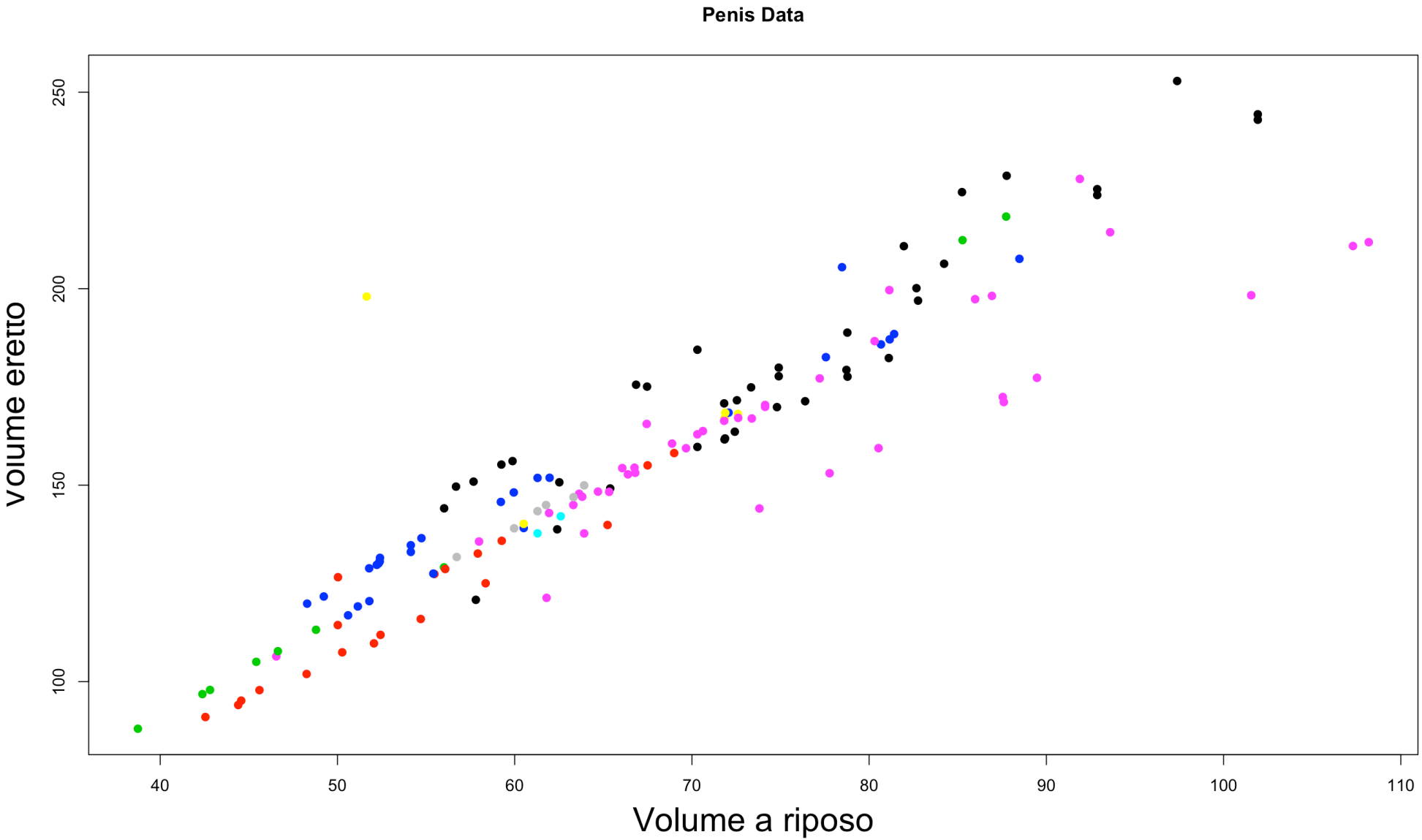
Esempio: Penis Data



ANCOVA

Uso delle variabili Dummy nei modelli di regressione.

Diventano fondamentali i termini di interazione tra variabili quantitative (di cui vogliamo verificare la stazionarietà) e le Dummy, dei codici qualitativi che indicano l'appartenza o meno ad un determinato gruppo.

$$y_{i,k} = X_{i,k}^T \beta_k + \varepsilon_{i,k} \quad \text{con} \quad \mathrm{V}\left(\varepsilon_{i,k}\right) = \sigma_k^2$$

Introducendo l'insieme G di variabili indicatrici di appartenenza al gruppo:

$$y_i = \left(X \otimes G\right)_i^T \hat{\beta} + e_i$$

Esempio: Penis Data

```
regns <- lm(penis$volume_erect ~ penis$volume_flaccid * penis$Region)
summary(regns)

Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                        -26.9086    11.7279   -2.29    0.023 *
penis$volume_flaccid                                 2.6818     0.1503   17.85  < 2e-16 ***
penis$RegionAsia                                     9.5378    18.8056    0.51    0.613
penis$RegionAustralia                               31.2830   409.9474    0.08    0.939
penis$RegionCentral America/Caribean                24.4189    18.1889    1.34    0.182
penis$RegionCentral Asia                           -38.4037   551.1668   -0.07    0.945
penis$RegionEurope                                  64.1677    13.9220    4.61  1.0e-05 ***
penis$RegionNorth America                          248.0752    31.5202    7.87  1.6e-12 ***
penis$RegionPacific Islands                         17.7591    87.6547    0.20    0.840
penis$RegionSouth America                           28.3575    17.2356    1.65    0.102
penis$RegionSouth Asia                              40.9529    40.2477    1.02    0.311
penis$RegionSoutheast Asia                          22.5866    30.3317    0.74    0.458
penis$RegionWestern Asia                             9.7389    19.9927    0.49    0.627
penis$volume_flaccid:penis$RegionAsia               -0.1538     0.3083   -0.50    0.619
penis$volume_flaccid:penis$RegionAustralia          -0.2428     4.7392   -0.05    0.959
penis$volume_flaccid:penis$RegionCentral America/Caribean  -0.3252  0.2463   -1.32    0.189
penis$volume_flaccid:penis$RegionCentral Asia        0.6312     8.8964    0.07    0.944
penis$volume_flaccid:penis$RegionEurope             -0.9642     0.1796   -5.37  3.9e-07 ***
penis$volume_flaccid:penis$RegionNorth America      -3.4860     0.4666   -7.47  1.3e-11 ***
penis$volume_flaccid:penis$RegionPacific Islands    -0.1999     1.4270   -0.14    0.889
penis$volume_flaccid:penis$RegionSouth America      -0.0938     0.2313   -0.41    0.686
penis$volume_flaccid:penis$RegionSouth Asia         -0.7001     0.7307   -0.96    0.340
penis$volume_flaccid:penis$RegionSoutheast Asia     -0.2875     0.6249   -0.46    0.646
penis$volume_flaccid:penis$RegionWestern Asia        0.1099     0.3255    0.34    0.736
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 8.24 on 122 degrees of freedom
Multiple R-squared:  0.953,     Adjusted R-squared:  0.944
F-statistic:  108 on 23 and 122 DF,  p-value: <2e-16
```

