1] 50
## [1] 50
```

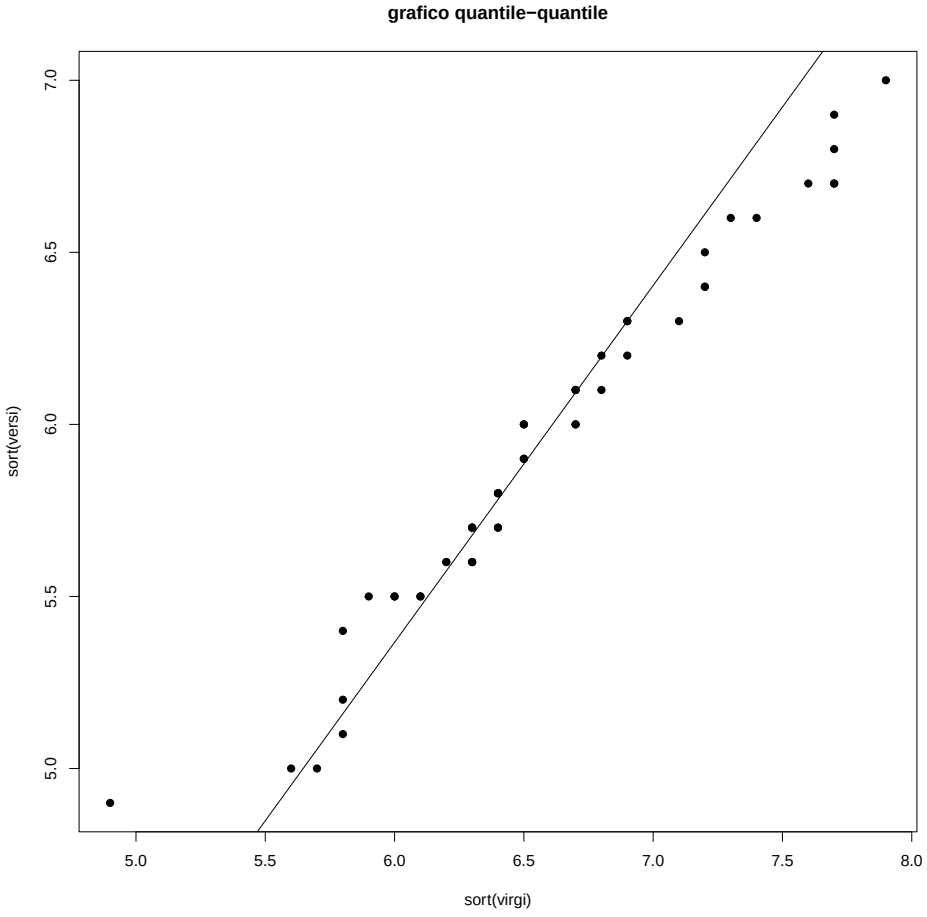
Se le due distribuzioni fossero uguali dovrebbere esssere disposti lungo la bisettrice. Si vede che non lo sono e in effetti i sepali della specie virginica sono più lunghi.

Tuttavia i dati sembrano disposti lungo una retta che è quasi parallela alla

bisettrice. Questo indica che i dati sui sepali della specie virginica pur essendo spostati a destra hanno una distribuzione simile.

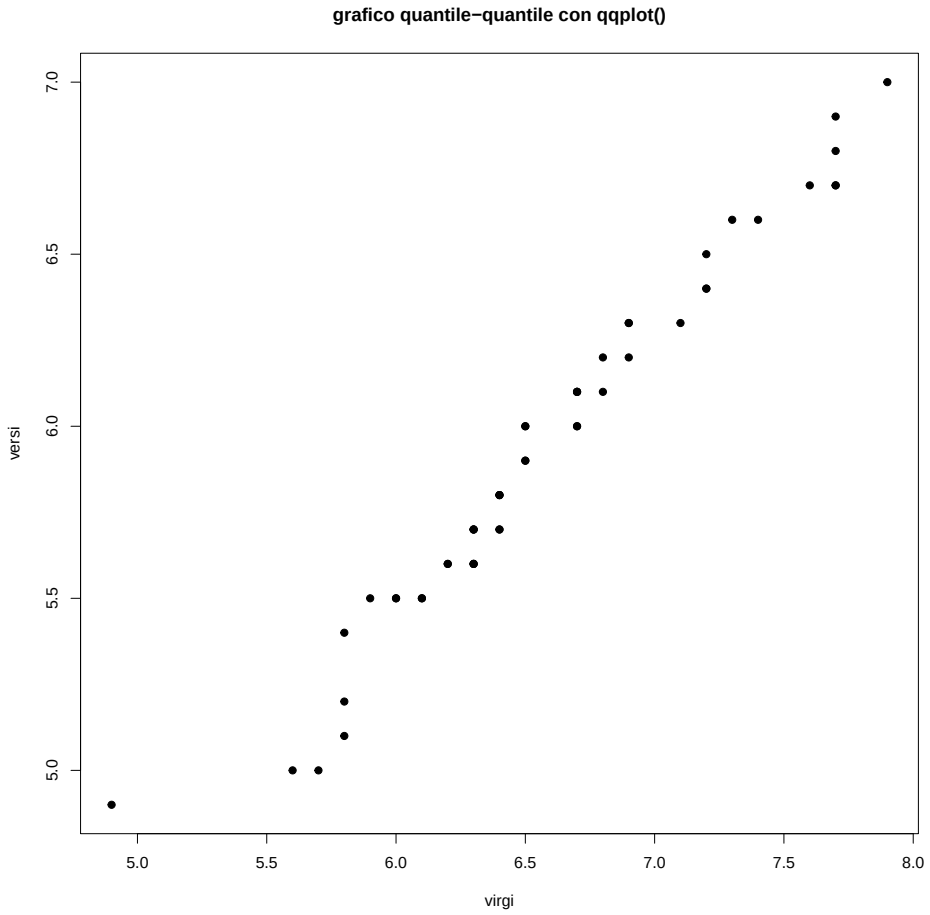
Si potrebbe considerare invece della bisettrice una linea che passi fra i rispettivi quartili.

```
qvirgi<-quantile(virgi,probs=c(.25,.75))
qversi<-quantile(versi,probs=c(.25,.75))
s<-(qversi[2]-qversi[1])/(qvirgi[2]-qvirgi[1])
int<-qversi[1]-s*qvirgi[1]
plot(sort(virgi),sort(versi), pch=19,
     main="grafico quantile-quantile")
abline(int,s)
```



Tale linea di riferimento permette di cogliere lo scostamento che esiste fra i due gruppi. Ovviamente quello che si è fatto è quanto si otterrebbe con l'uso della funzione `qqplot()`.

```
qqplot(virgi,versi, pch=19,
     main="grafico quantile-quantile con qqplot()")
```

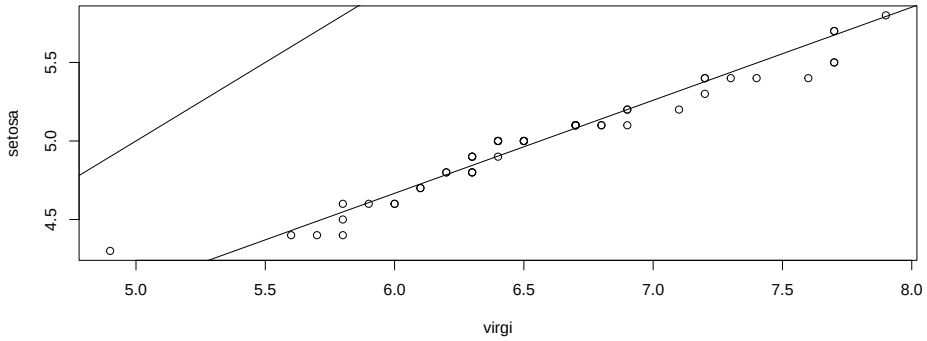


In sostanza:

- se i punti del qqplot si dispongono su una linea retta le due distribuzioni sono simili come forma.
  1. se tale retta è parallela alla bisettrice allora le due distribuzioni differiscono in media (ma hanno la stessa forma);
  2. se tale retta non è parallela allora hanno stessa forma ma diverse media e varianza;
- se i punti **non** si dispongono lungo una retta allora i due insiemi differiscono anche come forma.

```
setosa<-iris$Sepal.Length[iris$Species=="setosa"]
qqplot(virgi, setosa )
abline(0,1)
qsetosa<-quantile(setosa,probs=c(.25,.75))
```

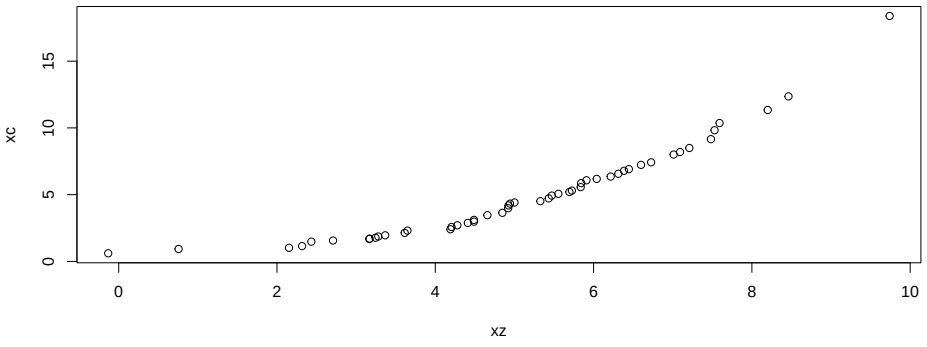
```
s<-(qsetosa[2]-qsetosa[1])/(qvirgi[2]-qvirgi[1])
int<-qsetosa[1]-s*qvirgi[1]
abline(int,s)
```




```
# in questo caso i punti sono lungo una linea ma non parallele alla
# bisettrice. Le distribuzioni sono simili ma con diversa media e varianza
```

Ovviamente il qq-plot può essere ottenuto anche in presenza di gruppi di numerosità diversa. In questo caso vengono calcolati i quantili corrispondenti a una sequenza di valori di p (che potrebbe corrispondere a quelli dell'insieme meno numeroso)

```
xz<-rnorm(50, 5, 2)
xc<-rchisq(100, 5)
qqplot(xz,xc)
```




```
# In questo caso le due distribuzioni sono marcatamente diverse
```

5.3 Analisi statistica bivariata

L'analisi congiunta di una coppia di variabili, due colonne di un data frame, permette di arricchire enormemente l'analisi dei dati. Agli aspetti già sottolineati con l'analisi di una singola variabile (ovvero sintetizzare sia numericamente che

graficamente l'informazione raccolta su un insieme di unità statistiche), se ne aggiungono di nuovi e di maggiore interesse.

In particolare, quello che si vorrebbe valutare è la presenza di relazioni fra le due variabili. Le due variabili possono infatti essere associate nel senso che alcune specifiche coppie di valori (o di modalità) delle due variabili tendono a presentarsi con maggiore frequenza e regolarità. Ad esempio, a valori elevati di una prima variabile quantitativa potrei trovare spesso associati valori elevati di un'altra variabile quantitativa. Oppure, potrei osservare, per i dati di cui si dispone, che la modalità di una variabile qualitativa è spesso associata a una specifica modalità di una seconda variabile qualitativa. Questi aspetti possono emergere solo dall'analisi congiunta di due variabili. Quando questo accade si dirà che le due variabili presentano una qualche forma di associazione.

La presenza di una **relazione** (associazione) fra le due variabili è potenzialmente di grande interesse perchè la conoscenza di una delle due variabili consente di descrivere o sientetizzare più accuratamente la seconda variabile fino al punto di poter prevedere con minore incertezza i valori (o le modalità) che essa può assumere. L'assenza di relazione (di associazione) fra le variabili, viceversa, è parallelo al concetto di **indipendenza**, già incontrato nel calcolo delle probabilità, e implica che l'analisi congiunta di due variabili non aggiunge nulla di nuovo e interessante a quello che si era imparato dall'analisi delle variabili prese singolarmente.

In realtà il tema della analisi congiunta di due variabili è stato già introdotto nelle sezioni precedenti quando si è mostrato come si possa ottenere una **tabella a doppia entrata** analizzando congiuntamente due variabili categoriali (fattori, quindi, o trasformate in fattori attraverso una aggregazione in classi). Ed è stato di fatto introdotto anche quando si è parlato del confronto della stessa variabile quantitativa in più gruppi (ad esempio, i boxplot multipli).

Al fine di considerare gli strumenti per l'analisi di due variabili converrà suddividere il tema a seconda della natura delle variabili. Seppure alcuni concetti generali sono comuni, gli strumenti di analisi e di rappresentazione grafica potranno essere diversi e, come nel caso di una singola variabile, più complessi quando sono coinvolte variabili quantitative.

5.3.1 Due variabili categoriali

Come già illustrato nella prima parte, la sintesi più elementare dei dati per una coppia di variabili è la tabella a doppia entrata.

Riprendiamo l'esempio già visto per i dati `AutoBi`.

```
library(insuranceData)
data("AutoBi")
AutoBi$MARITAL<- factor(AutoBi$MARITAL)
levels(AutoBi$MARITAL)<-c("married", "single", "previouslymarried", "previouslymarried
AutoBi$LOSSclass<-cut(AutoBi$LOSS,breaks=c(0,0.5,2,4,8,1100))
```

```
AutoBi$ATTORNEY<- factor(AutoBi$ATTORNEY)
levels(AutoBi$ATTORNEY) = c("yes", "no")
AutoBi$CLMSEX<-factor(AutoBi$CLMSEX)
levels(AutoBi$CLMSEX) <- c("M", "F")
tab1 <- table(AutoBi$ATTORNEY,AutoBi$CLMSEX)
tab1
```

```
##
##          M   F
##   yes 325 352
##   no  261 390
```

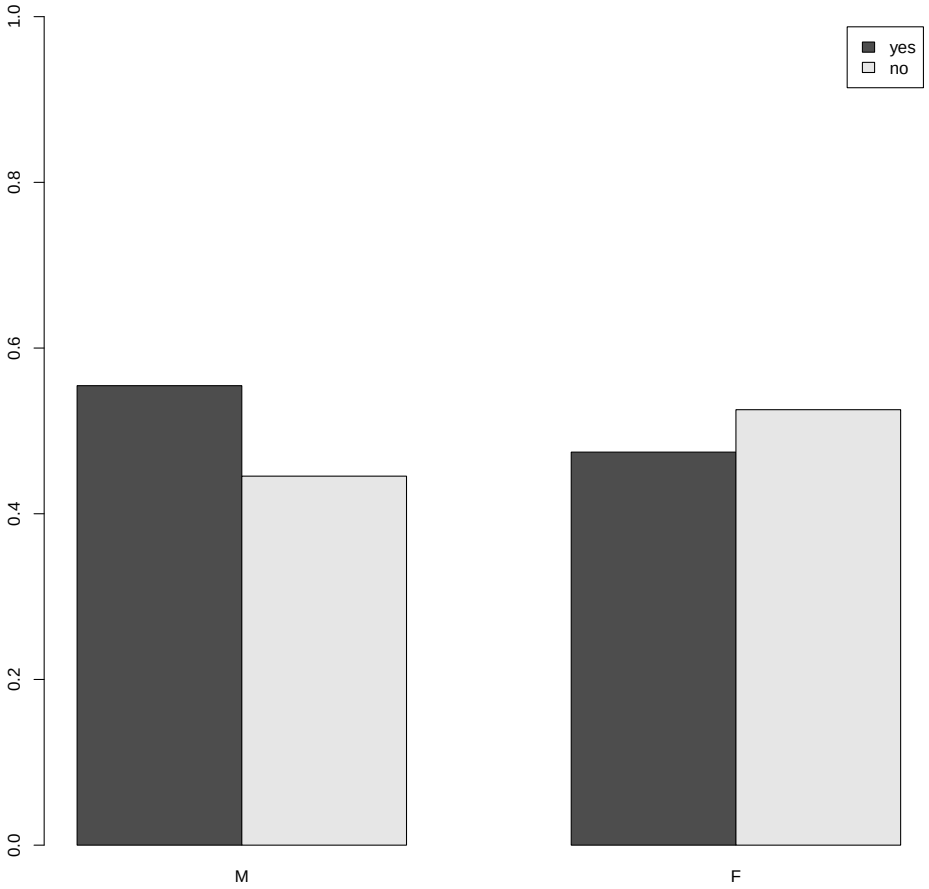
```
rtab<-prop.table(tab1, 1)
ctab<-prop.table(tab1, 2)
rtab
```

```
##
##                M         F
##   yes 0.4800591 0.5199409
##   no  0.4009217 0.5990783
```

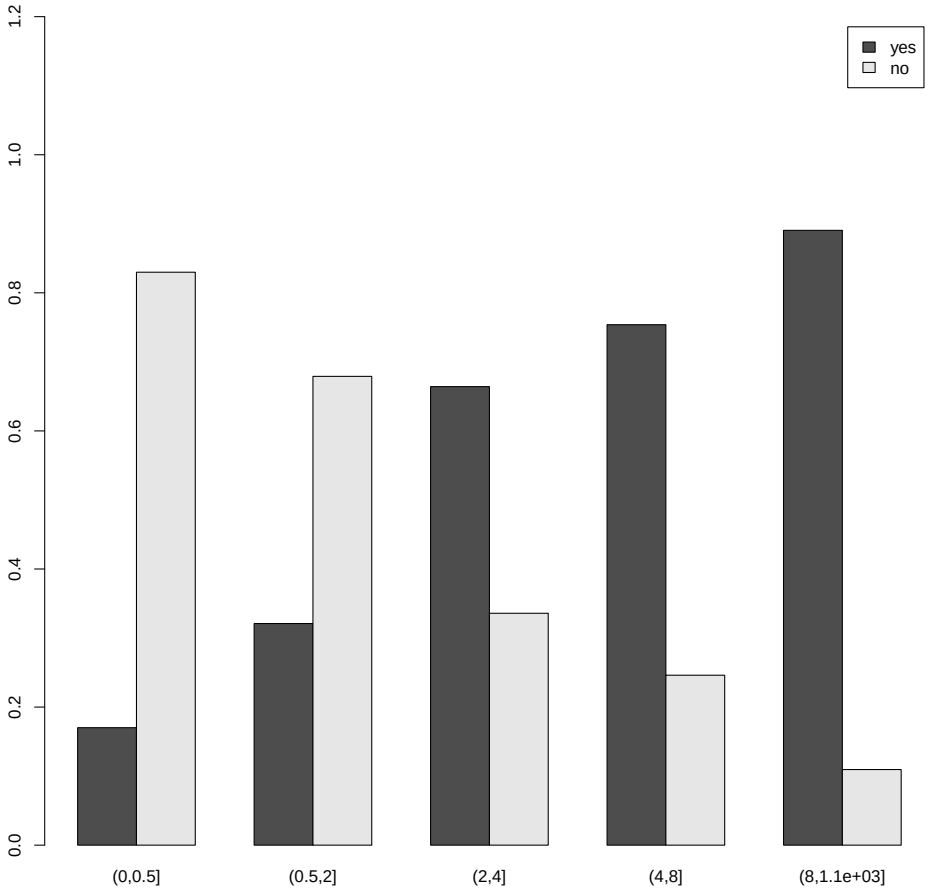
```
ctab
```

```
##
##                M         F
##   yes 0.5546075 0.4743935
##   no  0.4453925 0.5256065
```

```
barplot(ctab,legend=T,beside=TRUE, ylim=c(0,1))
```




```
# In questo caso vengono affiancate le due distribuzioni del ricorso
# all'avvocato condizionatamente al sesso

barplot(prop.table(table(AutoBi$ATTORNEY,AutoBi$LOSSclass),2), beside=T,
        legend=T, ylim=c(0,1.2))
```




```
# In questo caso vengono affiancate le due distribuzioni del ricorso
# all'avvocato condizionatamente al danno (trattato in classi come fattore)
```

Come si è già detto, la lettura di una tabella a doppia entrata va spesso fatta avendo in mente il ruolo delle due variabili. In effetti, è utile assegnare, nell'analisi che si svolge, a una delle due variabili il ruolo di variabile **risposta** (o dipendente): si tratta della variabile di principale interesse. L'altra variabile ha il ruolo di fattore esplicativo ed è detto variabile **indipendente** o covariata. Di solito questo ruolo è chiaro dal contesto dello studio. In questo caso è utile, ai fini interpretativi, rappresentare la distribuzione della variabile *risposta* condizionatamente alla variabile esplicativa.

In questo esempio, potrebbe essere interessante capire se vi sono delle tipicità nel ricorso all'avvocato: accade più per gli uomini o è più tipico delle donne? Quindi la variabile ATTORNEY è quella dipendente e verifichiamo se essa cambia condizionatamente al valore assunto dall'altra variabile. Si tratta quindi di confrontare semplici distribuzioni di una variabile dicotomica (Attorney=si, Attorney=no) per i diversi sottoinsiemi identificati dalla modalità di un'altra

variabile. Questo è quello che si è fatto con i grafici precedenti dove si confrontavano le distribuzioni del ricorso all'avvocato condizionate all'essere maschio o femmina oppure ad avere avuto una determinata classe di danno.

Se le distribuzioni condizionate sono diverse allora conoscere la variabile indipendente ci consente di *prevedere* più accuratamente la variabile risposta.

5.3.1.1 La statistica X^2

Il principio fondamentale da tenere a mente è:

Se le distribuzioni condizionate di una variabile rispetto all'altra sono uguali allora le due variabili sono indipendenti.

Se vi è indipendenza far due variabili non è utile condurre un'analisi congiunta delle stesse poichè tutta l'informazione è contenuta nella distribuzioni marginali.

Il concetto è evidentemente mutuato dal concetto di indipendenza del calcolo delle probabilità e in particolare dalla definizione di indipendenza tra due variabili aleatorie.

Ovviamente, nella pratica sarà piuttosto difficile che si presenti una situazione di perfetta uguaglianza delle distribuzioni condizionate. Per cui risulta utile ottenere delle statistiche che misurino la distanza fra la situazione ideale di indipendenza e quella osservata.

La più nota misura è la statistica X^2. Essa si propone di valutare la distanza fra la tabella a doppia entrata effettivamente osservata e con quello che ci si sarebbe atteso di osservare se vi fosse indipendenza.

Una *tabella a doppia entrata* per una coppia (X, Y) di variabili le cui modalità sono indicate per riga (la X) e per colonna (la Y), ha la seguente forma generale

X / Y	y_1	y_2	...	y_k	...	y_K	*Tot.*
x_1	n_{11}	n_{12}	...	n_{1k}	...	n_{1K}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2k}	...	n_{2K}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	...	
x_h	n_{h1}	n_{h2}	...	n_{hk}	...	n_{hK}	$n_{h.}$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	...	
x_H	n_{H1}	n_{H2}	...	n_{Hk}	...	n_{HK}	$n_{H.}$
Tot.	$n_{.1}$	$n_{.2}$	...	$n_{.k}$	...	$n_{.K}$	n

Le celle, definite dall'incrocio delle righe con le colonne, contengono il numero di casi (n_{hk}) che presentano le corrispondenti modalità delle due variabili. I valori n_{hk} sono le frequenze congiunte delle coppie (x_h, y_k) al variare di $h(= 1, \dots, H)$ e $k(= 1, \dots, K)$.

Le somme per riga, $n_{h.} = \sum_{k=1}^{K} n_{hk}$, e per colonna, $n_{.k} = \sum_{h=1}^{H} n_{hk}$, sono dette anche frequenze marginali di x_h (di riga) per $h = 1, \dots, H$ e, rispettivamente, y_k

(di colonna) per $k = 1, \ldots, K$; infine, la numerosità totale è denotata mediante $n = \sum_{h,k} n_{hk}$.

Se vi fosse indipendenza le distribuzioni marginali (di riga) dovrebbero essere tutte uguali alle distribuzioni condizionate della variabile di riga (ovviamente lo stesso varrebbe per la marginale di colonna). Per cui se vi fosse indipendenza, data una generica cella h, k, deve essere $\forall h, k$

$$\frac{n_{hk}}{n_{h.}} = \frac{n_{.k}}{n}$$

quindi la frequenza $\hat{n}_{hk}$ in caso di indipendenza sarebbe pari a

$$\hat{n}_{hk} = \frac{n_{h.} n_{.k}}{n}$$

.

La statistica X^2 si basa sul confronto fra questi valori e quelli effettivamente osservati. Ovviamente più piccola è tale statistica maggiore è l'indizio che vi sia indipendenza e che le discrepanze osservate debbano attribuirsi solo a scostamenti dovuti al caso.

$$X^2 = \sum_{\forall \text{cella}} \frac{(n_{hk} - \hat{n}_{hk})^2}{\hat{n}_{hk}}$$

Una valutazione di quanto tale discrepanza sia da ritenersi poco rilevante, e compatibile con fluttuazioni casuali, può essere condotta nell'ambito della statistica inferenziale dove si formula e si valuta rigorosamente una ipotesi di indipendenza fra le due variabili.

In attesa di affrontare tale argomento in modo formale in corsi successivi, si introduce qui, in modo informale, un criterio che consente di giudicare quanto le differenze fra quanto si osserva e quanto ci si aspetta di osservare in caso di indipenedenza siano da ritenersi solo effetto di fluttuazioni casuali:

se effettivamente le due variabili fossero indipendenti, allora il valore della statistica X^2 dovrebbe esser assimilabile a un valore tratto da un particolare tipo di variabile aleatoria nota come variabile χ^2_g. Essa ha una distribuzione che dipende dal parametro g che in questo caso è $g = (H-1)(K-1)$.

Con R si può valutare quanto il valore osservato nella tabella, X^2_{oss}, sia una valore che è plausibile provenga da qualla distribuzione di probabilità. A tal fine si calcola la probabilità di ottenere valori più grandi di X^2_{oss}. Questa probabilità, che è detta p-value, può essere usato per valutare quanto i dati sono conformi all'ipotesi di indipendenza: se essa è molto piccola (diciamo molto sotto a 0.01) allora non vi sono elementi nei dati a sostegno della indipendenza fra le due variabili.

Ad esempio:

```
                    # riconsideriamo la tabella per sesso e ricorso
tab1                          # all'avvocato
tot<-marginSums(tab1)
mr<- margin.table(tab1,1) # Marginale di riga (variabile avvocato)
mc<- margin.table(tab1,2) # Marginale di colonna (variabile sesso)
hatn<- mr%*%t(mc)/tot     # tabella delle frequenze teoriche di indipenedenza
hatn
X2<-sum((tab1-hatn)^2/hatn) # calcolo X^2
X2
pchisq(X2,1, lower.tail = FALSE) # calcolo la probabilità di ottenere un valore maggio
                              # quello ottenuto per X^2 in un chi-quadrato con 1 gdl
```

```
##
##         M   F
##   yes 325 352
##   no  261 390
##
##              M        F
##   yes 298.7364 378.2636
##   no  287.2636 363.7364
## [1] 8.430048
## [1] 0.003690706
```

```
# Esiste la funzione R che svolge questi calcoli

chisq.test(tab1, correct=FALSE)

# Come si vede  fornisce gli stessi risultati. Il parametro correct=FALSE
# evita che venga applicata una correzione al calcolo della differenze fra valori
# osservati e valori di indipendenza nota come correzione per continuità
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab1
## X-squared = 8.43, df = 1, p-value = 0.003691
```

```
# si provi anche con la tabella ricorso all'avvocata e danno accertato (in classi)

tab2<-table(AutoBi$ATTORNEY,AutoBi$LOSSclass)
tab2
chisq.test(tab2)
```

```
##
##       (0,0.5] (0.5,2] (2,4] (4,8] (8,1.1e+03]
##   yes      49     104   263   147          122
##   no      239     220   133    48           15
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab2
## X-squared = 342.89, df = 4, p-value < 2.2e-16
```

La probabilità, che denominata p-value, nei due esempi mostrati è abbastanza piccola, e in particolare nel caso della relazione fra ricorso all'avvocato e danno esso è così piccola da poter sostenere che se vi fosse indipendenza è del tutto implausibile osservare quei dati e pertanto essi non danno alcun sostegno a tale ipotesi.

Anche nel caso del sesso e ricorso all'avvocato, il $p - value$ è sufficientemente piccolo da poter concludere che i dati non danno molto sostegno all'ipotesi di indipendenza. Tuttavia, non è estremamente piccolo come nel caso precedente: la associazione fra le due variabile è presente ma è meno marcata che nel caso precedente. Questo è in accordo con quanto si osserva nei grafici.

5.3.1.2 Le tabelle 2×2 e l'odds ratio

Gli **odds-ratio** forniscono una misura semplice e immediatamente comprensibile riguardo al presenza di associazione in una tabella di frequenze congiunte quando le due variabili sono dicotomiche (tabelle 2×2).

Consideriamo due variabili dicotomiche: $Y = \{occupato, disoccupato\}$ e $X = \{laureato, nonlaureato\}$ osservate per un gruppo di 1000 persone. La tabella delle frequenze congiunte sia

Y	occupato	non occupato
X		
laureato	210	105
non laureato	353	332

In termini di odds-ratio (rapporto fra gli odds) la tabella si può sintetizzare con il rapporto tra laureati e non laureati tra gli occupati $210/353 = 0.59$ e laureati e non laureati tra i non occupati $105/332 = 0.31$, il cui risultato è 1.88. Gli odds sono espressioni simili ai rapporti di scommessa (del tipo il cavallo *fulmine* è dato 1 a 3) per cui si potrebbe dire, in questo caso, che il rapporto fra laureati e non laureati (tra i non occupati) e circa di 1 a 3 (1 laureato ogni 3 non laureati).

Si noti che il valore dell'odds-ratio si ottiene anche facendo il rapporto fra il prodotto dei valori che nella tabella sono sulla diagonale principale e il prodotto dei valori sulla diagonale secondaria $(210 \times 332)/(105 \times 353)$. Esso è infatti anche detto **rapporto dei prodotti incrociati**.

Un rapporto vicino ad 1 indica una scarsa associazione tra le variabili, (infatti se il rapporto è esattamente 1 ciò implica che le distribuzioni condizionate sono

uguali e vi è quindi indipendenza fra i caratteri in gioco). Quanto più il valore è distante da 1 tanto maggiore è l'associazione fra le due variabili. Un risultato maggiore di 1 come nel nostro caso indica una associazione positiva tra la modalità occupato e quella laureato (ovvero la frequenza osservata per questa coppia di modalità è maggiore della frequenza che si osserverebbe se vi fosse indipendenza).

Chiaramente, il reciproco di 1.88 (pari circa a 0.53) misura un'associazione che è esattamente della stessa forza (tale valore si otterrebbe difatti se si permutassero le righe della medesima tabella). L'allontanamento da 1 (cioè dall'indipendenza) dell'odds ratio va letto quindi in maniera diversa se il valore è maggiore o minore di 1 poiché evidentemento esso non può assumere valori negativi. In sostanza in una tabella ove risulti un odds ratio pari a 5 (che indica una forte associazione) è presente un'associazione del tutto equivalente ad una tabella per cui risulti un odds ratio pari a 0.2. Si noti infine che è possibile definire correttamente il rapporto solo se tutte le caselle della tabella hanno frequenze superiori a 0.

Per rendere più leggibili queste quantità spesso conviene passare ai logaritmi: il **log-odd-ratio** è quindi 0 se le variabili sono indipendenti, ed assume comunque lo stesso valore assoluto comunque si permutino righe e colonne della stessa tabella.

Consideriamo ancora i dati già analizzati su Ricorso all'avvocato e Genere, e calcoliamo odds-ratio e log-odds-ratio.

