

Dati multi-fonte e analisi territoriali

Marco Tosi, Irene Barbiera, e Federico Gianoli

Dipartimento di Scienze Statistiche

Approcci Tradizionali o Deterministici

- Si tratta di una **sostituzione** dei valori mancanti con valori prodotti “**artificialmente**” (es. la media) che consente di riprodurre un dataset completo sul quale utilizzare strumenti di analisi standard.
- I valori imputati tendono ad essere considerati come quelli osservati. Perciò la **variabilità delle stime associata alla non-risposta** viene trascurata (e SE sottostimati).
- Notare come cambia la distribuzione della variabile imputata e le conseguenze che può avere sull’associazione tra variabili (che dipende da N. missing).

Approcci Moderni (Metodi di Imputazione Multipla)

- **Imputazione Multipla** (Rubin 1978, 1987): procedura attraverso la quale imputiamo i valori mancanti diverse volte per produrre diverse stime plausibili. Consiste nella ripetizione del **processo di imputazione m volte ($m > 2$)** e nella generazione di un insieme di **m data-set completi**.
 - **Univariata**: imputiamo i valori di una sola variabile sulla base di altre osservate / complete tenendo conto della mancata risposta come sorgente di incertezza.
 - **Multivariata**: imputiamo i valori mancanti di più variabili sulla base di quelle osservate e sulla base delle imputazioni precedenti (**ricorsività, concatenazione**).
- I risultati delle analisi saranno così svolti su m dataset e verranno combinati in modo da tenere in conto dell'**incertezza causata dalla presenza di dati mancanti** (stimata dalla variabilità tra dagli m data-set imputati).
- **Imputazione Singola** quando il processo di imputazione viene fatto una volta su $m=1$ in base a dataset completi

Caratteristiche dell'Imputazione Multipla

- **Imputazione Multipla** (Rubin 1978, 1987) si basa su modelli statistici parametrici o semi-parametrici.
- **Trasparenza** del modello di imputazione (es., quali variabili e perchè).
- **Replicabilità** (es., settare i valori iniziali)
- Ha lo scopo di **preservare le proprietà dei dati osservati** (medie, varianze, e covarianze)
- Dipende dallo **scopo di una ricerca** o dalle relazioni tra variabili che vogliamo studiare (es. scelta delle variabili) e potenzialmente coinvolgere un set di variabili più ampio rispetto a quelle utilizzate in fase analitica.
- Il modello di imputazione scelto si basa spesso su una precisa distribuzione.

Caratteristiche dell'Imputazione Multipla

- **Imputazione Multipla** nel senso che il processo di imputazione viene **ripetuto m volte**. In teoria l'inferenza dei dati mancanti è efficiente quando $M = \infty$. Tuttavia quando $M=5, 10, 20$ per ripetizioni indipendenti del processo di imputazione otteniamo una inferenza efficiente dei dati mancanti.
- Mantiene le caratteristiche multivariate dei dati attraverso una sequenza di **imputazioni condizionate** ai dati osservati, oppure attraverso metodi la **catene Monte Carlo di Markov (MCMC)** che estraggono/ simulano i valori mancanti dalla distribuzione a posteriori del modello multivariato.

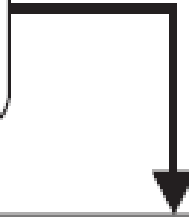
Tre fasi dell'imputazione multipla

- 1- Imputazione:** i valori mancanti sono sostituiti dai valori stimati e viene creato un dataset completo. Questo processo di «sostituzione» o «completamento» viene ripetuto m volte.
- 2- Analisi:** ogni dataset è analizzato utilizzando il metodo di interesse (es, regressione)
- 3- Pooling (integrazione):** le stime dei parametri (coefficienti e errori standard) ottenuti dalle analisi di tutti i dataset sono combinati per l'inferenza.

Step 1—Create 5 to 10 data sets
using data augmentation

Step 2—Estimate the model
(e.g., regression, logistic
regression, SEM) separately for
each of the 5 to 10 data sets
using data augmentation

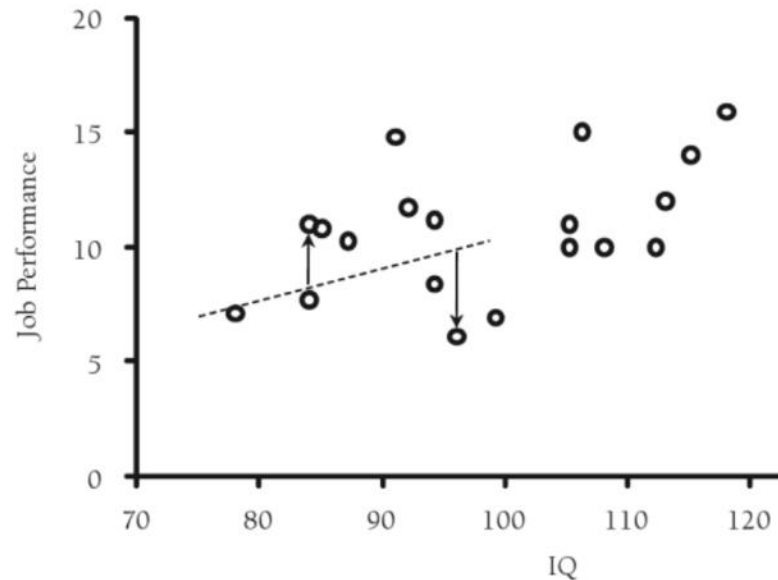
Step 3—Compute pooled
estimates of the parameters and
standard errors using the 5 to 10
solutions



Fase 1: Come avviene la fase 1 di Imputazione.