Misture finite di regressioni

Misture finite di S regressioni gaussiane con modelli concomitanti variabili sono dati da:

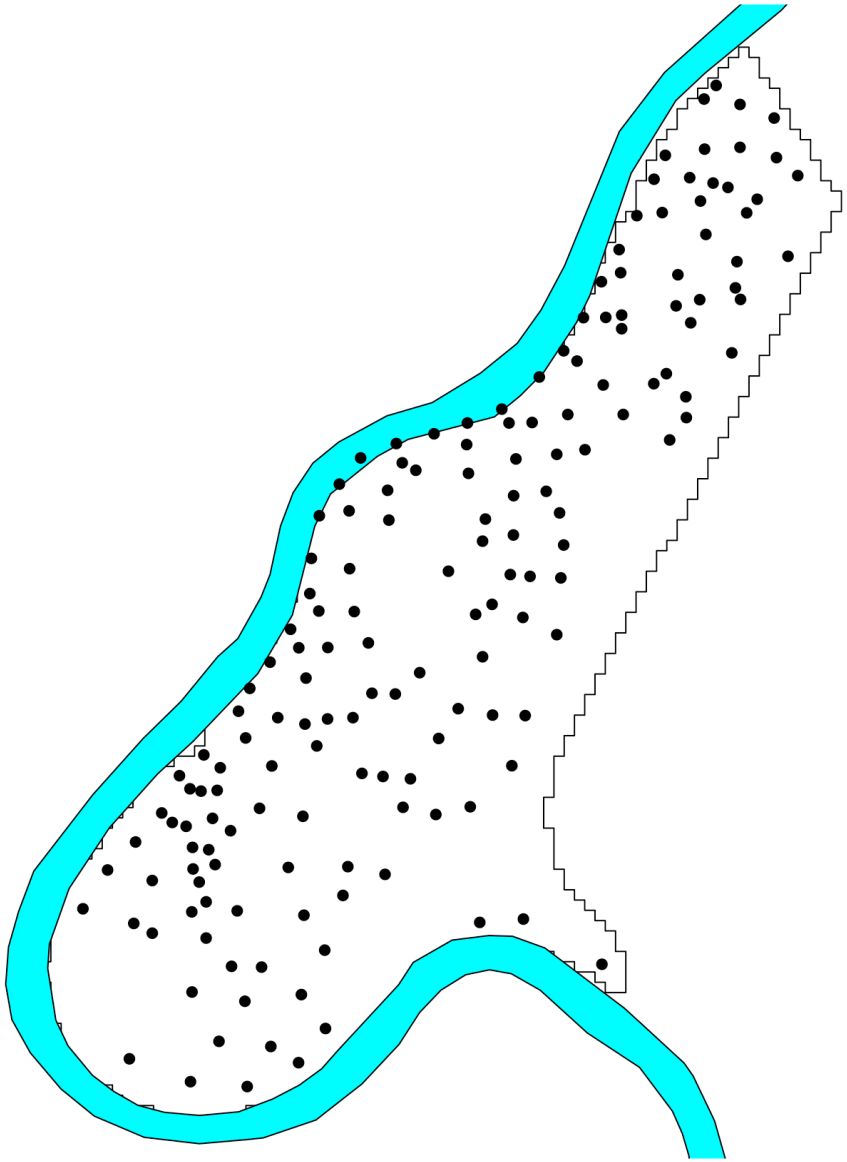
$$H(y \mid \boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\Theta}) = \sum_{s=1}^{S} \pi_s(\boldsymbol{w}, \boldsymbol{\alpha}) \mathrm{N}(y \mid \mu_s(\boldsymbol{x}), \sigma_s^2),$$

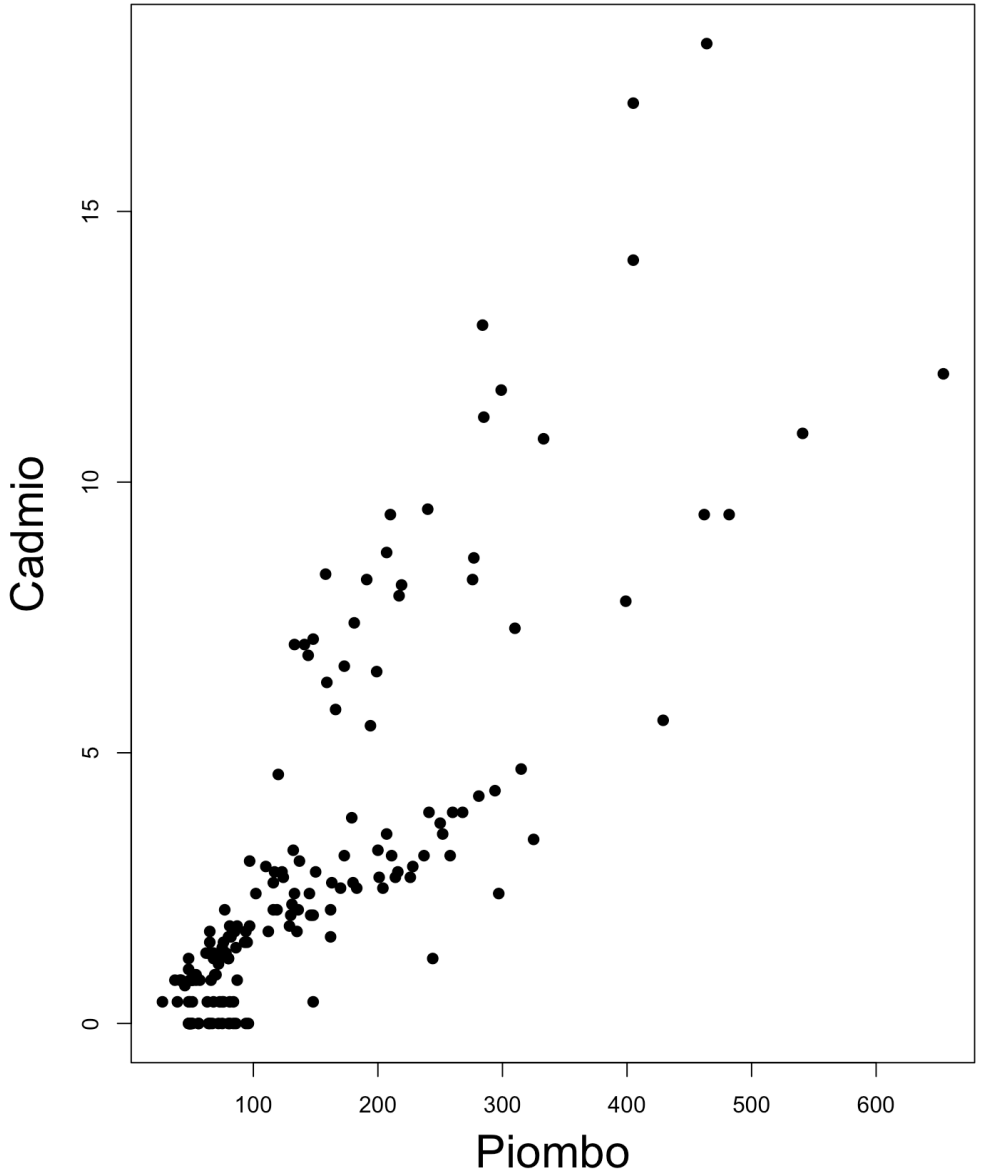
dove $N(\cdot \mid \mu_s(x), \sigma_s^2)$ è la distribuzione gaussiana con media $\mu_s(x) = x^T\beta_s$ e varianza σ_s^2. Θ indica il vettore di tutti i parametri della distribuzione della mistura e la variabile dipendente è y, x l'indipendente e la concomitanza è w. La stima viene fatta alternando la massimizzazione delle verosiglianza delle regressioni gaussiane ed i pesi di ciscuna regressione, seguendo la logica dell'algoritmo EM.