```
tab1
```

```
##
##         M   F
##   yes 325 352
##   no  261 390
```

```
# calcoliamo direttamente il rapporto dei prodotti incrociati

OR<-(tab1[1,1]*tab1[2,2])/(tab1[1,2]*tab1[2,1])

LOR<- log(OR)
```

Si noti come il valore di OR e di LOR sono abbastanza diversi, rispettivamente, da 1 e da 0, e forsniscono indicazione della assenza di indipendenza.

5.3.2 Una variabile quantitativa condizionata a un fattore

Come si è già detto, è spesso opportuno valutare il ruolo delle variabili coinvolte e chiedersi se, nello studio congiunto, si ritiene che si possa distinguere una delle due variabili come variabile risposta da analizzare condizionatamente a una seconda. Nel caso in questione si supporrà che la variabile risposta sia quella quantitativa. La seconda variabile, categoriale, è quindi la variabile di condizionamento ed essendo un fattore qualitativo esso identifica dei gruppi in

corrispondenza delle diverse modalità del fattore. In questo caso si tratta quindi di studiare le distribuzioni della variabile risposta in ciascuno di questi gruppi e verificare se differiscano e quanto e come differiscano.

Resta fermo quanto già osservato, ovvero:

se le distribuzioni della variabile quantitativa per ciascun gruppo definito dal fattore (le distribuzioni condizionate a ciascuna modalità del fattore) sono simili allora l'analisi bivariata non fornisce elementi di interesse.

Se le due variabili coinvolte fossero indipendenti allora ci aspetteremmo che le distribuzioni condizionate siano uguali.

Si denoti con Y la variabile risposta quantitativa e con X il fattore qualitativo, con G modalità, che definisce i gruppi di valori: si vogliono analizzare la variabili $Y|X = x_g$ con $g = 1, 2, ..., G$.

Poichè Y è quantitativa noi abbiamo molti modi per studiare somiglianze e differenze fra le distribuzioni condizionate.

1. Si potrebbe ritenere che le distribuzioni sono tutte uguali tra loro come forma e che si differenzino solo per quanto riguarda la tendenza centrale (misurata, ad esempio, dalla media o dalla mediana).
2. Le distribuzioni potrebbero avere la stessa forma (ad esempio, simmetrica e con curtosi simile a quella della gaussiana) ma sono diverse sia la tendenza centrale che la variabilità.
3. Le due distribuzioni potrebbero essere diverse tra loro in modo più radicale (in un gruppo abbiamo distribuzioni simmetriche in un altro sono asimmetriche).

Alcune delle rappresentazioni grafiche già introdotte sono idonee a valutare in una analisi esplorativa quale delle situazioni sopra elencate possa essere più plausibile. Nei diversi casi si può ricorrere a strumenti diversi per la sintesi dei dati.

Se in particolare il fattore ha solo due modalità (e definisce quindi solo due gruppi di dati per la variabile Y) strumenti grafici come il confronto della funzione di ripartizione empirica dei due gruppi o il grafico quantile-quantile possano rivelarsi utili.

Se X ha più di due modalità e vi sono più gruppi risulta utile l'analisi grafica con box-plot multiplo o il confronto degli istogrammi (o delle funzioni di densità empirica).

Resta da aggiungere, infine, la possibilità di ricorrere a strumenti analitici per rappresentare le diverse situazioni o a strumenti tipici della inferenza statistica (in particolare, i cosiddetti test statistici) per verificare se le differenze riscontrate nelle distribuzioni condizionate siano da ritenersi solo dovute a oscillazioni causali e siano quindi compatibili con l'ipotesi che tali distribuzioni siano uguali

tra loro. In questo ultimo caso il ricorso a analisi condizionata non aggiunge informazioni ed sufficinete guardare alla variabile marginale Y.

Ciascuno degli strumenti statistici ha delle funzioni in `R` che consentono facilmente di ottenere le informazioni rilevanti. Prima di presentarne alcuni conviene introdurre alacune proprietà elementari che riguardano l'analisi di una variabile quantitativa osservata in più gruppi.

5.3.3 La scomposizione della devianza

Un'importante proprietà della media aritmetica è l'associatività: *"la media complessiva di una variabile osservata in più gruppi non cambia se a ciascun gruppo di dati si sostituisce la loro media"*.

Si consideri una matrice di dati, esemplificata nello schema seguente, in cui n osservazioni su una variabile quantitativa Y sono suddivise, secondo le modalità di una variabile categoriale X, in G gruppi di numerosità $n_1, \dots, n_G$ in relazione alle modalità di X.

	x_1	x_2	$\cdots$	x_g	$\cdots$	x_G
1	$y_1^{(1)}$	$y_1^{(2)}$	$\cdots$	$y_1^{(g)}$	$\cdots$	$y_1^{(G)}$
2	$y_2^{(1)}$	$y_2^{(2)}$	$\cdots$	$y_2^{(g)}$	$\cdots$	$y_2^{(G)}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
i	$y_i^{(1)}$	$y_i^{(2)}$	$\cdots$	$y_i^{(g)}$	$\cdots$	$y_i^{(G)}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
n_1	$y_{n_1}^{(1)}$	$\vdots$		$\vdots$		$\vdots$
$\vdots$		$\vdots$		$\vdots$		$\vdots$
n_G		$\vdots$		$\vdots$		$y_{n_G}^{(G)}$
$\vdots$		$\vdots$		$\vdots$		
n_2		$y_{n_2}^{(2)}$		$\vdots$		
$\vdots$				$\vdots$		
n_g				$y_{n_g}^{(g)}$		

Si indichi con m la media complessiva degli $n = \sum_{g=1}^{G} n_g$ valori, cioè la media di tutti i dati in cui ignoro la suddivisione in gruppi,

$$m = \frac{1}{\sum_{g=1}^{G} n_g} \sum_{g=1}^{G} \sum_{i=1}^{n_g} y_i^{(g)},$$

sia poi $m_g = \frac{1}{n_g} \sum_{i=1}^{n_g} y_i^{(g)}$ la media specifica per il gruppo g (è la media condizionata $M(Y|X = x_g)$), si ha allora

$$m = \sum_{g=1}^{G} \frac{n_g}{n} m_g, \tag{5.1}$$

infatti

$$m = \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g} y_i^{(g)} = \sum_{g=1}^{G}\frac{n_g}{n}\frac{\sum_{i=1}^{n_g} x_i^{(g)}}{n_g} = \sum_{g=1}^{G}\frac{n_g}{n}m_g.$$

Vale la pena di osservare che questa proprietà si estende anche alle medie che possono essere espresse come medie aritmetiche di opportune trasformazioni dei dati. Sono esempi la media quadratica, che è media aritmetica dei dati trasformati secondo $f(y) = y^2$ e, in generale, le medie di ordine s.

Lievemente più complicato è esprimere la devianza in funzione della devianza all'interno dei gruppi.

Partendo dalla definizione di devianza DEV_{TOT} della totalità dei dati, che possiamo scrivere

$$\begin{aligned} DEV_{TOT} &= \sum_{g=1}^{G}\sum_{i=1}^{n_g}(y_i^{(g)} - m)^2 \\ &= \sum_{g=1}^{G}\sum_{i=1}^{n_g}(y_i^{(g)} - m_g + m_g - m)^2. \end{aligned}$$

Sviluppando il quadrato si ha

$$DEV_{TOT} = \left[\sum_{g=1}^{G}\sum_{i=1}^{n_g}(y_i^{(g)} - m_g)^2 + \sum_{g=1}^{G}\sum_{i=1}^{n_g}(m_g - m)^2 + 2\sum_{g=1}^{G}(m_g - m)\underbrace{\sum_{i=1}^{n_g}(y_i^{(g)} - m_g)}_{=0\ \forall g}\right]$$

(dove $\sum_{i=1}^{n_g}(y_i^{(g)} - m_g) = 0$ per una proprietà della media aritmetica) e quindi

$$DEV_{TOT} = \underbrace{\sum_{g=1}^{G}\sum_{i=1}^{n_g}(y_i^{(g)} - m_g)^2}_{DEV_{\text{within}}} + \underbrace{\sum_{g=1}^{G}\frac{n_g}{n}(m_g - m)^2}_{DEV_{\text{between}}}. \tag{5.2}$$

La Devianza totale DEV_{TOT} (complessiva) è cioè pari alla devianza all'interno dei gruppi (devianza **within** o **entro**) a cui va sommata la media pesata dei quadrati degli scostamenti delle medie di gruppo dalla media generale (devianza **between** o **fra**).

La prima parte è una misura di quanto siano dispersi i valori all'interno di ciascun gruppo, la seconda di quanto siano dispersi i gruppi tra di loro. Ciò risulta più chiaro se si guarda ai due casi limite: se $DEV_{TOT} = DEV_{\text{within}}$ allora $m_g = m$ per ogni $g = 1, \ldots G$, cioè non vi è differenza in media tra i

gruppi; mentre $DEV_{TOT} = DEV_{\text{between}}$ quando $v_g = 0$ per ogni $g = 1, ...G$, cioè quando tutte le osservazioni all'interno di uno stesso gruppo coincidono.

Si noti che data la relazione

$$DEV_{TOT} = DEV_{within} + DEV_{between}$$

Si può valutare quanto i gruppi spieghino della devianza complessiva guardando al rapporto

$$0 \leq \frac{DEV_{between}}{DEV_{TOT}} = \eta^2 \leq 1$$

L'indice η^2 (detto rapporto di correlazione) è quindi un indice che se vicino a 1 indica che i dati nei diversi gruppi definiti dalle modalità del fattore sono poco variabili e che la variabilità è in gran parte legata alle differenze fra le medie dei gruppi.

5.3.4 L'uguaglianza dei valori centrali dei diversi gruppi (ANOVA)

La scomposizione della devianza vista sopra è alla base di un'importante statistica che si usa per verificare, usando i metodi dell'inferenza statistica, l'ipotesi che le medie dei diversi gruppi siano uguali.

In questo caso, per usare i metodi dell'inferenza, è necessario sepcificare un opportuno modello statistico che si suppone abbia generato i dati e l'ipotesi riguarderà i parametri i tale modello.

In particolare si suppone che i dati provengano da gaussiane indipendenti $\mathbb{N}(\mu_g, \sigma^2)$ ove $g = 1, 2, ..., G$ identifica i gruppi identificati da un fattore categoriale con $G \geq 3$ modalità. Si dispone quindi di campioni casuali semplici di numerosità n_g da ogni gruppo con $n = \sum_{g=1}^{k} n_g$. Si assume anche che in ciascun gruppo la varianza sia sempre la stessa e pari a σ^2 (assunzione di omoschedasticità).

Si tratta di valutare le diverse medie dei gruppi siano sufficientemente diverse da poter escludere che questo accada per effetto del caso.

A tal fine si calcola una statistica costituita dal rapporto tra la $Dev_{between}$ e la DEV_{TOT} divise per rispettivamente per $G-1$ e $n-G$ ovvero

$$S = \frac{DEV_{between}/(K-1)}{DEV_{within}/(n-K)}$$

Tale statistica, dovrebbe avere valori bassi se le medie sono uguali tra loro e le differenze sono imputabili al caso. Se questo è vero tale statistica equivale a un

numero casuale proveniente da una distribuzione di probabilità nota come F di Fisher. In particolare è distribuita come una $F_{G-1,n-G}$ in cui $G-1$ e $n-G$ sono i parametri di tale distribuzione (e sono detti gradi di liberta rispettivamente del numeratore e del denominatore).

Se i valori di tale statistica sono molto elevati allora è poco plausibile che i dati provengano da tale distribuzione e questo implica che i dati diano poco supporto alla ipotesi che le medie siano tutte uguali tra loro. A tal fine, si può quindi calcolare la $Pr(F > S)$, questa probabilità è il cosidetto il p-value. Quanto più questa probabilità è piccola tanto meno sarà plausibile che i dati osservati siano coerenti con l'ipotesi che tutte le medie sono uguali tra loro (i dati tendono a segnalarci che almeno una media è diversa dalle altre).

In inferenza statistica tale procedura è nota come **Analisi della Varianza** o ANOVA. Essa è di grande rilievo in numerosissimi contesti applicativi. Esistono versioni di tale metodo che richiedono assunzioni meno restrittive, ma si rinvia per esse, come per i dettagli su l atecnica inferenziale dell'ANOVA, a corsi di statistica successivi, più avanzati su temi di statistica inferenziale.

Può essere però interessante vedere come usare tale procedura con `R` usando la funzione `aov()`.

```
#Consideriamo ancora i dati di iris. In essi la variabile Species definiva 3 gruppi
data(iris)

testanova<-aov(Sepal.Length~Species, data=iris) # usiamo la notazione già vista con la variabile
                                                # di interesse come variabile dipendente dal fattore
summary(testanova)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  63.21  31.606   119.3 <2e-16 ***
## Residuals   147  38.96   0.265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La seconda colonna (Sum Sq) contiene le due devianze rilevanti (la devianza between e quella within)

```
# Si noti che la loro somma è pari alla devianza della variabile di interesse, infatti
# calcoliamo la varianza totale dell lunghezza del sepalo

n<-length(iris$Sepal.Length)
Devtot<-var(iris$Sepal.Length)*(n-1)
Devtot
```

```
## [1] 102.1683
```

5.4 Analisi di due variabili quantitative

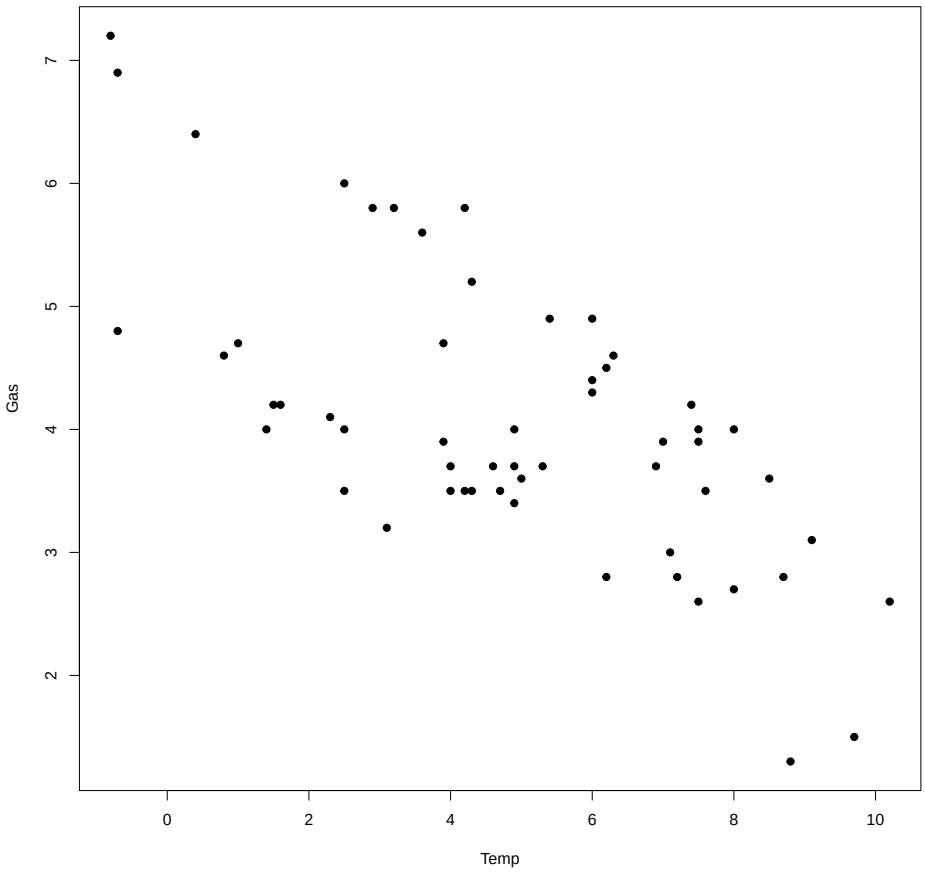
L'analisi congiunta di due variabili quantitative, diciamo (X, Y), si propone ancora di valutare se siano associate ed eventualmente di capire che forma assume tale associazione.

5.4.1 Il diagramma di dispersione

Dal punto di vista della rappresentazione grafica, i dati sulle due variabili si presenteranno come coppie di valori numerici (x_i, y_i) (con $i = 1, 2, ..., n$). Essi possono essere rappresentati direttamente come punti con coordinate x_i, y_i, possiamo usare quindi in `R` il comando `plot` per ottenere un grafico che in questo contesto viene denominato **diagramma di dispersione** (in inglese si usa il termine *scatterplot*).

Si considerino i dati nel dataframe `whiteside` nel package `MASS`. Contiene tre variabili: Gas, Temp e Insul. Si riferiscono a dati raccolti su abitazioni e Temp è la temperatura esterna (in gradi), Gas (in mc) è il consumo di gas per il riscaldamento dell'abitazione, mentre Insul è un fattore che registra se i dati si riferisco a misurazioni prese prima o dopo un intervento di isolamento termico della struttura.

```
library(MASS)
data(whiteside)
attach(whiteside)
plot(Temp,Gas, pch=19)
```




```
# si noti l'orientamento della nuvola di punti che denota una relazione
# inversa
```

Come era facile attendersi la nuvola di punti è orientata chiaramente e delinea una relazione negativa fra le due variabili.

5.4.2 Le misure di relazione lineare

Per valutare la tendenza delle due variabili a fornire valori che siano associati (positivamente o negativamente) così che, nel caso di relazione positiva, se la prima variabile mostra valori piccoli, ad esempio, anche la seconda tendenzialmente avrà valori piccoli, e lo stesso vale per valori elevati.

È piuttosto evidente nei grafici visti sopra che la nuvola di punti tenda a disporsi lungo una retta così da configurare una tendenziale relazione lineare fra le due variabili. Questa tendenza è misurata dalla **covarianza** definita come:

$$cov(X,Y) = \sum_{i=1}^{n} \frac{(x_i - M(x))(y_i - M(y))}{n}$$

Il numeratore dell'espressione è detto **codevianza**.

Vale per la covarianza lo stesso discorso fatto a suo tempo per la varianza: se essa viene calcolata nell'ambito di un approccio inferenziale, ovvero essa non è semplicemente un indice descrittivo di relazione lineare fra le due variabili, ma assume il ruolo di stimatore della varianza di una popolazione infinita da cui è tratto un campione casuale, si conviene di usare la divisione per $n-1$. Vedremo che in `R` la specifica funzione per calcolare la covarianza, utilizza la divisione per $n-1$

Si noti che la covarianza sopra definita, se divisa per n, può essere anche calcolata come: $cov(X,Y) = M(XY) - M(X) * M(Y)$

La covarianza assume valori positivi o negativi a seconda che la nuvola di punti tenda a disporsi lungo una retta orientata positivamente o negativamente. Può essere nulla se tale tendenza è assente. Si noti inoltre che $cov(X,Y) = cov(Y,X)$.

Non è facile valutare il grado della covarianza in quanto il suo valore assoluto dipende dall'unità di misura delle variabili. Per tale motivo è utile ricordare che, in conseguenza della disuguaglianza di Cauchy-Schwartz, si ha

$$cov(X,Y)^2 \leq var(X)var(Y)$$

che equivale a

$$\begin{array}{c} -\sqrt{(var(X)var(Y} \leq cov(X,Y) \leq \sqrt{(var(X)var(Y))} = \\ -sqm(X)sqm(Y) \leq cov(X,Y) \leq sqm(X)sqm(Y) \end{array} \tag{5.3}$$

Se quindi si dividono tutti i membri della doppia disuguaglianza sopra per $sqm(X)sqm(Y)$ si ha

$$-1 \leq \frac{cov(X,Y)}{sqm(X)sqm(Y)} = r(X,Y) \leq 1$$

L'espressione

$$\frac{cov(X,Y)}{sqm(X)sqm(Y)}$$

è quindi la covarianza standardizzata (si noti che si arriva alla stessa espressione se si calcola la covarianza tra le due variabili standardizzate). Tale quantità è nota come **coefficiente di correlazione lineare** fra le due variabili X e Y di solito denotato con $r(X,Y)$.

Esso è di più agevole interpretazione rispetto alla covarianza: infatti se esso è pari a 1 (o a -1) fra le due variabili vi è una perfetta relazione lineare (la nuvola di punti è perfettamente allineata lungo una retta). In `R` le funzioni `cov()`e `cor()` calcolano ripettivamente le covarianze e il coefficiente di correlazione lineare. La covarianza è calcolata come codevianza diviso per $n-1$.

```
# proviamo a calcolare la covarianza e il coeffficiente di correlazione per i dati whiteside
n<-length(Temp)
codev<-sum((Temp-mean(Temp)*(Gas-mean(Gas))))
covar<-codev/n
print(paste("covarianza calcolata dividendo per n", covar))
correl<-codev/(sd(Gas)*sd(Temp))
print(paste("coefficiente di correlazione lineare",correl))
```

```
## [1] "covarianza calcolata dividendo per n 4.875"
## [1] "coefficiente di correlazione lineare 85.017532033662"
```

```
# usiamo ora le funzioni di R
cov(Temp,Gas) # si noti che è diverso da quello sopra

# esso è infatti uguale a
codev/(n-1)
```

```
## [1] -2.194
## [1] 4.963636
```

5.4.3 La regressione lineare semplice

La covarianza e la correlazione misurano la tendenza delle due variabili ad essere associate linearmente. Tuttavia, come già discusso nel caso di due variabili categoriali, risulta di maggiore interesse il caso in cui le due variabili coinvolte hanno nell'analisi un ruolo diverso. Una delle due variabili, la variabile risposta, viene analizzata condizionatamente all'altra (la covariata o variabile indipendente).

L'interesse è quindi nel riuscire a fornire sintesi più efficaci della variabile risposta, diciamo Y, quando si conosca il valore della covariata X. Visto in altri termini, si tratta di capire se conoscere una variabile possa aiutare a prevedere l'altra con minore incertezza.

Per sintetizzare una variabile è spesso utile considerare la sua media e, in tal caso, una misura dell'incertezza è data dalla varianza (si ricordi che la media è il valore che rende minima la somma dei quadrati degli scarti da essa).

Si immagini quindi di voler ottenere per ogni possibile valore di $X = x$ la media di Y. Essa è la media di Y condizionata ad x, la si denoti con $M(Y|x)$.

Quest'ultima è in definitiva una funzione di x e $M(Y|x) = f(x)$ è detta **funzione di regressione**.

Per poter valutare tale funzione si dovrebbe disporre di un insieme di dati su Y in corrispondenza di ogni possibile valore x. Essendo X una variabile quantitativa essa assume presumibilmente molti valori distinti e quindi nella realtà di solito non si osservano più valori di Y per ogni x quindi non si è in grado di calcolare $M(Y|x)$ a partire dai dati (a meno che non si trasformi X in un fattore con un raggruppamento in classi con il consueto problema di perdere informazione).

Una semplice strategia per ottenere la funzione di regressione $M(Y|x) = f(x)$ è quella di supporre che tale funzione sia molto semplice: ad esempio, si può ipotizzare che le medie condizionate di Y per ogni x siano lungo una retta. Si assume quindi che la funzione di regressione sia lineare,cioè:

$$M(Y|x) = f(x) = \beta_0 + \beta_1 x$$

Per cui cerchiamo una retta, con intercetta pari a β_0 e coefficiente angolare β_1, lungo la quale si immagina siano collocate le medie condizionate.

I valori delle due variabili in generale si discosteranno dalla retta, per cui $\forall i$ si osserverà, con riferimento alla variabile Y, lo scostamento di un valore y_i da quella che dovrebbe essere la media condizionata $M(Y|x_i) = \beta_0 + \beta_1 x_i$.

Sembra sensato cercare quindi una retta, ovvero due valori β_0 e β_1, per i quali tali scostamenti siano nel complesso piccoli. Vi sono molti modi di misurare lo scostamento complessivo fra valori osservati e retta. Tuttavia una scelta ragionevole, avendo immaginato che sulla retta ci sono le medie, sarebbe quello di richiedere che per gli scostamenti valga la proprietà già enunciata per una media aritmetica: cerchiamo la retta per cui la somma degli scostamenti al quadrato risulti più piccola possibile, ovvero che sia minima $\sum_{i=1}^{n}(y_i - f(x_i))^2$.

5.4.3.1 La retta di regressione dei minimi quadrati

Definiamo con $\widehat{y}_i$ i valori su una retta in corrispondenza dei dati osservati x_i per la variabile indipendente. I valori $y_i - \widehat{y}_i$ sono detti residui.

Il principio appena enunciato, detto dei minimi quadrati, richiede quindi che definita la funzione

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - f(x_i))^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

che rappresenta la somma dei quadrati dei residui, e che si cerchino i valori β_0 e β_1 per cui risulti minima $L(\beta_0, \beta_1)$. Tali valori forniscono i coefficienti della **retta di regressione dei minimi quadrati**.

Le soluzioni possno esser reperite facilmente considerando la soluzione del seguente sistema di 2 equazioni (dette equazioni normali):

$$\frac{\partial L(\beta_0, \beta_1}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \tag{5.4}$$

$$\frac{\partial L(\beta_0, \beta_1}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \tag{5.5}$$

La soluzione del sistema, dopo alcuni passaggi, conduce ai seguenti valori per $\hat{\beta}_0$ e $\hat{\beta}_1$ che rappresentano quindi le soluzioni dei minimi quadrati:

$$\hat{\beta}_0 = M(y) - \hat{\beta}_1 M(x)$$

$$\hat{\beta}_1 = \frac{cov(XY)}{var(X)}$$

Ove $M(y)$ e $M(x)$ denotano le medie calcolate per i valori osservati delle due variabili Y e X.

Nel caso siano coinvolte solo due variabili si parla di **regressione lineare semplice**.

Di solito l'interesse maggiore è sul coefficinte angolare della retta $\hat{\beta}_1$ che è detto coefficiente di regressione. Se esso fosse nullo la retta è parallela all'asse orizzontale e le medie condizionate $M(Y|x)$ sarebbero tutte uguali tra loro. Non c'è quindi dipendenza lineare fra le variabili: si noti che in effetti il coefficiente di regressione è 0 se la covarianza è zero.

Si noti anche che il coefficiente di regressione angolare può anche essere scritto in funzione del coefficiente di correlazione fra le due variabili come segue

$$\hat{\beta}_1 = r(X, Y) \frac{sd(Y)}{sd(X)}$$

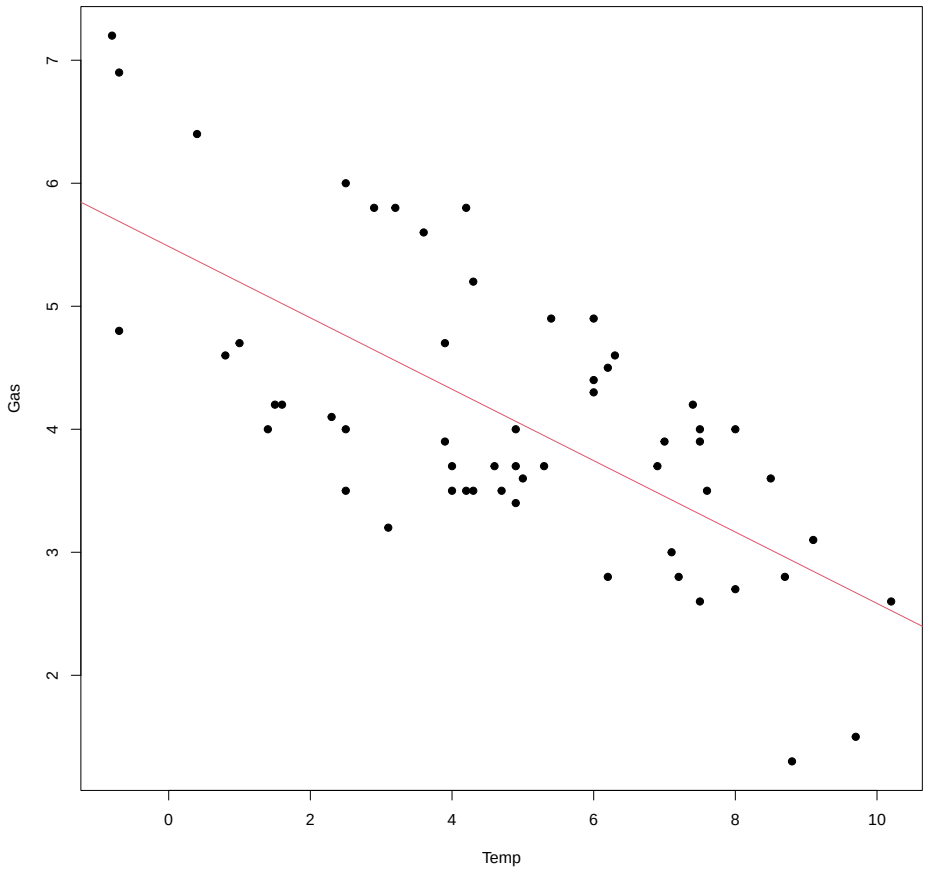
.

Si possono quindi calcolare i valori sulla retta dei minimi quadrati $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ e quindi i residui che corrispondono alla soluzione dei minimi quadrati $y_i - \hat{y}_i = r_i$ e ottenere, infine, la somma dei quadrati dei residui $SQE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

Si noti peraltro che in virtù della prima equazione normale è $\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$.

```
# proviamo  per i dati whiteside a ottenere la retta di regressione dei minimi quadrati
dev=sum((Temp-mean(Temp))^2)
b1<- cov(Gas,Temp)/var(Temp)
b1
print(paste("coefficiente di regressione",b1))
b0<- mean(Gas)-b1*mean(Temp)
b0
```

```
print(paste("intercetta",b0))
# otteniamo anche i valori sulla retta e la somma dei quadrati dei residui
Gasprev<-b0+b1*Temp
res<-Gas-Gasprev
SQE=sum(res^2)

print(paste("Devianza dei residui",SQE))

# e aggiungiamo la retta al diagramma di dispersione
plot(Temp,Gas, pch=19)
abline(b0,b1, col=2)
```




```
lm(Gas~Temp)
```

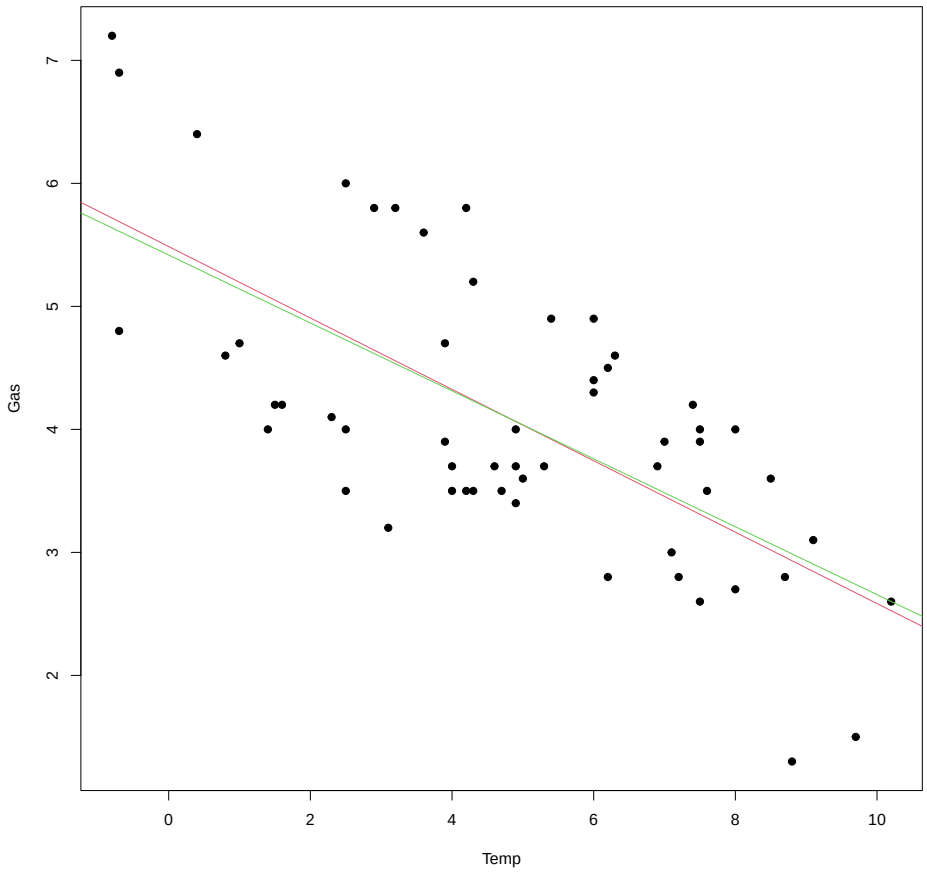
```
## [1] -0.2902082
## [1] "coefficiente di regressione -0.290208150455141"
## [1] 5.486193
## [1] "intercetta 5.48619330489738"
```

```
## [1] "Devianza dei residui 39.9948681988638"
##
## Call:
## lm(formula = Gas ~ Temp)
##
## Coefficients:
## (Intercept)         Temp
##      5.4862      -0.2902
```

Si potrebbe cercare una approssimazione dei punti con una retta scelta secondo altri criteri che potrebbero essere sensati, ma in tal caso se si decide che la misura della qualità dell'approssimazione fornita dalla retta sia misurata dalla somma dei quadrati delle distanze dei punti dalla retta (SSE), tale misura risulterà più elevata.

Vediamo un esempio. Si ottenga una retta che approssima i punti con un altro criterio. Consideriamo la retta che passa fra i rispettivi quartili delle due variabili e chiediamo poi che essa passi per il punto di coordinate $(M(y), M(x))$.

```
# otteniamo la retta che passa per i quartili per i dati visti sopra
b1q<- -(quantile(Gas,0.75)-quantile(Gas,0.25))/(quantile(Temp,0.75)
                                                -quantile(Temp,0.25))
# imponiamo che passi per i punti medi
b0q<- mean(Gas)-b1q*mean(Temp)

plot(Temp,Gas, pch=19)
abline(b0,b1, col=2)
abline(b0q,b1q,col=3)
```




```
# aggiungiamo anche tale retta al diagramma di dispersione
# otteniamo  i valori su tale retta e calcoliamo
# la somma dei quadrati dei residui
Gasprevq<-b0q+b1q*Temp
resq<-Gas-Gasprevq
SQE2=sum(resq^2)

print(paste("Devianza dei residui",SQE2))
# si noti che è più grande di quello già ottenuto con i minimi quadrati
```

```
## [1] "Devianza dei residui 40.077939784819"
```

5.4.3.2 Misurare la qualità della regressione: il coefficiente di determinazione lineare

La variabilità complessiva della quantità di interesse Y se non si disponesse (o se non si utilizza) anche l'informazione su x è misurata dalla devienza

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 \qquad \text{Somma dei quadrati dei residui totale (SQT)}$$

Si noti che questa è una misura dell'incertezza su Y se utilizzassimo come sintesi della variabile la sua media.

- La **devianza totale SQT** di Y dopo aver considerato la regressione dei minimi quadrati, può essere scomposta come segue:
  - **devianza di regressione SQR** che è la parte di variabilità di Y spiegata dal modello di regressione:

$$\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \qquad \text{devianza di regressione SQR}$$

  - **devianza dei residui SQE** che rappresenta la variabilità dell'errore che commettiamo utilizzando $\hat{y}_1$:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \text{variabilità dell'errore (devianza dei residui) SQE}$$

- Si può verificare facilmente che per la regressione dei minimi quadrati:

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}_{\text{SSR}}$$

Si noti che tale scomposizione è parallela a quella già discussa nel caso nell'analisi di una variabile quantitativa analizzata per gruppi.

Se la retta di regressione interpreta bene i dati allora:

- SSE, devianza dei residui, è piccola rispetto alla SST (devianza totale)
- SSR, la devianza spiegata dalla regressione, è la parte principale di SST

Possiamo quindi introdurre il seguente **coefficiente di determinazione (lineare)**:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

- $0 \leq R^2 \leq 1$ e indica la quota di variabilità che è spiegata dalla regressione
- Più basso è R^2 peggiore è l'adattamento del modello ai dati
- Se fosse pari a 1 indicherebbe un perfetto allineamento dei dati
- Si può infine dimostrare che $R^2 = r(X, Y)^2$, cioè è pari al quadrato del coefficiente di correlazione lineare fra le due variabili.

Si noti tuttavia che se il coefficiente di correlazione lineare, e di conseguenza il coefficiente di determinazione lineare, è da ritenersi nullo (molto vicino a 0), concluderemo che non c'è relazione lineare fra le variabili.