Il caso più facile: Imputazione multipla univariata

- Regredendo X su Y sui dati osservati/ completi otteniamo la distribuzione per ogni valore mancante di Y: $\hat{y} = \hat{\beta}_0 + X_{\text{mis}}\hat{\beta}_1 + \epsilon$
- Dove $\hat{\beta}_0$ and $\hat{\beta}_1$ sono stimati dal modello di regressione e ϵ è un valore casuale della distribuzione normale standardizzata $\epsilon \sim N(0, \hat{\sigma}^2)$



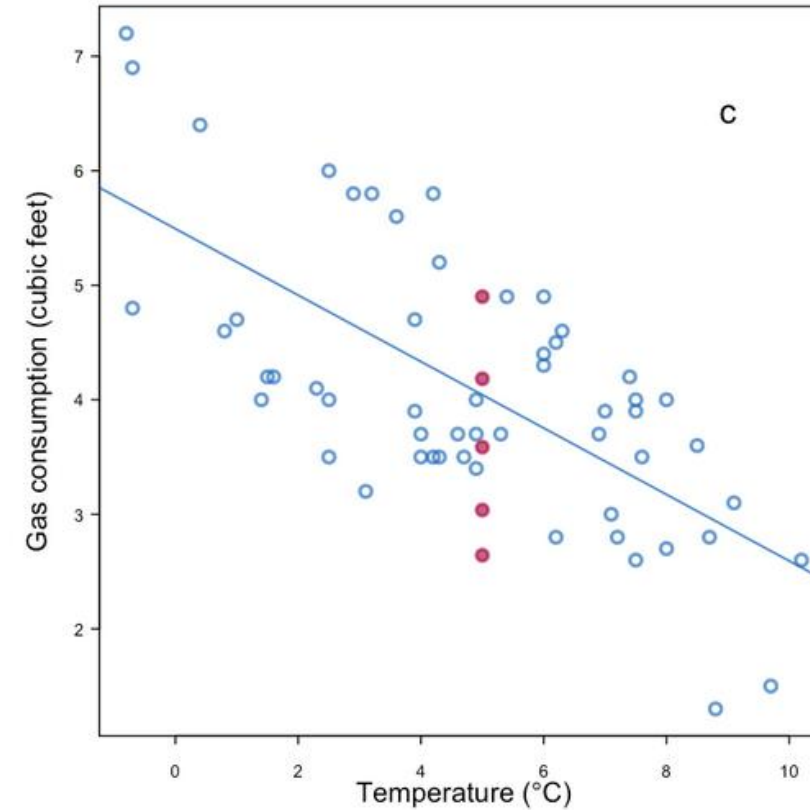
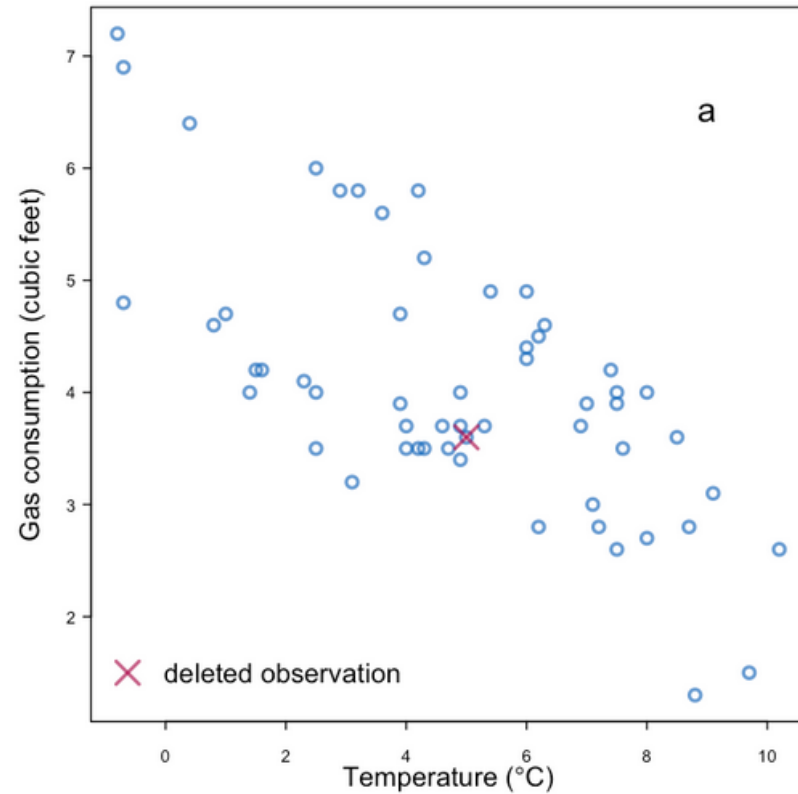
Regressione
stocastica di
imputazione

Fase 1: Imputazione univariata

Utilizziamo le tecniche tipiche dell'imputazione singola, ossia modelli di regressione per predire le variabili incomplete con i valori mancanti (Y) partendo dalle variabili complete (X).

