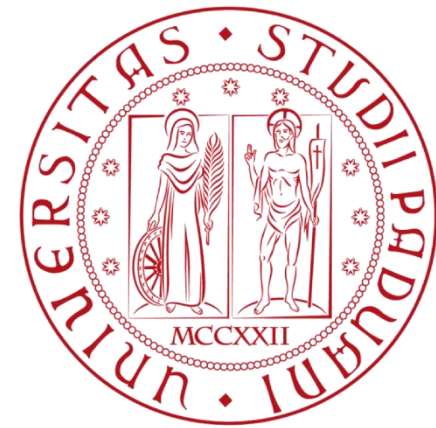




Dati multi-fonte e analisi territoriali

Marco Tosi, Irene Barbiera, e Federico Gianoli

Dipartimento di Scienze Statistiche

Imputazione Multipla Multivariata

- Abbiamo alcune variabili (>1) con valori mancanti (potenzialmente con **una natura diversa**: qualitative categoriali, ordinali, quantitative metriche e di conteggio). E' complicato stimare la distribuzione a posteriori dei parametri con i metodi standard dell'Imputazione univariata. 3 tipi:
 - 1- imputazione per pattern monotono
 - 2- imputazione per distribuzioni normali
 - 3- imputazione tramite equazioni concatenate

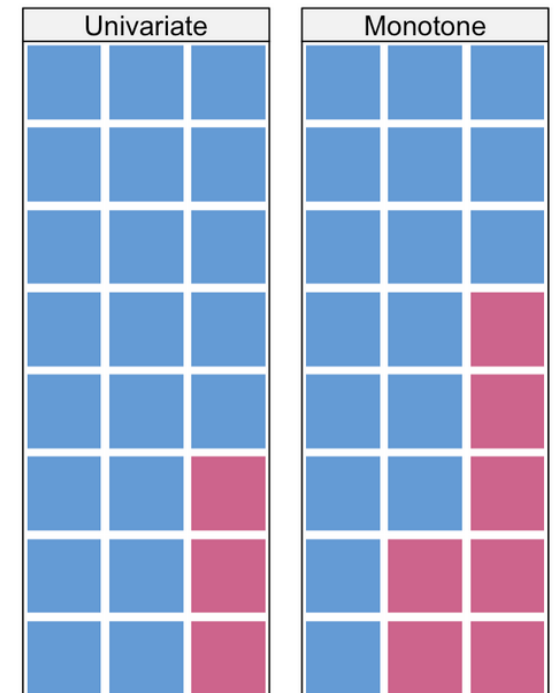
Pattern dei missing

- Il pattern monotono dei dati mancanti può essere dovuto alle cadute nei dati longitudinali oppure (come in SHARE) a rispondenti che non si sottopongono a determinati stimoli (misurazione fisica della salute).
- Si usa una sequenza di imputazioni univariate: Y_1 viene imputato da un set di variabili complete X ignorando Y_p incomplete.

Missing-value patterns
(1 means complete)

Percent	Pattern	
	1	2
63%	1	1
31	0	0
5	1	0
100%		

Variables are (1) chronicw1 (2) maxgrip



1- Imputazione Multivariata per pattern monotono

Si può sintetizzare con $P(Y_j^{\text{mis}}|X, Y_1, \dots, Y_{p-1}, \phi_j)$, where ϕ_j rappresenta il parametro sconosciuto dell'imputazione univariata condizionata ai valori osservati in Y e X e quelli imputati alla variabile precedente Y_{p-1} .
Vantaggi: convergenza veloce, no iterazioni (cambiano i valori iniziali m volte), la natura delle variabili e la conseguente funzione può essere di diverso tipo (logistica, normale, etc..).