Attenzione: questo non significa che possa esserci una relazione fra le variabili di natura diversa. Ad esempio, i punti nel diagramma di dispersione potrebbero disporsi lungo una parabola e suggerire la presenza di una relazione fra le variabili non di tipo lineare. Quindi l'assenza di relazione lineare non implica l'indipendenza delle variabili.

5.4.3.3 la funzione `lm()` e il modello lineare

Dato il rilievo che ha l'analisi di regressione lineare non è sorprendente che in `R` vi sia una funzione che svolge tale analisi. Tale funzione, che come si vedrà può essere generalizzabile al caso dell'analisi con più variabili, fornisce tutte le informazioni rilevanti.

Tuttavia prima di introdurre la funzione `lm()` e per comprendere al meglio alcuni suoi risultati occorre introdurre il **modello di regressione lineare**.

Il modello di regressione lineare (lm sta appunto per *linear model*) è un modello statistico che è formulato per l'ambito inferenziale: i dati osservati sono visti come determinazioni di variabili aleatorie sulle quali si fanno alcune specifiche assunzioni. Come in ogni modello statistico vi sono dei parametri ignoti che caratterizzano le variabili aleatorie e su questi parametri si fa inferenza (secondo le procedure di stima, intervalli di confidenza, verifica di ipotesi) avendo osservato un campione casuale di dati.

Il modello lineare ha la seguente specificazione:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

ϵ_i sono determinazioni di una variabile aleatoria sulle quali si fanno le seguenti assunzioni: - $E(\epsilon_i) = 0$ - $Var(\epsilon_i) = \sigma^2$ (omoschedasticità ovvero stessa varianza $\forall i$). - $E(\epsilon_i, epsilon_j) = 0 \quad \forall i \neq j$ ovvero i residui sono incorrelati.

Inoltre è molto comune fare l'ulteriore assunzione di gaussianità per ϵ_i, cioè $\epsilon_i \sim \mathbb{N}(0, \sigma^2)$

Con le seguenti assunzioni quindi i dati osservati $Y_i \sim \mathbb{N}(\beta_0 + \beta_1 x_i, \sigma^2)$ sono determinazioni di gaussiane la cui media varia con x e i dati x_i sono non stocastici.

La determinazione dei parametri del modello $(\beta_0, \beta_1, \sigma^2)$, usando i metodi dell'inferenza statistica, può essere affrontata in vari modi. Si osservi però che se si accetta l'assunzione di normalità della componente aleatoria ϵ_i il metodo più adeguato per ottenere tali parametri è ancora il criterio dei minimi quadrati già introdotto. Inoltre la varianza dei residui $\hat{\sigma}^2$ viene ottenuta considerando la somma dei quadrati dei residui divisa per $n-2$ cioè $\hat{sigma}^2 = SQE/(n-2)$.

In ambito inferenziale, come discusso per i precedenti esempi si può valutare se il coefficiente angolare della retta $\hat{\beta}_1$ è abbastanza piccolo così da ritenere che sia nullo e che lo scostamento rispetto a 0 è da ritenersi dovuto solo al caso. Se, ad esempio, possiamo ritenere che β_1 è in effetti pari a 0 allora non vi è relazione lineare e X non è utile a prevedere Y secondo questo modello. Tuttavia non possiamo semplicemente guardare al valore di β_1 perchè il suo ordine di grandezza dipende dalla scala di misura delle due variabili coinvolte e dalla qualità della rappresentazione della relazione lineare (ovvero da quanto i punti si discostano dalla retta di regressione).

Tuttavia, anche in tal caso si può associare al valore di $\hat{\beta}_1$ ottenuto a partire dai dati, una probabilità, il $p - value$, che è in relazione a quanto si può ritenere plausibile osservare tale valore (o valori ancora più estremi cioè relativi a casi in cui il coefficiente angolare è ancora più distante da 0) se in effetti non vi fosse alcuna relazione lineare, cioè se $\beta_1 = 0$. In questo caso, si può calcolare $\frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2_{(\beta_1)}}}$.

La quantità al denominatore dell'ultima espressione, che dipende da SQT e dalla varianza di X, consente di standardizzare il valore di β_1 e di fare riferimento a una particolare distribuzione di probabilità (che verrà introdotta nei corsi di inferenza statistica ed è denominata t di student). Con riferimento ad essa si potrà calcolere il $p - value$. E, ancora una volta, quanto più questo valore è prossimo a 0 tanto più i dati non danno sostegno all'ipotesi che **non** esista una relazione lineare fra le variabili.

Illustriamo quindi l'uso della funzione `lm()` e i dei risultati principali che fornisce usando ancora i dati whiteside.

```
# otteniamo la retta che passa per i quartili per i dati visti sopra

mod1<- lm(Gas~Temp, data=whiteside)
# imponiamo che passi per i punti medi
summary(mod1)
# come si vede in corrispondenza del parametro di regressione
# vi è un p-value molto piccolo (praticamente 0)
# I dati danno quindi non danno supporto all'ipotesi che non esista una relazione lineare fra le
```

```
##
## Call:
## lm(formula = Gas ~ Temp, data = whiteside)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6324 -0.7119 -0.2047  0.8187  1.5327
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.4862     0.2357  23.275  < 2e-16 ***
```

```
## Temp         -0.2902     0.0422  -6.876 6.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8606 on 54 degrees of freedom
## Multiple R-squared:  0.4668, Adjusted R-squared:  0.457
## F-statistic: 47.28 on 1 and 54 DF,  p-value: 6.545e-09
```

5.4.4 Il *lisciamento* della curva di regressione

Assumere che le medie condizionate $M(Y|x)$ siano allineate lungo una retta (o un'altra curva semplice come la parabola) rende agevole l'analisi ma i dati potrebbero suggerire un comportamento molto diverso. Un approccio alternativo è quindi quello di esplorare metodi che consentano di tracciare la curva di regressione liberamente lasciando che siano i dati stessi a suggerire la forma della curva.

Un semplice espediuente potrebbe esser quello di suddividere la variabile X in classi e calcolare la media dei valori y_i per i dati che sono in quella classe. A quel punto si potrebbe calcolare la media condizionata e porla costante all'interno di ogni classe. Questo porterebbe a una funzione costante a tratti la cui forma è suggerita dai dati.

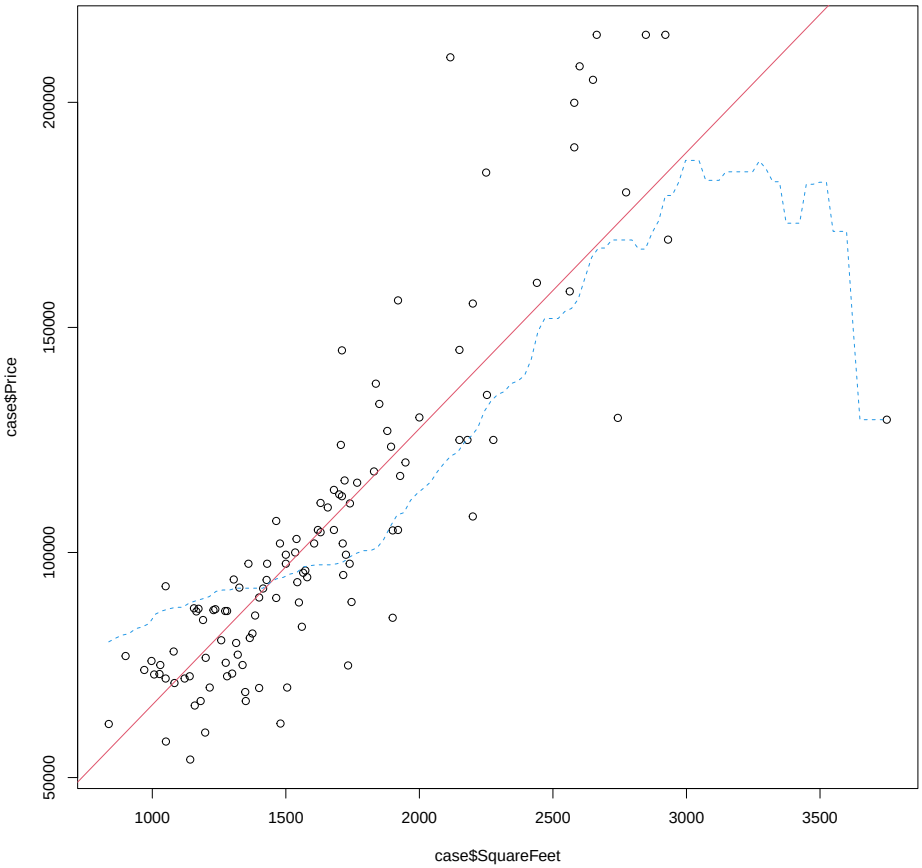
Tuttavia sarebbe desiderabile, anche per una visualizzazione dei dati sintetica, che la funzione di regressione sia *liscia* senza sbalzi. In tal senso il problema è simile a quello affrontato nel caso del *lisciamento* della funzione di densità come alternativa all'istogramma.

La funzione di regressione descrive il comportamento della media della variabile dipendente condizionata ai valori in corrispondenza dei valori di una variabile esplicativa. Anche se l'idea che tali medie stiano lungo funzioni semplici come la retta è spesso utile, se guardiamo al diagramma di dispersione si ha spesso l'impressione che le cose possano essere più complicate. Potrebbe rivelarsi interessante che i dati stessi ci suggeriscano l'andamento delle medie di Y condizionatamente ai valori di x.

Il grafico che segue contiene la nuvola di punti relativa a due variabili: il prezzo delle case vendute in un certo periodo ad Albuquerque (New Mexico), y, e la loro dimensione (in piedi quadrati), x. Che fra le due variabili ci sia una relazione è scontato, che il prezzo medio delle case vari linearmente al variare della dimensione lo è un po' meno.

```
detach(whiteside)
# leggiamo i dati dal file case.csv
case<-read.csv("case.csv")

plot(case$SquareFeet,case$Price)
modl<-lm(Price~SquareFeet, data=case) # otteniamo i parametri di un modello lineare
```

```
abline(modl, col=2)
# usiamo la funzione ksmooth per ottenere un lisciamento col metodo del nucleo
# la funzione vuole le due variabili inserite come in plot
# ed è specificato il parametro baandwidth che regola il lisciamento
lines(ksmooth(case$SquareFeet,case$Price, bandwidth=1400), col=4, lty=2)
```



Nel grafico è riportata la retta dei minimi quadrati e una curva (più o meno) liscia che tende a seguire con meno rigidità l'andamento della nuvola; peraltro, ci si potrebbe, ad esempio, ragionevolmente attendere che aumenti della metratura per case molto grandi si tradurranno in aumenti di prezzo inferiori a quelli che si osservano per le case piccole (ci si aspetta che la derivata decresca). In realtà, per i dati americani questo sembra non succedere.

La funzione che passa tra i punti è una versione della curva di regressione ottenuta non parametricamente con il metodo del nucleo. In particolare si è utilizzato il cosidetto metodo delle medie locali. Esso consiste nel calcolare in corrispondenza di ciascun valore x (sia esso osservato o meno) la media delle coordinate y per le unità che sono vicine ad x (si calcola cioè una media locale). Questo

si potrebbe realizzare prendendo una "striscia", più o meno stretta, attorno ad x e considerando la media delle coordinate y solo per le unità nella striscia. Si può tuttavia fare di meglio: per calcolare la funzione $m(x_0) = Media(y|x_0)$ in corrispondenza di un dato valore x_0, si possono considerare tutte le unità statistiche e fare una media ponderata di tutti i valori y dando però pesi maggiori a quelle unità che hanno valori di x vicini a x_0.

La funzione che determina questi pesi è una funzione nucleo $K()$ che ha le stesse caratteristiche viste nel caso della determinazione della funzione di densità. Si ottiene la seguente media ponderata

$$m(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}. \tag{5.6}$$

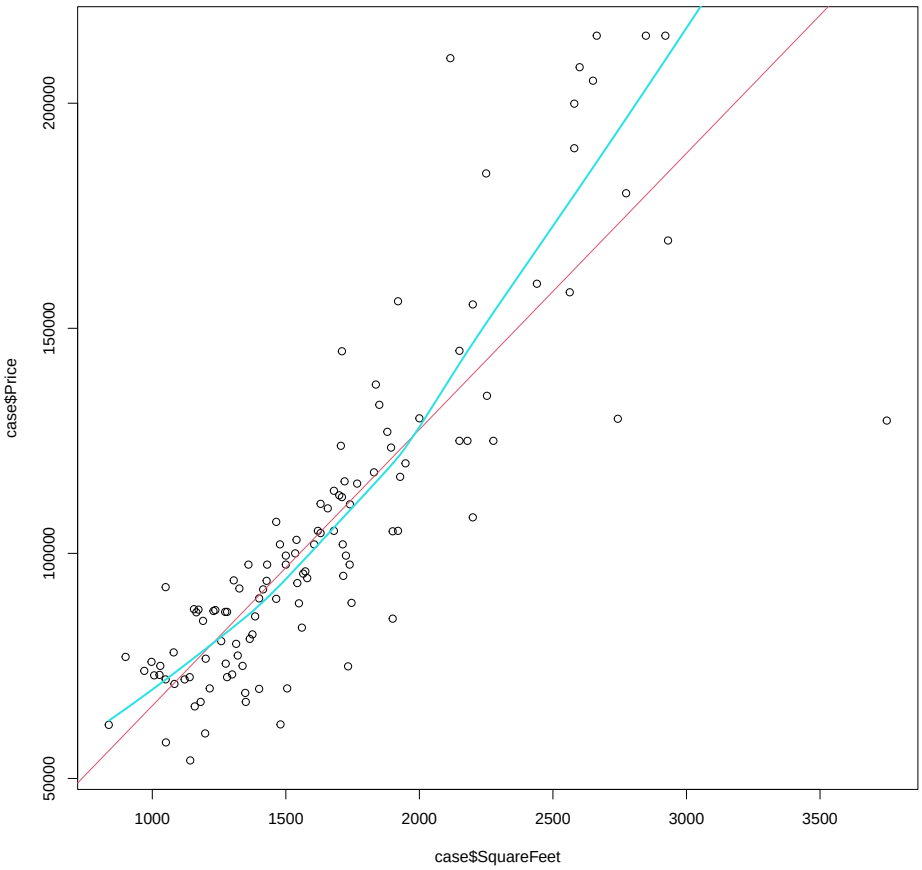
.

Si tratta quindi ancora del metodo del nucleo, di solito $K()$ è una normale standard, e ancora una volta è determinante il valore di h: se troppo piccolo accade che la curva tende a seguire troppo da vicino ciascun punto ovvero adattarsi troppo ad ogni particolarità dei dati, se troppo grande praticamente tutte le osservazioni hanno lo stesso peso per ogni x e quindi si ottiene semplicemente una retta parallela all'asse X all'altezza della media di Y. Questo comportamente è un esempio del cosiddetto trade-off tra bias e varianza: ovvero fra una funzione liscia che ha molto bias, cioè rischia di essere diversa dalla funzione che misura la relazione far le variabili, e una molto accidentata che è molto variabile.

Nel caso della regressione si può fare ancora di meglio: ad esempio, si possono usare polinomi locali invece che medie locali, oppure si può tenere conto della densità locale dei punti x o, ancora, utilizzare metodi più resistenti (che risentono meno di valori anomali). Un metodo che fa tutto questo è il metodo "Lowess" (per il quale esiste la funzione `lowess()` in R).

Per avere solo un'idea di cosa verrebbe fuori utilizzando il Lowess si veda la figura sopra (si noti come il valore a destra nelle figura abbia, con il Lowess, meno rilevanza nel determinare la funzione).

```
# leggiamo i dati dal file case.csv
case<-read.csv("case.csv")


plot(case$SquareFeet,case$Price)
modl<-lm(Price~SquareFeet, data=case) # otteniamo i parametri di un modello lineare
abline(modl, col=2)
# usiamo ora la funzione lowess per ottenere un lisciamento col metodo dei
# polinomi locali in una versione resistente ai valori anomali
#

lines(lowess(case$Price~case$SquareFeet, f=.8), col=5, lwd=2)
```



Esistono altre funzioni ad esempio nel pacchetto `KernSmooth` per ottenere la funzione con il metodo del nucleo.

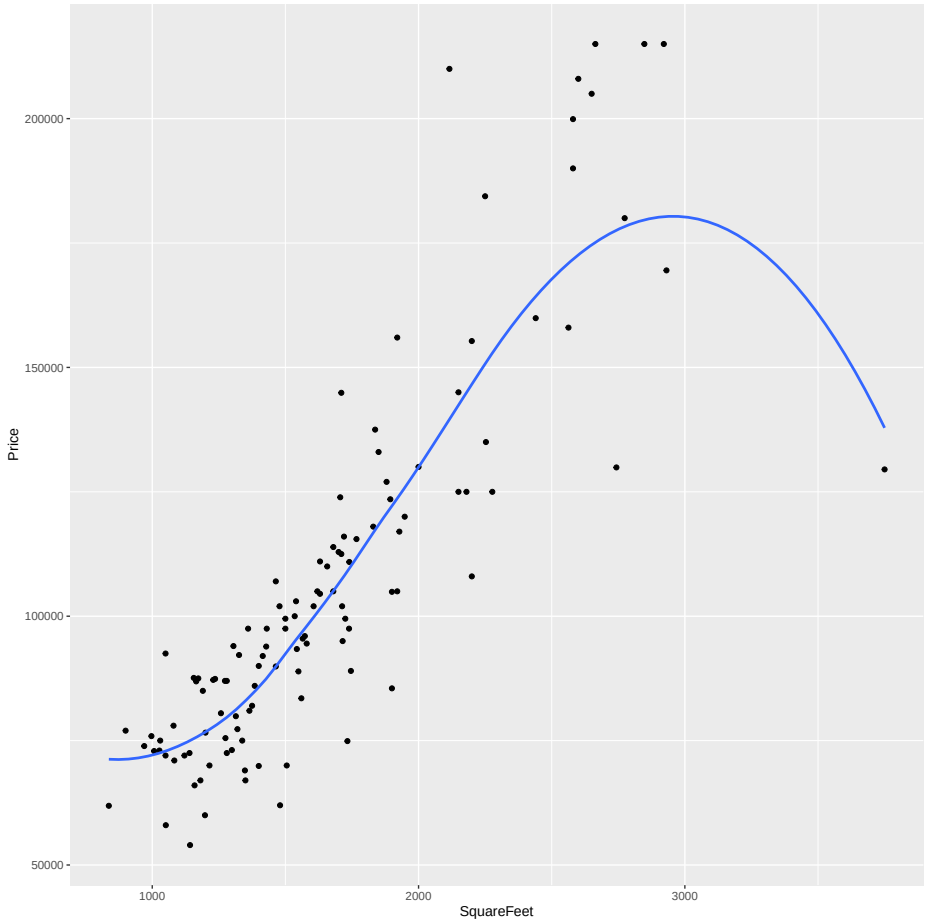
Si noti che il lisciamento con il metodo del nucleo risulta più *robusto* in quanto include automaticamente un metodo per dare meno peso a valori anomali (ad esempio la curva risente meno del punto a destra).

L'uso di versioni lisciate della curva di regressione al fine di visualizzare i dati su due variabili esplicative è diventato ormai consueto.

Pertanto non stupisce che nei sistemi di visualizzazione, come ad esempio ggplot2, sia molto semplice ottenere le curve tipo lowess (o loess).

```
library(ggplot2)
ggplot(mapping = aes(x = SquareFeet, y = Price), data = case) +
geom_point() +
geom_smooth(aes(),  se=F)
```

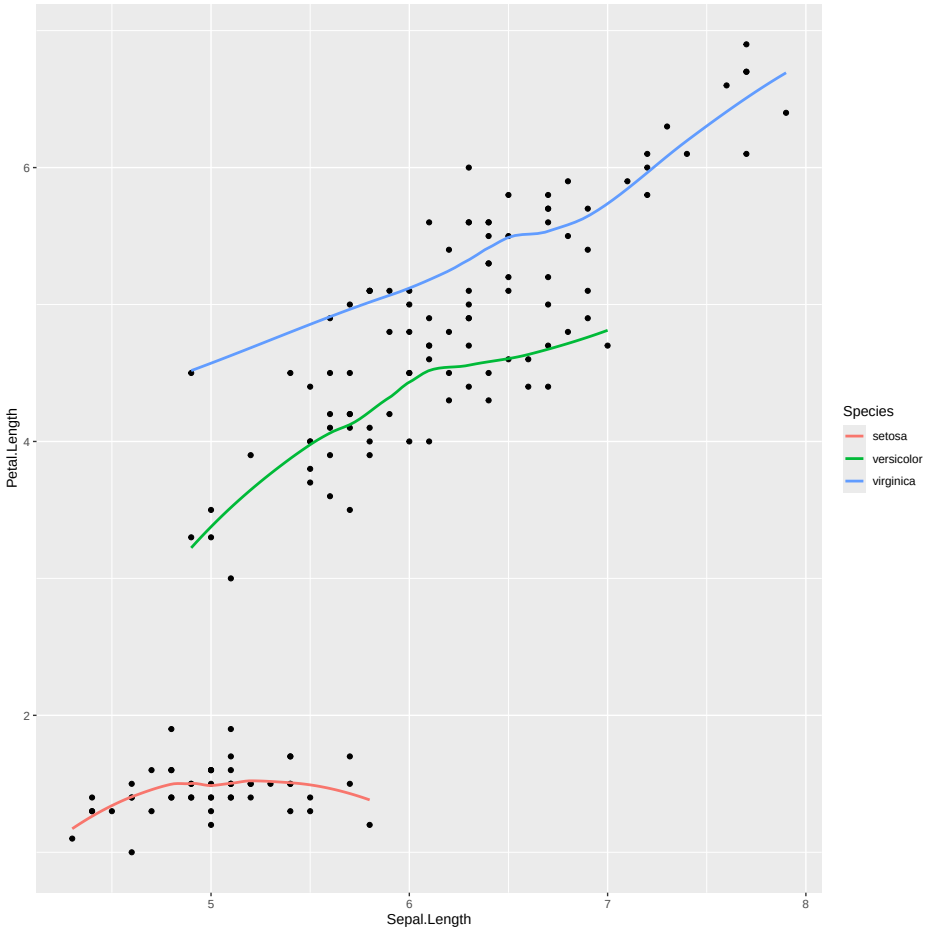
```
## `geom_smooth()` using method =
## 'loess' and formula = 'y ~ x'
```



Sotto si riporta anche un esempio con i dati iris in cui le curve loess sono mostrate per sottogruppi di dati.

```
ggplot(mapping = aes(x = Sepal.Length, y = Petal.Length), data = iris) +
geom_point() +
geom_smooth(aes(colour=Species),  se=F)
```

```
## `geom_smooth()` using method =
## 'loess' and formula = 'y ~ x'
```


Petal.Length
6
4
2
5
6
7
8
Sepal.Length
Species
setosa
versicolor
virginica

Capitolo 6

Analisi statistica multivariata

Seppure aspetti di grande interesse possono già emergere da una corretta analisi di coppie di variabili, l'analisi congiunta di un insieme più ampio delle variabili disponibili (che a volte possono essere anche molto numerose) consente di cogliere relazioni complesse presenti nei dati.

Nel seguito quindi si introdurranno elementi che riguardano l'analisi di un insieme variabili $x_1, x_2, \ldots, x_p$ (con $p \geq 3$) misurate su n unità. Si parla in tal caso di insieme di dati multivariato e di analisi dei dati multivariata.

Le variabili coinvolte possono essere di diversa natura: quantitative o categoriali. La trattazione fatta precedentemente per l'analisi bivariata distingueva i casi in cui entrambe le variabili erano categoriali, una variabile era quantitativa e una categoriale ed entrambe quantitative. Sono stati illustrati strumenti di analisi e di visualizzazione adatti ai diversi casi che richiedono scelte tecniche adeguate.

Nel caso di un insieme di dati multivariato sarà ancora più frequente trovarsi nella situazione in cui i dati sono di tipo misto (sia quantitative che categoriali) e converrà sempre tenere a mente tale distinzione. Vale sempre, se si utilizza `R` la regola aurea in di definire come fattore (`factor`) le variabili categoriali.

Si inizia con l'introduzione di primi semplici strumenti di analisi per i casi in cui tutte le variabili sono della medesima natura. Successivamente, si introdurranno due particolari tematiche tipiche dell'analisi di dati multivariati.

Il primo tema è l'estensione dell'analisi di regressione semplice (che interessava due sole variabili) al caso in cui fra le informazioni disponibili si individua una variabile risposta (output o target) quantitativa e l'obiettivo è quello di spiegare (o prevedere) tale variabile utilizzando un insieme di variabili esplicative (regressori, input) più ampio. Tale analisi è detta di **regressione multipla** e

costituisce uno delle principali tecniche di analisi statistica: essa verrà ripresa in altri corsi, a partire da quelli di machine learning, ma se si vuole approfondire realmente, e in modo serio, tale fondamentale tematica il consiglio vivissimo è di seguire il corso di **Modelli Statistici**, da inserire come scelta nel piano di studi.

Il secondo tema che verrà introdotto cerca di utilizzare le misure multivariate ottenute sulle diverse unità per cercare pattern nei dati individuando le unità simili tra di loro con riferimento alle variabili osservate. Non c'è quindi in questo caso una variabile risposta e l'obiettivo non è quello di prevedere o spiegare una variabile ma è quello di cercare strutture nei dati multivariati che non sono immediatamente riconoscibili. Si fa riferimento a tali problemi con il termine analisi di raggruppamento o **cluster analysis**.

I due approcci sono paradigmatici di una serie molto più ampia di tecniche di analisi statistiche oltre che dei due approcci all'apprendimento statistico (o automatico, *machine learning*) detti **apprendimento supervisionato** (la regressione ne è l'esempio più semplice e immediato) e **apprendimento non supervisionato** (per il quale la cluster analysis rappresenta una delle tecniche di maggior rilievo).

6.1 L'analisi di più variabili categoriali

Nel caso di due sole variabili categoriali l'analisi è apparentemente molto semplice. Gli strumenti principali sono la rappresentazione in una tabella di frequenze a doppia entrata, la rappresentazione mediante barplot affiancati e la costruzione di misure di associazione come l'indice X^2 o, nel caso di tabelle 2×2, il rapporto dei prodotti incrociati (**odds ratio**).

Non appena le variabili che vogliamo analizzare congiuntamente sono tre, o anche più di tre, l'analisi diviene più complessa e difficile da gestire. Questo anche perché l'associazione congiunta fra più variabili può essere caratterizzata da aspetti che coinvolgono forme di indipendenza (è il caso, ad esempio, dell'indipendenza condizionale) che non sono semplicemente l'indipendenza mutua completa fra tutte le variabili.

Come già detto lo strumento di base per l'analisi di più variabili categoriali è semplicemente la tabella incrociata a più entrate.

Un'analisi approfondita della tematica relativa alla valutazione delle relazioni fra più variabili categoriali va ben oltre questa trattazione introduttiva. E' tuttavia di qualche interesse fornire un esempio del concetto di indipendenza condizionale quando a essere coinvolte sono tre variabili categoriali.

6.1.1 L'associazione marginale, l'associazione condizionale e il paradosso di Simpson

Quando si considera la relazione fra una coppia di variabili, categoriali in questo caso, isolatamente, senza cioè tenere conto della presenza di ulteriori altre variabili, si parla di relazione **marginale**: se quindi c'è indipendenza fra tali due variabili dovremmo più correttamente parlare di **indipendenza marginale** o se ci fosse associazione di **associazione marginale**.

Se tuttavia si osservano altre variabili, nel caso in questione ancora categoriali, si potrebbe considerare la relazione fra le due variabili **condizionatamente** alla terza variabile. La relazione fra le due variabili, quando si tenga conto della terza variabile, può essere diversa da quella osservata a livello marginale.

Potrebbe quindi accadere che una relazione di associazione (o di indipendenza) marginale non sia più presente a livello condizionale o addirittura possa invertirsi (in tali caso si parla di paradosso di Simpson). Vale la pena di ricordare che l'associazione marginale fra due variabili potrebbe essere illusoria e pertanto non bisognerebbe mai trarre conclusioni sulle relazioni fra variabili solo basandosi sull'associazione marginale. In particolare: **la presenza di una associazione marginale fra due variabili non implica che vi sia una relazione causale**, questo in quanto non si controllano, di solito, tutte le altre variabili in gioco.

6.1.1.1 Un esempio

Per esemplificare quanto detto sopra si considerino i seguenti dati relativi alle ammissioni ai vari dipartimenti all'università di Berkeley nel 1973 contenuti nella tabella a tre entrate:`UCBAdmissions` (che è di fatto un array a tre dimensioni)

Le variabili coinvolte sono 3:

1. Ammissione (due modalità: Admitted, Rejected)
2. Genere (due modalità: Male, Female)
3. Dipartimento (6 dipartimenti).

I dati sono riportati di seguito:

DIPARTIMENTO A

	Male	Female
Admitted	512	89
Rejected	313	19

DIPARTIMENTO B

	Male	Female
Admitted	353	17
Rejected	207	8

DIPARTIMENTO C

	Male	Female
Admitted	120	202
Rejected	205	391

DIPARTIMENTO D

	Male	Female
Admitted	138	131
Rejected	279	244

DIPARTIMENTO E

	Male	Female
Admitted	53	94
Rejected	138	299

DIPARTIMENTO F

	Male	Female
Admitted	22	24
Rejected	351	317

Si consideri dapprima la tabella a doppia entrata per le prime due variabili (la tabella marginale bivariata) ottenuta sommando rispetto ai diversi Dipartimenti

```
marg<-margin.table(UCBAdmissions,c(1,2))
kable(marg)
```

	Male	Female
Admitted	1198	557
Rejected	1493	1278

Per valutare l'associazione fra Ammissione e Genere: ad esempio, calcoliamo il rapporto dei prodotti incrociati **OR**.

```
OR<-marg[1,1]*marg[2,2]/(marg[1,2]*marg[2,1])
OR
```

```
## [1] 1.84108
```

come si vede il valore ottenuto è abbastanza lontano da 1 (nel caso fosse vicina a 1 si propenderebbe per l'indipendenza fra i caratteri), e in effetti osservando i dati sembrerebbe che ci sia un *bias* nell'ammissione per cui la proporzione di ammissione è più alta per gli uomini che per le donne. Si sarebbe tentati di concludere che gli uomini risultino più preparati e vengano ammessi più spesso.

Nelle tabelle, per le due variabili prese condizionatamente al dipartimento (cioè separatamente per ogni dipartimento) la situazione però appare diversa. Ad esempio, nel dipartimento A, si ha la seguente tabella a doppia entrata (distribuzione bivariata fra Ammssione e Genere condizionata al Dipartimento A)

```
cond<-UCBAdmissions[,,1]
kable(cond)
```

	Male	Female
Admitted	512	89
Rejected	313	19

Si vede subito che nel dipartimento A non è presente la stessa relazione osservata a livello marginale, infatti sembra che in questo caso la relazione sia opposta. Se calcoliamo anche per questa l'odds-ratio:

```
ORA<-cond[1,1]*cond[2,2]/(cond[1,2]*cond[2,1])
ORA
```

```
## [1] 0.349212
```

Il rapporto dei prodotti incrociati è di segno opposto rispetto a quello marginale e in questo Dipartimento si verifica un'ammissione preferenziale per le donne.

Potremmo fare lo stesso per tutti i dipartimenti e calcolare i 6 odds-ratio condizionali

```
nn<-dim(UCBAdmissions)[3]
ordip<-1:nn
for (i in 1:nn)
{
ordip[i]<-UCBAdmissions[1,1,i]*UCBAdmissions[2,2,i]/(UCBAdmissions[1,1,i]*UCBAdmissions[2,1,i])
ordip
}
ordip
```

```
## [1] 0.06070288 0.03864734 1.90731707 0.87455197 2.16666667 0.90313390
```

Come si vede la relazione marginale non è uguale alle relazioni condizionali e in 4 dipartimenti su 6 la relazione è invertita sono le femmine ad essere ammesse più spesso. L'apparente associazione marginale tra ammissione e sesso, a favore dei maschi, è il risultato del fatto che vi è diversa propensione di maschi e femmine a candidarsi all'ammissione ai singoli dipartimenti. Le candidate femmine, in effetti, facevano prevalentemente domanda di ammissione ai dipartimenti con tassi di rifiuto più elevati (che sono il C, E e D) mentre i maschi fanno domanda a Dipartimenti in cui è facile essere ammessi. Quando si sommano i dati per tutti i dipartimenti appare quindi complessivamente che i maschi sono ammessi più spesso ma da questo non possiamo concludere che i maschi siano più bravi o che vi sia una preferenza a ammettere i maschi. Questo è appunto un esempio di quello che è detto **paradosso di Simpson**.

6.2 L'analisi di piu variabili quantitative

Quando si analizzano più variabili, tutte quantitative, si può sempre partire dall'esame delle relazione fra coppie di variabili. Si può inizialmente guardare a un indice che misuri la forza della relazione lineare fra esse: ovvero la covarianza o il coefficiente di correlazione lineare.

6.2.1 La matrice di varianza-covarianza e la matrice di correlazione

Se sono disponibili p variabili $X_1, X_2, ... X_p$ è possibile quindi calcolare la covarianza o il coefficiente di correlazione lineare fra ciascuna coppia di variabili ottenendo $\frac{p \times (p-1)}{2}$ covarianze (o correlazioni) distinte. Di solito tali indici vengono organizzati in una matrice di dimensione p.

Nel caso delle covarianze, si definisce la **matrice di varianze-covarianze** S il cui generico elemento si denota $s(i,j) = cov(X_i, X_j)$.

Tale matrice risulterà essere simmetrica poichè $s(i,j) = s(j,i)$. Inoltre sulla diagonale di tale matrice saranno presenti i valori di $cov(X_i, X_i) = \frac{\sum_k^n (X_i k - M(X_i))^2}{n-1} = var(X_i)$ ovvero le varianze delle variabili.