Esempio: Penis Data

```
library("flexmix")
Model <- FLXMRglm(~ volume_flaccid)
regmis <-
stepFlexmix(volume_erect~1,model=Model,nrep=3,k=5,data=penis,concomitant=FLXPmultinom(formula=~1))
regmis <- relabel(regmis, "model", "volume_flaccid")
summary(refit(regmis))
$Comp.1        Estimate Std. Error z value Pr(>|z|)
(Intercept)      72.856     23.393    3.11   0.0018 **
volume_flaccid    1.220      0.294    4.15  3.4e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
$Comp.2        Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.9553     1.1318    0.84      0.4
volume_flaccid   2.2863     0.0188  121.62   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
$Comp.3        Estimate Std. Error z value Pr(>|z|)
(Intercept)      10.008      1.067    9.38   <2e-16 ***
volume_flaccid    2.298      0.018  127.72   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
$Comp.4        Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.5818     2.4950    0.63     0.53
volume_flaccid   2.5883     0.0336   77.08   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
$Comp.5        Estimate Std. Error z value Pr(>|z|)
(Intercept)    -28.3081     5.0486   -5.61  2.1e-08 ***
volume_flaccid   2.7249     0.0749   36.38  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

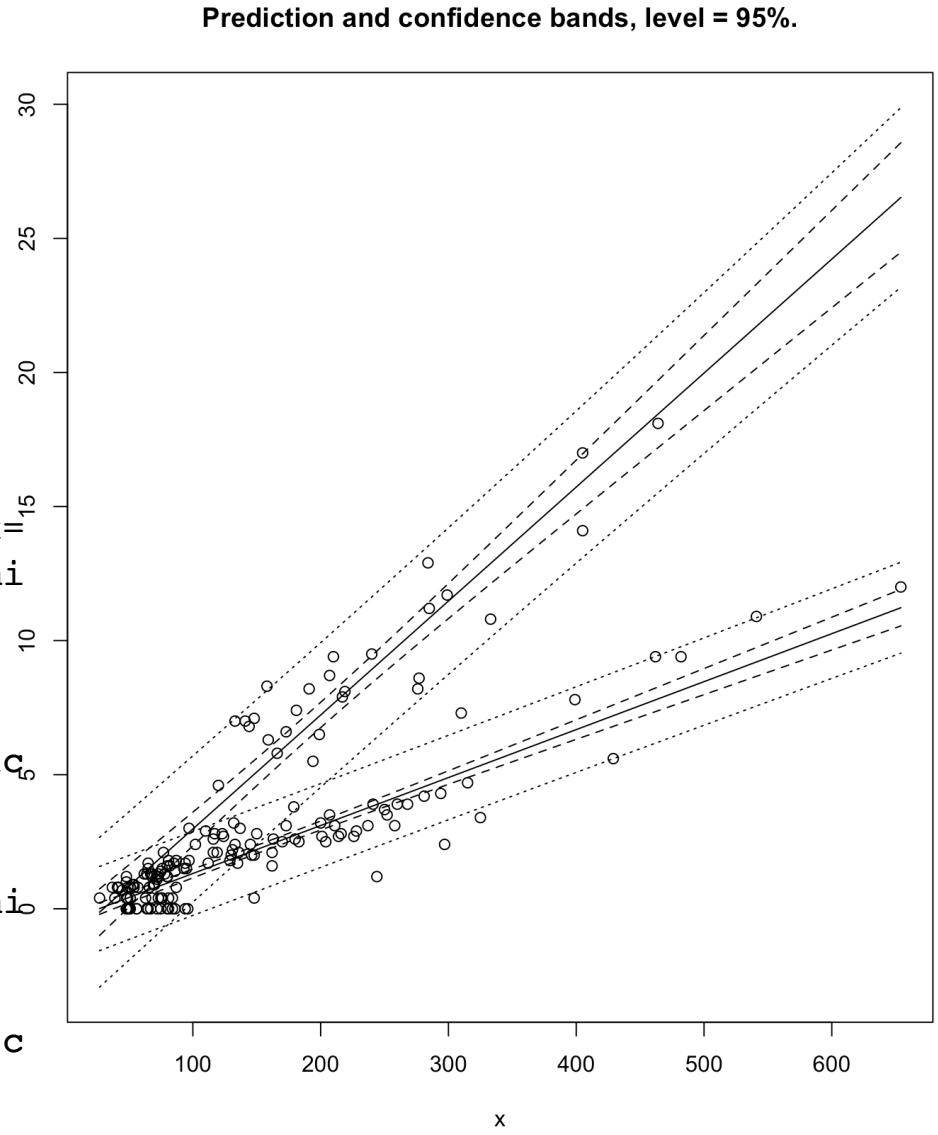
Esempio: Meuse river data





Esempio: Meuse river data

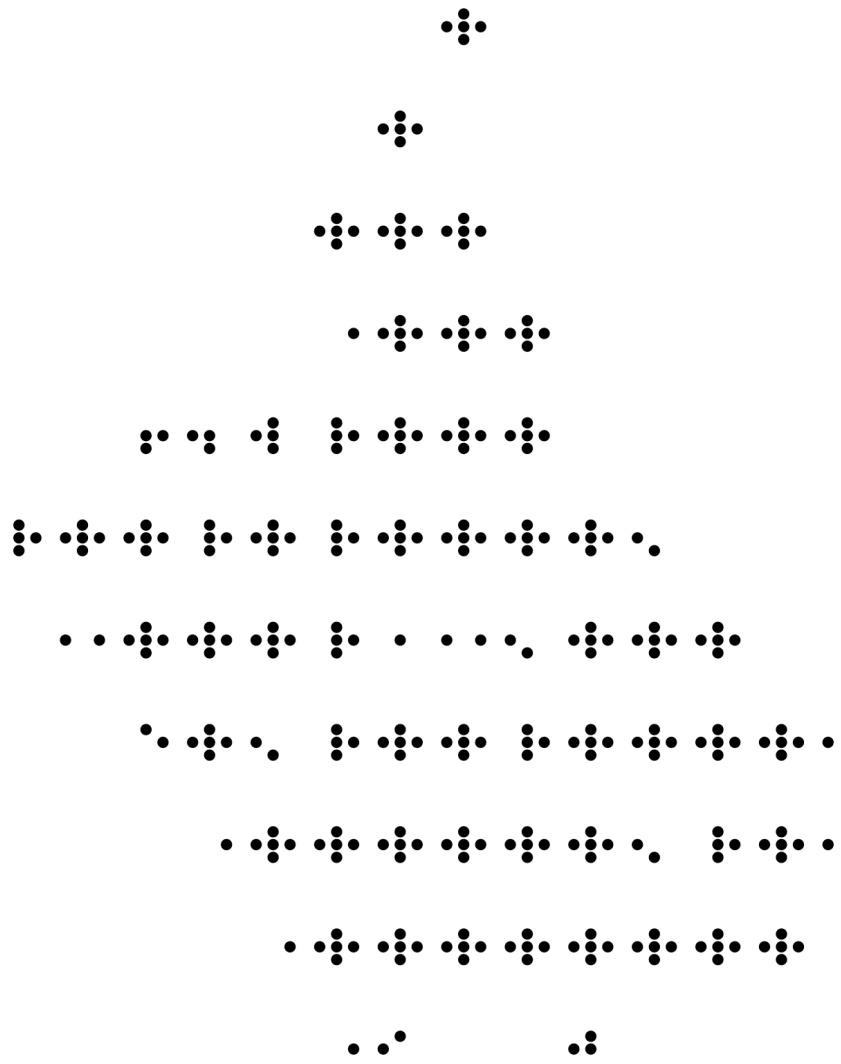
```
library(gstat)
data(meuse.riv)
data(meuse.area)
par(mar=c(1,1,1,1),mfrow=c(1,2))
plot(meuse.area, type = "l", asp =
1,axes=F)
polygon(meuse.riv,col=5)
points(meuse.all[,2:3],cex=1,pch=19)
par(mar=c(4.5,4.5,1,1))
plot(meuse.all$lead,meuse.all$cadmium,cex=
1,pch=19,main="",xlab="Piombo",ylab="Cadmi
o",cex.lab=2)
library(mixreg)
mixr <-
mixreg(meuse.all$lead,meuse.all$cadmium,nc
omp=2)
cvmr <-
covmix(mixr,meuse.all$lead,meuse.all$cadmi
um)
cbdr <-
cband(mixr,cvmr,meuse.all$lead,meuse.all$c
admium)
plot(cbdr)
```



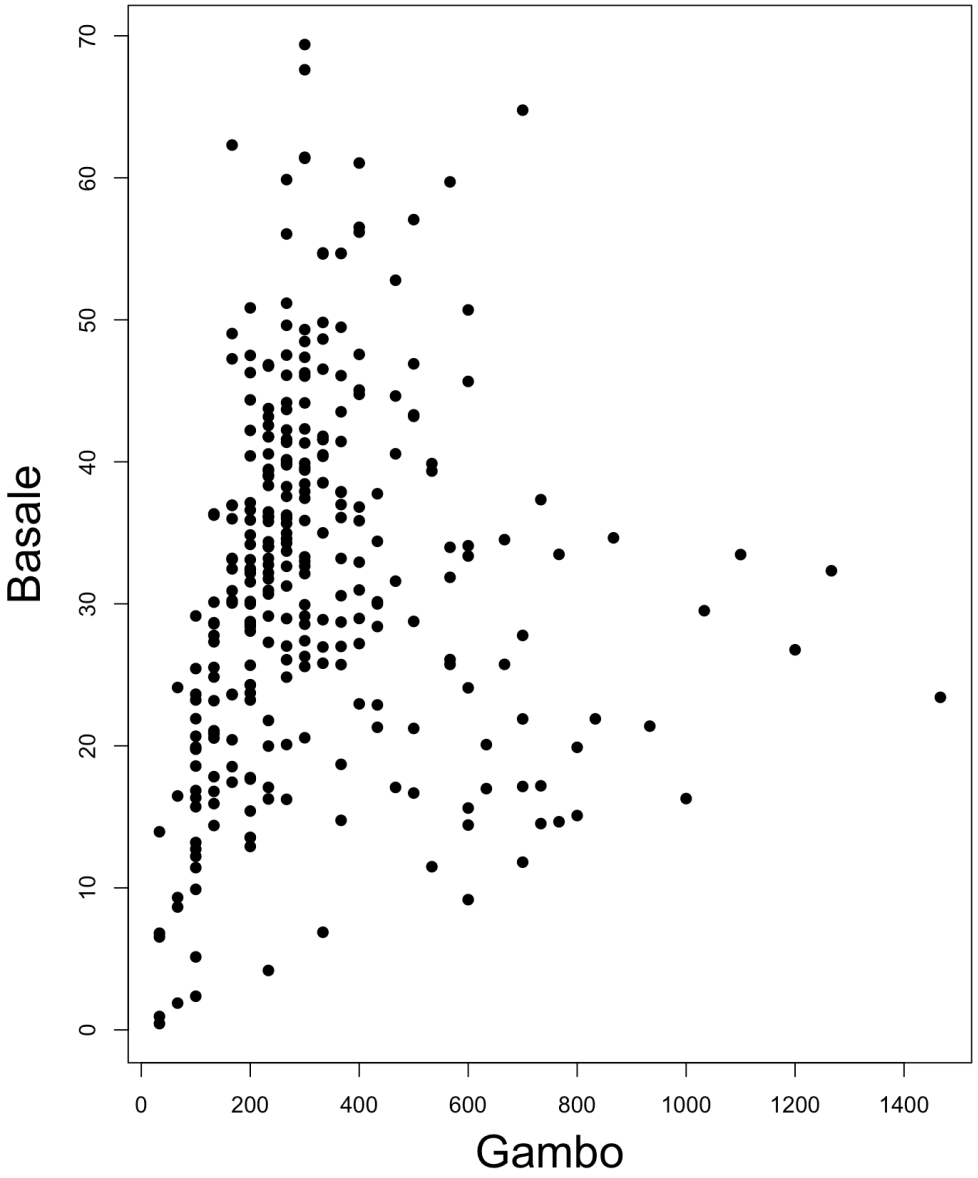
Esempio: Zberg allometric measure

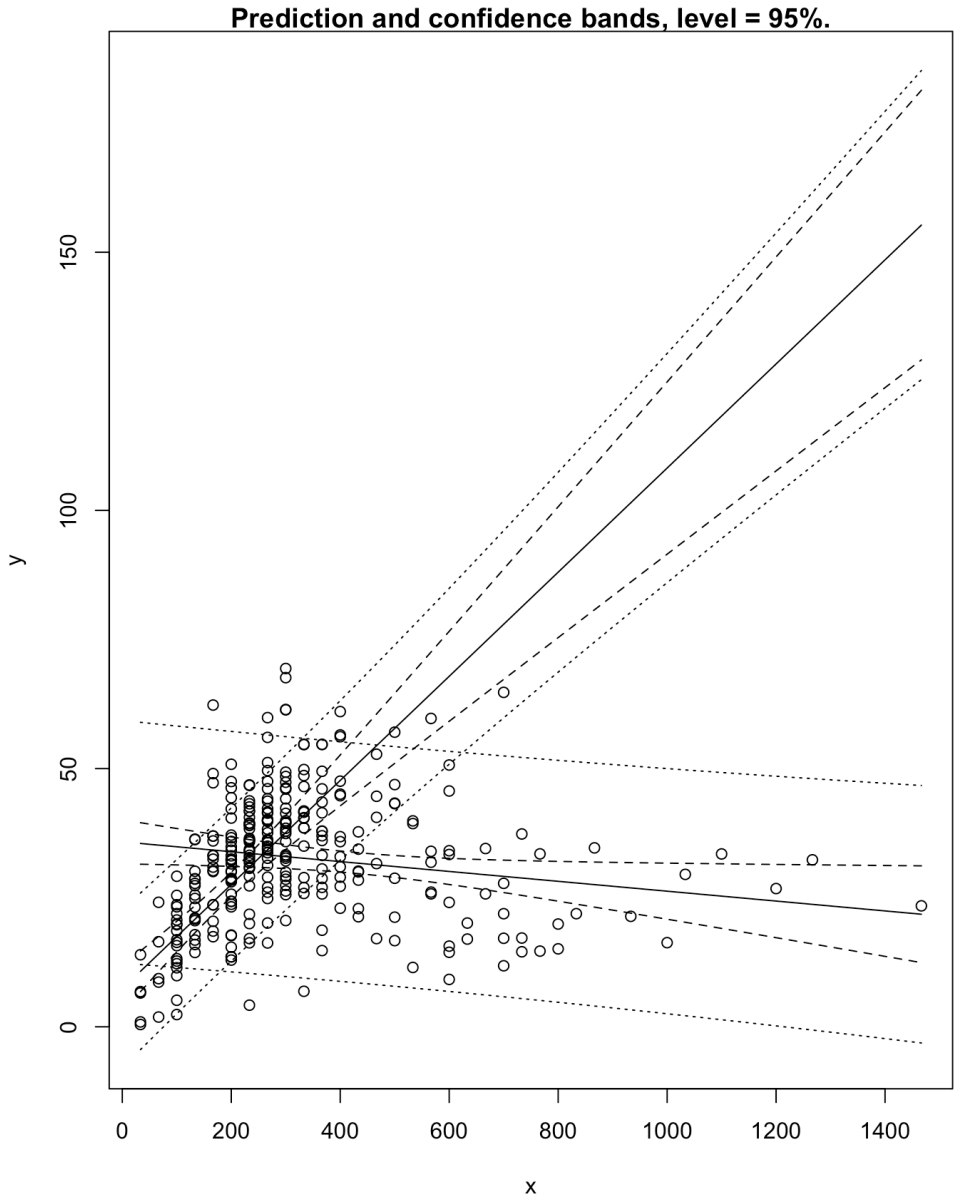
Selezione campionaria di punti in una regione vicino a Zurigo per una indagine forestale in cui vengono effettuate svariate misure quantitative e qualitative della pianta.

L'allometria è quella scienza che studia la dimensione di un organo o di una parte di un organismo rispetto a tutto il corpo dell'organismo



Esempio: Zberg allometric measure





Il trattamento dei dati mancanti

La mancata disponibilità dei dati è un errore che rientra fra gli errori non campionari e si manifesta nella non partecipazione alla rilevazione di unità statistiche appartenenti al campione o nell'assenza di alcuni valori

Si distingue quindi fra 2 tipi di dati mancanti

- **Mancate risposte TOTALI**

- **Mancate risposte PARZIALI**

Le mancate risposte "totali" (unit non response)

Si parla di **mancata risposta "totale"** quando mancano tutte le informazioni relative ad una unità statistica, così che nel database ci si trova di fronte ad un record di meno, ossia di una riga di meno nella matrice dei dati

La presenza delle mancate risposte totali nei dati è un problema comune a tutte le basi di dati, comunque costituite

Tutti gli strumenti adottabili per la prevenzione di tale fenomeno possono ridurne l'intensità, ma non riescono ad eliminarne del tutto la presenza

Il rischio è che, come spesso accade, siano degli individui con determinate caratteristiche a non partecipare, inducendo una distorsione nei dati raccolti

SELECTION BIAS

Che fare ?

Il trattamento delle mancate risposte avviene a 3 livelli:

1. in fase di rilevazione: *richiami*, *sostituzioni, …*
2. in fase di editing: *imputazione*
3. in fase di stima: *riponderazione*, *post-stratificazione …*

Le soluzioni 1 sono impensabili, anche se non hanno molto a che vedere con i metodi statistici ma principalmente con gli aspetti organizzativi;

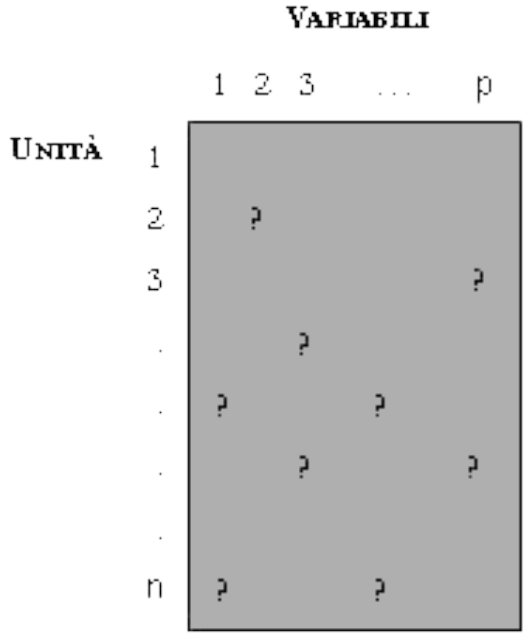
Le soluzioni 2 sono le tecniche maggiormente utilizzate nel caso di mancate risposte parziali o quando si lavora su dati aggregati;

Le soluzioni 3 sono le tecniche maggiormente utilizzate nel caso di mancate risposte totali, specialmente nelle indagini campionarie.

Le mancate risposte "parziali" (item non response)

Si parla di **mancate risposte parziali** quando si è presenza di un record con molti campi pieni, ma non tutti, oppure se tra i valori presenti risulta che alcuni di essi sono non corretti o non validi.

Questo causa una cella vuota nella matrice dei dati



L'imputazione di dati mancanti