- 1- Modello di regressione sui dati completi
- 2- Dal modello calcoliamo i valori predetti e l'errore standard delle stime. I valori predetti per i missing sono basati sulle altre variabili inserite nel modello.
- 3- Aggiungiamo variabilità random ai valori che vogliamo imputare, ossia moltiplichiamo l'errore standard con una variabile casuale normalmente distribuita,

Fase 1: Imputazione univariata

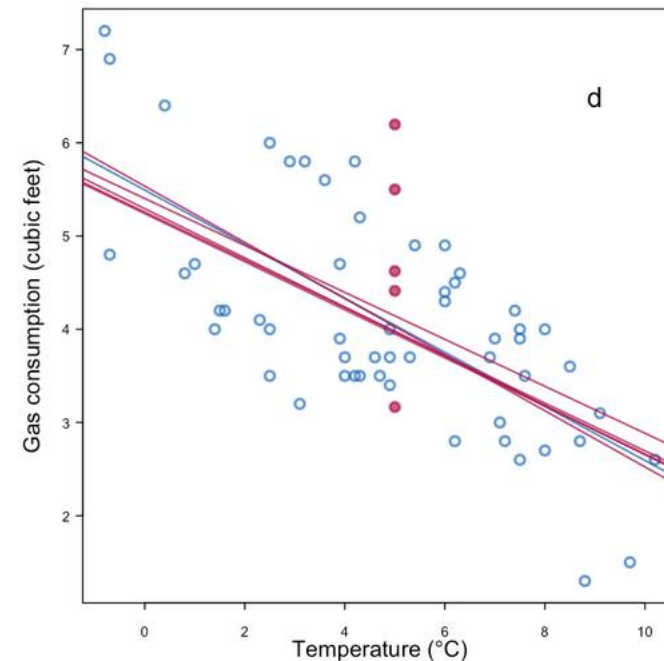


Variabilità di intercetta

Van Buuren (2021) Flexible Imputation of Missing Data

Fase 1: Imputazione Univariata

- **Nell'Imputazione Multipla Univariata** utilizziamo le previsioni di un modello di regressione aggiungendo un elemento di variabilità casuale. Abbiamo tuttavia una parte fissa, ossia assumiamo che l'intercetta, la pendenza, e la deviazione standard dei residui siano conosciuti.
- Aggiungiamo un elemento di incertezza sui parametri basato sulla distribuzione a posteriori, secondo $\hat{y} = \hat{\beta}_0 + X_{\text{mis}}\hat{\beta}_1 + \epsilon$ dove ϵ è $\epsilon \sim N(0, \hat{\sigma}^2)$ e $\hat{\sigma}$ un valore casuale della distribuzione a posteriori.



Fase 1: E' Multipla

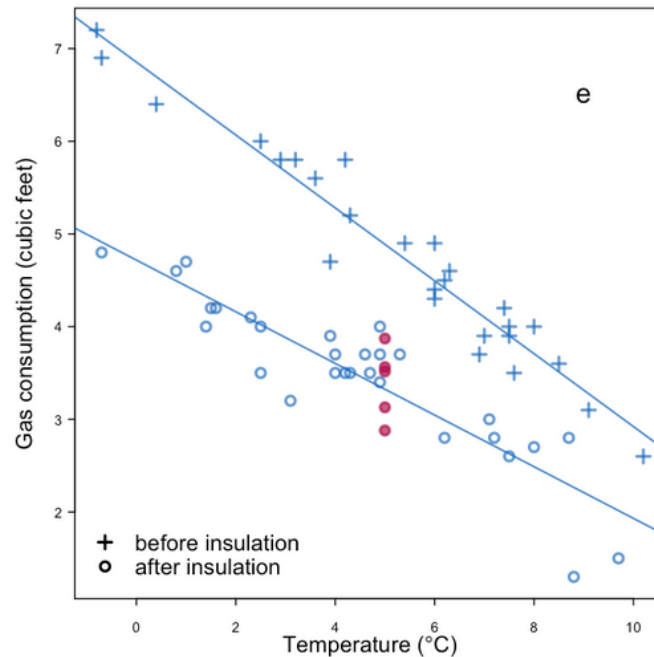
- Abbiamo quindi una **Imputazione Multipla** quando il processo viene ripetuto m volte per m dataset ($m > 2$; es, $m = 3$). Queste ripetizioni sono basate sulla distribuzione a posteriori dei parametri condizionata ai dati osservati.

Subject	Data		Imputation 1		Imputation 2		Imputation 3		Imputation 4	
	Y	X	Y	X	Y	X	Y	X	Y	X
1	1.1	3.4	1.1	3.4	1.1	3.4	1.1	3.4	1.1	3.4
2	1.5	3.9	1.5	3.9	1.5	3.9	1.5	3.9	1.5	3.9
3	2.3	2.6	2.3	2.6	2.3	2.6	2.3	2.6	2.3	2.6
4	3.6	1.9	3.6	1.9	3.6	1.9	3.6	1.9	3.6	1.9
5	0.8	2.2	0.8	2.2	0.8	2.2	0.8	2.2	0.8	2.2
6	3.6	3.3	3.6	3.3	3.6	3.3	3.6	3.3	3.6	3.3
7	3.8	1.7	3.8	1.7	3.8	1.7	3.8	1.7	3.8	1.7
8	?	0.8	0.2	0.8	0.8	0.8	0.3	0.8	2.3	0.8
9	?	2.0	1.7	2.0	2.4	2.0	1.8	2.0	3.5	2.0
10	?	3.2	2.7	3.2	2.5	3.2	1.0	3.2	1.7	3.2