Nel caso dei coefficienti di correlazione lineare, si definisce la **matrice di correlazione** R il cui generico elemento si denota con $r(i,j) = cor(X_i, X_j)$. Tale matrice è anch'essa simmetrica essendo $r(i,j) = r(j,i)$. Inoltre sulla diagonale di tale matrice saranno presenti valori unitari essendo pari a 1 la correlazione fra una variabile e se stessa. La matrice di correlazione è spesso preferibile in prima istanza essendo più agevole avere un'indicazione fra l'intensità del legame lineare fra le coppie di variabili.

Le funzioni `cov()` e `cor()`, già introdotte nel caso di due variabili, permettono di ottenere la matrice di covarianze se in input viene fornita una matrice di dati (o un data frame) di variabili numeriche. Ad esempio.

```
irisnum<- iris[,-5] # elimino l'ultima variabile del dataframe che è un fattore

# Matrice di varianza/covarianza
kable(round(cov(irisnum),2))
```

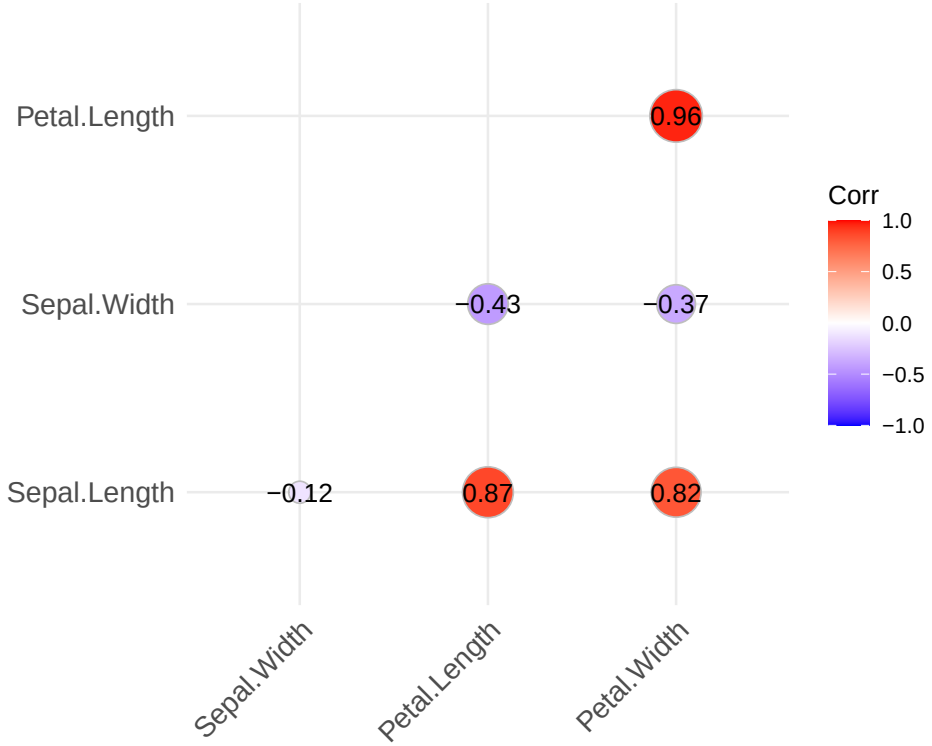
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	0.69	-0.04	1.27	0.52
Sepal.Width	-0.04	0.19	-0.33	-0.12
Petal.Length	1.27	-0.33	3.12	1.30
Petal.Width	0.52	-0.12	1.30	0.58

```
# Matrice di correlazione
kable(round(cor(irisnum),2))
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.00	-0.12	0.87	0.82
Sepal.Width	-0.12	1.00	-0.43	-0.37
Petal.Length	0.87	-0.43	1.00	0.96
Petal.Width	0.82	-0.37	0.96	1.00

`ggplot2` consente di usare alcuni strumenti grafici per visualizzare la matrice di correlazione, ad esempio il package `ggcorrplot` (che va installato).

```
library(ggcorrplot)
library(ggplot2)
ggcorrplot(cor(iris[,-5]),type="lower", method="circle", lab=TRUE)
```

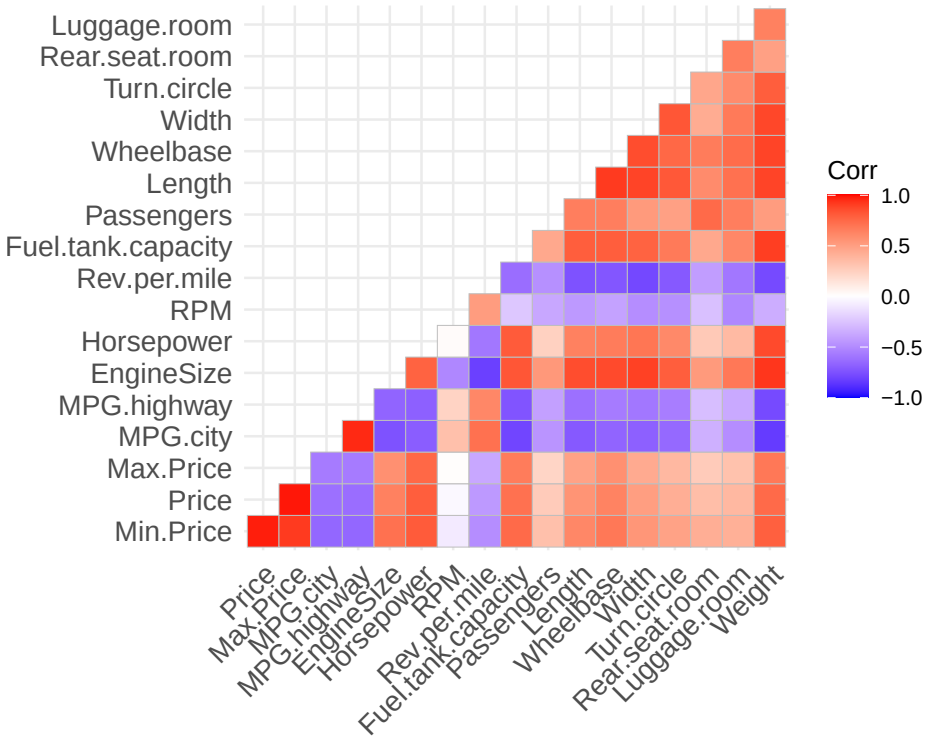


Si noti che la funzione richiede in input una matrice di correlazione. Vi sono vari parametri che permettono di abbellire il grafico. negli esempi che seguono se ne usano alcuni ma si consiglia di guardare i tutorial presenti in rete sull'uso della funzione `ggcorrplot`.

Un secondo esempio, con il data frame delle auto già usato in precedenza (ii noti il primo comando che sfrutta le potenzialità di `dplyr` per selezionare solo le variabili numeriche) è fornito di seguito:

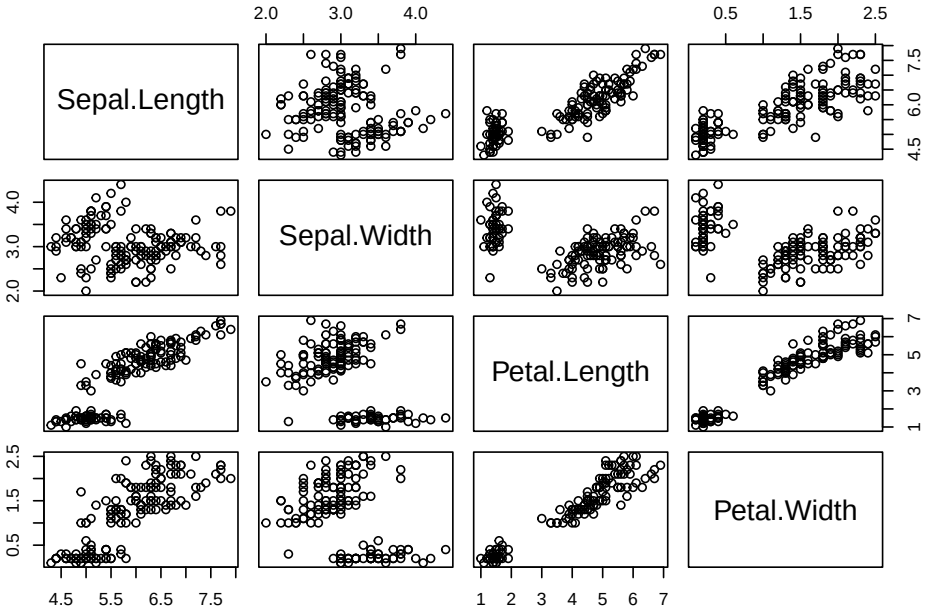
```
library(dplyr)
Cars93num<- Cars93 %>% select_if(is.numeric)
```

```
corrcar<-round(cor(Cars93num, use = "complete.obs", method="pearson"),2)
ggcorrplot(corrcar,type="lower")
```



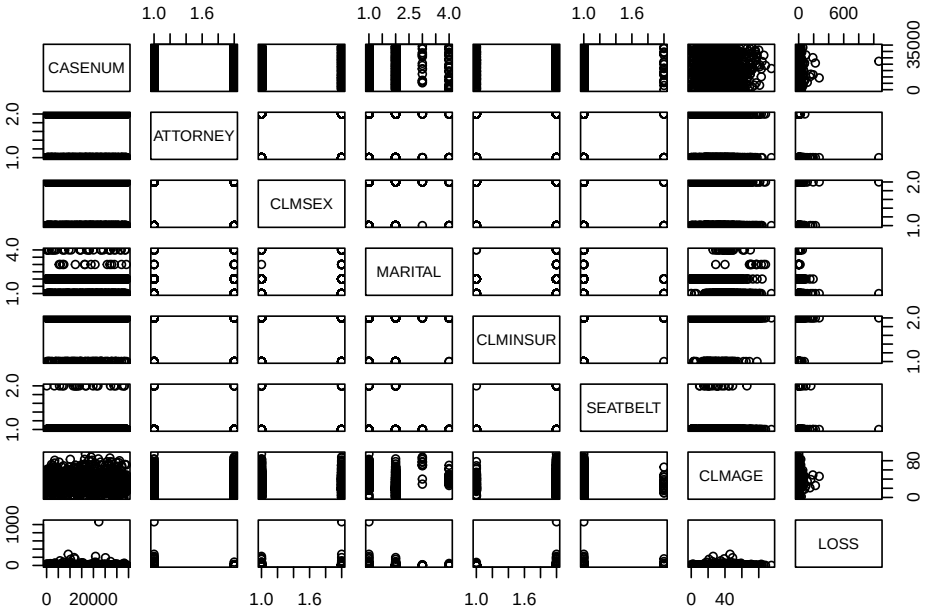
Un altra utile funzione per verificare le relazioni fra tutte le variabili, senza calcolare il coefficiente di correlazione ma rappresentando come in una matrice di correlazione tutti gli scatterplot delle coppie, è la funzione `pairs()`

```
pairs(iris[,-5])
```



Si osservi che `pairs` funzionerebbe anche se le variabili coinvolte non sono numeriche. Ad ogni modalità del fattore vengono associati livelli numerici e si ottiene un grafico che tuttavia in molti casi risulterà privo di significato o di non facile interpretazione. Si osservi ad esempio cosa accade se utilizziamo `pairs` per il data frame `AutoBi` che conteneva sia variabili numeriche che fattori.

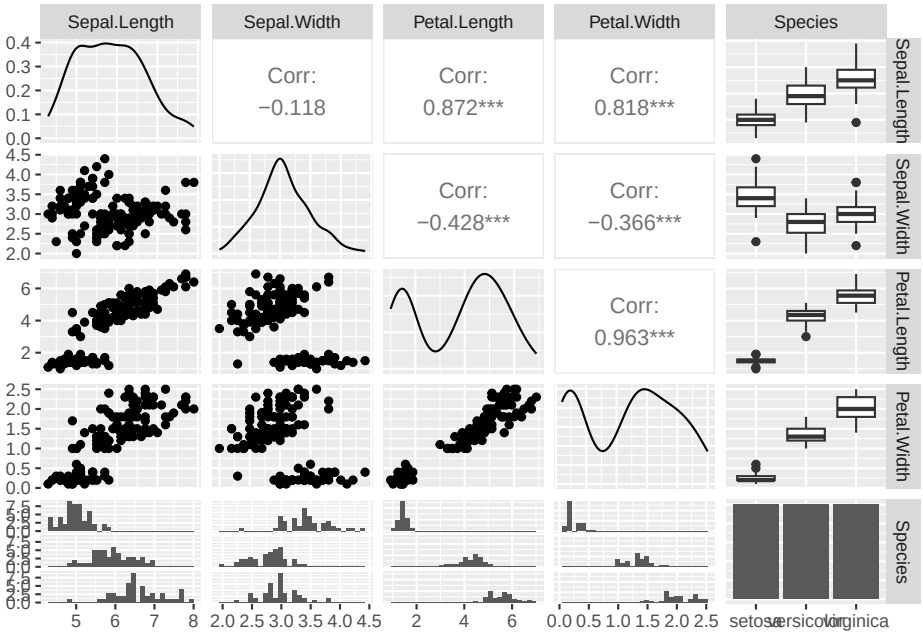
```
data(AutoBi)
pairs(AutoBi)
```



I grafici per alcune coppie sono o poco informativi o non molto sensati.

Tuttavia, se vi sono nel data frame, variabili sia numeriche che fattori esistono altre funzioni molto utili. Si illustra, ad esempio, la funzione `ggpairs` nel package `GGally` di ggplot.

```
library(GGally)
ggpairs(iris)
```



Come si vede in questo caso si adotta la rappresentazione grafica più appropriata data la natura delle variabili.

6.2.2 La correlazione parziale

Nel caso dell'analisi tra due variabile categoriali si era già introdotta l'idea che, se fossero presenti più variabili, non era sufficiente guardare all'associazione marginale fra esse e che poteva esser di intersse guardare all'associazione fra esse condizionatamente a queste ulteriori variabili. In particolare, si voleva mettere in guardia dal non interpretare l'associazione marginale come indicazione di un legame (anche causale fra le due variabili) perchè esso poteva non essere presente quando si tenga conto delle altre variabili.

Una simile attenzione va posta anche nel caso delle variabili quantitative. Quando si calcola il coefficiente di correlazione lineare fra una coppia di variabili, è a volte difficile resistere alla tentazione di interpretare la presenza di un alto livello di correlazione come indicativo di una relazione di "causa-effetto" fra esse. Tuttavia, specie in ambito osservazionale, tale conclusione è all'origine di una fallacia da evitare: occorre ricordare quindi sempre che:

> **Association is not causation**
>
> La presenza di un elevato livello di associazione (lineare) (misurata dal coefficiente di correlazione lineare) non può, in genere, essere considerato una prova di un legame causale fra le variabili.

Anche in questo caso, occorre individuare uno strumento che consenta di misurare l'associazione fra le variabili tenendo conto anche delle interazioni con ulteriori variabili osservate.

E' possibile ottenere misure di correlazione fra coppie di variabili tenendo conto anche di altre variabili introducendo il concetto di **correlazione parziale**.

Può essere utile considerare un semplice esempio per illustrare sia l'idea di correlazione parziale che le misure che ne derivano.

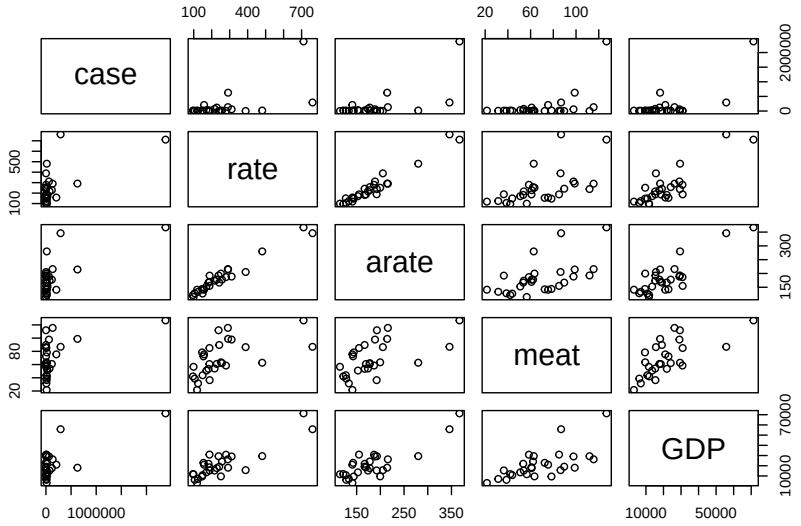
6.2.2.1 Il consumo di carne causa il cancro?

I dati nel file `cancer.csv` contengono alcune variabili quantitative relative a dati del 2021 per i paesi delle Americhe.

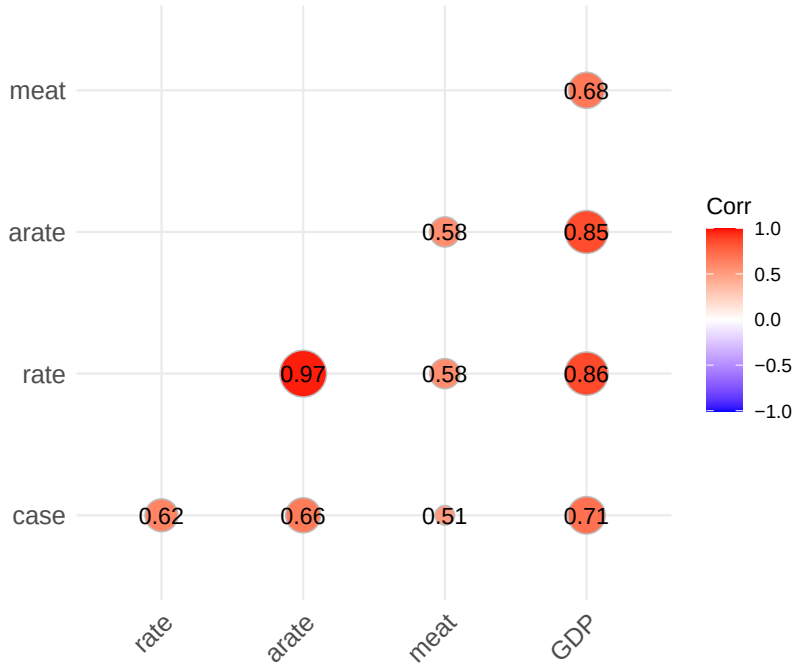
```
america<-read.csv("spurious/cancerdata.csv", header=T)
kable(america[,-c(1,6,9)])
```

country	case	rate	arate	meat	GDP
Argentina	133420	290.0	215.8	115.47999	26300.273
Bahamas	955	238.4	192.7	111.89726	29340.473
Barbados	1120	388.9	205.3	86.37877	15683.885
Belize	409	99.2	114.7	56.61901	11677.461
Bolivia	17579	146.6	143.8	78.36997	9454.198
Brazil	627193	291.2	214.4	98.84000	18075.705
Canada	292098	760.9	345.9	86.85000	55781.700
Chile	59876	311.0	188.7	97.78003	29105.102
Colombia	117620	228.3	177.6	60.94000	17700.060
Costa Rica	13325	257.1	177.7	62.17001	24140.434
Dominican Republic	20171	182.4	167.0	53.52998	21912.260
Ecuador	30888	170.5	152.7	50.91001	13539.711
El Salvador	9799	149.6	127.1	43.85003	10809.714
Guatemala	17801	95.8	121.8	42.08999	11828.330
Guyana	1225	154.3	142.5	72.41991	22865.932
Haiti	13860	118.7	141.1	21.54999	3134.950
Honduras	10815	105.8	128.0	38.75001	6202.981
Jamaica	7500	251.3	199.6	63.40998	9577.164
Mexico	207154	157.5	140.9	75.42000	21031.709
Nicaragua	8409	124.0	133.5	31.52997	7086.623
Panama	8353	187.8	154.9	85.04012	30932.764
Paraguay	13783	188.7	192.2	36.43999	15406.167
Peru	72827	216.2	173.8	54.01001	15282.273
Saint Lucia	448	242.0	166.8	89.60352	19101.166
Suriname	1119	187.5	169.5	60.79010	18458.234
Trinidad and Tobago	3931	279.5	186.7	58.49958	30826.209
United States	2380189	710.9	367.0	126.83000	71318.305
Uruguay	16817	481.0	279.9	62.76991	29441.477

```
daticanc<-america[,c(3,4,5,7,8)]
corcanc<-round(cor(daticanc),2)
pairs(daticanc)
```




```
ggcorrplot(corcanc, type="lower", method="circle", lab=TRUE)
```

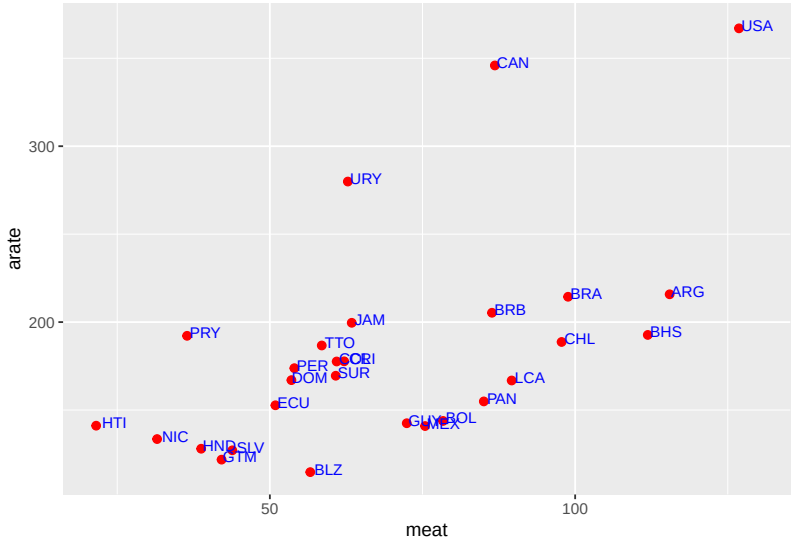


Vi sono correlazioni elevate fra le variabili coinvolte. In particolare, si osservi la correlazione fra la variabile `meat`che è il consumo di carne pro-capite nei diversi paesi e `arate` che è il tasso di tumori al colon-retto nei diversi paesi (aggiustato per la diversa struttura di età).

Il coefficiente di correlazione lineare è 0.68, elevato e positivo. Vale la

pena di osservare anche il diagramma di dispersione più in dettaglio.

```
ggplot(america, aes(x=meat, y=arate)) +
  geom_point(colour="red", size=2) + # Show dots
  geom_text(
    label=america$Code, color="blue", size=3,
    nudge_x = 3, nudge_y = 1.75,
    check_overlap = F
  )
```



Vi è quindi una associazione fra il consumo di carne e il numero di tumori all'apparato intestinale. Ma è questo un indizio che il consumo di carne possa essere uno dei fattori che causano i tumori al colon-retto? L'evidenza contenuta nei dati sulle due variabili non consetirebbe di trarre tale conclusione: si tratta di dati osservazionali e non sperimentali e quindi l'associazione potrebbe esser dovuta all'effetto di altre variabili.

Ad esempio,in questo caso uno può osservare che entrambe le variabili, `meat` e `arate`, sono anche fortemente correlate con la variabile `GDP` che rappresenta il Prodotto Interno Lordo (PIL in Italia, in inglese GDP, Gross Domestic Product) pro-capite nei diversi paesi.

In effetti, il cancro è una malattia degenerativa e si osserva una maggiore incidenza della malattia in paesi più ricchi nei quali la vità è più lunga e nei quali è meno frequente che si contraggano anche in età giovane malattie che si rilevano fatali come accade in paesi meno ricchi. In alto a destra nel grafico vedete rappresentati i paesi più "ricchi".

Dovremmo potere quindi misurare la correlazione fra `meta`e `arate` avendo eliminato l'influenza della terza variabile che è un fattore comune che determina sia alti livelli di casi di tumore che alti livelli di consumo della carne: si tratterebbe di valutare la relazione fra le due variabili **al netto** della ricchezza, cioè della variabile `GDP`.

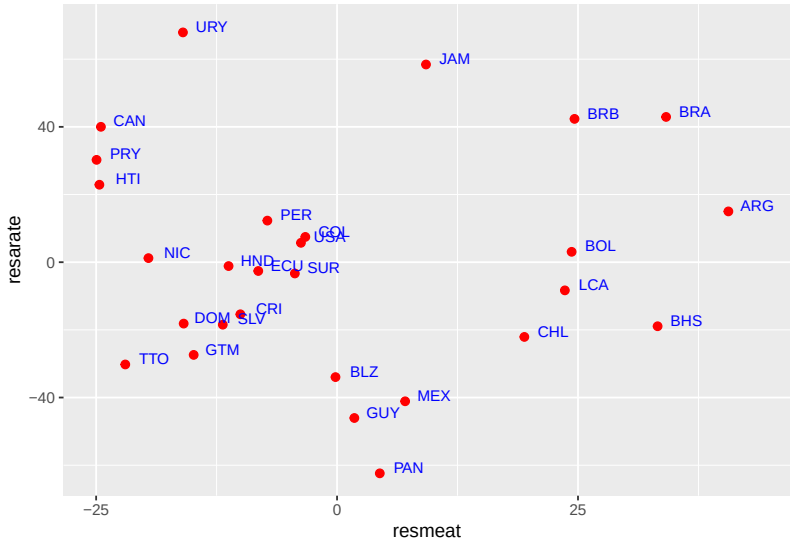
Questo è possibile misurando la **correlazione parziale** fra due variabili al netto di una terza. L'idea è di calcolare la correlazione fra le variabile avendo prima eliminato l'influenza lineare della terza. Per fare questo possiamo ricorrere a una semplice idea:

1. si consideri la regressione lineare della variabile `meat`su `GDP` e otteniamo i residui di questa. Si ottiene la variabile consumo di carne `meat` dalla quale si è eliminata l'influenza lineare del reddito.
2. si passa poi a fare la stessa cosa con la variabile `arate`; anche in questo caso si calcolano i residui della regressione del tasso di tumori sul reddito: tali valori non risentiranno più della relazione lineare con il reddito.
3. si calcola infine la correlazione lineare fra i residui delle due regressioni che rappresentano le due variabili al netto dell'influenza comune del livello di reddito.

```
resarate<-residuals(lm(arate~GDP,data=america))
resmeat<-residuals(lm(meat~GDP,data=america))
cor(resarate,resmeat)
```

```
## [1] 0.004255976
```

```
# si noti che il coefficiente è ora quasi pari a zero
# evidenziando assenza di correlazione lineare