Imputazione = sostituzione dei valori mancanti o errati di un dato record con alternative coerenti e plausibili ottenute dai dati stessi (campionari e non), da fonti esterne o dalla combinazione di entrambi, in conformità a regole e metodi prestabiliti.

Pro, contro e considerazioni

- Le procedure d'imputazione hanno l'obiettivo di ridurre le distorsioni introdotte dalla presenza di dati mancanti e di offrire, maggiori garanzie sulla coerenza dei risultati derivati dalle analisi applicate, ma al contempo non confermano la generalità di tale riduzione, addirittura amplificando in taluni casi le distorsioni esistenti
- Mentre i produttori di dati hanno spesso la necessità "istituzionale" di imputare i valori per fornirli agli utenti, indipendentemente dall'uso che questi ne faranno, sono stati sviluppati metodi statistici che non assegnano esplicitamente valori, ma includono il dato mancante all'interno della stessa analisi

Meccanismo di generazione dei dati mancanti

È importante capire se i dati mancanti e i dati osservati hanno strutture comuni

Si distinguono così diversi tipi di dati mancanti, ponendo come elemento cruciale da valutare, nel loro trattamento, se è possibile assumere che il **meccanismo** che gli ha **generati** sia **trascurabile**, oppure comporti delle distorsioni.

È utile capire se i dati mancanti e i dati osservati seguono o meno un certo andamento, e quindi se è possibile definire pattern che descriva i meccanismi che li determinano: ciò è utile per individuare le procedure più adatte per trattare quindi i dati mancanti, richiedendo i diversi approcci un differente impegno computazionale

I pattern possono dipendere tanto dalla natura delle variabili quanto dall'aspetto del fenomeno che con esse si vuole investigare; è possibile considerare tre specifici pattern di dati:

Meccanismo di generazione dei dati mancanti

Si parla di :

- **unit missing** quando mancano interi blocchi di risposte

  ad esempio, *nel corso di studi di tipo longitudinale, i dati mancanti seguono questo modello quando parte dei soggetti decide di abbandonare l'indagine prematuramente, causando uno squilibrio nella matrice dei dati*

- **pattern monotono** quando in uno stesso set di dati, l'insieme delle variabili y_j ha più valori registrati dell'insieme delle variabili y_{j+1} (queste ultime sono cioè maggiormente soggette a non risposta)

- **pattern di dati** quando alcune caratteristiche vengono rilevate solo per un sottoinsieme di casi, si parla allora di **missing by design**

Missing at Random

La probabilità di risposta ad Y (variabili target) dipende:

- dalla variabile X (variabili strutturali, ausiliarie) e dalla variabile Y
- da X ma non da Y
- da X ed eventualmente anche da Y

Se la probabilità di risposta è indipendente sia da X che da Y si dice che i dati mancanti sono **missing at random** e che i dati osservati sono **observed at random**, o più semplicemente che i dati mancanti sono **Missing Completely At Random** (MCAR):

in questo caso i valori osservati della variabile Y formano un sottocampione casuale dei valori già campionati

Missing at Random

Se la probabilità di risposta dipende da X ma non da Y, si dice che i dati mancanti sono **missing at random (MAR)**:

il pattern che definisce il meccanismo di non risposta è ricostruibile o prevedibile dalle altre variabili coinvolte nell'indagine (piuttosto che dalla specifica variabile per la quale mancano alcune determinazioni), dette **mechanism variables.**

Es.: studio sul livello d'ansietà di alcuni individui nel tempo: i partecipanti con un basso livello di autostima saranno meno propensi ad essere coinvolti nelle successive sessioni di ricerca, così il ricercatore può utilizzare tale parametro per prevedere il modello di non risposta

Missing at Random

Se la probabilità di risposta dipende da X ma non da Y, si dice che i dati mancanti sono **missing at random (MAR)**:

il pattern che definisce il meccanismo di non risposta è ricostruibile o prevedibile dalle altre variabili coinvolte nell'indagine (piuttosto che dalla specifica variabile per la quale mancano alcune determinazioni), dette **mechanism variables.**

Se la probabilità di risposta dipende da entrambe le variabili il meccanismo che causa la non risposta non è trascurabile e si parla di **missing not at random (NMAR)**

Le principali soluzioni per il trattamento dei dati mancanti

1. **Procedure basate sull'analisi delle unità completamente registrate**

   non si considerano nell'analisi le unità per le quali manca, in alcuni campi, la registrazione dei valori

PRO facile, piccole quantità di dati mancanti

CONTRO rischio di distorsione, riconducibile ad un *selection bias*