1. Sort the data Y_j^{obs} with $j = 1, \dots, p$ according to their missingness.

Ordinare da Y_1 con meno dati mancanti a Y_p con più dati mancanti

2. Draw $\dot{\phi}_1 \sim P(Y_1^{\text{obs}}|X)$.

3. Impute $\dot{Y}_1 \sim P(Y_1^{\text{mis}}|X, \dot{\phi}_1)$.

4. Draw $\dot{\phi}_2 \sim P(Y_2^{\text{obs}}|X, \dot{Y}_1)$.

5. Impute $\dot{Y}_2 \sim P(Y_2^{\text{mis}}|X, \dot{Y}_1, \dot{\phi}_2)$.

1.1- Esempio: Imputazione per pattern monotono

```
. mi impute monotone (poisson) chronicw1 (regress) maxgrip = gender age_int age2 i.country, add(10) noisily replace
```

```
Conditional models:
    chronicw1: poisson chronicw1 gender age_int age2 i.country , noisily
    maxgrip: regress maxgrip chronicw1 gender age_int age2 i.country , noisily
```

Running **poisson** on observed data:

```
Iteration 0: log likelihood = -45823.702
Iteration 1: log likelihood = -45823.66
Iteration 2: log likelihood = -45823.66
```

```
Poisson regression
Log likelihood = -45823.66
Number of obs = 29000
LR chi2(14) = 4905.84
Prob > chi2 = 0.0000
Pseudo R2 = 0.0508
```

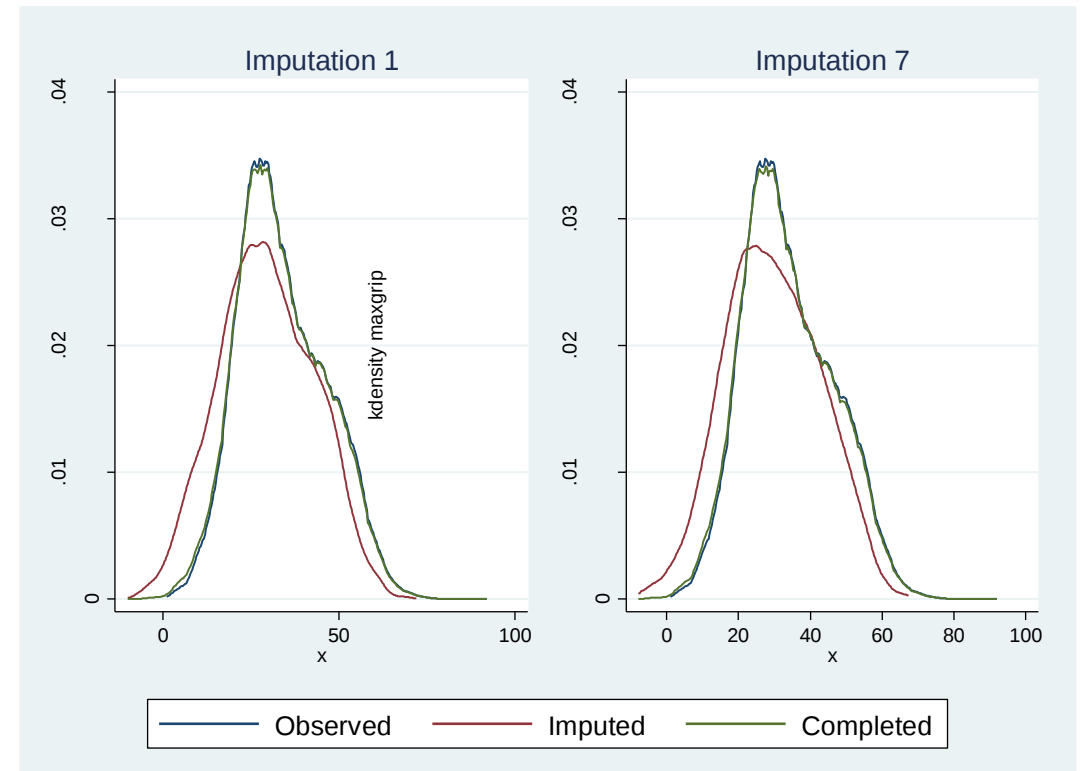
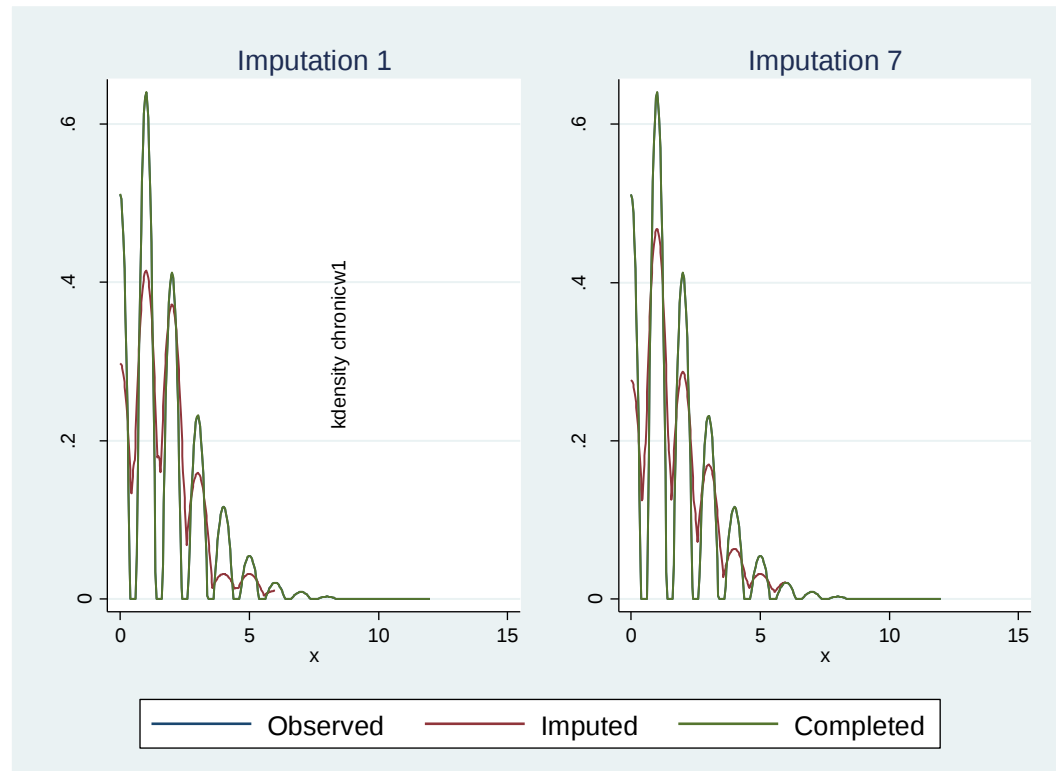
chronicw1	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
gender	.1364891	.0095341	14.32	0.000	.1178027	.1551756
age_int	.1379382	.0056337	24.48	0.000	.1268964	.1489799
age2	-.0008077	.0000408	-19.79	0.000	-.0008878	-.0007277
country						