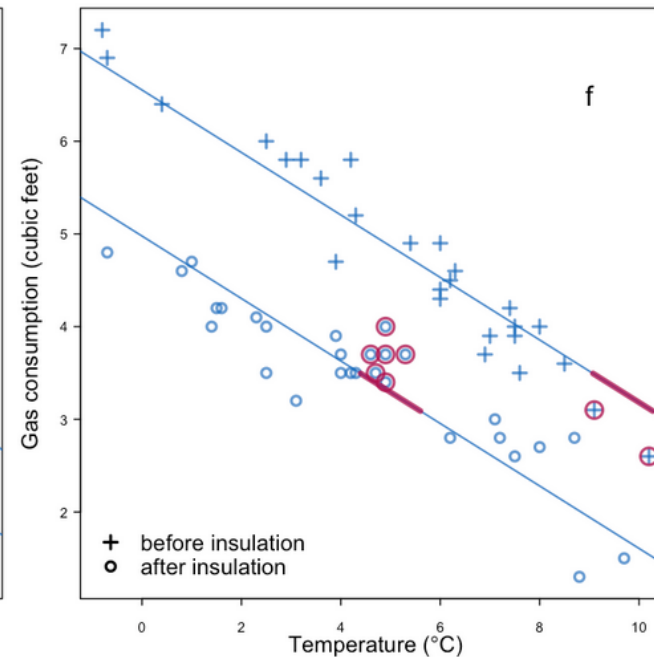
Fase 1: Variabili nel modello di imputazione

- Solitamente nel modello di imputazione includiamo più variabili per imputare i valori mancanti in modo più **accurato** (variabilità diminuisce e produce stime migliori). Qui un esempio con una terza variabile dummy.

Regressione OLS



PMM predictive mean matching



Fase 1: Variabili nel modello di imputazione

- **La scelta delle variabili non** è limitata a quelle che hanno valori mancanti da imputare e a quelle che useremo nel modello analitico. Solitamente utilizziamo anche variabili che possano predire la variabile da imputare e quelle che potrebbero predire la generazione dei missing.
- Se non includiamo tutte le variabili del modello analitico introduciamo una distorsione nelle associazioni studiate modello deve essere **appropriato** (Rubin 1996). Se X è correlato a Y ma non viene usato nel modello di imputazione, abbiamo valori imputati di Y indipendenti da X, così la relazione tra X e Y è distorta verso lo 0.

Fase 1: Variabili nel modello di imputazione

- La scelta delle variabili introduce un altro problema (che vedremo nei Modelli di imputazione Multivariata): quando includiamo tante variabili di diversa natura (quantitative continue, di conteggio, qualitative categoriali, ecc...) che possono includere dati mancanti (*item non-response*) è praticamente **impossibile specificare la forma della distribuzione a posteriori** congiunta di queste variabili.
- Le **Catene Monte Carlo di Markov (MCMC)** sono un metodo per approssimare le estrazioni partendo da **una distribuzione a posteriori congiunta sconosciuta**.

Fase di Imputazione in Stata

- Dataset con 2 variabili a (regular), b (imputed). Creiamo 3 dataset (con anche $c = a + b$, come variabile passiva).

```
mi set mlong
```

```
mi register imputed maxgrip
```

```
6109 m=0 obs. now marked as incomplete)
```

```
mi register regular mergeid country wave gender age_int hhsize adl
```

a	b
1	2
4	.



$m=0$:

a	b	c
1	2	3
4	.	.

$m=1$:

a	b	c
1	2	3
4	4.5	8.5

$m=2$:

a	b	c
1	2	3
4	5.5	9.5

Stili diversi

- Wide

	a	b	c	_1_b	_2_b	_1_c	_2_c	_mi_miss
1.	1	2	3	2	2	3	3	0
2.	4	.	.	4.5	5.5	8.5	9.5	1

- Flong

	a	b	c	_mi_miss	_mi_m	_mi_id
1.	1	2	3	0	0	1
2.	4	.	.	1	0	2
3.	1	2	3	.	1	1
4.	4	4.5	8.5	.	1	2
5.	1	2	3	.	2	1
6.	4	5.5	9.5	.	2	2

- Mlong

	a	b	c	_mi_miss	_mi_m	_mi_id
1.	1	2	3	0	0	1
2.	4	.	.	1	0	2
3.	4	4.5	8.5	.	1	2
4.	4	5.5	9.5	.	2	2

mi set mlong

Modello di imputazione

```
mi impute regress maxgrip i.country wave gender age_int hhsz adl, add(20) rseed(2232) no  
isily
```

unning regress on observed data:

Source	SS	df	MS
Model	5697907.85	19	299889.887
Residual	3324389.32	61443	54.1052572
Total	9022297.17	61462	146.794721

```
Number of obs = 61463  
F( 19, 61443) = 5542.71  
Prob > F      = 0.0000  
R-squared     = 0.6315  
Adj R-squared = 0.6314  
Root MSE     = 7.3556
```

maxgrip	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
country						
Germany	.1070684	.1822272	0.59	0.557	-.2500973	.4642341
Sweden	-.4831797	.1803796	-2.68	0.007	-.8367242	-.1296351
Netherlands	-.9025835	.1813169	-4.98	0.000	-1.257965	-.547202
Spain	-6.417636	.1898281	-33.81	0.000	-6.7897	-6.045573