2. **Procedure basate su modelli**

   questa classe generale di procedure prevede che sia generato un modello per i dati parzialmente mancanti, il procedimento inferenziale è quindi basato sulla verosimiglianza sotto quel particolare modello

PRO non ha i rischi dei casi precedenti

CONTRO non esplicita un valore imputato, ma lo ingloba implicitamente

3. **Procedure basate sull'imputazione** singola o multipla

   tecniche d'imposizione di codici plausibili in modo da creare un set di dati completo che può poi essere analizzato con le tecniche standard

PRO apparentemente elimina il problema

CONTRO rischio di errore ulteriore: il codice forzato è plausibile (anche se è considerato "vero")

Analisi dei soli dati disponibili

Si consideri la matrice dei dati (n, k): si eliminano le unità per le quali mancano alcuni valori, e l'analisi si effettua solo sulle $m < n$ unità complete (complete-case analysis): si riduce l'informazione, ma non si modificanoi dati rilevati per svolgere analisi standard

1. **listwise deletion:** si escludono le unità per cui manca anche un solo valore (è il *default* dai più diffusi *software* statistici)
2. la **pairwise deletion**, più complessa, prevede l'inclusione di un'unità solo se vengono registrati i valori relativi ad una predeterminata coppia di variabili delle quali si vuole stimare la correlazione: tale soluzione fornisce la miglior stima per ogni correlazione, e utilizza tutte le informazioni disponibili per quell'obiettivo.

Uso dei modelli di dato mancante

I metodi basati sui modelli non si pongono l'obiettivo di identificare un opportuno valore da assegnare al record con valori mancanti, ma piuttosto di utilizzare tutta l'informazione disponibile per preservare le strutture di associazione e correlazione presente nei dati

Uno degli strumenti più noti è:

L'algoritmo EM (Expectation-Maximization)

Consente di effettuare stime di Massima Verosimiglianza dei parametri di interesse su set di dati incompleti, come se fossero completi

Uso dei metodi di imputazione

Le procedure d'imputazione possono favorire la riduzione delle distorsioni introdotte dalla presenza di dati mancanti, costruendo valori coerenti, ma al contempo non confermano l'universalità di tale riduzione, addirittura amplificando in taluni casi le distorsioni esistenti

L'immissione di dati effettuata ricorrendo a **codici casuali**, o ottenuti formulando **congetture**, comporta notevoli rischi: il sistema migliore per trattare i dati mancanti, secondo alcuni studiosi, è quello di utilizzare il **codice "nessuna risposta"**, rimandando la valutazione del fenomeno alla fase d'interpretazione dei risultati (Es. SPAD, di *default*)

Es.: imputazione della MEDIA
imputazione della MEDIANA
HOT DECK IMPUTATION

Principali problemi: attenzione alla riduzione di variabilità!!!

Algoritmo EM per imputazione

Dato un modello statistico, un insieme di dati osservati Y_{obs}, un insieme di dati mancanti Y_{mis} e un vettore di parametri ignoti θ, partendo da una stima iniziale $\theta_{(0)}$ (*starting guess*), l'algoritmo consiste, ad ogni iterazione *t* di 2 passi:

E-step calcolo del valore atteso $H(\theta, \theta_{(t-1)})$ della verosimiglianza $L(\theta \mid Y_{obs})$ rispetto alla distribuzione dei dati mancanti Y_{mis} condizionatamente ai dati osservati e alle stime correnti dei parametri :

$$H(\theta , \theta_{(t-1)}) \equiv \int L(\theta \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}) f(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs}, \theta_{(t-1)}) d\boldsymbol{Y}_{mis}$$

M-step massimizzazione di $H(\theta , \theta_{(t)})$ rispetto a θ.

L'algoritmo genera una successione $\{\theta (t)\}$ *t=1,2,..* che, sotto alcune ipotesi di regolarità, si dimostra (Dempster et al., 1977) convergere alla stima di massima verosimiglianza di θ.

Imputazione HOT DECK

La ragione primaria per ricorrere a procedure hot deck è la necessità di ridurre le distorsioni prodotte dal verificarsi del fenomeno delle mancate risposte

L'imputazione hot deck utilizza **processi di classificazione** di tutte le unità in gruppi distinti e quanto più omogenei all'interno, in base a caratteristiche precise stabilite di volta in volta in relazione al contesto dell'analisi:

per ogni valore mancante si imputa un valore presente nello stesso gruppo, sotto l'assunzione che all'interno dei gruppi definiti i non rispondenti seguono la stessa distribuzione dei rispondenti

Imputazione HOT DECK: varianti

- **hot deck imputation within adjustment cells**

  una volta costruite le celle d'aggiustamento, si imputano i valori mancanti in ogni singola cella scegliendo tra i valori registrati nelle stesse