Running **regress** on observed data:

Source	SS	df	MS	Number of obs = 26696		
Model	2588734.3	15	172582.287	F(15, 26680) = 3006.45		
Residual	1531539.3	26680	57.4040218	Prob > F = 0.0000		
Total	4120273.61	26695	154.346267	R-squared = 0.6283		
				Adj R-squared = 0.6281		
				Root MSE = 7.5765		

maxgrip	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
chronicw1	-.7792959	.0346246	-22.51	0.000	-.8471618	-.7114299
gender	-16.62528	.0933534	-178.09	0.000	-16.80826	-16.4423
age_int	.153978	.0597172	2.58	0.010	.0369291	.2710269
age2	-.0042891	.0004452	-9.63	0.000	-.0051617	-.0034166
country						

1.2- Esempio: diagnostica



2- Imputazione Multivariata per distribuzioni Normali (MVN)

- MVN utilizza un approccio basato sulla **distribuzione normale multivariata** dei parametri e quindi utilizza una unica modellizzazione per le variabili da imputare.
- E' un metodo pensato per variabili continue distribuite normalmente ma può dare risultati robusti anche per variabili ordinali e dicotomiche (Allison, 2001).
- **Pattern** di dati mancanti è **arbitrario** ed è quindi difficile stimare il parametro sconosciuto dalla distribuzione delle variabili osservate. Ricorriamo quindi ad un algoritmo chiamato *Data augmentation*.