Risultati del Modello di imputazione

Univariate imputation
Linear regression
Imputed: m=1 through m=20

Imputations = 20
added = 20
updated = 0

Variable	Observations per m			Total
	Complete	Incomplete	Imputed	
maxgrip	61463	6109	6109	67572

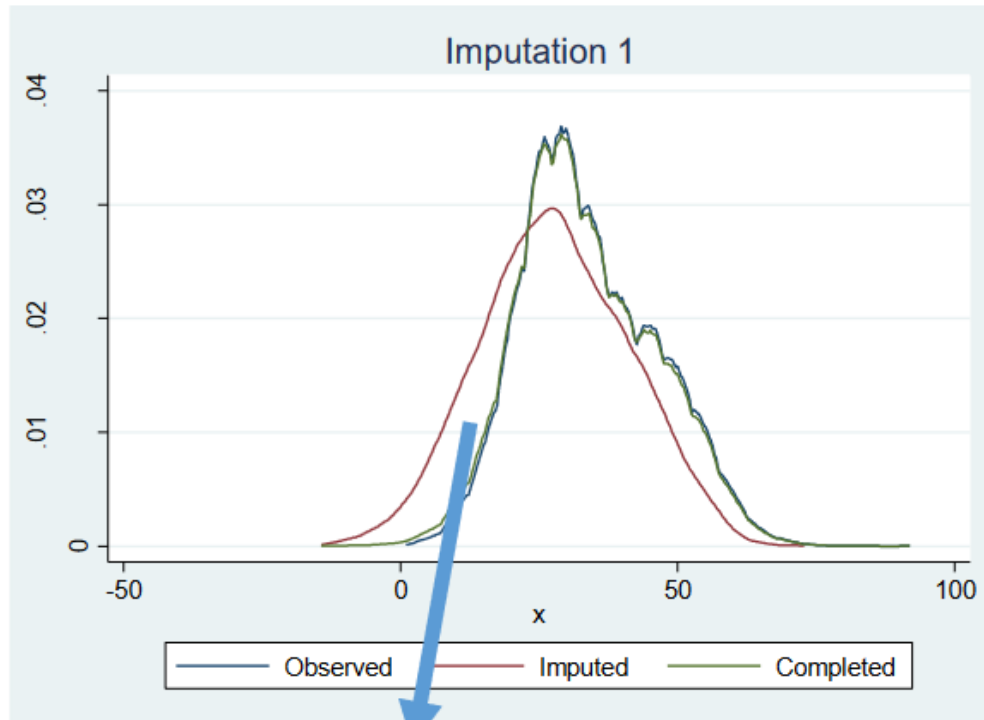
(complete + incomplete = total; imputed is the minimum across m of the number of filled-in observations.)

```
. list maxgrip age_int _mi_id _mi_miss _mi_m if _mi_id ==
```

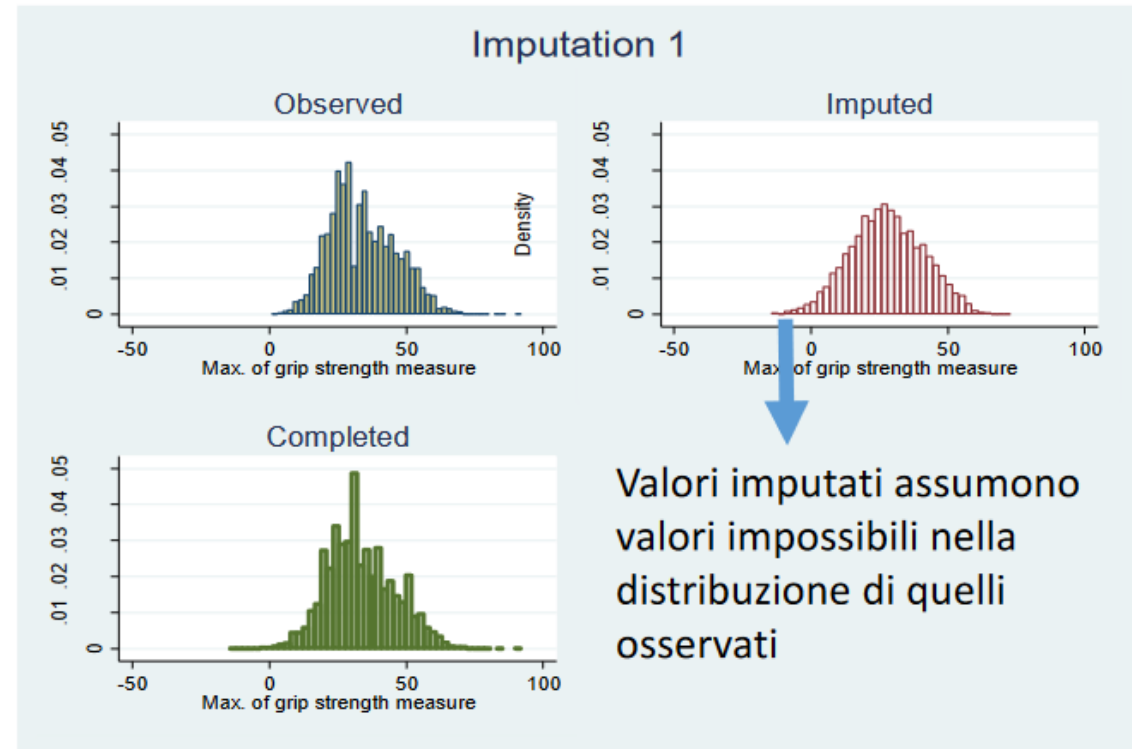
	maxgrip	age_int	_mi_id	_mi_miss	_mi_m
8.	.	70	8	1	0
21.796093	70	8	.	1	
28.755367	70	8	.	2	
22.112566	70	8	.	3	
15.874813	70	8	.	4	
19.866873	70	8	.	5	
23.204582	70	8	.	6	
12.764587	70	8	.	7	
16.141899	70	8	.	8	
11.882469	70	8	.	9	
18.148088	70	8	.	10	
22.607437	70	8	.	11	
26.27165	70	8	.	12	
21.346783	70	8	.	13	
31.968353	70	8	.	14	