- **nearest neighbor hot deck imputation**

  prevede la definizione di una metrica per misurare la distanza tra le diverse unità, e scegliendo quindi come valori da imputare quelli relativi alle unità rispondenti più vicine a quelle che sono invece affette da incompletezza

Imputazione multipla

L'idea di base del metodo di imputazione multipla (proposta da Rubin nel 1987) è quella di generare più di un valore (*m* >2) da imputare per ogni dato mancante campionando da un'opportuna distribuzione, in modo che i data set completati siano *m*. Su ciascuno di essi sono quindi effettuate le analisi statistiche pianificate utilizzando software standard. I risultati delle *m* analisi vengono poi combinati con regole tali che il risultato inferenziale finale tenga conto dell'incertezza causata dalla presenza di dati mancanti, stimata dalla variabilità tra gli *m* risultati indipendenti.

Supponiamo che *Q* sia il parametro incognito di interesse; al termine della procedura di inferenza sugli *m* data set, sono disponibili *m* coppie di valori composte dalla stima puntuale del parametro di interesse $\widehat{Q_i}$ e dalla stima della varianza dello stimatore, $\widehat{U_i}$ (*i*=1, 2, ..., *m*). In accordo alla procedura d'imputazione multipla, la stima puntuale è data dalla media delle singole stime calcolate sulle *m* matrici dei dati completate:

$$\bar{Q} = \frac{1}{m}\sum_{1=1}^{m} \widehat{Q_i}$$

La varianza di questa stima sarà la somma di una componente di variabilità entro imputazione (l'unica di cui terremmo conto in una procedura d'imputazione singola) e da una componente di variabilità tra imputazioni.

Imputazione multipla

La varianza entro imputazione, può essere stimata come media delle varianze $\widehat{U_i}$:

$$\bar{U} = \frac{1}{m}\sum_{1=1}^{m} \widehat{U_i}.$$

La varianza tra imputazioni, *B*, è invece calcolata:

$$B = \frac{1}{m-1}\sum_{1=1}^{m}\left(\widehat{Q_i} - \bar{Q}\right)^2.$$

Combinando queste due componenti si ottiene complessiva:

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B.$$

L'imputazione multipla ha una naturale interpretazione secondo il paradigma bayesiano, in base al quale i valori da imputare sono estratti dalla distribuzione predittiva a posteriori dei dati mancanti, dati i dati osservati, $P(y_{mis}|y_{obs})$. In generale, un campione da questa distribuzione può essere ottenuto per via numerica utilizzando algoritmi MCMC.

Esempio: modello di Solow

Nell'ambito della teoria della crescita in economia, il modello di Solow, o modello neoclassico di crescita, prende il nome dal Premio Nobel Robert Solow, che lo sviluppò nel 1956.

Il modello studia la dinamica della crescita economica di un paese nel lungo periodo. In particolare, nel suo modello Solow rilassa l'ipotesi di costanza del rapporto capitale-prodotto (o intensità di capitale) e, sulla base degli assunti neoclassici, introduce la sostituibilità tra fattori produttivi e dunque la possibilità di aggiustamenti nel lungo periodo del rapporto.

L'introduzione dell'ipotesi di sostituibilità tra lavoro e capitale ha come conseguenza che, nel modello di Solow l'equilibrio di crescita del sistema economico è stabile e la crescita del prodotto pro-capite nel lungo periodo risulta funzione del solo progresso tecnico.

Questo modello tradizionale di crescita ha implicito nel suo framework teorico l'ipotesi di convergenza.

Esempio: modello di Solow

Convergenza significa che le diverse economie devono approssimarsi nel corso del tempo verso un comune livello di reddito di stato stazionario.
In realtà soprattutto negli anni ’60 e ’70 i paesi ricchi sono cresciuti più dei paesi poveri.
L’unica eccezione è stato il caso del Giappone. Oggi osserviamo altri processi di *catching up* che riguardano alcune economie emergenti.
Il fatto stilizzato più forte però sembra l’esistenza di un processo di convergenza all’interno di singoli paesi o tra gruppi di paesi che presentano le stesse caratteristiche.
Questi fatti hanno condotto a nuovi concetti di convergenza:

- Convergenza in termini di tassi di crescita o convergenza in termini di livelli di reddito
- β *convergence* o σ *convergence*
- Convergenza assoluta o convergenza condizionale (diversi parametri, diverse strutture istituzionali etc..)
- Convergenza globale o club *convergence* (o convergenza locale)

Esempio: modello di Solow

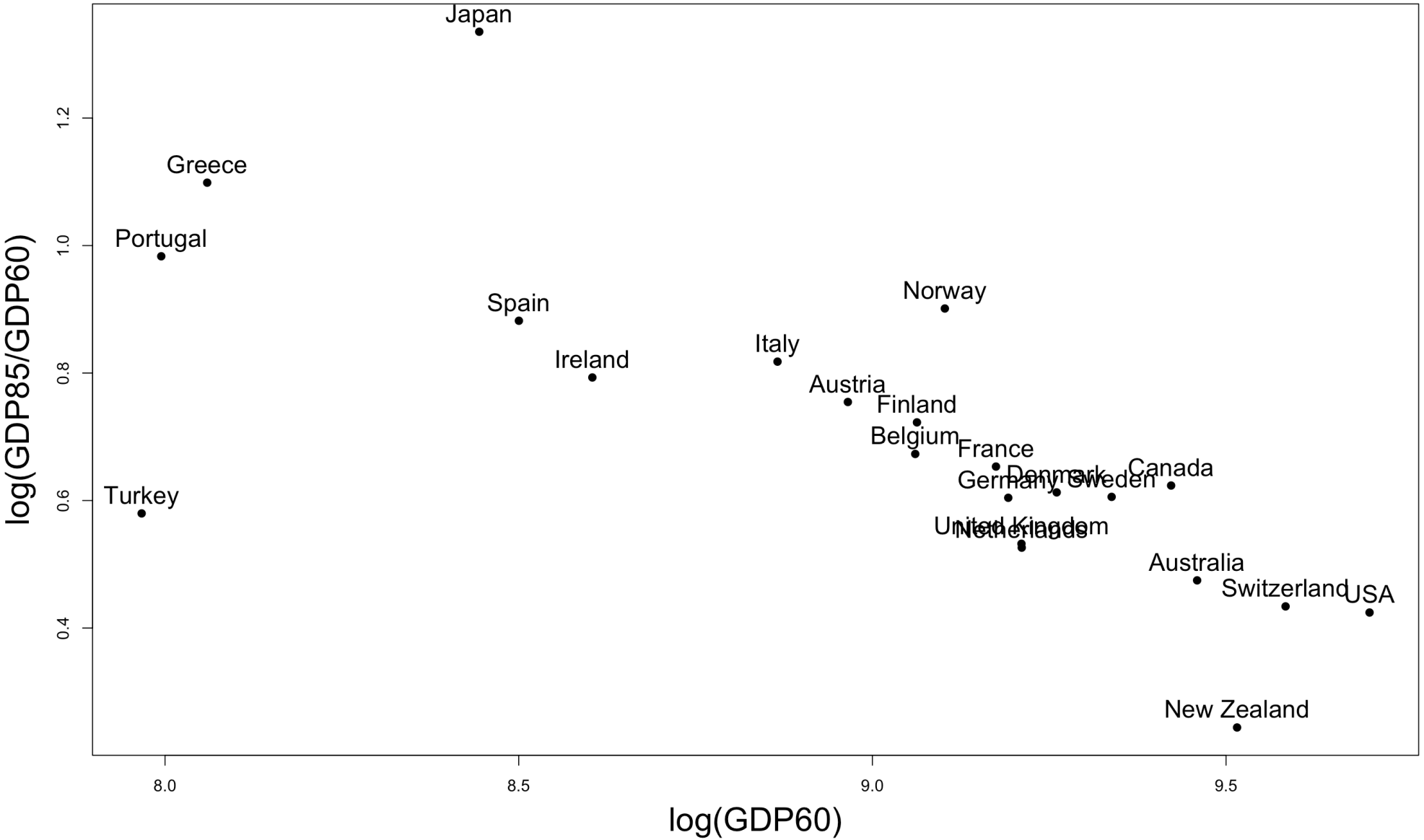
L'equazione da stimare è la seguente :

$$\frac{1}{T} log\left(\frac{y_{i,t}}{y_{i,0}}\right) = \alpha + \beta log\left(y_{i,0}\right) + \gamma X_{i,t} + \varepsilon_{i,t}$$