2.1- Imputazione Multivariata per distribuzioni Normali (MVN)

- **Data augmentation (DA)** che appartiene alla famiglia di algoritmi del Monte Carlo Markov Chain (MCMC). Si assume che la distribuzione multivariata dei missing condizionata alle variabili osservate/ complete sia approssimabile alla normale.
- Algoritmo Expectation Maximization (**EM**) per i valori iniziali:
 - 1- **Fase di Previsione (Expectation)** utilizza le medie e la matrice di covarianza per costruire un insieme di equazioni di regressione e prevedere i valori incompleti partendo dalle variabili osservate.
 - 2 - **Fase di Massimizzazione (Maximization)** utilizza i dati appena "creati" nella fase E per aggiornare le stime del vettore delle medie e della matrice di varianza e covarianza. Le nuove stime dei parametri vengono utilizzate nel passo E successivo.

2.1 – Esempio MVN

Abbiamo 4 variabili da imputare che assumiamo siano distribuite normalmente, anche se abbiamo visto precedentemente la distribuzione di Casp e Maxgrip (skewed). Abbiamo inoltre Adl2 come dummy e Sphus come variabile ordinale. In questo caso non siamo tanto interessati a riprodurre la distribuzione univariata tra variabili (soprattutto per Adl2 e Sphus) ma vogliamo **ridurre il bias nell'associazione tra variabili** considerando le mancate risposte come elemento di incertezza.

```
mi impute mvn maxgrip adl2 casp sphus = i.country wave hhsiz gender age_int,  
rseed (53421) add(5) noisily
```

```
Performing EM optimization:  
note: 249 observations omitted from EM estimation because of all imputation variables missing
```

2.2 – Esempio MVN

EM converge in 17 iterazioni per la stima dei valori iniziali (procedura già vista nell'imputazione multipla univariata).

L'algoritmo è più lento a convergere se più osservazioni vengono escluse dalla stima. Prior si riferisce alla distribuzione a priori (uniform, Jeffreys [distribuzione non-informativa], ridge [basata su df]). Qui (uniform) tutti i valori dei parametri sono ugualmente probabili.

```
Expectation-maximization estimation      Number obs      =      67313
                                         Number missing   =      20443
                                         Number patterns    =         14
Prior: uniform                          Obs per pattern: min =         1
                                         avg =      4808.071
                                         max =      49360
```

```
Observed log likelihood = -247577.56 at iteration 17
```

	maxgrip	adl2	casp	sphus
Coef				
11b.country	0	0	0	0
12.country	.1481132	-.0020214	.0097676	.2252113
13.country	-.3733106	-.0118922	.7133972	-.3718959
14.country	-.7868113	-.0155719	1.477563	-.0392693
15.country	-6.477767	.0227605	-2.618414	.3691808

2.2 – Esempio MVN

Dopo aver ottenuto i valori iniziali:

```
Performing MCMC data augmentation ...
```

```
Multivariate imputation
```

E imputazione dei 5 dataset:

```
Multivariate imputation          Imputations =      5
Multivariate normal regression      added =      5
Imputed: m=1 through m=5          updated =      0

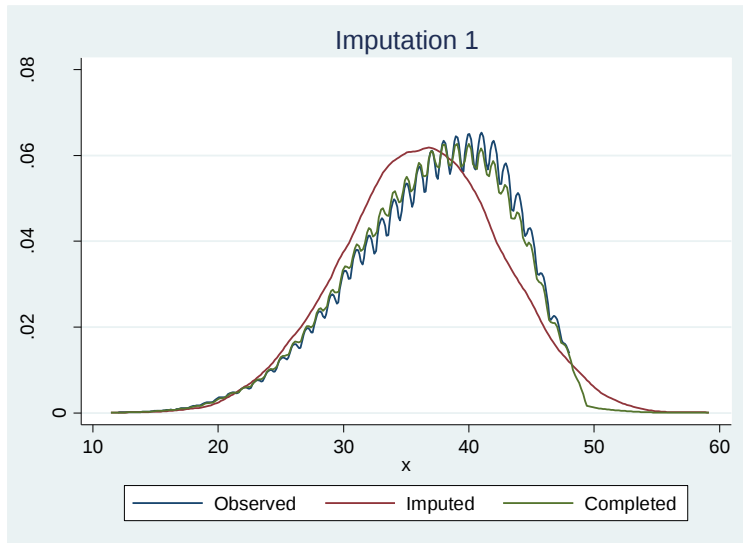
Prior: uniform                    Iterations =     500
                                   burn-in =     100
                                   between =     100
```