Diagnostica Sulla distribuzione

- Modello di imputazione: Regressione lineare OLS



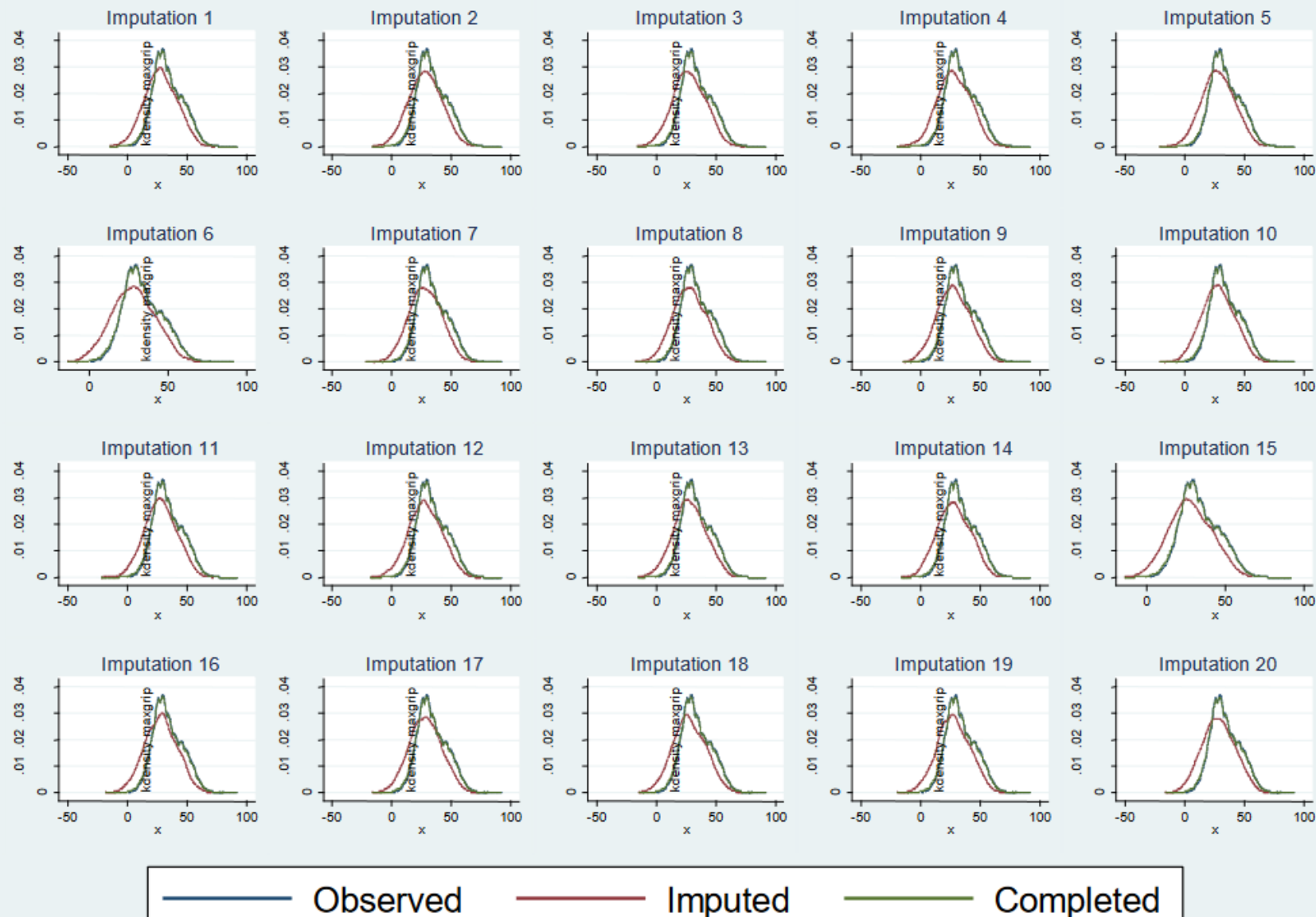
Assunzioni sulla distribuzione normale



Diagnostica

20 imputazioni diverse
(anche se simili).