```
library(AER)
data(OECDGrowth)
solow_lm <- lm(log(gdp85/gdp60) ~
log(gdp60)+log(invest)+log(popgrowth+.05)+log(school)+log(randd),data=OECD
Growth)
summary(solow_lm)
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.67593    1.42052   3.996  0.00104 **
log(gdp60)           -0.51594    0.08501  -6.069 1.62e-05 ***
log(invest)           0.50798    0.18403   2.760  0.01393 *
log(popgrowth + 0.05) -0.44630    0.27742  -1.609  0.12722
log(school)           0.18022    0.11304   1.594  0.13044
log(randd)            0.11771    0.05871   2.005  0.06217 .

Multiple R-squared:  0.8277,  Adjusted R-squared:  0.7739
```

Esempio: modello di Solow



Esempio: modello di Solow

```
pp <- (1/OECDGrowth$gdp60)/sum(1/OECDGrowth$gdp60)
set.seed(4000)
OECDGrowth_miss <- OECDGrowth
OECDGrowth_miss[sample(1:nrow(OECDGrowth),3,replace=F,prob=pp),3] <- NA
OECDGrowth_miss[sample(1:nrow(OECDGrowth),3,replace=F,prob=pp),4] <- NA
OECDGrowth_miss[sample(1:nrow(OECDGrowth),3,replace=F,prob=pp),5] <- NA
OECDGrowth_miss[sample(1:nrow(OECDGrowth),3,replace=F,prob=pp),6] <- NA
solow_lm_miss <- lm(log(gdp85/gdp60) ~
log(gdp60)+log(invest)+log(popgrowth+.05)+log(school)+log(randd),data=OECDGrowth_miss)
summary(solow_lm_miss)
library(miceadds)
md.pattern(OECDGrowth_miss, plot = T)
OECDGrowth_imp <- mice(OECDGrowth_miss, m = 15, seed = 4000)
lmimp <- with(OECDGrowth_imp, lm(log(gdp85/gdp60) ~
log(gdp60)+log(invest)+log(popgrowth+.05)+log(school)+log(randd)))
clmimp <- summary(pool(lmimp))
library(mvdalab)
# OECDGrowth_miss <- introNAs(OECDGrowth, percent = 25)
OECDGrowth_EM <- imputeEM(OECDGrowth_miss,impute.ncomps = 1)
solow_lm_EM <- lm(log(gdp85/gdp60) ~
log(gdp60)+log(invest)+log(popgrowth+.05)+log(school)+log(randd),data=as.data.frame(OECDGrowth_E
M$Imputed.DataFrames))
OECDGrowth_BS <- imputeBasic(OECDGrowth_miss)
solow_lm_BS <- lm(log(gdp85/gdp60) ~
log(gdp60)+log(invest)+log(popgrowth+.05)+log(school)+log(randd),data=OECDGrowth_BS$Imputed.Data
Frame)
compare <-
cbind(summary(solow_lm)$coefficients[,1],summary(solow_lm_miss)$coefficients[,1],summary(solow_l
m_BS)$coefficients[,1],summary(solow_lm_EM)$coefficients[,1],clmimp$estimate)
compstd <-
cbind(summary(solow_lm)$coefficients[,2],summary(solow_lm_miss)$coefficients[,2],summary(solow_l
m_BS)$coefficients[,2],summary(solow_lm_EM)$coefficients[,2],clmimp$std.error)
```

Esempio: modello di Solow