ggplot(america, aes(x=resmeat, y=resarate)) +
  geom_point(colour="red", size=2) + # Show dots
  geom_text(
    label=america$Code, color="blue", size=3,
    nudge_x = 3, nudge_y = 1.75,
    check_overlap = F
  )
```



Il coefficente di correlazione appena ottenuto è detto coefficiente di correlazione parziale fra `meat` e `arate` al netto di `GDP` e evidenzia come la correlazione esistente fra le due variabili prese isolatamente potrebbe esser dovuta al fatto che esse condividono una forte relazione lineare con la ricchezza che le determina entrambe: paesi più ricchi hanno sia un alto consumo di carne che un tasso di tumori più elevato (la popolazione è più anziana).

Esistono formule dirette per calcolare i coefficienti di correlazione lineare parziale fra due variabili al netto di una terza in funzione dei coefficienti di correlazione lineare semplice che sono detti anche *coefficienti di correlazione totale*.

Si usa denotare il coefficiente di correlazione parziale (o netto) fra due variabili x e y al netto di una terza z con $r_{xy.z}$ ed esso può essere calcolato nel modo visto o direttamente, attraverso i coefficienti di correlazione lineare semplici usando la formula:

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}}$$

Analogamente si potrebbe definire un coefficiente di correlazione parziale al netto di più variabili, riprenderemo tale concetto in seguito.

6.3 La Regressione lineare multipla

L'analisi di regressione lineare con due variabili (detta regressione semplice) può essere estesa al caso in cui si vogliano inserire fra le variabili esplicative altre variabili.

In particolare quindi, nei modelli di regressione lineare multipla viene definita una variabile risposta (rigorosamente quantitativa), per distinguerla dalle altre variabili la denoteremo con Y e sarà Y_i il valore osservato per l'i-esima unità. Possono essere però inserite molte variabili fra le esplicative (e queste come vedremo possono essere sia qualitative che quantitative). Tali variabili verranno denotate in generale con X, Z, etc., o, meglio con $X_1, X_2, \dots, X_p$.

L'idea di fondo resta la stessa: vogliamo usare una semplice funzione (di più variabili) per descrivere la media della variabile risposta. Quindi in generale la funzione di **regressione multipla** ha la forma $M(Y|x_1, x_2, \dots, x_p) = f(x_1, x_2, \dots, x_p)$ ove $f()$ è la funzione sulla quale si trovano le medie condizionate della variabile Y in corrispondenza dei valori delle esplicative.

Nell'analisi di **regressione lineare multipla** la funzione $f()$ è assunta essere una funzione lineare ovvero una combinazione lineare delle variabili esplicative con coefficienti β da determinare secondo qualche criterio avendo a disposizione dati sulle variabili $(Y, x_1, x_2, \dots, x_p)$ quindi si porrà

$$M(Y|x_1, x_2, \dots, x_p) = f(x_1, x_2, \dots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Si noti che le variabili x_k che inseriamo possono essere trasformazioni delle variabili: ad esempio, considerando il logaritmo o qualche potenza della variabile originalmente osservata.

Questo implica che nel modello di regressione lineare multipla si può specificare una funzione come:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

Ottenendo così un allontanamento dalla forma lineare della funzione e quindi immaginando che le medie di Y siano collocate lungo una parabola invece che lungo una retta. Si potrebbe pensare anche a una cubica o, più in generale, alla regressione polinomiale. Questo implica che l'analisi di regressione lineare (multipla) è in realtà molto più flessibile di quanto possa apparire a prima vista.

In questa sede vedremo si fornirà una semplice e informale introduzione all'analisi di regressione (multipla) e ai suoi usi. Non verranno approfonditi gli aspetti che richiedono elementi di statistica inferenziale anche se si farà riferimento ad alcuni quantità solo a scopo interpretativo essendo forniti dai principali software. Come già detto, l'analisi di regressione multipla è uno dei principali strumenti della statistica (e non solo) per cui si rinnova il consiglio di seguire un corso ad esso dedicato come quello di **Modelli Statistici** in cui verranno chiariti gli aspetti formali e inferenziali e verrà approfondito il suo uso.

6.3.1 La funzione di regressione dei minimi quadrati

Data una variabile risposta (target) Y e p predittori $x_1, \dots, x_p$, osservati su un campione di n soggetti, la regressione lineare multipla specifica che:

$$M(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$$

Le medie condizionate di Y sono supposte come giacenti su un iperpiano nello spazio $p+1$-dimensionale. L'equazione di regressione sopra contiene i parametri $\beta_0, \beta_1, \beta_2 \cdots + \beta_p$ ed essi vanno determinati utilizzando i dati a disposizione.

Per determinare l'equazione di regressione si può ricorrere ancora al **criterio dei minimi quadrati**, per cui si cercano i valori dei parametri β_j per cui è minima la funzione

$$L(\beta_0, \beta_1, \beta_2 \dots \beta_p) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots + \beta_p x_{ip})^2$$

Per ottenere la soluzione di questo problema di minimo conviene ancora ricorrere alla notazione matriciale per cui si può scrivere la somma sopra come

$$Y - X\beta$$

ove In particolare, si ha

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p-1} \\ \vdots & & & \vdots \\ 1 & x_{n1} & \cdots & x_{np-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

con:

- $\mapsto$ y = vettore $(n \times 1)$ della variabile risposta;
- $\mapsto$ X = matrice $(n \times p+1)$ di regressione contenente i valori delle variabili esplicative, è detta anche *matrice disegno*;
- $\mapsto$ β = vettore $(p+1 \times 1)$ dei parametri (coefficienti) di regressione;

La funzione $L(\beta_0, \beta_1, \beta_2 \cdots + \beta_p)$ può quindi esser riscritta usando le matrici introdotte come

$$L(\beta) = (y - X\beta)^T (y - X\beta)$$

per cui se si deriva rispetto a β e si uguaglia al vettore 0 si ottiene

$$\frac{\partial L(\beta)}{\partial \beta} = (y - X\beta)^T (y - X\beta) = -X^T y + X^T X\beta = 0$$

da cui risolvendo si ottiene

$$\widehat{\beta} = (X^T X)^{-1} X^T y$$

Con $\widehat{\beta}$ si denotano i valori dei coefficienti di regressione ottenuti come soluzione del sistema di equazioni dei minimi quadrati. Si noti che la soluzione esiste se la matrice quadrata $(X^T X)$ è invertibile. Tale condizione si realizza se le variabili (ovvero le colonne) nella matrice X sono linearmente indipendenti così che la matrice quadrata abbia rango pieno.

Problemi potrebbero emergere, ad esempio, se alcune fra le variabili $x_1, x_2, ..., x_p$ fossero perfettamente correlate. In realtà sarebbe bene evitare anche di inserire fra le varibili esplicative variabili con correlazione molto elevata.

6.3.2 Il controllo della qualità della funzione di regressione

6.3.2.1 La significatività dei singoli coefficienti di regressione multipla

Anche in questo caso una variabile introdotta nel modello è di interesse se il coefficientre β ad essa associato ha un valore diverso da 0. Se il coefficiente fosse 0, la variabile non avrebbe rilievo per prevedere la Y e quindi andrebbe omessa. Come sempre a partire dai dati non si otterrà un valore esattamente pari a 0 e si tratta di giudicare se si possa considerare sufficientemente grande da poter escludere che esso sia diverso da zero solo per effetto del caso e dei particolari valori osservati nei nostri dati.

Una risposta rigorosa a tale quesito può essere fornita ricorrendo ai metodi dell'inferenza statistica; essi consentono di ottenere per ogni coefficiente il cosiddetto $p-$ value che può essere interpretato col medesimo criterio introdotto nel caso dell'analisi con due sole variabili: esso, sotto opportune ipotesi relative a un modello probabilistico che generi i dati, consente di valutare la probabilità che, se effettivamente fosse pari a 0 il coefficiente, si osservi un valore ancora più estremo (quindi lontano dallo 0) di quello ottenuto: se tale probabilità e molto piccola allora possiamo ritenere che i dati non forniscono supporto alla congettura che il coefficiente sia pari a zero. La variabile quindi è da ritenersi rilevante (significativa) per prevedere il valore di Y.

Il modello generatore che si adotta per fare inferenza prevede che i valori Y_i osservati derivino da

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i$$

dove ϵ_i sono valori tratti da variabili aleatorie per le quali valgono alcune assunzioni (ad esempio, tali variabili aleatorie hanno valore atteso pari a 0, hanno varianza costante per ogni i, sono incorrelate $\forall i \neq j$ e, talvolta, si assume che siano variabili gaussiane). L'affidabilità delle valutazioni sulla significatività dipende, in genere, da quanto sono realistiche tali assunzioni.

6.3.2.2 Il coefficiente di determinazione lineare R^2

Per l'analisi di regressione multipla vale la scomposizione della varianza già introdotta nel caso della regressione con due variabili. Pertanto:

$$Dev(Y) = DEV(residua) + DEV(spiegata) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - M((Y)))^2$$

e quindi si può ancora proporre l'indice

$$R^2 = \frac{DEV(spiegata)}{Dev(Y)} = 1 - \frac{DEV(residua)}{Dev(Y)}$$

Quanto più $R^2% è vicino a 1 tanto migliore è l'approssimazione della funzione ai dati osservati. Si noti che tale indice aumenterà sempre se si introducono nuove variabili, al più resta costante. Per tale motivo spesso si considera anche R^2_C detto R^2 corretto o aggiustato (*adjusted*).

$$R^2_C = R^2\text{corretto} = 1 - (1 - R^2) * \frac{n-1}{n-p}$$

potrebbe anche decrescere se una nuova variabile inserita non è significativa.

R^2_C avrà quindi un valore sempre inferiore o uguale a quello di R^2. A differenza di R^2, l'R^2_C aumenta solo quando l'aumento di R^2 (a seguito dell'inclusione di una nuova variabile esplicativa) è maggiore di quanto ci si aspetterebbe di vedere per caso. Se un insieme di variabili esplicative con una gerarchia di importanza predeterminata venissero introdotte in una regressione una alla volta, calcolando ogni volta l'R^2 aggiustato, il livello al quale R^2_C aggiustato raggiunge un massimo, e poi potrebbe diminuire; quando ciò accade l'analisi di regressione realizza il compromesso ideale fra capacità esplicativa e complessità del modello evitando di aggiungere variabili esplicative superflue.

L'R^2_C corretto risolve il compromesso (trade off) bias-varianza. Quando consideriamo le prestazioni di un modello, un errore inferiore, misurato in questo caso dalla riduzione della devianza residua, rappresenta una prestazione migliore. Quando il modello

diventa più complesso, esso diviene tuttavia più instabile e meno affidabile nel fare previsione per nuovi dati (c'è il cosidetto rischio di sovra adattamento): aumenta cioè la varianza legata al modello.

Un modello più semplice invece sarà spesso più distante dalla realtà, cioè sarà più distorto (biased). L'errore totale, quando si usa il modello per fare previsioni, dipende dalla somma di queste due componenti. Un R^2 elevato indica un errore di distorsione inferiore perché il modello può prevedere meglio la variazione per i valori di Y osservato se si usano numerosi predittori. Nel frattempo, il modello tende ad essere più complesso e sarà più difficile che si adatti a nuove situazioni (e prevedere con nuovi dati). Guardare a R^2_C corretto è quindi opportuno perchè di fatto si penalizza la complessità della equazione di regressione se le nuove componenti introdotte non aggiungono molta capacità predittiva.

Inoltre spesso per decidere fra modelli alternativi evitando l'eccessiva complessità, si fa ricorso a degli indici detti *criteri di informazione*: un esempio è l'**AIC**. Si tratta di una misura che tiene conto del compromesso tra la bontà dell'adattamento e la complessità del modello (trade-off bias-varainza) ma esso richiede però alcune assunzioni sul modello generatore: più piccolo è l'AIC, migliore è il modello.

6.3.2.3 Selezione (automatica) del modello

Esistono vari metodi per rendere più agevole la scelta del modello quando sono presenti molte potenziali variabili di input. Fra questi ricordiamo quelli basati sull'inserimento o la eliminazione progressiva di regressori nell'equazione. Ad esempio:

- **regressione backward**: inizia con tutti i predittori
  1. determina i parametri dell'equazione di regressione
  2. se sono presenti coefficienti non significativi, rimuovere il predittore corrispondente con il $p-$value più alto e tornare al punto 1
- **regressione forward**: inizia con la regressione con la sola intercetta (detto modello nullo)
  1. aggiunge un nuovo candidato predittore (scegliendo in prima istanza quello più correlato)
  2. determina i parametri dell'equazione di regressione
  3. se il coefficiente associato alla variabile è significativo la si mantiene nell'equazione e si prova a introdurre un nuovo predittore.
- **regressione stepwise**: una combinazione delle precedenti. Si inizia come nella regressione forward ma ad ogni nuova inclusione di predittori l'equazione di regressione viene controllata ed eventualmente i predittori non significativi vengono rimossi, co-

me nell'eliminazione backward. Quest'ultima procedura è utile (è presente inoltre un'apposita opzione nei software perl'analisi di regerssione), ma è da usare con cautela.

Esiste inoltre una versione di tali procedure per passi (stepwise appunto) che procede analogamente basandosi però sulla valutazione dell'AIC.

6.3.2.4 L'analisi dei residui

I residui da un'analisi di regressione sono definiti come

$$\begin{aligned} r_i &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x) \\ &= y_i - \hat{y}_i \quad i = 1, \dots n \end{aligned}$$

Essi danno preziose indicazioni su aspetti che potrebbero esser migliorati.

Se l'analisi funziona adeguatamente, essi dovrebbero:

- essere vicini a 0
- mostrare sia valori positivi che negativi (si ricordi che la loro somma è pari a 0).

Di solito si guarda a uno scatterplot in cui si mettono in ordinata i valori dei residui e in ascissa i valori di $\hat{y}_i$, cioè r_i contro $\hat{y}_i$.

In tale grafico i valori dovrebbero:

- "ballare" attorno a 0 e distribuirsi in un modo che potremmo definire casuale senza mostrare alcuna regolarità
- non mostrare quindi pattern caratteristici (ad esempio, piu piccoli nella parte destra del grafico e più grandi nella parte sinistra, con una forma a imbuto)

Altri grafici dei residui interessanti sono:

- r_i contro x_i
- QQ-plot di r_i contro la gaussiana (qqnorm)
- il grafico dei residui standardizzati, delle azioni di leva (leverage) e delle distanze di Cook: si tratta di indici che evidenziano la presenza di valori anomali e/o influenti (ovvero la cui omissione cambierebbe in modo non irrilevante la determinazione dell'equazione di regressione)

6.3.2.5 La natura delle variabili coinvolte e l'interpretazione dei parametri di regressione

La variabile Y, come detto, deve essere una variabile quantitativa perchè l'analisi abbia senso, di tipo continuo (o anche discreta ma con un numero molto elevata di valori distinti)

Le variabili x_j possono invece essere: + quantitative + categoriali (fattori qualitativi)

Se x_j è quantitativa allora il coefficiente β_j associato ha la stessa interpretazione che per il modello lineare semplice: rappresenta di quanto in media cambia Y se aumento di una unità la variabile X_j. Si faccia attenzione però che nel modello di regerssione multipla sono coinvolte altre variabili per cui il coefficiente β_j misura l'impatto della variabile x_j sulla media di Y immaginando di tenere ferme le altre variabili (al netto delle altre variabili) incluse nel modello. Il coefficiente β_j è infatti, per tale motivo, detto coefficiente di regressione parziale: ed esso non sarà lo stesso coefficiente che otterrei in un modello di regressione semplice che includa la sola variabile x_j.

Se invece x_j è categoriale, un fattore con K livelli $l_1, \dots, l_K$ può essere codificata con $k-1$ variabili indicatrici:

$$d_{ik} = \begin{cases} 1 & \text{quando } x_{ij} = l_k \\ 0 & \text{altrimenti} \end{cases}$$

for $k = 1, \dots, K-1$.

Se quindi x_j è un fattore categoriale scrivere

$$\begin{aligned} y_i &= \beta_0 + \beta_j x_j + \epsilon_i \qquad \text{non avrebbe senso, mentre si può scrivere} \\ &= \beta_0 + \alpha_1 d_{i1} + \dots + \alpha_k d_{ik} + \dots + \alpha_{K-1} d_{iK-1} + \epsilon_i \\ &= \beta_0 + \alpha_k + \epsilon_i \text{ quando } x_j = l_k \text{for } k = 1, \dots K-1 \\ &= \beta_0 \text{ quando } x_j = l_K \text{ è assunto come livello di riferimento} \end{aligned}$$

Il parametro α_k va interpretato come il cambiamento medio y associato al passaggio di x_i dal livello l_K, assunto come riferimento, al livello l_k.

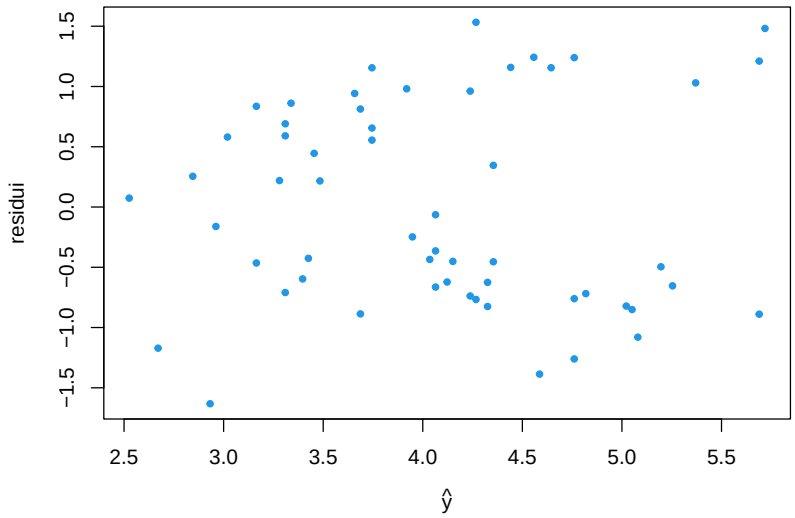
6.3.3 Un esempio con l'aggiunta di una seconda variabile esplicativa categoriale

Si riconsiderino i dati `whiteside`. I dati su Gas e temperature sono stati raccolti in due condizioni, prima e dopo l'intervento di isolamento della casa (questa informazione è nella variabile `Insul` che presenta due modalità "Before" e "After").

Si riprenda l'analisi di regressione svolta e si calcolino i residui

```
library(MASS)
data(whiteside)
linm <- lm(Gas~Temp, data=whiteside)
res <- residuals(linm)
yhat <- fitted(linm)
```

```
par(mar=c(4,4,2,1))
plot(yhat, res, pch=20, col=4, xlab=expression(hat(y)), ylab = "residui")
```



Sebbene non appaiono grosse anomalie, il grafico dei residui suggerisce alcuni problemi

1. La variabilità dei residui non è costante
2. Sembra che ci siano due gruppi di residui

- I residui positivi aumentano all'aumentare di $\hat{y}$
- I residui negativi diminuiscono all'aumentare di $\hat{y}$

Esistono due gruppi di osservazioni nei quali la relazione presente tra consumo di riscaldamento e la temperatura esterna è diversa da quella precedentemente ottenutata (i residui presentano un andamento in cui sono ancora funzione della temperatura tramite $\hat{y}$) e non dovrebbero visto che la variabile temperatura è stata già introdotta.

Si è tralasciato qualcosa di rilevante?

Si consideri la terza variabile, quindi si ha:

- `Gas` (la variabile risposta che denoteremo con Y)
- `Temperature` (che denoteremo con x)
- `Insul` (che denoteremo con z)

I dati disponibili sono quindi

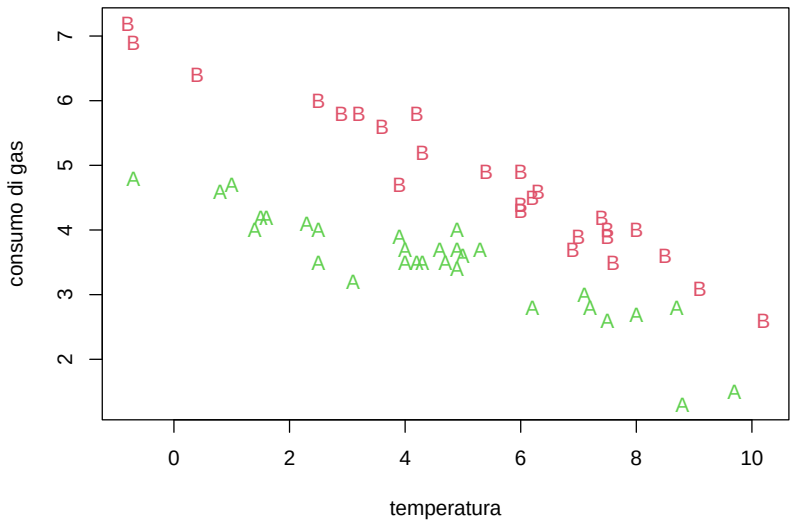
$$(x_1, z_1, y_1), (x_2, z_2, y_2), \dots, (x_i, z_i, y_i), \dots, (x_n, z_n, y_n)$$

con $z_i =$ "Before" for $i = 1, \dots, 26$, and $z_i =$ "After" for $i = 27, \dots, 56$.

È ragionevole aspettarsi che l'intervento di isolamento abbia un impatto sul consumo medio di gas per riscaldamento e che dopo l'intervento il consumo di gas sia inferiore rispetto a prima.

Si crei la variabile z a partire da $Insul$

```
z <-  whiteside[,1]
par(mar=c(4,4,2,1))
plot(whiteside$Temp, whiteside$Gas, pch=as.character(z),
     col=as.numeric(z)+1, xlab="temperatura", ylab="consumo di gas")
```



Considerando la terza variabile il nuovo modello può essere genericamente specificato come segue:

$$\begin{aligned} \text{Consumo di gas} &= g(\text{temperatura}, \\ &\quad \text{prima/dopo l'intervento}) \\ y_i &= g(x_i, z_i; \epsilon_i) \end{aligned}$$

Essendo presenti ora 32 variabili esplicative si è passati a un'analisi di regressione lineare multipla

$$M(Gas_i) = \beta_0 + \beta_1 Temp + \beta_2 Insul_i$$

Quindi la media condizionata di Y_i è ora una funzione lineare di due variabili.

Tuttavia la variabile aggiuntiva che è stata inserita $Insul$ è qualitativa con modalità "Before" e "After" e il modello scritto come sopra non avrebbe senso in quanto $Insul$ non ha valori numerici. L'approccio standard per superare il problema è quello di introdurre una

variabile ausiliaria (indicatrice) definita come:

$$d_i = \begin{cases} 0 & se \quad Insul_i = \text{"Before"} \\ 1 & se \quad Insul_i = \text{"After"} \end{cases}$$

- la precedente equazione di regressione allora diventa:

$$\begin{aligned} M(Gas_i) &= \beta_0 + \beta_1 Temp_i + \beta_2 d_i \\ &= \beta_0 + \beta_1 x_i \quad \text{prima dell'intervento} \\ &= (\beta_0 + \beta_2) + \beta_1 Temp_i + \quad \text{dopo l'intervento} \end{aligned}$$

In altre parole l'introduzione della variabile fittizia dà origine a due rette parallele, una per ogni valore di $Insul_i$

La determinazione dei parametri del modello tramite i minimi quadrati si estende facilmente al seguente modello di regerssione lineare multipla:

$$min_{\beta_0,\beta_1,\beta_2} \sum_{i=1}^{n} (Gas_i - (\beta_0 + \beta_1 Temp_i + \beta_2 d_i))^2$$

Per ottenere i valori dei minimi quadrati per i tre parametri si può ancora usare la funzione `lm()`. La sintassi è la stessa e la variabile `Insul` viene trasformata automaticamente nella variabile ausiliaria vista prima.

```
whiteside$Insul <- relevel(whiteside$Insul, ref="Before")
# possiamo
mlm <- lm(Gas~Temp+Insul, data=whiteside)
beta0 <- coef(mlm)[1]
beta1 <- coef(mlm)[2]
beta2 <- coef(mlm)[3]

beta0 # l'intercetta
```

```
## (Intercept)
##    6.551329
```

```
beta1 # il coefficiente lineare della regressione di Gas su Temp
```

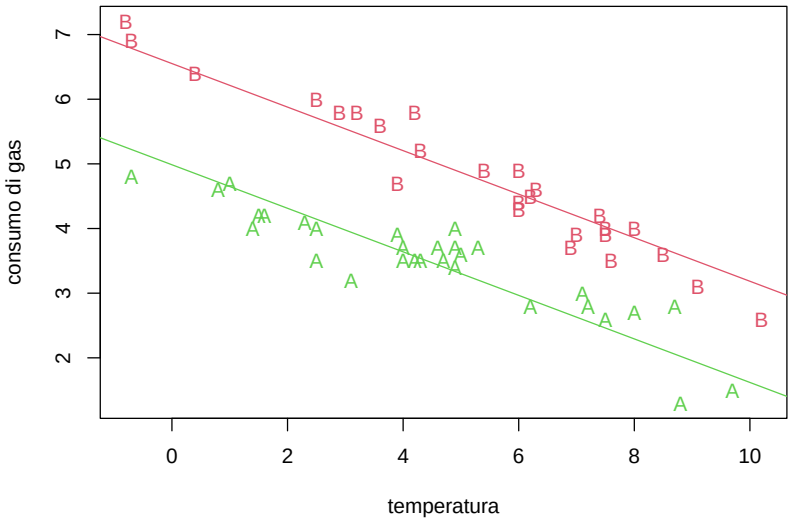
```
##      Temp
## -0.336697
```

```
beta2 # misura l'effetto della terza variabile
```

```
## InsulAfter
##  -1.565205
```

Come interpretare β_2?

```
options(width=60)
par(mar=c(4,4,2,1))
plot(whiteside$Temp,whiteside$Gas, pch=as.character(z), col=as.numeric(z)+1,
     xlab="temperatura", ylab="consumo di gas")
abline(beta0, beta1, col=2)
abline(beta0+beta2, beta1, col=3)
```



Si consideri, l'interpretazione dei parametri ottenuti con i minimi quadrati (indicati col "cappelletto"): $\hat{\beta}_0$: Consumo medio di Gas quando tutti predittori sono pari a 0. In questo caso ha un senso interpretarlo perchè 0 è un valore che fa parte del range dei valori dei predittori. Spesso l'intercetta non ha un'interpretazione: se la temperatura è 0 gradi, e siamo prima dell'isolamento ($d_i = 0$), il consumo medio di gas sarebbe 6.6 metri cubi.

$\hat{\beta}_j$ ($j > 1$): in generale i valori dei coefficienti associati alla $j-$ma variabile misurano la variazione media di y quando la $j-$ma variabile esplicativa è incrementata di 1 unità e tutti le altre variabili predittive sono tenute costanti. Quindi se la temperatura esterna aumenta di un grado, il consumo medio di gas decresce di circa 0.34 metri cubo, indipendentemente dal fatto che si tratti di una misura ottenutra prima o dopo l'isolamento. Infine, se le case vengono isolate (d_i passa da 0, cioe da "Before", a 1, cioè "After"), il consumo medio di gas decresce di circa 1.57 metri cubi indipendentemente dalla temperatura.

Ovviamente, possiamo usare la funzione ottenuta per ottenere una previsione del consumo di Gas per qualsiasi valore delle variabili predittrici (possibilmente limitandosi al range dei valori osservati)

$$\hat{Gas} = M(y) = \hat{\beta}_0 + \hat{\beta}_1 Temp + \hat{\beta}_2 d$$

Ad esempio, qual è il consumo medio del gas quando la temperature è 5? - se la casa non è isolata

$$\begin{aligned} \hat{y} &= 5.486 - 0.2902x \\ &= 6.551 - 0.3367 \cdot 5 = 4.8675 \text{ metri cubi} \end{aligned}$$

- se la casa è isolata:

$$\begin{aligned} \hat{y} &= 5.486 - 0.2902x \\ &= 6.551 - 1.565 - 0.3367 \cdot 5 = 3.3025 \text{ metri cubi} \end{aligned}$$

Un ulteriore estensione del modello, sempre limitandoci alle sole due variabili esplicative disponibili, sarebbe quella di introdurre il prodotto fra le variabili `Temp`e la varabile indicatrice d che è legata all'isolamento.

L'equazione di regressione diviene:

$$M(Gas_i) = \beta_0 + \beta_1 Temp_i + \beta_2 Insul_i + \beta_3 Insul_i * Temp_i$$

Si noti che in questo caso l'equazione ottenuta equivarrebbe alla seguente situazione

$$\begin{aligned} M(Gas_i) &= \beta_0 + \beta_1 Insul_i + \beta_2 d_i \\ &= \beta_0 + \beta_1 x_i \quad \text{prima dell'intervento} \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) Insul_i + \quad \text{dopo l'intervento} \end{aligned}$$

Ci sono cioè due diversi effetti della temperatura nei due regimi (prima e dopo l'isolamento). Questo può essere specificato in `R` introducendo l'**interazione** fra le due variabili `Temp` e `Insul`.

```
mlmint <- lm(Gas~Temp+Insul+Temp*Insul, data=whiteside)
summary(mlmint)
```

```
##
## Call:
## lm(formula = Gas ~ Temp + Insul + Temp * Insul, data = whiteside)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97802 -0.18011  0.03757  0.20930  0.63803
```

```
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.85383    0.13596  50.409  < 2e-16 ***
## Temp            -0.39324    0.02249 -17.487  < 2e-16 ***
## InsulAfter      -2.12998    0.18009 -11.827 2.32e-16 ***
## Temp:InsulAfter  0.11530    0.03211   3.591 0.000731 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.323 on 52 degrees of freedom
## Multiple R-squared:  0.9277, Adjusted R-squared:  0.9235
## F-statistic: 222.3 on 3 and 52 DF,  p-value: < 2.2e-16
```

```
beta0 <- coef(mlmint)[1]
beta1 <- coef(mlmint)[2]
beta2 <- coef(mlmint)[3]
beta3 <- coef(mlmint)[4]

beta0 # l'intercetta
```

```
## (Intercept)
##    6.853828
```

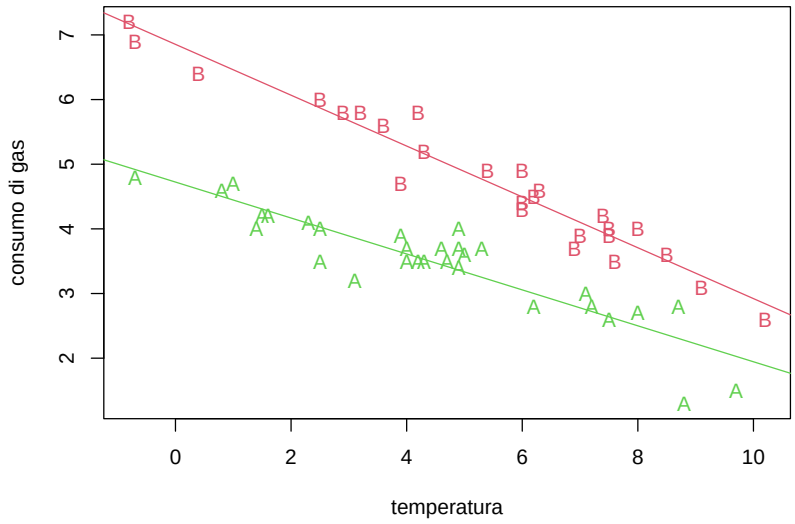
```
beta1 # il coefficiente lineare della regressione di Gas su Temp
```

```
##       Temp
## -0.3932388
```

```
# beta2 e beta3 # misurano le variazioni che l'introduzione
#                della terza variabile comporta nell'inetrecetta e nel
#                coefficiente angolare
```

Si noti che tale modello migliora le equazioni precedenti (si vedano i coefficienti e la loro significatività). Graficamente equivale a:

```
options(width=60)
par(mar=c(4,4,2,1))
plot(whiteside$Temp,whiteside$Gas, pch=as.character(z), col=as.numeric(z)+1,
     xlab="temperatura", ylab="consumo di gas")

abline(beta0, beta1, col=2)
abline(beta0+beta2, beta1+beta3, col=3)
```



6.3.4 Regressione multipla: un esempio guidato con R

In questo esempio, si esaminano i dati della Survey of Consumer Finances (SCF), indagine condotta in USA su un campione di 275 cittadini che hanno acquistato una polizza vita e per i quali si raccolgono informazioni sul loro reddito e sulle loro caratteristiche demografiche.

i dati sono disponibili su https://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/data.html

Obiettivo dell'analisi è determinare quali caratteristiche influenzino la spesa per polizze vita. Per quanto riguarda le polizze temporanee in caso morte (term life insurance), l'ammontare che viene rilevato è FACE, ovvero quanto la compagnia pagherà in caso di morte dell'assicurato.

Le informazioni disponibili sono riassunte nella seguente tabella:

VARIABILE	DESCRIZIONE
AGE	Age of the survey respondent
MARSTAT	Marital status of the respondent (single/not single)
EDUCATION	Number of years of education of the survey respondent
NUMHH	Number of household members
INCOME	Annual income of the family
FACE	Amount that the company will pay in the event of the death of the named insured

Col seguente comando carichiamo i dati sul workspace di R:

```
TL <- read.csv("TL.csv", header=TRUE, sep=",", row.names=1)
```

Consideriamo la struttura dell'oggetto TL.

```
str(TL)
```

```
## 'data.frame':    275 obs. of  6 variables:
##  $ AGE      : int  30 50 39 43 34 29 72 51 58 73 ...
##  $ MARSTAT  : chr  "not single" "not single" "not single" "not single" ...
##  $ EDUCATION: int  16 9 16 17 11 16 17 16 14 12 ...
##  $ NUMHH    : int  3 3 5 4 4 3 2 4 1 2 ...
##  $ INCOME   : int  43000 12000 120000 40000 28000 100000 112000 15000 32000 250
##  $ FACE     : int  20000 130000 1500000 50000 220000 600000 100000 2500000 2500
```

```
# conviene trasformare in fattore la variabile MARSTAT
TL$MARSTAT<-factor(TL$MARSTAT)
```

La variabile risposta che consideriamo è `FACE`, e come prima possibile variabile esplicativa usiamo solo `INCOME` che è ragionevole pensare che sia legata alla prima.

Vediamo la sintesi delle due variabili

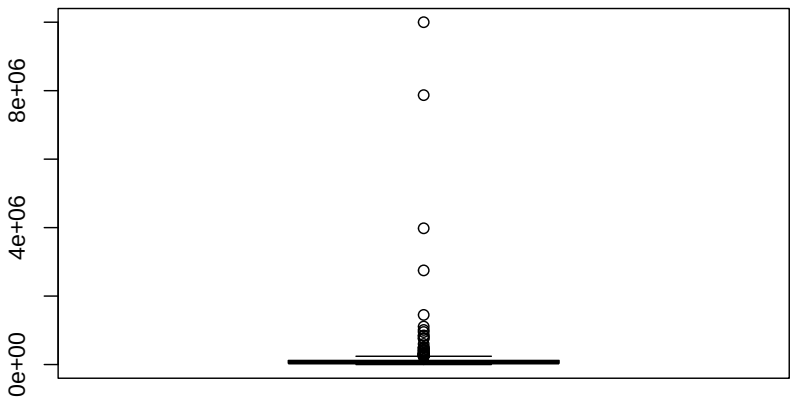
```
summary(TL$FACE)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##      800    50000   150000   747581   590000 14000000
```

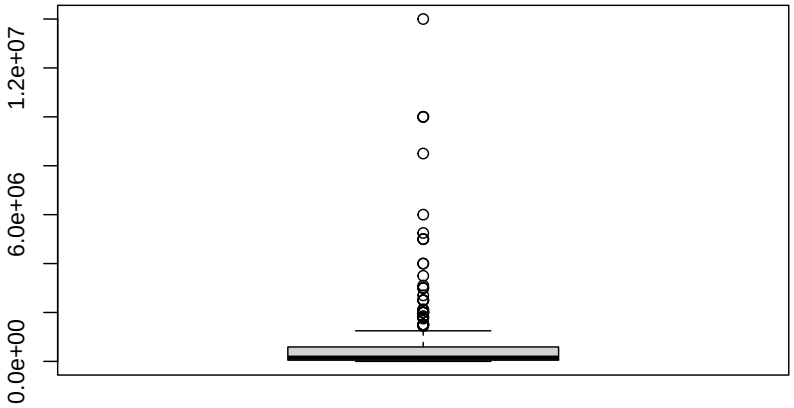
```
summary(TL$INCOME)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##      260    38000    65000   208975   120000 10000000
```

```
boxplot(TL$INCOME)
```

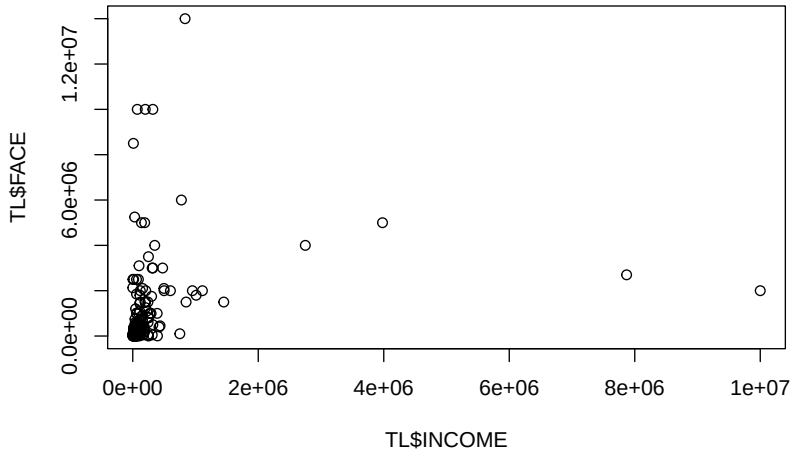



```
boxplot(TL$FACE)
```



Le due variabili sembrano avere una forte asimmetria positiva (si vedano media e mediana, ove la seconda è di molto minore della prima) e questo appare chiaramente anche dal diagramma di dispersione:

```
plot(TL$INCOME, TL$FACE)
```



la grande maggioranza dei valori sono in basso a sinistra mentre solo pochi valori, molto elevati appaiono in alto, 'e una conseguenza della forte asimmetria, e questo maschera la relazione fra le variabili.

6.3.5 Costruzione dell'equazione di regressione

Come primo banale esercizio, proviamo la regressione lineare semplice di `FACE`, vs `INCOME`, ignorando per ora il problema dell'asimmetria nella distribuzione delle due variabili. L'equazione di regressione ha la forma

$$M(FACE_i) = \beta_0 + \beta_1 INCOME_i$$

Utilizzare `lm()`

```
m0<- lm(TL$FACE~TL$INCOME)
```

Una sintesi dei risultati della stima dei parametri si può visualizzare con il comando generico `summary`:

```
summary(m0)
```

```
##
## Call:
## lm(formula = TL$FACE ~ TL$INCOME)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3069185  -628950  -551689  -167579 12976542
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.553e+05  1.019e+05   6.433 5.59e-10 ***
## TL$INCOME   4.414e-01  1.200e-01   3.677 0.000284 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1637000 on 273 degrees of freedom
## Multiple R-squared:  0.04718,    Adjusted R-squared:  0.04369
## F-statistic: 13.52 on 1 and 273 DF,  p-value: 0.0002843
```

Innanzitutto si può provare a dare un'interpretazione dei coefficienti ottenuti

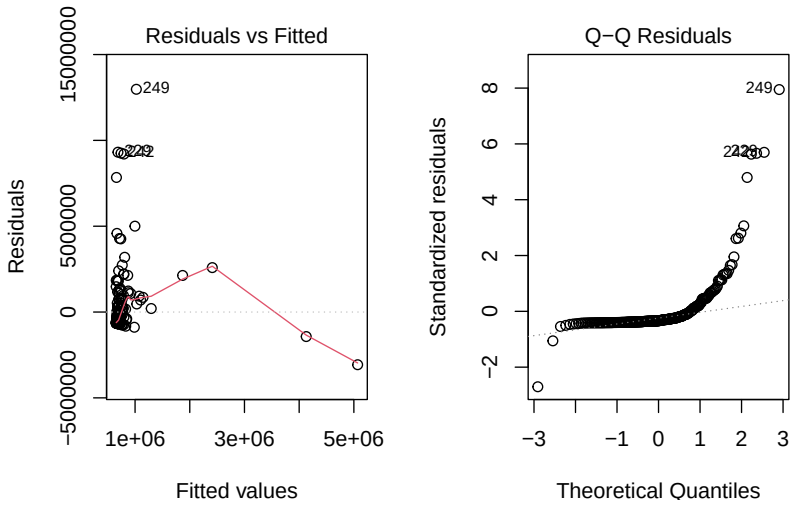
- $\hat{\beta}_0$: questo parametro rappresenta il valore atteso dell'ammontare del valore della polizza quando x=0, ed è pari a 655300. Non ha molto senso interpretarlo in questo caso poiché y ha sempre valori positivi.
- $\hat{\beta}_1$: è positivo, è indica una relazione positiva fra reddito e valore delle polizze acquistate. Ci dice di quanto aumenta il valore atteso di y se ho un incremento unitario del predittore 1. Quando "'INCOME"' aumenta di un dollaro, la media di "'FACE"' aumenta di 0.44.

La tabella consente di valutare agevolmente se i parametri sono **significativamente** differenti da zero. Si guardano ad esempio i $p-$values e si conclude che il parametro è significativamente diverso da 0 se essi sono molto piccoli (sotto 0.01 ad esempio) ovvero è molto bassa la probabilità di ottenere valori ancora più **estremi** del para-

metro nell'ipotesi che questo coefficiente sia in realtà pari a 0. In particolare, $\hat{\beta}_1$ è significativo e conferma che il reddito ha un effetto sull'ammontare delle polizze acquistate e non è plausibile pensare che questo sia dovuto al caso.

Tuttavia si noti che R^2 'e quasi 0. Inoltre, si può guardare ai residui

```
par(mfrow=c(1,2))
plot(m0, which=1:2)
```

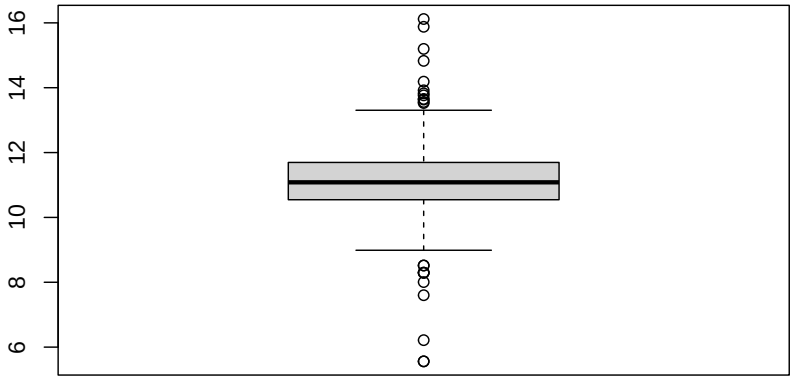


Tali grafici confermano che tale equazione di regressione non è del tutto convincente. I residui non seguono un pattern regolare e non sembrano adattarsi a una gaussiana (primi due grafici).

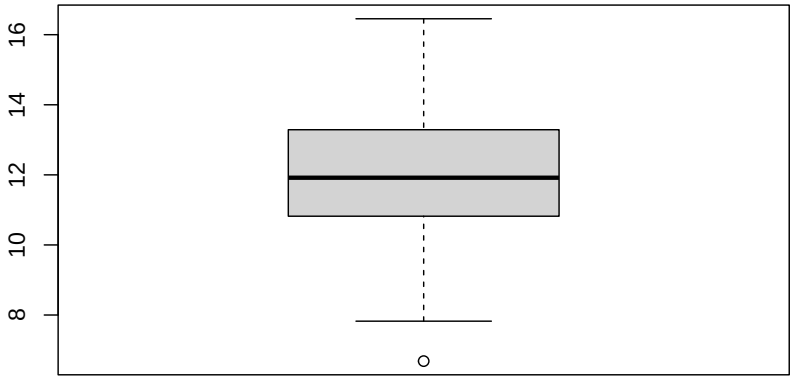
Una possibile causa del non soddisfacente adattamento è la presenza di variabili fortemente asimmetriche. In questo caso potrebbe rivelarsi più opportuno utilizzare una trasformazione delle variabili che riduca l'asimmetria.