Se guardiamo solo un
dataset imputato
potremmo avere
distribuzioni distorte
dovute al caso



Diagnostica

- M=0 dataset senza imputazione
- M=1 dataset 1 con imputazione
- Si noti il range della variabile maxgrip
- Si noti la media e SD della variabile maxgrip

m=0 data:

-> summarize maxgrip gender

Variable	Obs	Mean	Std. Dev.	Min	Max
maxgrip	61463	34.1811	12.11589	1	92
gender	67572	1.557228	.4967179	1	2

m=1 data:

-> summarize maxgrip gender

Variable	Obs	Mean	Std. Dev.	Min	Max
maxgrip	67572	33.60823	12.37259	-14.34454	92
gender	67572	1.557228	.4967179	1	2

m=10 data:

-> summarize maxgrip gender

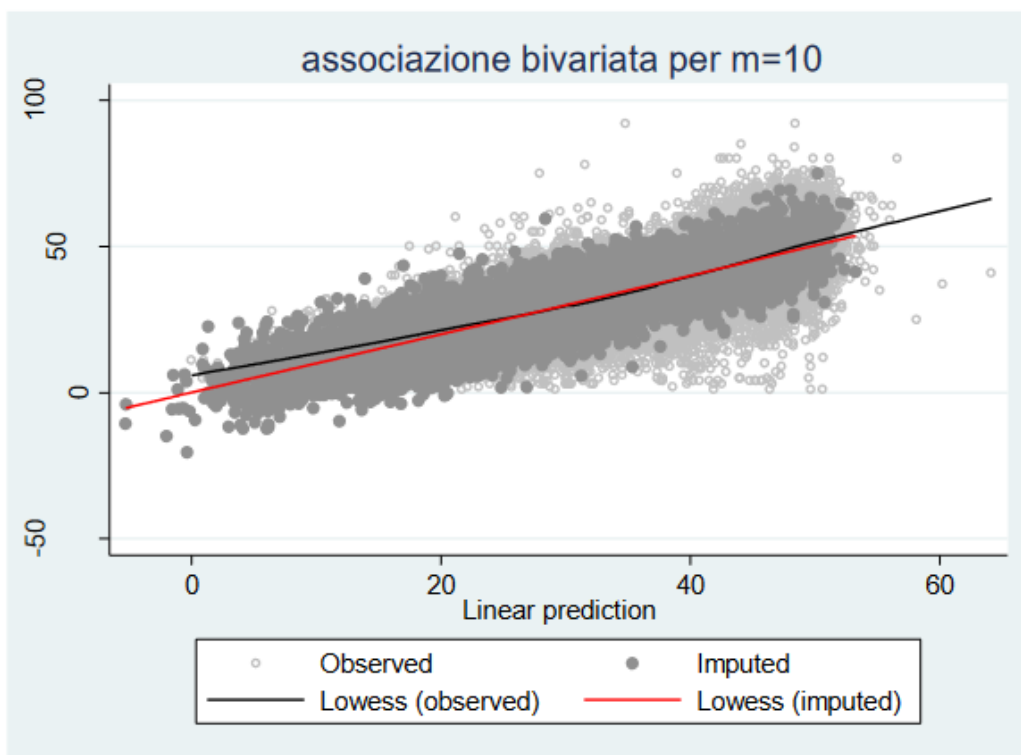
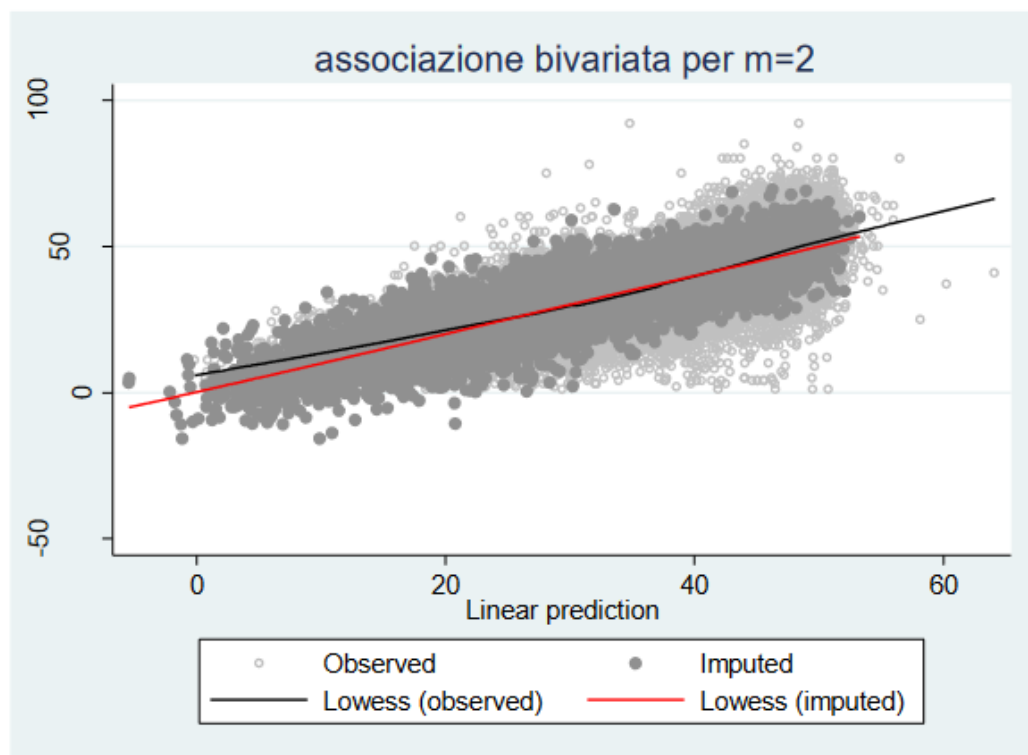
Variable	Obs	Mean	Std. Dev.	Min	Max
maxgrip	67572	33.60994	12.38274	-20.49694	92
gender	67572	1.557228	.4967179	1	2