Variable	Observations per m			
	Complete	Incomplete	Imputed	Total
maxgrip	61454	6108	6108	67562
adl2	67249	313	313	67562
casp	52893	14669	14669	67562
sphus	67213	349	349	67562

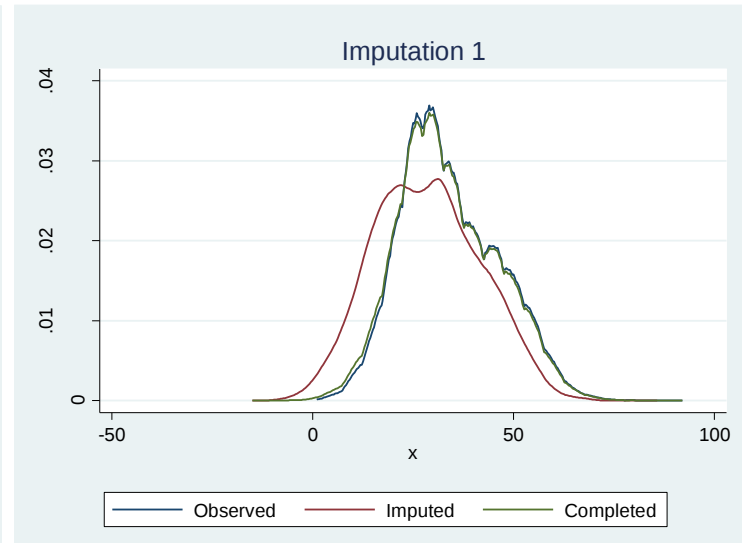
```
(complete + incomplete = total; imputed is the minimum across m
of the number of filled-in observations.)
```

2.3 – Esempio MVN: diagnostica

CASP



Maxgrip



Sphus

Proportions of sphus for m=1

Number of observed = 67213
Number of imputed = 349
Number of completed = 67562

Self-perceived health - us version	Observed	Imputed	Completed
-.11051963		0.003	0.000
.18370019		0.003	0.000
.53293246		0.003	0.000
.69702286		0.003	0.000
.81570554		0.003	0.000
.9382537		0.003	0.000
.96542102		0.003	0.000
Excellent	0.097	0.000	0.096
1.0002817		0.003	0.000

3- Equazioni concatenate (MICE)

- E' il metodo più comune per imputare i dati mancanti e consiste in **un mix di regressioni sequenziali per variabili di varia natura (continue, nominali, ordinali, di conteggio)**.
- Durante la prima iterazione, Y_1 , la variabile da imputare con il minore numero di missing, viene regredita su tutte le altre variabili complete. Nella seconda fase, la variabile Y_2 con il minor numero di missing viene regredita su tutte le variabili complete, più Y_1 imputata. Una iterazione consiste in un ciclo di imputazioni su tutte le variabili da imputare Y_j .
- L'imputazione avviene attraverso una estrazione casuale dei valori dalla distribuzione a posteriori ottenuta dal modello di imputazione.

3.1- Equazioni concatenate (MICE)

- Basate su un tipo di algoritmo MCMC. Le catene Monte Carlo di Markov (MCMC) approssimano l'estrazione (pseudo-random) dei valori da imputare da una distribuzione sconosciuta e multidimensionale. Tecnica iterativa di simulazione della distribuzione a posteriori per pattern arbitrari di dati mancanti.
- L'algoritmo procede così fino a che non viene fatto il modello di regressione per $Y_{j(m)}$ con il maggior numero di valori mancanti.
- Questa procedura viene ripetuta più volte (n=10 iterazioni) per **completare un ciclo e produrre un dataset imputato.**

3.2- Equazioni concatenate (MICE)

- In sostanza utilizziamo un modello di regressione a seconda della natura della variabile da imputare per stimare la distribuzione condizionata di Y_1 . Le imputazioni avvengono attraverso una estrazione stocastica da questa distribuzione. **Y_1 imputata serve a predire Y_2 fino a completare un ciclo di imputazione.**
- Questa procedura viene ripetuta più volte per avere M dataset imputati.
- Il modello analitico deriva dalla semplice media aritmetica dei parametri stimati nei singoli M dataset.