```
TL$LFACE <- log(TL$FACE)
TL$LINCOME <- log(TL$INCOME)
```

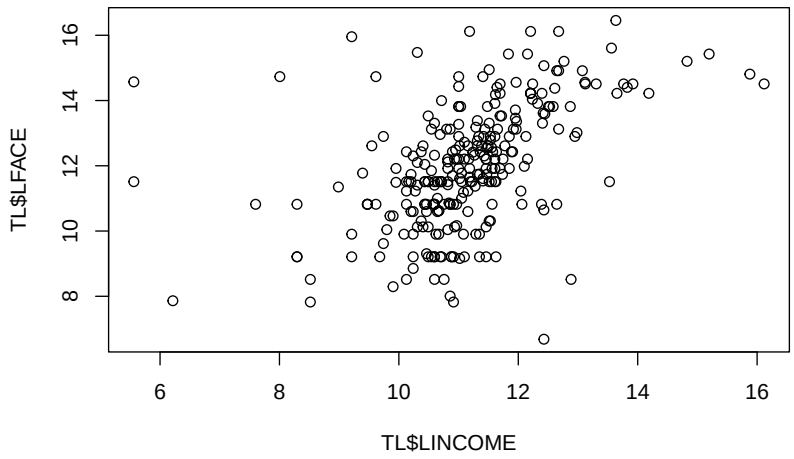
```
boxplot(TL$LINCOME)
```




```
boxplot(TL$LFACE)
```




```
plot(TL$LINCOME, TL$LFACE)
```



Prima di specificare una nuova analisi di regressione, può essere comodo rimpiazzare nella matrice dei dati (data frame), le variabili trasformate.

Si ottengano i parametri della seguente equazione di regressione lineare semplice dopo avere trasformato le variabili

$$M(\log(FACE_i)) = \beta_0 + \beta_1 \log(INCOME_i)$$

```
m1 <- lm(LFACE~LINCOME, data=TL)
```

```
summary(m1)
```

```
##
## Call:
## lm(formula = LFACE ~ LINCOME, data = TL)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1967 -0.8032 -0.0018  0.8954  6.4711
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.23003    0.85985   4.920  1.5e-06 ***
## LINCOME      0.69604    0.07661   9.086  < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.642 on 273 degrees of freedom
## Multiple R-squared:  0.2322, Adjusted R-squared:  0.2294
## F-statistic: 82.55 on 1 and 273 DF,  p-value: < 2.2e-16
```

Nell'interpretazione dei coefficienti è necessario ricordare che le variabili sono ora espresse nei logaritmi

- $\hat{\beta}_0$: Si noti che il predittore log(“‘INCOME“‘) è pari a 0 se “‘INCOME“‘=1. Tuttavia anche y è su scala logaritmica, e se si vuole una corretta interpretazione è necessario ritrasformare in dollari $\Rightarrow$ quando “‘INCOME“‘ = 1, il valore atteso di “‘FACE“‘ è $exp(4.23) \sim 69$ dollari. In questo caso l'interpretazione ha tuttavia più senso che nel caso del modello “‘m0“‘.
- $\hat{\beta}_1$: come ci si aspettava è positivo. Tuttavia ora quando log(“‘INCOME“‘) aumenta di 1, $\log(FACE)$ aumenta circa di 0.7. Ma se si valuta cosa accade nelle variabili originarie allora

  $$0.7 \simeq log(FACE)_{x+1} - log(FACE)_x = log\left(\frac{FACE_{x+1}}{FACE_x}\right) \Rightarrow \left(\frac{FACE_{x+1}}{FACE_x}\right) \simeq \exp(0.7) \simeq 2.01.$$

  Quindi un incremento unitario di log(“‘INCOME“‘) corrisponde a raddoppiare l'ammontare di polizza. <!–

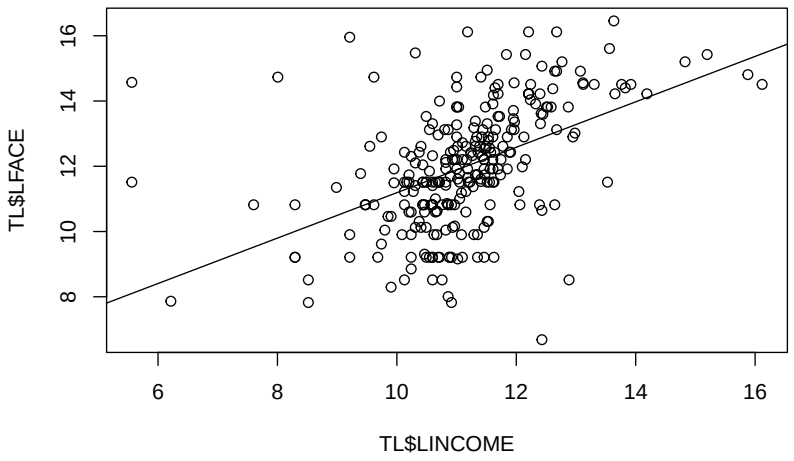
Ancora una volta i parametri ottenuti, e in particolare $\hat{\beta}_1$ sono *significativamente* diversi da zero ($p-$values sempre molto piccoli).

```
exp(11.41)
```

```
## [1] 90219.42
```

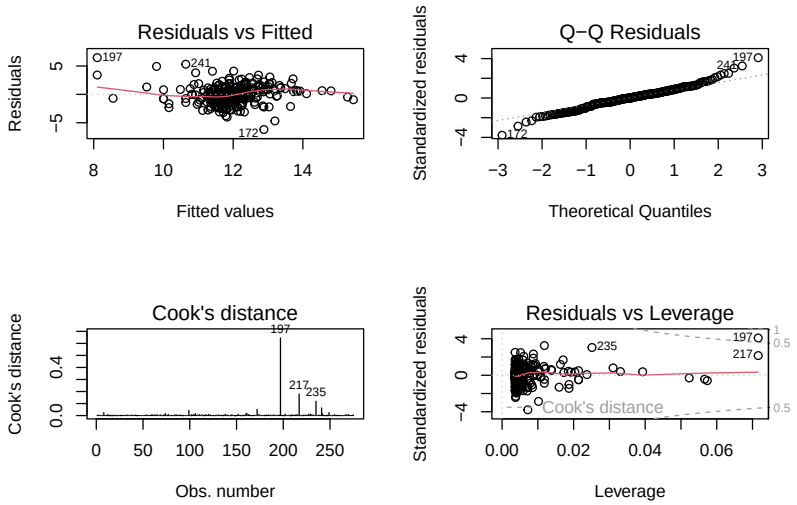
Si può sovrapporre la retta con i coefficienti stimati sul grafico come segue:

```
beta<- coef(m1)
plot(TL$LINCOME, TL$LFACE)
abline(beta[1], beta[2])
```



e dare uno sguardo ai residui

```
par(mfrow=c(2,2))
plot(m1, c(1,2,4,5))
```




```
par(mfrow=c(1,1))
```

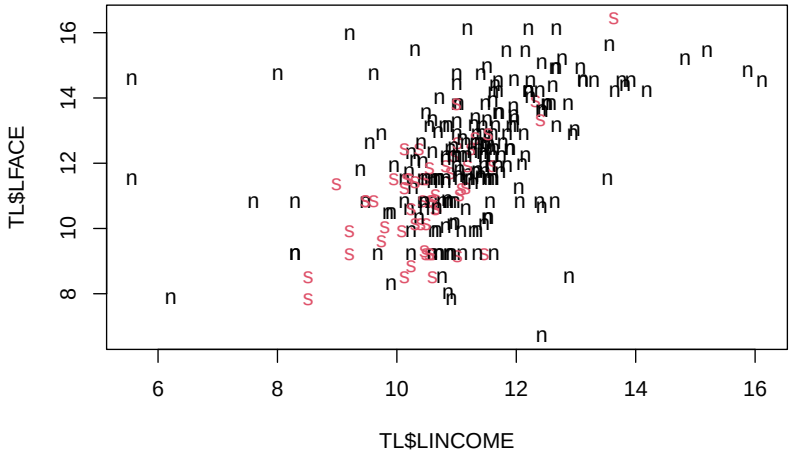
I grafici vanno meglio e R^2 è più elevato. Ci sono però ancora margini di miglioramento:

- il normal qq-plot mostra che le code della distribuzione dei residui non sembrano adattarsi alla normale;
- R^2 è ancora basso in assoluto;
- la linea che interpola lo scatterplot "'LFACE, vs "'LINCOME, non sembra soddisfacente,

6.3.5.1 L'inserimento di predittori categoriali

Il data frame contiene altre variabili di potenziale valore predittivo. Per esempio, essere sposati o meno potrebbe avere qualche rilievo. la variabile `MARSTAT` assume solo due valori: `single` e `not single`. Si può tracciare lo scatterplot `LFACE` vs `LINCOME` e usare simboli diversi per le due categorie della variabile `MARSTAT`.

```
plot(TL$LINCOME, TL$LFACE, pch=as.character(TL$MARSTAT), col=as.numeric(TL$MARSTAT
```



Le due nuvole di punti si sovrappongono ma si vede abbastanza bene che i single tendono a avere polizze meno elevate. Si vede anche che la relazione fra reddito e polizza è meno marcata per le coppie cio'e potrebbe esserci un'interazione fra `MARSTAT` e `LINCOME`. Si provi allora l'equazione di regressione

$$M(\log(FACE_i)) = \beta_0 + \beta_1 \log(INCOME)_i + \beta_2 MARSTAT + \beta_3 \log(INCOME)_i \cdot MARSTAT$$

```
m2 <- lm(LFACE~LINCOME+MARSTAT+LINCOME*MARSTAT, data=TL)
print(summary(m2), digits=3, signif.stars = FALSE)
```

```
##
## Call:
## lm(formula = LFACE ~ LINCOME + MARSTAT + LINCOME * MARSTAT, data = TL)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.215 -0.829  0.070  0.931  5.607
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
```

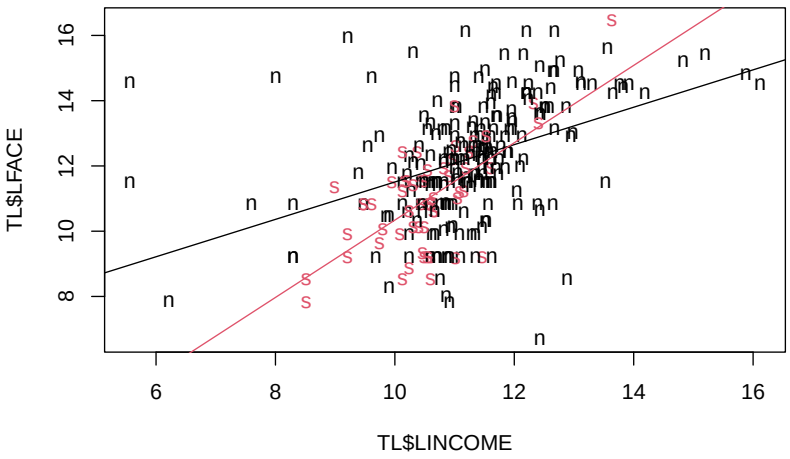
```
## (Intercept)             5.7790     0.9255    6.24  1.6e-09
## LINCOME                 0.5729     0.0812    7.05  1.5e-11
## MARSTATsingle          -7.2921     2.7422   -2.66   0.0083
## LINCOME:MARSTATsingle   0.6124     0.2576    2.38   0.0181
##
## Residual standard error: 1.6 on 271 degrees of freedom
## Multiple R-squared:  0.276,  Adjusted R-squared:  0.268
## F-statistic: 34.4 on 3 and 271 DF,  p-value: <2e-16
```

l'equazione risulta essere:

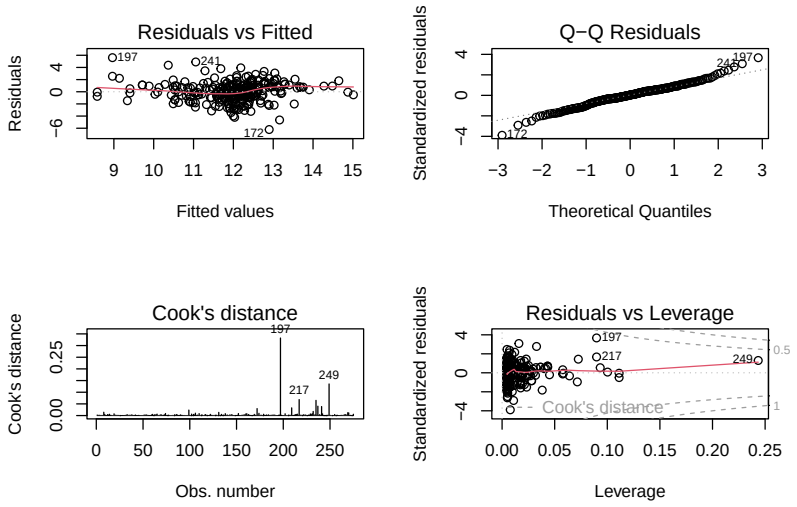
$$\begin{aligned}
\log(\hat{FACE}_i) &= \hat{\beta}_0 + \hat{\beta}_1 \log(INCOME)_i + \hat{\beta}_2 MARSTAT + \hat{\beta}_3 \log(INCOME)_i \cdot MARSTAT \\
&= \hat{\beta}_0 + \hat{\beta}_1 \log(INCOME)_i \qquad \text{when } i \text{ is not single} \\
&= 5.7790 + 0.5729 \log(INCOME)_i \\
&= (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3) \log(INCOME)_i \qquad \text{when } i \text{ is single} \\
&= (5.7790 - 7.2921) + (0.5729 + 0.6124) \log(INCOME)_i \\
&= -1.5131 + 1.1853 \log(INCOME)_i
\end{aligned}$$

In altre parole, i single comprano meno polizze per i redditi bassi ma se il reddito cresce l'ammontare delle polizze aumenta più che negli sposati e quindi per redditi alti i single comprano in media più polizze.

```
beta<- coef(m2)
plot(TL$LINCOME, TL$LFACE, pch=as.character(TL$MARSTAT), col=as.numeric(TL$MARSTAT
abline(beta[1], beta[2])
abline(beta[1]+beta[3], beta[2]+beta[4], col=2)
```




```
par(mfrow=c(2,2))
plot(m2, c(1,2,4,5))
```




```
par(mfrow=c(1,1))
```

–> –>

E' possibile a questo punto arricchire l'analisi con le altre variabili disponibili (`EDUCATION` e `NUMHH` e `AGE`) e verificare se si abbia un ulteriore miglioramento della capacità predittiva.

```
m3 <- lm(LFACE~LINCOME+MARSTAT+LINCOME*MARSTAT+NUMHH+EDUCATION+AGE, data=TL)
print(summary(m3), digits=3, signif.stars = FALSE)
```

```
##
## Call:
## lm(formula = LFACE ~ LINCOME + MARSTAT + LINCOME * MARSTAT +
##     NUMHH + EDUCATION + AGE, data = TL)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.718 -0.790  0.166  0.879  4.619
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.09101    1.00424    4.07  6.1e-05
## LINCOME                0.40596    0.08131    4.99  1.1e-06
## MARSTATsingle         -7.27081    2.58856   -2.81  0.00534
## NUMHH                  0.25345    0.07429    3.41  0.00075
## EDUCATION              0.20370    0.03839    5.31  2.3e-07
## AGE                   -0.00464    0.00795   -0.58  0.55996
## LINCOME:MARSTATsingle  0.63895    0.24426    2.62  0.00940
##
## Residual standard error: 1.5 on 268 degrees of freedom
```

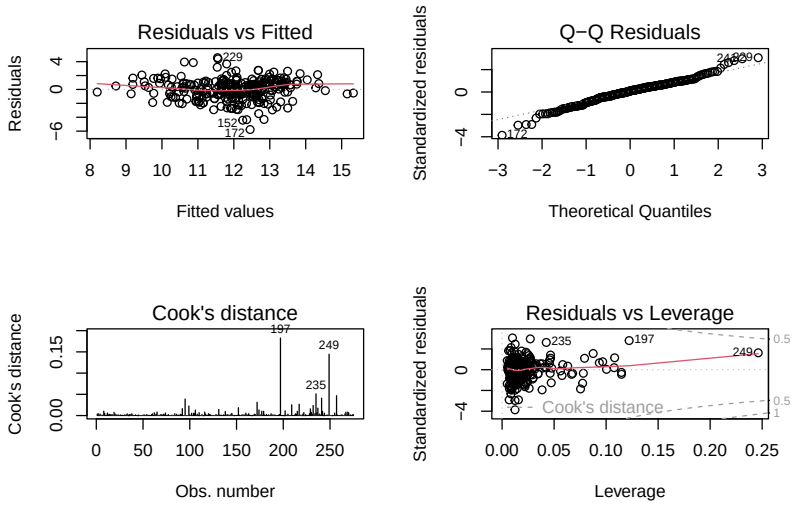
```
## Multiple R-squared:  0.369,  Adjusted R-squared:  0.355
## F-statistic: 26.1 on 6 and 268 DF,  p-value: <2e-16
```

si noti che l'età non sembra aver un effetto significativo per cui essa può essere esclusa dall'equazione di regressione.

```
m4 <- lm(LFACE~LINCOME+MARSTAT+LINCOME*MARSTAT+NUMHH+EDUCATION, data=TL)
print(summary(m4), digits=3, signif.stars = FALSE)
```

```
##
## Call:
## lm(formula = LFACE ~ LINCOME + MARSTAT + LINCOME * MARSTAT +
##     NUMHH + EDUCATION, data = TL)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.769 -0.762  0.140  0.919  4.572
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.8589     0.9210    4.19  3.8e-05
## LINCOME                 0.4036     0.0811    4.98  1.2e-06
## MARSTATsingle          -7.2195     2.5839   -2.79  0.00558
## NUMHH                   0.2689     0.0694    3.88  0.00013
## EDUCATION               0.2027     0.0383    5.29  2.5e-07
## LINCOME:MARSTATsingle   0.6364     0.2439    2.61  0.00959
##
## Residual standard error: 1.5 on 269 degrees of freedom
## Multiple R-squared:  0.368,  Adjusted R-squared:  0.356
## F-statistic: 31.3 on 5 and 269 DF,  p-value: <2e-16
```

```
par(mfrow=c(2,2))
plot(m4, c(1,2,4,5))
```



Il comando permette di ottenere ancora i grafici dei residui. Si vede che ora la situazione e molto migliorata. Tuttavia sono presenti alcuni valori anomali (quelli cui sono associati valori del residuo di Cook elevato e con effetto leva elevato). Anche se il modello potrebbe essere suscettibile di ulteriori ritocchi, esso appare ora accettabile.

L'equazione finale ottenuta 'e quindi:

$$\begin{aligned}
\log(\hat{FACE}_i) &= \hat{\beta}_0 + \hat{\beta}_1 \log(INCOME)_i + \hat{\beta}_2 MARSTAT_i + \\
&\quad + \hat{\beta}_3 \log(INCOME)_i \cdot MARSTAT_i + \hat{\beta}_4 EDUCATION_i + \hat{\beta}_5 NUMHH_i \\
&= \hat{\beta}_0 + \hat{\beta}_1 \log(INCOME)_i + \hat{\beta}_4 EDUCATION_i + \hat{\beta}_5 NUMHH_i \\
&= 3.86 + 0.40 \log(INCOME)_i + 0.20 EDUCATION_i + 0.27 NUMHH_i \\
&\qquad \text{quando } i \text{ non è single} \\
&= (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3) \log(INCOME)_i + \hat{\beta}_4 EDUCATION_i + \hat{\beta}_5 NUMHH_i \\
&= (3.86 - 7.21) + (0.40 + 0.64) \log(INCOME)_i + 0.20 EDUCATION_i + 0.27 NUMHH_i \\
&= -3.35 + 1.04 \log(INCOME)_i + 0.20 EDUCATION_i + 0.27 NUMHH_i \\
&\qquad \text{quando } i \text{ è single}
\end{aligned}$$

6.4 Introduzione all'analisi di raggruppamento

Si consideri un insieme di variabili $x_1, x_2, \ldots, x_p$ misurate su n unità. Piuttosto che considerare la relazione fra una variabile risposta e un insieme di esplicative, come, ad esempio, per la regressione multipla,

siamo principalmente interessati a scoprire altri aspetti relativi alle variabili osservate.

Uno di questi riguarda l'individuazione di sottogruppi di osservazioni (o talvolta di variabili) che sono 'omogenee' secondo un determinato criterio.

Le tecniche di questo tipo appartengono a quelle denominate anche di **apprendimento non supervisionato** e includono, fra le principali, l'analisi di raggruppamento **(cluster analysis)** e l'analisi delle componenti principali.

L'analisi di raggruppamento, in particolare, consiste in un'ampia classe di metodi per individuare sottogruppi nei dati non noti a priori. Il problema di base del clustering può essere affermato come segue:

Dato un insieme di dati vogliamo suddividerli in un insieme di gruppi tali che i dati in ciascun gruppo sono il più possibile simili tra loro e i dati di gruppi diversi siano il più possibile diversi tra loro.

I metodi per affrontare tale problema possono condurre a risultati diversi a seconda dell'approccio, per la specifica definizione utilizzata per definire la somiglianza tra due unità, per l'algoritmo usato per derivare i cluster (metodi gerarchici e non gerarchici) e in relazione allo specifico tipo di dati (alcune tecniche sono applicabili solo a dati quantitativi, alcuni altri metodi possono essere utilizzati su dati qualitativi o anche su dati misti).

6.4.1 Un semplice esempio con dati simulati

L'esempio che segue utilizza dati simulati: essi sono costituiti dal numero medio di volte in cui il cibo fuori casa viene acquistato dai ristoranti (mensile) e dalle spese per il cibo consumato a casa (annuale) per un insieme di 100 famiglie. Prima si leggano i dati dal file `food_spending.csv`, che contiene le colonne `FreqAway` e `Home`, e poi si ottenga il grafico di dispersione dei dati:

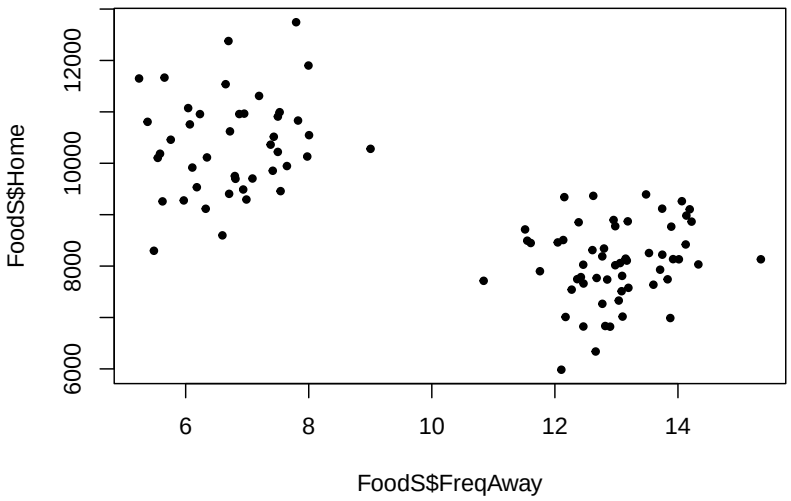
```
FoodS<-read.csv("food_spending.csv", header=TRUE)
head(FoodS)
```

```
##    FreqAway      Home
## 1  6.951280 10965.524
## 2  7.415451  9854.342
## 3 12.462907  6824.662
## 4 14.136932  8982.304
## 5  5.969204  9276.950
## 6  7.976161 10131.322
```

```
dim(FoodS)
```

```
## [1] 100   2
```

```
plot(FoodS$FreqAway, FoodS$Home, pch=20)
```



I dati del grafico sembrano raggrupparsi in due gruppi ben separati, il gruppo con basse spese per il cibo a casa e alta frequenza di cibo fuori casa, e il gruppo di famiglie con una grande spesa per il cibo a casa e una bassa media dei pasti fuori casa.

In un problema di raggruppamento si mira ad identificare gruppi (cluster) distinti sulla base delle misurazioni ottenute su un insieme di variabili. I gruppi risultanti sono tali che i membri di uno stesso gruppo sono omogenei rispetto alle variabili coinvolte, mentre le osservazioni in gruppi diversi sono abbastanza diverse l'una dall'altra. Per rendere questo concreto, dobbiamo definire tuttavia cosa significa che due osservazioni qualsiasi siano 'simili' o 'dissimili', per caratterizzare i diversi gruppi.

6.4.2 Misure di dissomiglianza

In `R`, la libreria `cluster` può essere utilizzata per calcolare le dissomiglianze (distanze) a coppie tra le osservazioni nel set di dati tramite la funzione `daisy` inoltre contiene, come si vedrà, diversi metodi per condurre l'analisi dei cluster.

```
library(cluster)  # load cluster library
```

Come primo esempio, si calcoleranno le **distanze euclidee** tra i record (righe) del set di dati `FoodS` (opzione di default per l'argomento `metric` è la distanza euclidea)

```
dist1<-daisy(FoodS)
?daisy
```

Data una matrice numerica o un dataframe con n righe e p colonne, le dissomiglianze vengono in questo caso calcolate tra le righe, risultando in una matrice simmetrica $n \times n$, con valore nella cella (i, j) dato da

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

In questo esempio si hanno due variabili ($p = 2$) e $n = 100$ misurazioni. Le prime righe/colonne della matrice di dissomiglianza per i dati simulati possono essere ottenute come

```
as.matrix(dist1)[1:5, 1:5]
```

```
##          1         2        3         4         5
## 1    0.000 1111.1824 4140.866 1983.2338 1688.5747
## 2 1111.182    0.0000 3029.684  872.0644  577.3939
## 3 4140.866 3029.6844    0.000 2157.6424 2452.2967
## 4 1983.234  872.0644 2157.642    0.0000  294.7596
## 5 1688.575  577.3939 2452.297  294.7596    0.0000
```

Si noti che la scala di misura in genere influisce in modo determinante sui risultati dell'analisi di raggruppamento. Per questo motivo, in alcuni casi, si consiglia di standardizzare i dati prima di eseguire una procedura di clustering, in modo che tutte le variabili siano su una scala comparabile. Questo può essere fatto, ad esempio, trasformando le variabili in modo che abbiano tutte media zero e deviazione standard uno. Nel caso in cui le variabili siano tutte misurate nelle stesse unità di misura, si potrebbe scegliere di non standardizzarle.

Si noti che la funzione `daisy` può procedere alla standardizzazione specificando l'argomento `stand=TRUE` ma viene utilizzata la deviazione media assoluta anziché la deviazione standard:

```
dist2<-daisy(FoodS, stand=TRUE)
```

L'oggetto che si ottiene appartiene alla classe "dist", tuttavia può essere trasformato in una matrice come segue

```
dist2m<-as.matrix(dist2)
dist2m[1:5, 1:5]
```

```
##           1         2        3        4         5
## 1 0.0000000 0.9536446 3.945398 2.891414 1.4663512
## 2 0.9536446 0.0000000 3.053032 2.321756 0.6808883
```

```
## 3 3.9453977 3.0530320 0.000000 1.908494 2.9730049
## 4 2.8914145 2.3217564 1.908494 0.000000 2.6862840
## 5 1.4663512 0.6808883 2.973005 2.686284 0.0000000
```

Una funzione `R` alternativa per il calcolo della matrice delle distanze è `dist()`, che usa la metrica specificata dall'argomento `method` per calcolare le distanze tra le righe di una matrice di dati (si possono scegliere tra metodi "euclideo", "massimo", "manhattan", "canberra", "binario" o "minkowski").

La funzione `dist()` per calcolare la matrice 100X100 della distanza euclidea tra le osservazioni dell'esempio:

```
D <- as.matrix(dist(FoodS))
D[1:5, 1:5]
```

```
##          1         2        3         4         5
## 1    0.000 1111.1824 4140.866 1983.2338 1688.5747
## 2 1111.182    0.0000 3029.684  872.0644  577.3939
## 3 4140.866 3029.6844    0.000 2157.6424 2452.2967
## 4 1983.234  872.0644 2157.642    0.0000  294.7596
## 5 1688.575  577.3939 2452.297  294.7596    0.0000
```

Per standardizzare le variabili prima di eseguire il calcolo della dissomiglianza con `dist()`, si può usare la funzione `scale()` che standardizza però dividendo per lo scarto quadratico medio.

```
D <- as.matrix(dist(scale(FoodS)))
D[1:5, 1:5]
```

```
##           1         2        3        4         5
## 1 0.0000000 0.7916579 3.369431 2.631689 1.2216515
## 2 0.7916579 0.0000000 2.639903 2.178157 0.6049498
## 3 3.3694312 2.6399030 0.000000 1.598605 2.6515809
## 4 2.6316888 2.1781567 1.598605 0.000000 2.5489766
## 5 1.2216515 0.6049498 2.651581 2.548977 0.0000000
```

```
# si noti la differenza da quanto ottenuto con daisy
```

La gestione di variabili di tipo misto (ad es. dati binari nominali, ordinali e (a)simmetrici) può essere ottenuta utilizzando il **coefficiente di Gower**, impostando `metric = "gower"` nella funzione `daisy`.

Si utilizza ora il set di dati simulati sui reclami che contiene le seguenti variabili:

- "litig": se c'è un contenzioso
- "soft_injury": se la lesione è stata una lesione dei tessuti molli
- "emergency_tr": se c'è stato un trattamento di emergenza

- "NumProv": numero di provider (numerico)
- "NumTreat": numero di cure mediche (numerico)

Le prime tre sono variabili categoriali, assumendo valore zero per "no" e uno per "sì". Per prima cosa leggiamo i dati e convertiamo le variabili categoriali in fattori

```
d<-read.csv(file="simclust2.csv", header = TRUE)

d$litig<-as.factor(d$litig)
d$soft_injury<-as.factor(d$soft_injury)
d$emergency_tr<-as.factor(d$emergency_tr)

str(d)
```

```
## 'data.frame':    1000 obs. of  6 variables:
##  $ ID          : int  1 1 1 1 1 1 1 1 1 2 ...
##  $ litig       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
##  $ soft_injury : Factor w/ 2 levels "0","1": 2 1 1 1 2 2 1 1 1 2 ...
##  $ emergency_tr: Factor w/ 2 levels "0","1": 1 2 1 1 2 2 2 1 2 2 ...
##  $ NumProv     : int  2 2 0 2 1 0 1 2 2 5 ...
##  $ NumTreat    : int  2 1 2 4 5 3 4 3 1 8 ...
```

Si può utilizzare la metrica di Gower per calcolare la dissomiglianza tra i ricorrenti sulla base di dati misti. La "distanza" tra due unità è la media ponderata dei contributi di ciascuna variabile, dove

- il contributo di una variabile nominale o binaria alla dissomiglianza totale è 0 se entrambi i valori sono uguali, 1 altrimenti
- il contributo delle variabili ordinali codificate come `ordered factor` è la differenza in valore assoluto
- il contributo delle altre variabili è la differenza assoluta di entrambi i valori, divisa per l'intervallo totale di quella variabile. Alle variabili ordinali codificate come `ordered factor` con K livelli, si associa un intero (da 0 a K-1) ad ogni livello secondo l'ordine definito.
- nella formula originale di Gower è possibile assegnare un peso per ogni variabile.

La dissomiglianza basata solo su variabili categoriali è calcolata di seguito:

```
head(d)
```

```
##   ID litig soft_injury emergency_tr NumProv NumTreat
## 1  1     0           1            0       2        2
## 2  1     0           0            1       2        1
## 3  1     0           0            0       0        2
```

```
## 4  1     0           0           0       2        4
## 5  1     0           1           1       1        5
## 6  1     0           1           1       0        3
```

```
d3var<-d[ , c("litig", "soft_injury", "emergency_tr")]
as.matrix(daisy(d3var, metric= "gower"))[1:5, 1:5]
```

```
##           1         2         3         4         5
## 1 0.0000000 0.6666667 0.3333333 0.3333333 0.3333333
## 2 0.6666667 0.0000000 0.3333333 0.3333333 0.3333333
## 3 0.3333333 0.3333333 0.0000000 0.0000000 0.6666667
## 4 0.3333333 0.3333333 0.0000000 0.0000000 0.6666667
## 5 0.3333333 0.3333333 0.6666667 0.6666667 0.0000000
```

Si noti che il peso assegnato a ciascuna variabile è 1 nel nostro esempio. Si calcoli ora la matrice di dissomiglianza sull'intero set di dati (la prima colonna contiene le vere etichette del cluster, quindi è stata scartata).