```
compare
                              [,1]        [,2]        [,3]        [,4]        [,5]
(Intercept)             5.6759292  3.29266665  3.62226735  8.26401699  4.23141070
log(gdp60)             -0.5159424 -0.41494482 -0.38355409 -0.63312680 -0.38304074
log(invest)             0.5079847  0.54226470  0.31147982  0.23693699  0.41319516
log(popgrowth + 0.05) -0.4462998 -0.71823366 -0.58418128  0.06267753 -0.34289077
log(school)             0.1802157  0.03731932  0.23488456  0.15873753  0.16027536
log(randd)              0.1177132  0.03619187  0.03977282  0.22626482  0.03599816


compstd
                              [,1]       [,2]       [,3]       [,4]       [,5]
(Intercept)           1.42051819 1.91343783 2.05674117 1.76710297 2.23767795
log(gdp60)            0.08500675 0.09994658 0.10117052 0.09696506 0.10399471
log(invest)           0.18402562 0.20060979 0.30689191 0.23427083 0.35319937
log(popgrowth + 0.05) 0.27741663 0.34425421 0.54931880 0.44253353 0.56478535
log(school)           0.11303993 0.08964249 0.18246921 0.14563931 0.18357710
log(randd)            0.05870640 0.07269077 0.07734862 0.06148103 0.07984595
```