3.1 – Un esempio MICE

mi impute chained (reg) maxgrip (logit) adl2 (reg) casp (ologit) sphus
= i.country wave hhsize gender age_int, rseed (53421) replace noisily

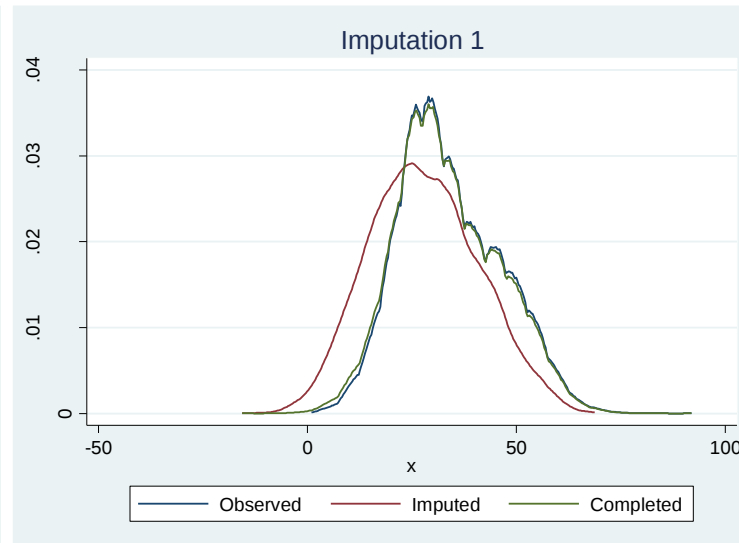
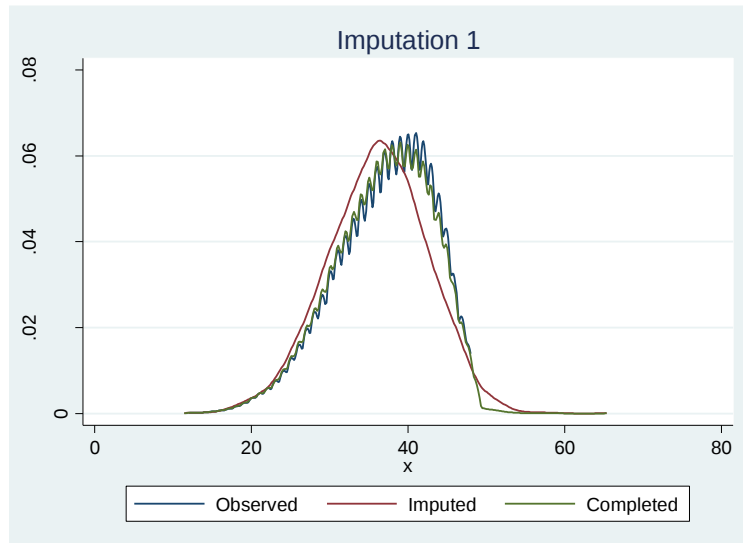
Running regress on data from iteration 8, m=1:

Adl2, Sphus, e Casp imputate precedentemente servono a predire maxgrip nell'iterazione 8 (ad esempio)

maxgrip	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
adl2						
1+ adl limitations	-2.191482	.115534	-18.97	0.000	-2.417929	-1.965035
sphus						
Very good	-.2361878	.112496	-2.10	0.036	-.4566803	-.0156953
Good	-.9974089	.1064887	-9.37	0.000	-1.206127	-.7886907
Fair	-2.279448	.1188525	-19.18	0.000	-2.512399	-2.046497
Poor	-4.125177	.1603997	-25.72	0.000	-4.43956	-3.810793
casp	.1361096	.0057744	23.57	0.000	.1247919	.1474274

3.2 – Un esempio MICE

```
mi impute chained (reg) maxgrip (logit) adl2 (reg) casp (ologit) sphus  
= i.country wave hhsiz gender age_int, rseed (53421) replace noisily
```



```
Number of observed = 67249  
Number of imputed  = 313  
Number of completed = 67562
```

	Observed	Imputed	Completed
1+ adl limitations			
No adl limitations	0.896	0.856	0.896
1+ adl limitations	0.104	0.144	0.104

Modello Analitico:

Coef. Svezia= -0.36 (dati completi); -0.04 (MVN);
-0.04 (MICE)