```
diss<-as.matrix(daisy(d[, -1], metric= "gower"))
diss[1:5, 1:5]
```

```
##           1         2          3          4         5
## 1 0.0000000 0.4142857 0.24000000 0.22857143 0.2628571
## 2 0.4142857 0.0000000 0.25428571 0.24285714 0.2771429
## 3 0.2400000 0.2542857 0.00000000 0.06857143 0.4628571
## 4 0.2285714 0.2428571 0.06857143 0.00000000 0.4342857
## 5 0.2628571 0.2771429 0.46285714 0.43428571 0.0000000
```

6.4.3 I diversi metodi di raggruppamento

Nell'ultimo esempio, oltre alle caratteristiche utilizzate, è possibile includere nel set di dati altre variabili di interesse (ad esempio un sinistro precedente, contraente di polizza) e quindi utilizzare l'analisi di raggruppamento per identificare gruppi di records simili tra loro.

Esiste un gran numero di metodi di raggruppamento. Se ne introdurranno brevemente alcuni che corrispondono a tre approcci per il *clustering* più comunemente utilizzati:

- tecniche di raggruppamento non gerarchico;
- metodi di clustering gerarchico;
- clustering basato sulla densità: l'algoritmo DBSCAN.

6.4.4 I metodi di raggruppamento non gerarchico

6.4.4.1 Il metodo delle K-medie (K-means)

La funzione `kmeans()` esegue il clustering di K-medie in `R`: richiede in input i dati e il numero di gruppi che si vogliono ottenere (centri).

Si consideri ancora il semplice esempio simulato che utilizza i dati in `food_spending.csv`, riguardanti le spese per il cibo a domicilio e la media mensile dei pasti fuori casa per 100 famiglie:

```
x<-read.csv("food_spending.csv", header=TRUE)
head(x)
```

```
##    FreqAway      Home
## 1  6.951280 10965.524
## 2  7.415451  9854.342
## 3 12.462907  6824.662
## 4 14.136932  8982.304
## 5  5.969204  9276.950
## 6  7.976161 10131.322
```

Il metodo di raggruppamento delle K-medie prevede che sia specificato il numero di cluster (si noti che in realtà esso di solito non è noto).

L'algoritmo è piuttosto semplice e procede come segue:

1. sia K il numero di gruppi prefissato (da cui il nome del metodo K-medie).
2. Si scelgono in modo casuale K punti (semi) fra i dati osservati con la sola condizione che non siano coincidenti. Questi fungono da **centroidi** provvisori attorno a cui costruire un'aggregazione dei dati;
3. si calcola la distanza di ogni dato da ogni centroide e il dato viene viene associato al centroide più vicino. Al termine di questa fase avremo K gruppi costituiti attorno ai centroidi;
4. per ogni gruppo, si calcola un nuovo centroide come il vettore di medie
$$M(x_k) = \sum_{\forall i \in C_k} x_i / n_k$$
(ove ove x_i è il vettore dei dati per l'i-esima unità e n_k è la numerosità del k-esimo gruppo) tutti i punti del cluster associato a tale centroide;
5. si itera dal punto 3 fino a quando la soluzione è stabile (nessun dato cambia di classe quando si calcola il nuovo centroide);
6. la funzione obiettivo da minimizzare è la somma delle distanze al quadrato dai centroidi di ciascun gruppo cui il dato è associato, ovvero

$$Dev(entro) = \sum_{k=1}^{K} \sum_{\forall i \in C_k} \|x_i - M(x_k)\|$$

Si noti che tale quantità decresce al crescere di K.

Si illustra ora il funzionamento dell'algoritmo con i dati simulati sopra.

Prima si standardizzano i dati e quindi si esegue il raggruppamento con le K-medie con $K = 2$ centri (la funzione `set.seed()` viene utilizzata per garantire che l'output delle K-medie sia completamente riproducibile così che i centroidi scelti casualmente all'inizio siano gli stessi a ogni prova).

```
x.scalato<-scale(x) # standardiziiiamo i dati
set.seed(4)
km.out <- kmeans(x.scalato, centers=2) # scegliamo una soluzione con 2 gruppi
km.out
```

```
## K-means clustering with 2 clusters of sizes 43, 57
##
## Cluster means:
##     FreqAway       Home
## 1 -1.1045122  0.9206786
## 2  0.8332285 -0.6945470
##
## Clustering vector:
##   [1] 1 1 2 2 1 1 1 1 2 2 1 1 2 2 1 2 2 1 2 1 1 1 2 2 1 2 2
##  [28] 1 2 2 2 2 1 1 2 1 2 2 1 1 1 1 2 1 1 1 2 2 2 1 2 2 2 1
##  [55] 1 1 1 2 2 2 2 1 1 1 2 1 2 1 1 2 2 2 1 2 1 2 2 1 2 2 2
##  [82] 2 2 1 2 2 2 2 2 1 2 2 2 2 1 2 2 1 2 2
##
## Within cluster sum of squares by cluster:
## [1] 21.96630 20.05714
##  (between_SS / total_SS =  78.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"
## [4] "withinss"     "tot.withinss" "betweenss"
## [7] "size"         "iter"         "ifault"
```

L'output mostra:

- le medie dei cluster ottenuti: una matrice KXP, nella quale per ogni gruppo (riga) sono riportate le medie delle variabili;

- il vettore con la attribuzione di gruppo (vettore di clustering): un vettore di interi (da 1 a K) che indica il cluster a cui è allocato ciascuna unità;
- il rapporto tra $\frac{\text{somma dei quadrati}}{\text{somma totale dei quadrati}}$ (cioè la parte della varianza spiegata attraverso il raggruppamento). Avendo raggruppato i dati si ricorda che è possibile calcolare la **devianza tra i gruppi** e la **devianza entro i gruppi** (la loro somma come sappiamo da la **devianza totale**). Il risultato di un algorimo di raggruppamento è tanto migliore quanto maggiore è la devianza entro i gruppi o, in alternativa, quanto più piccola è la *devianza tra i gruppi*. Una funzione obiettivo è quindi il rapporto tar *devianza tra i gruppi* e *devianza totale* che deve essere quanto più grande possibile.

Il vettore con la attribuzione di ogni unità ai cluster è in `km.out$cluster`:

```
str(km.out$cluster)
```

```
##  int [1:100] 1 1 2 2 1 1 1 1 2 2 ...
```

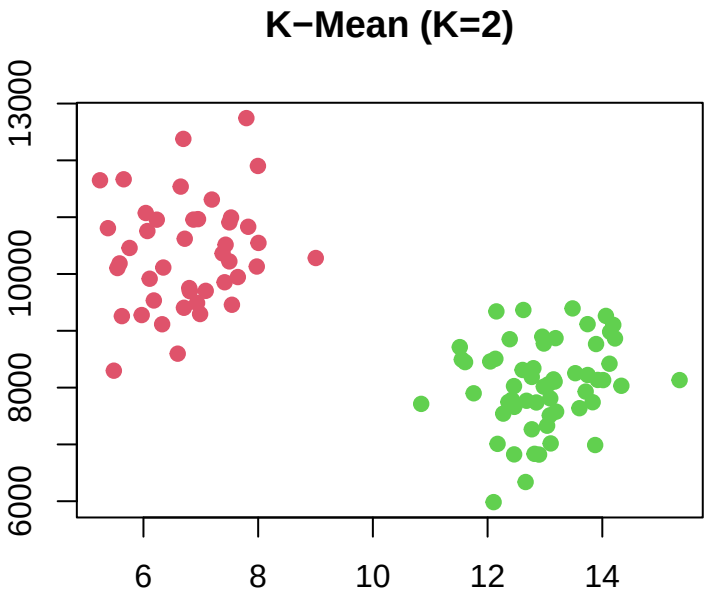
```
table(km.out$cluster)
```

```
##
##  1  2
## 43 57
```

I due cluster ottenuti contengono rispettivamente 56 e 44 osservazioni, le famiglie sono raggruppate in due cluster ben separati.

Poiché, in questo caso, vi sono solo due dimensioni, possiamo ottenere il diagramma di dispersione dei dati, usando colori diversi per i punti in ogni cluster

```
plot(x, col =(km.out$cluster+1),main="K-Mean (K=2)", xlab ="", ylab="",pch=19)
```



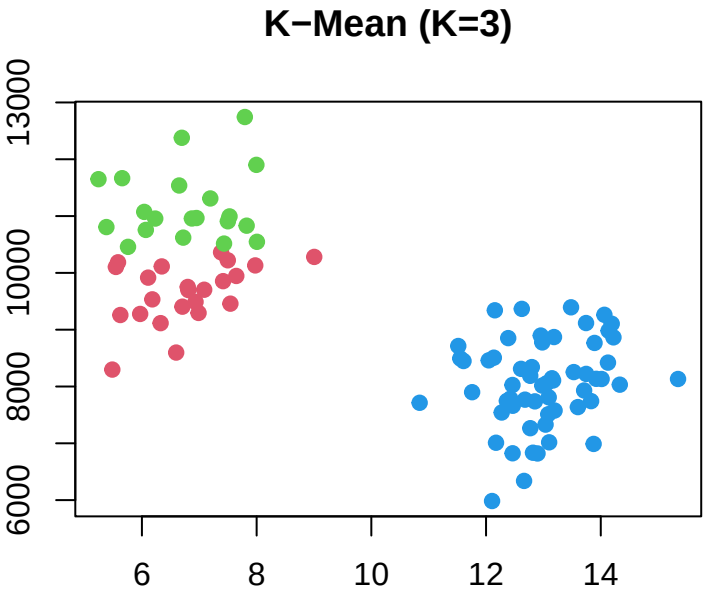
I risultati ottenuti dipenderanno dall'assegnazione iniziale (casuale) del cluster di ciascuna osservazione nel primo passaggio dell'algoritmo. Per questo motivo, una pratica comune consiste nell'eseguire l'algoritmo più volte da diverse centroidi iniziali casuali e scegliere la soluzione per cui la funzione obiettivo (in questo caso la funzione obiettivo è *between_SS/total_SS*) è più piccola. Per fare ciò, si può semplicemente impostare l'argomento `nstart` su un valore molto maggiore di uno (es. `nstart=10` o `nstart=20`), in tal caso `kmeans()` riporterà solo la soluzione migliore.

```
set.seed(5)
km.out<-kmeans(scale(x), 3)
km.out
```

```
## K-means clustering with 3 clusters of sizes 23, 20, 57
##
## Cluster means:
##     FreqAway       Home
## 1 -1.1065155  0.4232302
## 2 -1.1022083  1.4927443
## 3  0.8332285 -0.6945470
##
## Clustering vector:
##   [1] 2 1 3 3 1 1 1 1 3 3 2 2 3 3 2 3 3 1 3 1 2 1 3 3 2 3 3
##  [28] 2 3 3 3 3 1 1 3 2 3 3 1 1 2 1 3 2 1 1 3 3 3 2 3 3 3 1
##  [55] 1 1 2 3 3 3 3 2 1 2 3 1 3 2 2 3 3 3 1 3 1 3 3 2 3 3 3
##  [82] 3 3 2 3 3 3 3 3 2 3 3 3 3 1 3 3 2 3 3
##
```

```
## Within cluster sum of squares by cluster:
## [1]  4.671572  5.057885 20.057140
##  (between_SS / total_SS =  85.0 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"
## [4] "withinss"     "tot.withinss" "betweenss"
## [7] "size"         "iter"         "ifault"
```

```
plot(x, col =(km.out$cluster+1), main="K-Mean (K=3)", xlab ="", ylab="", pch=19)
```



Qui si confrontano le soluzioni usando `nstart=1`e `nstart=20` in termini di somma totale di quadrati delle distanze dai centroidi all'interno del cluster, che cerchiamo di ridurre al minimo eseguendo il clustering di K-medie

```
set.seed(6)
km.out<-kmeans(x, 3, nstart=1)
km.out$tot.withinss
```

```
## [1] 31535467
```

```
km.out<-kmeans(x, 3, nstart=20)
km.out$tot.withinss
```

```
## [1] 31385541
```

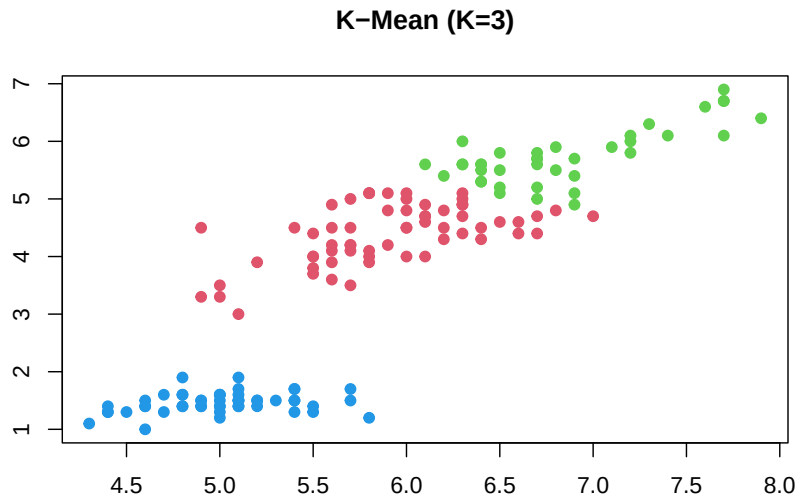
In questo esempio con dati simulati era noto che erano presenti due cluster. Tuttavia, per i dati reali non è noto il numero reale di

cluster. Si sarebbe invece potuto eseguire il clustering K-medie con K variabile da 2 a 10 e confrontare i risultati in termini del rapporto fra *devianza tra i gruppi* e *devianza totale* o del valore della *devianza entro i gruppi*.

Si consideri ora un altro esempio riprendendo i dati `iris`. Anche qui conosciamo il numero dei gruppi, corrispondente alle 3 specie: versicolor, virginica, setosa.

Vediamo se è possibile riconoscerli a partire dalle 4 variabili e, successivamente, calcolare soluzioni con K che va da 2 a 10.

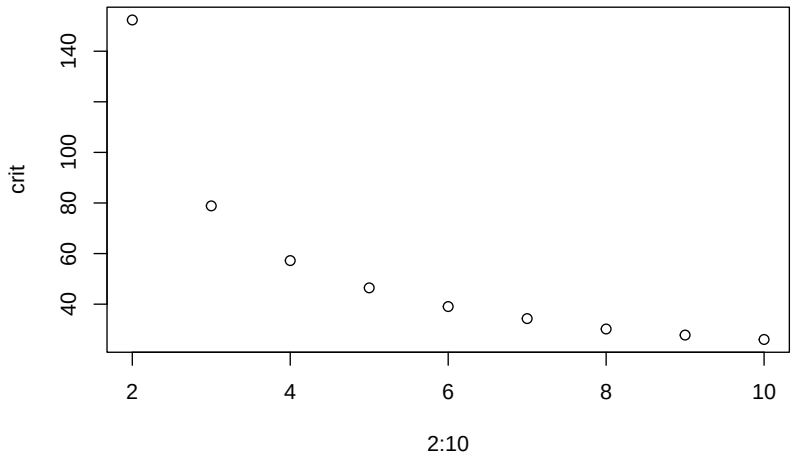
```
set.seed(7)
irisgroup<-kmeans(iris[,-5], 3, nstart=10)
plot(iris$Sepal.Length, iris$Petal.Length, col =(irisgroup$cluster+1),
     main="K-Mean (K=3)", xlab ="", ylab="", pch=19)
```




```
table(irisgroup$cluster,iris$Species)
```

```
##
##     setosa versicolor virginica
##   1      0         48        14
##   2      0          2        36
##   3     50          0         0
```

```
crit<-0
for (i in 2:10) {
set.seed(7)
irisgroup<-kmeans(iris[,-5], i, nstart=10)
crit[i-1]<-irisgroup$tot.withinss
}
plot(2:10, crit)
```



6.4.4.2 Partizionamento intorno ai Medoidi (PAM)

Sempre per l'esempio precedente, si introduce un altro metodo non gerarchico basato su **medoidi**. Si tratta di una versione più robusta di K-means e restituisce i K oggetti rappresentativi, uno per ogni cluster.

L'algoritmo di clustering è il seguente:

- si inizia con una selezione di k punti, detti **medoidi**, con un criterio tale da minimizzare una funzione obiettivo (ad esempio, la somma delle distanze entro i gruppi) e si associa ogni dato al medoide più simile;
- la similarità fra i punti è definita usando distanze come la distanza euclidea, la distanza di Manhattan o la distanza di Minkowski, ma si possono anche usare misure di dissimilarità più generali;
- si selezionano in modo casuale nuovi candidati come medoidi $O^{'}$;
- si calcola il costo totale S_i come la somma delle distanze dei singoli elementi dai corrispondenti medoidi iniziali e il costo totale S_f nel caso dei nuovi medoidi $O^{'}$ e se ne calcola la differenza $S^{'} = S_f - S_i$;
- se $S < 0$ allora si scambia il medoide iniziale con il nuovo (se $S < 0$ ci sarà quindi un nuovo insieme di medoidi);
- si ripetono i passi dal 2 al 5 sino a quando si hanno cambiamenti nell'insieme dei medoidi.

A differenza delle K-medie i medoidi sono sempre punti che corrispondono a dati osservati. Inoltre si possono usare i diversi tipi di distanze. Non c'è infine bisogno di iterare con diverse scelte dei semi iniziali.

La funzione `pam` nel pacchetto **cluster** può essere usata per eseguire il partizionamento attorno ai medoidi. Si noti che questa funzione opera la standardizzazione preliminare delle variabili (ogni colonna del database viene trasformata sottraendo la media e dividendo per lo scarto medio assoluto) se tra gli argomenti è specificato `stand = TRUE` (il default è FALSE). La metrica utilizzata per calcolare le dissomiglianze tra le osservazioni è euclidea (predefinita), l'alternativa è la distanza di Manhattan:

```
pam.out<-pam(x, 2, metric="euclidean", stand=TRUE)
pam.out
```

```
## Medoids:
##      ID  FreqAway      Home
## [1,] 42  7.380646 10362.114
## [2,] 53 13.059628  8058.363
## Clustering vector:
##   [1] 1 1 2 2 1 1 1 1 2 2 1 1 2 2 1 2 2 1 2 1 1 1 2 2 1 2 2
##  [28] 1 2 2 2 2 1 1 2 1 2 2 1 1 1 1 2 1 1 1 2 2 2 1 2 2 2 1
##  [55] 1 1 1 2 2 2 2 1 1 1 2 1 2 1 1 2 2 2 1 2 1 2 2 1 2 2 2
##  [82] 2 2 1 2 2 2 2 2 1 2 2 2 2 1 2 2 1 2 2
## Objective function:
##     build      swap
## 0.8438882 0.6608854
##
## Available components:
##  [1] "medoids"    "id.med"     "clustering" "objective"
##  [5] "isolation"  "clusinfo"   "silinfo"    "diss"
##  [9] "call"       "data"
```

La funzione `pam()` restituisce un oggetto i cui componenti includono i medoidi (oggetti che rappresentano i cluster) e il vettore di clustering. È possibile accedere a questi componenti come segue:

```
pam.out$medoids
```

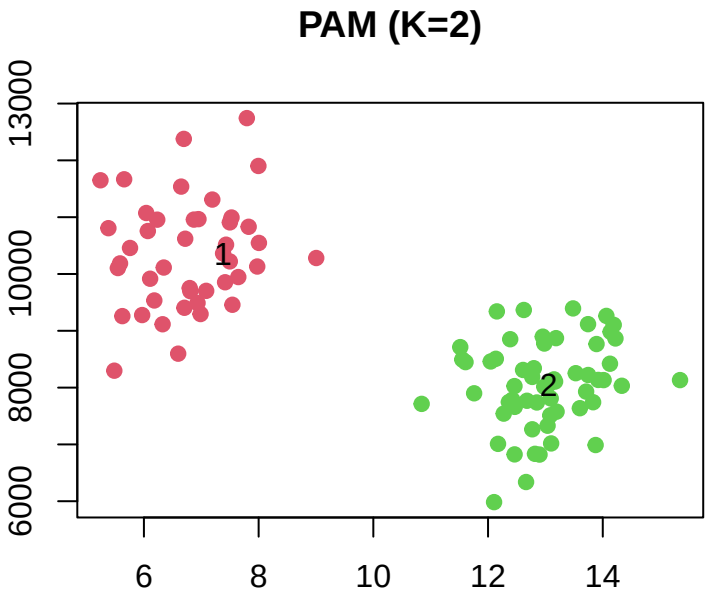
```
##       FreqAway      Home
## [1,]  7.380646 10362.114
## [2,] 13.059628  8058.363
```

```
head(pam.out$clustering)
```

```
## [1] 1 1 2 2 1 1
```

Il risultato del clustering può essere visualizzato sul data scatter plot, vengono visualizzati anche i medoidi:

```
plot(x, col =(pam.out$cluster+1), main="PAM (K=2)", xlab ="", ylab="", pch=19)
points(pam.out$medoids, pch=as.character(pam.out$cluster[pam.out$id.med]))
```

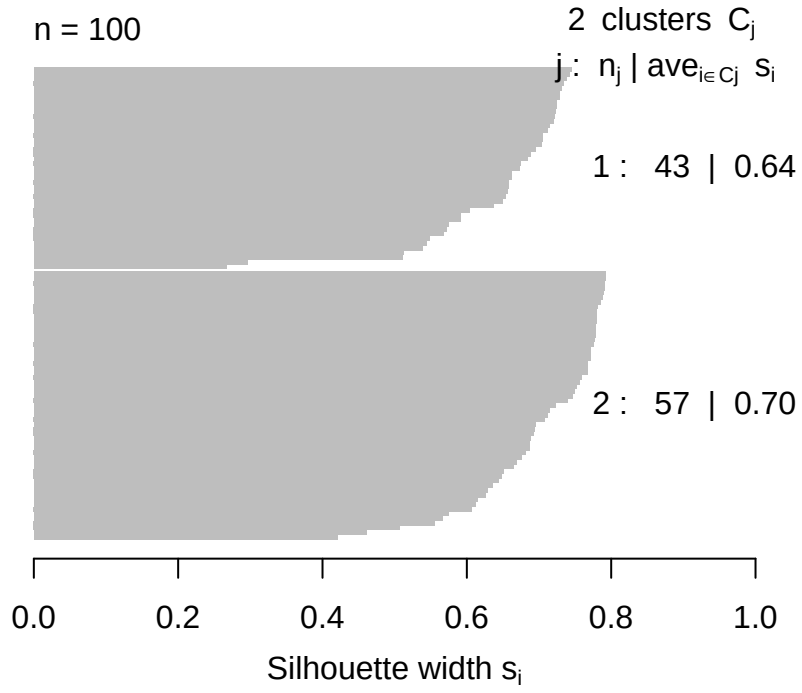


6.4.4.3 Valutare la qualità del raggruppamento: il grafico della silhouette

La statistica **silhouette** misura la somiglianza s_i dei membri all'interno di un cluster rispetto aquelli degli altri cluster (ovvero, il cluster successivo più vicino nel caso di più di due cluster). Il grafico delle silhouette è un grafico a barre delle sagome dei punti per ciascun cluster, ordinate per gruppo in ordine decrescente. Da questa trama, è facile identificare osservazioni problematiche. esso può esser calcolato nel caso si misurino le distanze con metriche come quella euclidea o quella di Manhattan.

Il codice seguente produce il **grafico della silhouette** per i risultati dell'algoritmo PAM:

```
plot(pam.out, which=2, main="") #silhouette plot
```



Le osservazioni con un s_i grande (quasi 1) sono molto ben raggruppate, un piccolo s_i (vicino a 0) significa che l'osservazione si trova in mezzo due cluster e le osservazioni con s_i negativo sono probabilmente posizionato nel cluster sbagliato.

Le informazioni sulla *larghezza media della silhouette* totale in base alle singole larghezze della silhouette possono fornire indicazioni utili per la convalida del clustering.

Nota Per set di dati di grandi dimensioni, `pam` può essere impegnativo dal punto di vista computazionale. Quindi `clara()` è preferibile, vedere la documentazione online se necessario.

Si noti che l'idea della silhouette è generale e può essere usata anche per valutare i raggruppamenti ottenuti con altri metodi.

6.4.4.4 PAM con matrice di dissomiglianza arbitraria

La funzione `pam` permette all'utente di fornire una matrice di dissomiglianza arbitraria, impostando `diss=TRUE`.

Ad esempio, si supponga di voler identificare i gruppi per i dati simulati nel file "simclust2.csv", che contiene informazioni sulle caratteristiche di 1000 reclami. In particolare, questo set di dati contiene tre

variabili binarie (contenzioso, lesione dei tessuti molli, trattamento di emergenza) e due variabili numeriche (numero di fornitori e numero di trattamenti).

Prima si leggono i dati

```
d<-read.csv(file="simclust2.csv", header = TRUE)
head(d)
```

```
##   ID litig soft_injury emergency_tr NumProv NumTreat
## 1  1     0           1            0       2        2
## 2  1     0           0            1       2        1
## 3  1     0           0            0       0        2
## 4  1     0           0            0       2        4
## 5  1     0           1            1       1        5
## 6  1     0           1            1       0        3
```

Successivamente, come già in precedenza, si definiscano come fattori le variabili categoriali e si crei un nuovo frame di dati per l'analisi dei cluster:

```
d$litig<-as.factor(d$litig)
d$soft_injury<-as.factor(d$soft_injury)
d$emergency_tr<-as.factor(d$emergency_tr)

ClusterDat<-d[ , -1]
str(ClusterDat)
```

```
## 'data.frame':    1000 obs. of  5 variables:
##  $ litig       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
##  $ soft_injury : Factor w/ 2 levels "0","1": 2 1 1 1 2 2 1 1 1 2 ...
##  $ emergency_tr: Factor w/ 2 levels "0","1": 1 2 1 1 2 2 2 1 2 2 ...
##  $ NumProv     : int  2 2 0 2 1 0 1 2 2 5 ...
##  $ NumTreat    : int  2 1 2 4 5 3 4 3 1 8 ...
```

La metrica di Gower viene quindi utilizzata per derivare la matrice di dissomiglianza

```
dclaims<-as.matrix(daisy(ClusterDat, metric= "gower"))
```

Infine, si esegua l'algoritmo PAM, usando $K = 2$:

```
pam.claims<-pam(dclaims, 2, diss=TRUE)
```

Di seguito si possono ottenere informazioni numeriche per ciascun cluster: la cardinalità, la massima e media dissomiglianza tra le osservazioni nel cluster e il medoide del cluster, la massima dissomiglianza tra due osservazioni del cluster (diametro), la minima dissomiglianza tra un'osservazione del cluster e un'osservazione di un altro gruppo (separazione), la larghezza media della silhouette

```
pam.claims$clusinfo
```

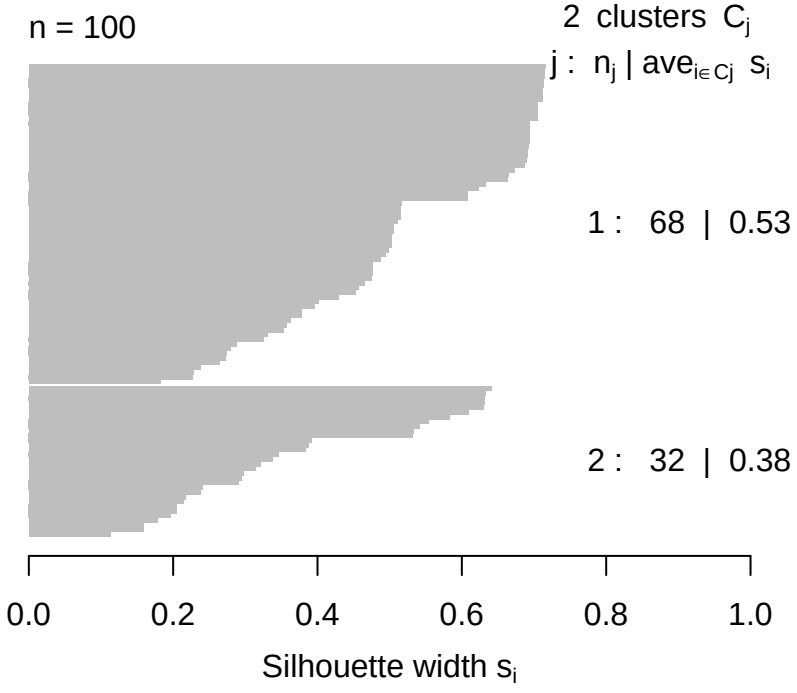
```
##      size  max_diss   av_diss  diameter separation
## [1,]  635 0.3885714 0.1601935 0.6571429        0.2
## [2,]  365 0.3771429 0.1744736 0.7085714        0.2
```

```
pam.claims$silinfo$clus.avg.widths
```

```
## [1] 0.5307724 0.4396967
```

Il grafico della silhouette viene visualizzato per un sottoinsieme delle osservazioni

```
# silhouette di 100 os. dei dati sui sinistri
plot(silhouette(pam.claims$clustering[1:100],
                as.matrix(dclaims)[1:100, 1:100]), main="")
```



Per confrontare i risultati del clustering con i gruppi effettivi si può usare la funzione `table()`

```
attuale<-d$ID
pred<-pam.claims$clustering
table(attuale, pred)
```

```
##        pred
## attuale   1   2
##       1 608  98
##       2  27 267
```

Si può notare che i cluster ottenuti utilizzando l'algoritmo PAM sono leggermente diversi dalle partizioni vere, dove la procedura alloca correttamente 91% del cluster più piccolo e 86% di quello più grande.

Per aggiungere le classificazioni dei punti ai dati originali, si puo utilizzare:

```
dd <- cbind(ClusterDat, pam.cluster=pam.claims$clustering)
head(dd, n = 3)
```

```
##   litig soft_injury emergency_tr NumProv NumTreat
## 1     0           1            0       2        2
## 2     0           0            1       2        1
## 3     0           0            0       0        2
##   pam.cluster
## 1           1
## 2           1
## 3           1
```

Infine, si può esaminare il riepilogo di ciascun cluster

```
pam_results <- split(dd, dd$pam.cluster)

summary(pam_results[[1]]) #gruppo 1
```

```
##  litig   soft_injury emergency_tr    NumProv
##  0:612   0:543       0:374        Min.   :0.000
##  1: 23   1: 92       1:261        1st Qu.:1.000
##                                   Median :2.000
##                                   Mean   :2.107
##                                   3rd Qu.:3.000
##                                   Max.   :8.000
##     NumTreat       pam.cluster
##  Min.   : 0.000   Min.   :1
##  1st Qu.: 2.000   1st Qu.:1
##  Median : 3.000   Median :1
##  Mean   : 3.074   Mean   :1
##  3rd Qu.: 4.000   3rd Qu.:1
##  Max.   :12.000   Max.   :1
```

```
summary(pam_results[[2]]) #gruppo 2
```

```
##  litig   soft_injury emergency_tr    NumProv
##  0:120   0: 45       0: 34        Min.   : 0.000
```

```
##  1:245   1:320       1:331        1st Qu.: 2.000
##                                   Median : 3.000
##                                   Mean   : 3.551
##                                   3rd Qu.: 5.000
##                                   Max.   :10.000
##     NumTreat       pam.cluster
##  Min.   : 0.000   Min.   :2
##  1st Qu.: 3.000   1st Qu.:2
##  Median : 5.000   Median :2
##  Mean   : 5.356   Mean   :2
##  3rd Qu.: 7.000   3rd Qu.:2
##  Max.   :14.000   Max.   :2
```

Si noti che esistono molti altri algoritmi per gestire dati di tipo variabile misto e relativi pacchetti R (si veda, ad esempio, il pacchetto **clustMixType** per eseguire il clustering di partizionamento k-prototipi di Huang, 1998).

6.4.5 Raggruppamento gerarchico

Ci sono diverse funzioni disponibili in R per il clustering gerarchico:

- Approccio agglomerativo (dal basso): `hclust()` (nel package **stats**), `agnes()` (pacchetto **cluster**).
- Approccio divisivo (top-down): `diana()` (sempre in **cluster**)

Il clustering gerarchico è particolarmente noto nella sua versione agglomerativa, che unisce cluster "simili", fino a quando tutte le unità sono raggruppate in un unico cluster. Come vedremo, il processo crea una gerarchia di cluster, rappresentata dal cosiddetto **dendrogramma**.

Le funzioni `hclust()` e `agnes()` sono abbastanza simili; tuttavia, con la funzione `agnes` si può anche ottenere il coefficiente agglomerativo, che misura la qualità della struttura di raggruppamento trovata (valori più vicini a 1 suggeriscono una struttura di raggruppamento forte).

Per iniziare con un semplice esempio, considereremo il dataset di `eurodist` R, contenente le distanze geografiche tra le città europee:

```
data(eurodist)

# poche righe e colonne
as.matrix(eurodist)[1:8, 1:8]
```

```
##            Athens Barcelona Brussels Calais Cherbourg
## Athens          0      3313     2963   3175      3339
## Barcelona    3313         0     1318   1326      1294
```

```
## Brussels     2963      1318         0    204       583
## Calais       3175      1326       204      0       460
## Cherbourg    3339      1294       583    460         0
## Cologne      2762      1498       206    409       785
## Copenhagen   3276      2218       966   1136      1545
## Geneva       2610       803       677    747       853
##            Cologne Copenhagen Geneva
## Athens        2762       3276   2610
## Barcelona     1498       2218    803
## Brussels       206        966    677
## Calais         409       1136    747
## Cherbourg      785       1545    853
## Cologne          0        760   1662
## Copenhagen     760          0   1418
## Geneva        1662       1418      0
```

6.4.5.1 I metodi agglomerativi

Il clustering gerarchico agglomerativo prevede che si parta dai singoli punti e si proceda aggregando questi in gruppi sempre più ampi fino ad avere tutti i punti in un unico gruppo.

1. Si parte quindi da una matrice di dissimilarità (ad esempio generata con `daisy()` o `dist()`) e si cerca in essa quei due punti che sono meno dissimili (ovvero con dissimilarità più bassa).
2. Si uniscono quindi tali due punti ch formano un primo gruppo e si ricalcola la distanza di questo dagli altri punti.
3. La distanza del nuovo gruppo da un altro punto (o da un altro gruppo) si può calcolare in diversi metodi (legami) usando una misura di dissimilarità $d(i, j)$ adeguata. Ne citiamo alcuni fra i più comuni (detti A e B gli insiemi dei punti di due cluster):

- legame singolo: la minima distanza fra gli elementi di ciascun cluster (single linkage) $min[d(i, j) : i \in Aej \in B)]$
- legame completo: la massima distanza fra gli elementi di ciascun cluster (complete linkage) $max[d(i, j) : i \in Aej \in B)]$
- legame medio: la ditanza media fra tutti gli elementi di ciascun cluster (average linkage)
- Ward: l'aumento in termini di varianza per il cluster che viene unito.

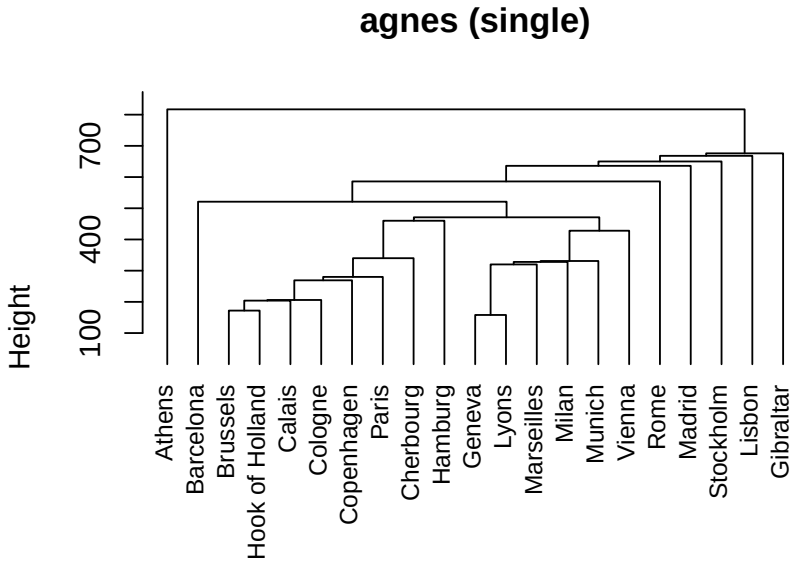
Eseguiamo il clustering gerarchico agglomerativi usando `agnes()` (all'interno del pacchetto **cluster**) sulle distanze **eurodist**, usando `method=single` per il collegamento singolo (se passiamo un oggetto `dist` come primo argomento allora non c'è bisogno di specificare altro)

```
#collegamento unico
agnes.single<-agnes(eurodist, method="single")
# coefficiente agglomerato
agnes.single$ac
```

```
## [1] 0.5115696
```

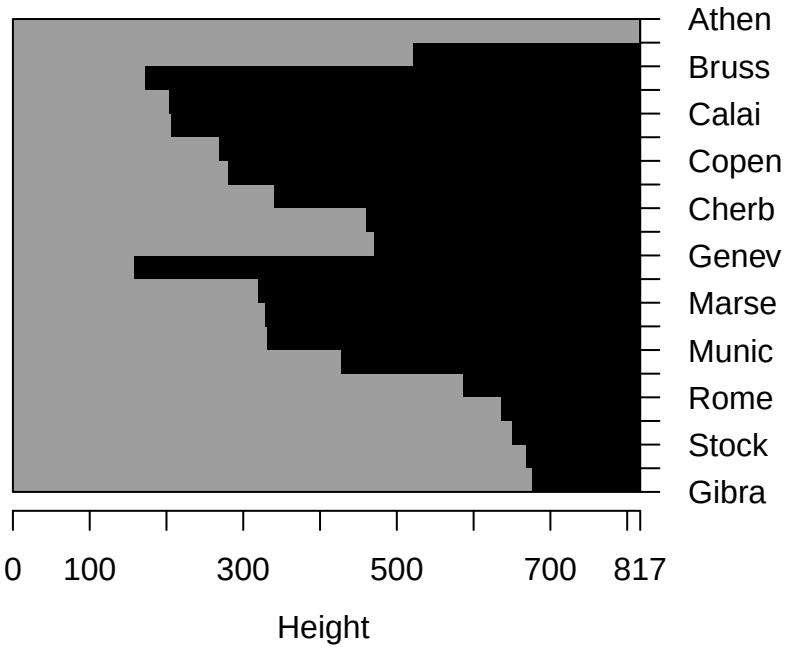
Per visualizzare il dendrogramma, si utilizza la funzione `pltree()`

```
# Metti le etichette alla stessa altezza: hang = -1
pltree(agnes.single, cex=0.8, hang = -1, main = "agnes (single)", xlab="", sub ="")
```



La funzione `agnes` fornisce una visualizzazione alternativa dell'aggregazione dei cluster:

```
plot(agnes.single, which=1, col=c(8,1), main="")
```



Se sono necessari grafici più flessibili, una soluzione alternativa è chiamare la funzione `plot()`, dopo la coercizione alla classe "hclust" e "dendrogram", come nel codice seguente

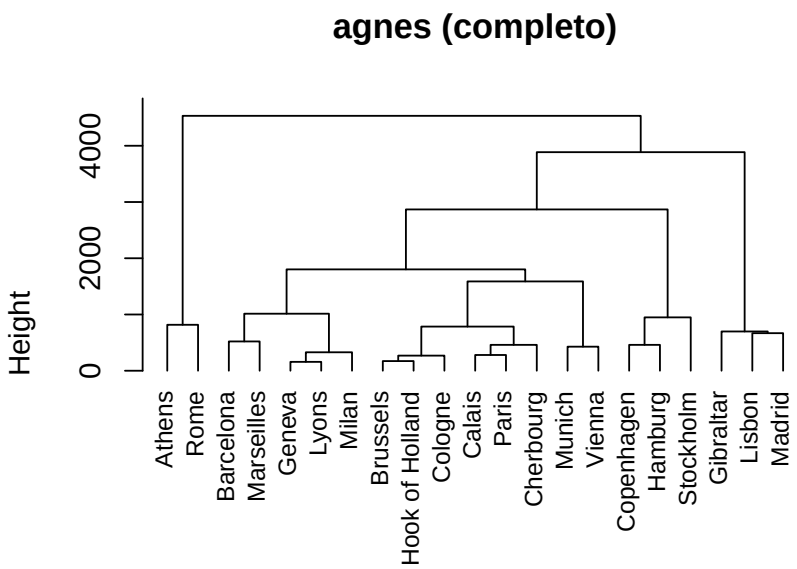
```
# Converti in un dendrogramma e traccia
eurocity.dend<-as.dendrogram(as.hclust(agnes.single))

# trama(eurocity.dend)
```

L'output del legame singolo spesso non da risultati molto soddisfacenti, a causa di un effetto catena. Si possono replicare i passaggi precedenti utilizzando gli altri legami, che porteranno a risultati diversi

```
#legame completo
agnes.comp<-agnes(eurodist,method="complete")
agnes.comp$ac
```

```
## [1] 0.8979532
```

```
pltree(agnes.comp, cex = 0.8, hang = -1, main = "agnes (completo)", xlab="", sub =
```

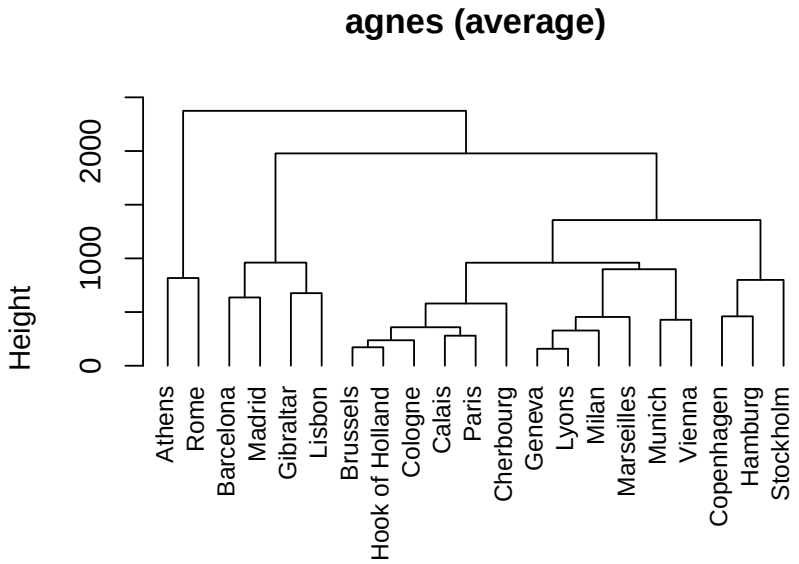


Il collegamento completo identifica bene i gruppi "estremi", che corrispondono al sud (Roma, Atene), alla penisola iberica (Madrid, Gibilterra e Lisbona) e a un gruppo di città del nord Europa (Amburgo, Copenaghen e Stoccolma).

```
#average linkage
agnes.ave<-agnes(eurodist,method="average")
agnes.ave$ac
```

```
## [1] 0.8063934
```

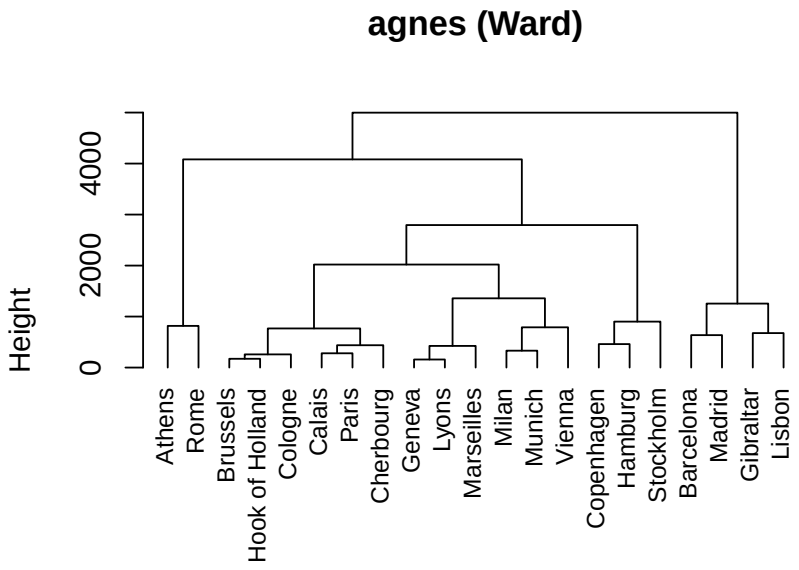
```
pltree(agnes.ave, cex = 0.8, hang = -1, main = "agnes (average)", xlab="", sub ="")
```




```
# Ward's method
agnes.Ward<-agnes(eurodist,method="ward")
agnes.Ward$ac
```

```
## [1] 0.905946
```

```
pltree(agnes.Ward, cex = 0.8, hang = -1, main = "agnes (Ward)", xlab="", sub ="")
```



Il legame medio e il metodo di Ward producono risultati simili: sono ancora identificati il cluster meridionale, la penisola iberica e una regione nord-orientale. Inoltre, possiamo identificare

- una regione centro-settentrionale continentale (Bruxelles, Hoek van Holland, Colonia)
- una regione del Centro (Ginevra, Lione, Marsiglia, Milano, Monaco e Vienna)
- il cluster della Francia settentrionale (Calais, Parigi, Cherbourg)

Per determinare le etichette dei cluster per ogni osservazione associata a un dato taglio del dendrogramma, si può usare la funzione `cutree()`:

```
hc<-cutree(agnes.ave, 4)
# hc4 <- cutree(as.hclust(agnes.ave), h = 1000)

cnames<-row.names(as.matrix(eurodist))
cnames[hc==1]
```

```
## [1] "Athens" "Rome"
```

```
cnames[hc==2]
```

```
## [1] "Barcelona" "Gibraltar" "Lisbon"    "Madrid"
```

```
cnames[hc==3]
```

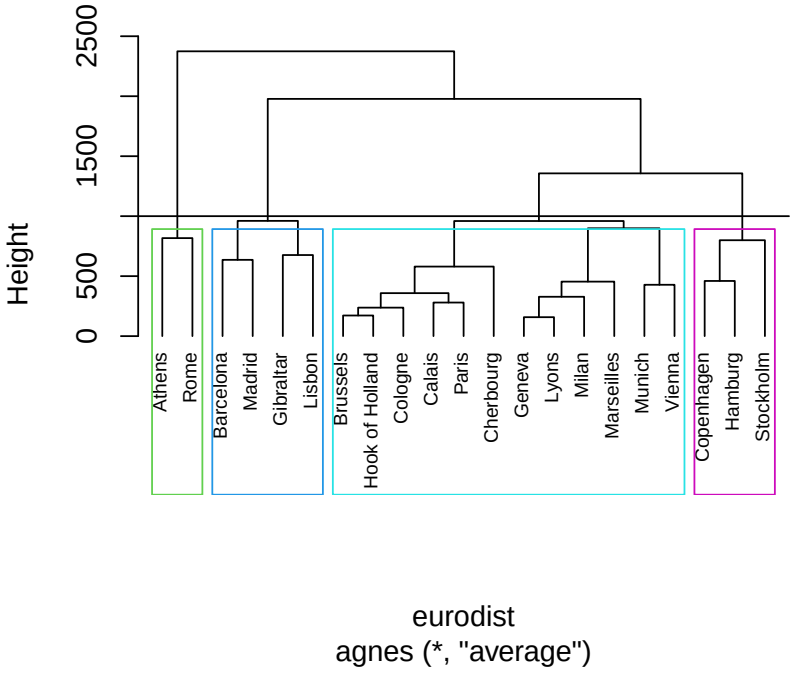
```
##  [1] "Brussels"        "Calais"          "Cherbourg"
##  [4] "Cologne"         "Geneva"          "Hook of Holland"
##  [7] "Lyons"           "Marseilles"      "Milan"
## [10] "Munich"          "Paris"           "Vienna"
```

```
cnames[hc==4]
```

```
## [1] "Copenhagen" "Hamburg"    "Stockholm"
```

La funzione `abline()` disegna una linea retta sopra ogni tracciato esistente in R. Il codice seguente traccia una linea orizzontale ad altezza 1000 sul dendrogramma, producendo i quattro cluster ottenuti sopra. È possibile anche visualizzare i cluster all'interno del dendogramma stesso inserendo i bordi come mostrato di seguito

```
pltree(agnes.ave, hang=-1, cex = 0.7, main="")
abline(h=1000)
rect.hclust(agnes.ave, k = 4, border = 3:7)
```



Si proceda ora a raggruppare gerarchicamente le capitali del mondo, utilizzando il database delle città del mondo nel pacchetto **mappe**, con l'obiettivo di visualizzare gruppi di città sulla base di informazioni geografiche.

```
library(maps)
```

```
##
## Attaching package: 'maps'

## The following object is masked from 'package:cluster':
##
##     votes.repub
```

```
data(world.cities)
str(world.cities) #vedi i componenti del set di dati
```

```
## 'data.frame':    43645 obs. of  6 variables:
##  $ name       : chr  "'Abasan al-Jadidah" "'Abasan al-Kabirah" "'Abdul Hakim" '
##  $ country.etc: chr  "Palestine" "Palestine" "Pakistan" "Kuwait" ...
##  $ pop        : int  5629 18999 47788 21817 2456 3434 9198 5492 22706 41731 ...
##  $ lat        : num  31.3 31.3 30.6 29.4 32 ...
##  $ long       : num  34.3 34.4 72.1 48 35.1 ...
##  $ capital    : int  0 0 0 0 0 0 0 0 0 0 ...
```

Per selezionare le maiuscole, si considera il sottoinsieme di righe per cui `capital==1`

```
d1<-subset(world.cities,capital==1)

# seleziona colonne con latitudini e longitudini
d2<-d1[ ,c(4,5)]
head(d2)
```

```
##        lat    long
## 26   31.95   35.93
## 265  24.48   54.37
## 280   9.18    7.17
## 308   5.56   -0.20
## 366 -25.05 -130.10
## 382   9.03   38.74
```

Dato il numero di città coinvolte, si possono rappresentare i raggruppamenti disegnando le città su una mappa utilizzando colori e simboli diversi per i diversi gruppi, anziché il dendrogramma. Di seguito si esegue il clustering agglomerativo e si confrontano le soluzioni per $k = 5$ e $k = 8$:

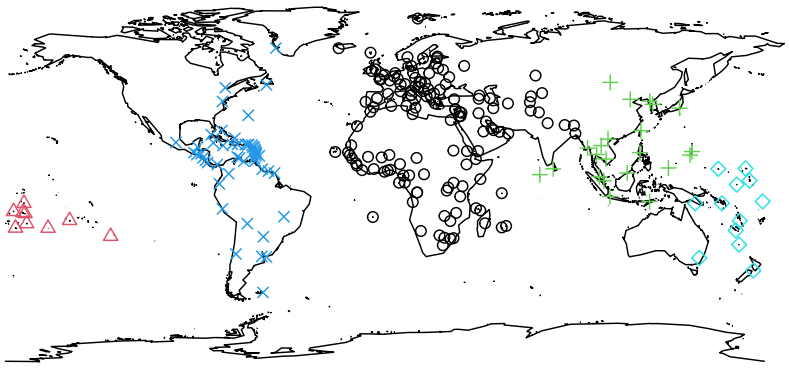
```
ris.agnes<-agnes(d2, method="average")

gruppi_5<-cutree(ris.agnes,k=5)
table(gruppi_5)
```

```
## gruppi_5
##   1   2   3   4   5
## 136   9  22  51  12
```

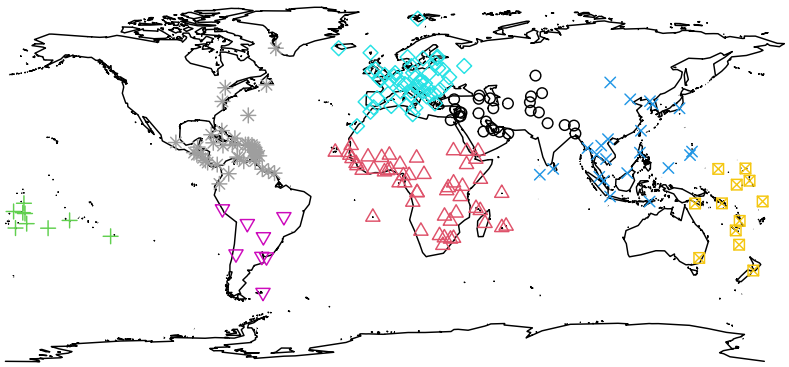
```
gruppi_8<-cutree(ris.agnes,k=8)
table(gruppi_8)
```

```
## gruppi_8
##  1  2  3  4  5  6  7  8
## 30 52  9 22 54  8 12 43
```

```
# plot on map with 5 clusters
par(mar=c(0,0,0,0))
map("world",interior=FALSE)
points(d2[,2],d2[,1],col=gruppi_5,pch=gruppi_5)
```



```
# plot on map with 8 clusters
par(mar=c(0,0,0,0))
map("world",interior=FALSE)
points(d2[,2],d2[,1],col=gruppi_8,pch=gruppi_8)
```



Infine, è semplice ottenere le silhouette da raggruppamenti gerarchici usando la funzione `silhouette` con `cutree()` e la distanza come input:

```
# Silhouette per un clustering gerarchico:
labels<-d1$name
sil<- data.frame(Name=labels, silhouette(gruppi_5, daisy(d2))[, 1:3])
head(sil)
```

```
##          Name cluster neighbor sil_width
## 1      'Amman       1        3 0.5763558
## 2   Abu Dhabi       1        3 0.3039018
## 3       Abuja       1        4 0.5128536
## 4       Accra       1        4 0.4079634
## 5   Adamstown       2        4 0.4662954
## 6 Addis Abeba       1        3 0.4906205
```

6.4.5.2 Clustering basato sulla correlazione

Finora, si è usata la distanza euclidea come misura di dissomiglianza. Questa è la radice quadrata della somma delle differenze quadrate. In alcune circostanze, potrebbero essere preferite altre misure di dissomiglianza.

Ad esempio, una distanza basata sulla correlazione viene utilizzata per concentrarsi sulle forme dei profili di osservazione piuttosto che sulle loro grandezze. Si presume che due righe della matrice di dati siano simili se le loro caratteristiche sono altamente correlate, producendo così una distanza uguale a zero quando sono perfettamente correlate:

$$d(x, y) = 1 - r_{xy}$$

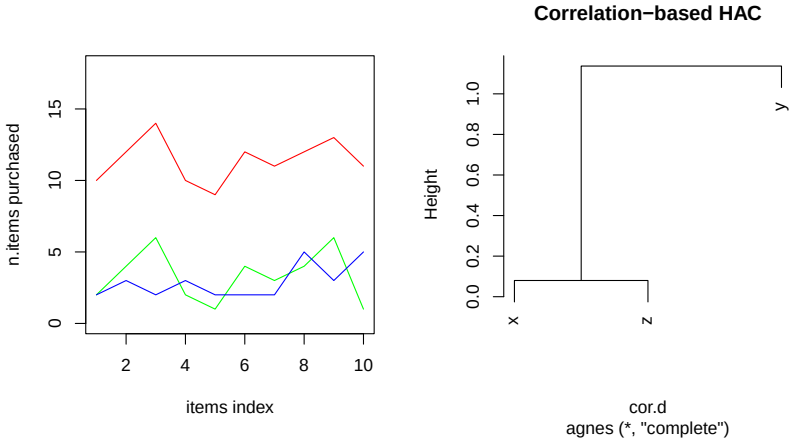
dove r_{xy} è il coefficiente di correlazione di Pearson. La matrice di distanza risultante può essere utilizzata come input per il clustering gerarchico. La correlazione di Spearman può essere utilizzata al posto della correlazione di Pearson, che corrisponde al calcolo di r_{xy} sui ranghi.

Esempi: La correlazione funziona bene per l'espressione genica nel raggruppamento di campioni e geni. Un altro esempio è il seguente.

Si consideri una matrice in cui le righe sono acquirenti e le colonne sono gli articoli disponibili per l'acquisto; gli elementi di matrice rappresentano il numero di volte in cui un determinato acquirente ha acquistato un determinato articolo. La distanza basata sulla correlazione viene utilizzata per identificare gli acquirenti con preferenze simili (ad es. acquirenti che hanno acquistato articoli A e B ma mai articoli C o D), indipendentemente dalle differenze di volume.

```
x<-c(2,4,6,2,1,4,3,4,6,1)
y<-c(2,3,2,3,2,2,2,5,3,5)
z<-c(10,12,14,10,9,12,11,12,13,11)

d<-rbind(x, y, z) # data matrix

par(mfrow=c(1,2))
plot(1:10, z, type="l", col="red", ylim=c(0,18), xlab="items index",
     ylab="n.items purchased")
points(x, type="l", col="green")
points(y, type="l", col="blue")

# compute Pearson correlation between rows
r<-cor(t(d))
# distance
```

```
cor.d<-as.dist(1-r)
#dendrogram
pltree(agnes(cor.d, method ="complete"), main="Correlation-based HAC")
```



I clienti "x" e "z" hanno valori abbastanza diversi per ciascuna variabile, ma sono altamente correlati, quindi c'è una piccola distanza basata sulla correlazione tra loro, quindi sono raggruppati insieme.

Nota: il clustering basato sulla correlazione viene utilizzato anche nelle applicazioni dei mercati finanziari, dove è opportuno determinare gruppi di titoli che si muovono insieme nel tempo. In tale contesto, i coefficienti di correlazione r_{ij} vengono convertiti in distanze, ad esempio utilizzando $d_{ij} = \sqrt{(2(1 - r_{ij}))}$. Approcci alternativi utilizzano misure di associazione/concordanza come la correlazione di Kendall e Spearman, nonché varie misure di dipendenza dalla coda.

6.4.6 Metodi di raggruppamento basati sulla densità

6.4.6.1 Il metodo DBSCAN

L'obiettivo dell'algoritmo DBSCAN è identificare i cluster come regioni dense, che possono essere misurate dal numero di oggetti "vicini" a un dato punto.

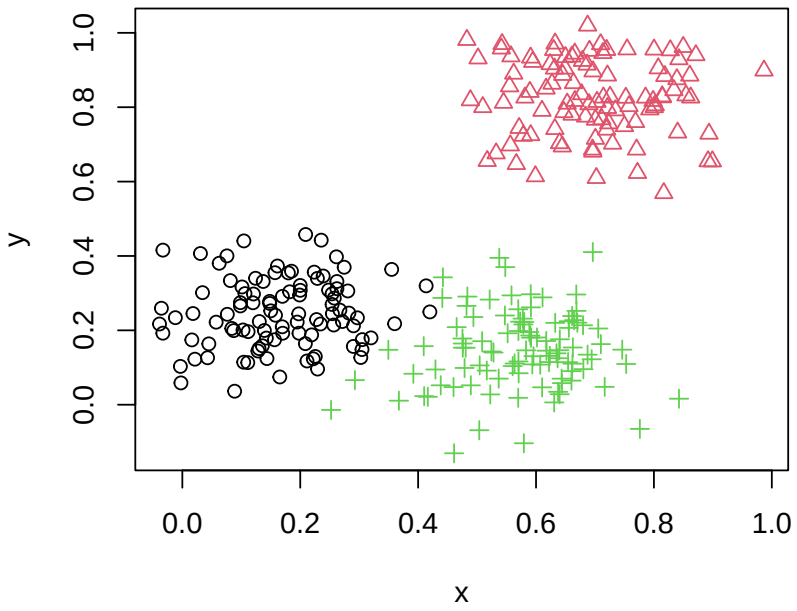
La funzione R per implementare il metodo DBSCAN è la funzione `dbscan` nel pacchetto **dbscan**. Per illustrare come funziona l'algoritmo, simuleremo prima tre cluster nello spazio bidimensionale con 100 punti ciascuno:

```
set.seed(2)
n <- 300
```

```
d <- cbind(x = runif(3, 0, 1) + rnorm(n, sd = 0.1),
           y = runif(3, 0, 1) + rnorm(n, sd = 0.1))
head(d)
```

```
##               x          y
## [1,] 0.08869306 0.03626618
## [2,] 0.86083718 0.88495681
## [3,] 0.67011472 0.22061473
## [4,] 0.19744429 0.24429363
## [5,] 0.63138780 0.90385668
## [6,] 0.48210192 0.26570588
```

```
true_cl <- rep(1:3, 100)
plot(d, col = true_cl, pch = true_cl)
```

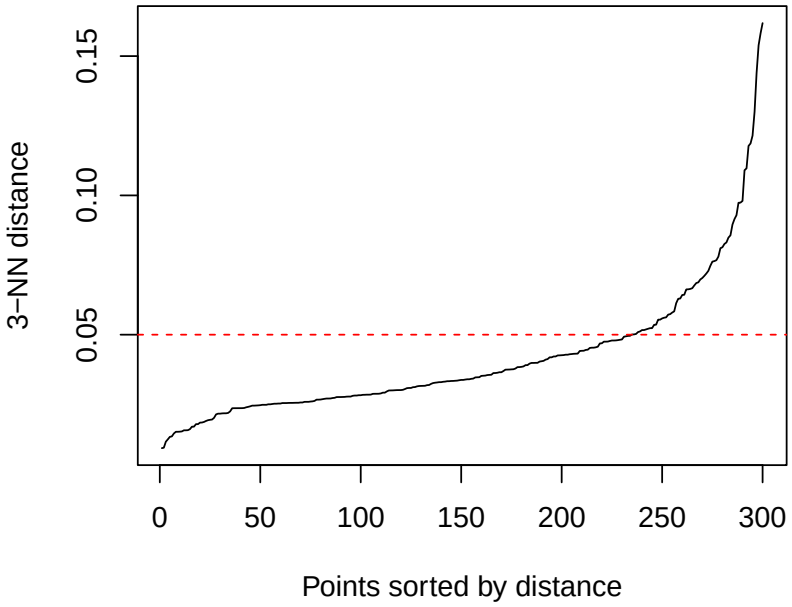


Per DBSCAN sono necessari due parametri importanti:

- il parametro *eps* definisce il raggio dell'intorno attorno ad un punto x.
- Il parametro *MinPts*, il numero minimo di punti entro la distanza *eps*, necessario per produrre un nuovo cluster

Il valore predefinito per *MinPts* nella funzione `dbscan` è 5 (usa almeno il numero di dimensioni del set di dati più uno). Per *eps*, possiamo trovare un valore adatto esplorando un diagramma K-NN dei punti (cioè il diagramma delle distanze al k-esimo vicino più vicino in ordine decrescente) e cercare un angolo acuto nel diagramma:

```
library(dbscan) # carica la libreria

# trova il parametro eps adatto usando un diagramma k-NN
kNNdistplot(d, k = 3)
abline(h=.05, col = "red", lty=2)
```



Un gomito" può essere identificato a una distanza di circa 3-NN di 0.05. Successivamente, possiamo eseguire il clustering con i parametri scelti:

```
db.out <- dbscan(d, eps = 0.05, minPts = 3)
db.out
```

```
## DBSCAN clustering for 300 objects.
## Parameters: eps = 0.05, minPts = 3
## Using euclidean distances and borderpoints = TRUE
## The clustering contains 5 cluster(s) and 30 noise points.
##
##  0  1  2  3  4  5
## 30 90 89 81  4  6
##
## Available fields: cluster, eps, minPts, metric,
##                   borderPoints
```

Vengono identificati i tre cluster più grandi, insieme a due piccoli gruppi di 4 e 6 punti.

Per confrontare i cluster identificati con i gruppi effettivimante esistenti possiamo utilizzare il cosiddetto **Adjusted Rand Index** (ARI), che è un *indice esterno* per la valutazione del clustering, misurando la somiglianza tra due classificazioni degli stessi oggetti dal proporzioni degli accordi tra le due partizioni. La funzione `adj.rand.index()` nel pacchetto **pdfCluster** può essere utilizzata per questo compito:

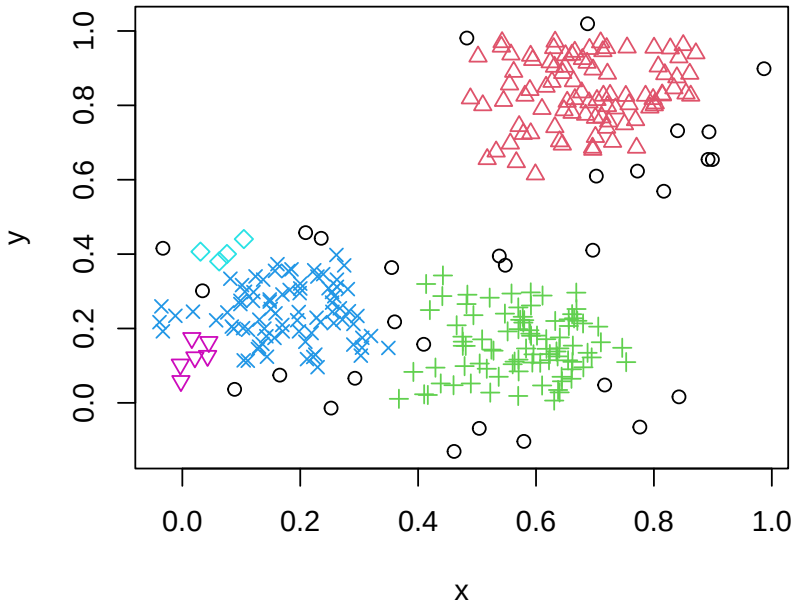
```
library(pdfCluster)
adj.rand.index(true_cl, db.out$cluster)
```

```
## [1] 0.7692038
```

Come altri indici di validità del cluster, l'ARI assume un valore compreso tra 0 e 1 e maggiore è il valore dell'indice, migliore è la qualità del clustering se comparata con la vera classificazione (che però di solito non è nota come in questo caso in cui avevamo una simulazione).

Infine, le informazioni sull'assegnazione dei cluster possono essere utilizzate per tracciare i dati con i cluster identificati da etichette e colori diversi

```
plot(d, col = db.out$cluster + 1, pch = db.out$cluster + 1)
```



La funzione `predict` può essere utilizzata per prevedere le appartenenze ai cluster per nuovi punti dati.

6.4.6.2 Raggruppamento basato sulla ricerca di mode: `pdfCluster`.

Un altro metodo per ottenere cluster basati sulla densità è disponibile con la funzione `pdfCluster`. Esso è basato su una stima non parametrica della funzione di densità immaginando che la distribuzione abbia più regioni a alta densità che sono identificabili come mode. È quindi basato sulla ricerca di mode nella distribuzione dei dati (che si immagina siano quantitativi).