Diagnostica

- 20 dataset con stime diverse per la variabile imputata (maxgrip).
- Caso 5 è il primo che abbiamo imputato
- Età è costante per i dataset visto che non è stata imputata

	maxgrip	age_int	_mi_id	_mi_miss	_mi_m
5.	.	70	5	1	0
67573.	21.796093	70	5	.	1
73682.	28.755367	70	5	.	2
79791.	22.112566	70	5	.	3
85900.	15.874813	70	5	.	4
92009.	19.866873	70	5	.	5
98118.	23.204582	70	5	.	6
104227.	12.764587	70	5	.	7
110336.	16.141899	70	5	.	8
116445.	11.882469	70	5	.	9
122554.	18.148088	70	5	.	10
128663.	22.607437	70	5	.	11
134772.	26.27165	70	5	.	12
140881.	21.346783	70	5	.	13
146990.	31.968353	70	5	.	14

Diagnostica sul modello



Modello di imputazione 2

OLS con range

- Regressione OLS «troncata» con range (e per gruppi)
- Diagnostica: miglioriamo il range dei valori, ma distribuzione peggiore

m=0 data:

-> summarize maxgrip

Variable	Obs	Mean	Std. Dev.	Min	Max
maxgrip	61463	34.1811	12.11589	1	92

m=1 data:

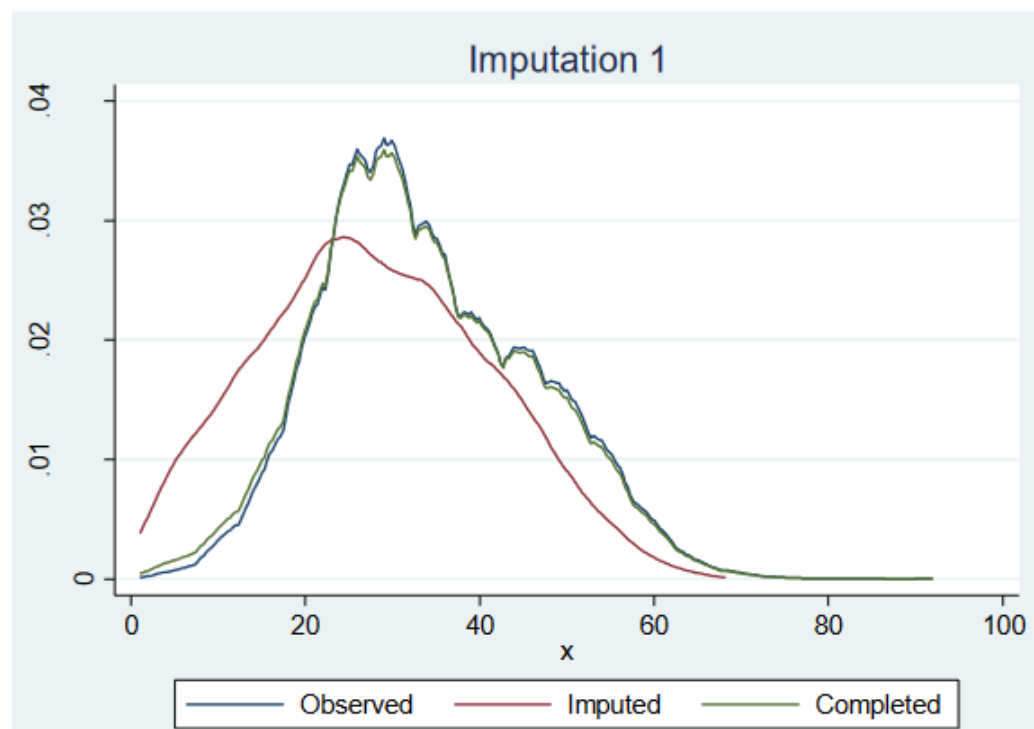
-> summarize maxgrip

Variable	Obs	Mean	Std. Dev.	Min	Max
maxgrip	67572	33.62091	12.34432	1	92

m=10 data:

-> summarize maxgrip

Variable	Obs	Mean	Std. Dev.	Min	Max
maxgrip	67572	33.64115	12.33044	1	92



FASE 2 modello analitico e FASE 3 Pooling

- Fase 2: stimiamo il modello analitico (che si basa su ciò che vogliamo studiare) includendo la variabile imputata X .
- Fase 3: ripetiamo il modello analitico m volte per tutti i dataset creati e combiniamo (media aritmetica) tutte le stime in un solo set di stime e errori standard.

FASE 2 e FASE 3

```
. mi estimate, dots: regress adl maxgrip i.country wave gender age_int hhsize
```

```
Imputations (20):
```

```
.....10.....20 done
```

```
Multiple-imputation estimates
```

```
Linear regression
```

```
Imputations = 20
```

```
Number of obs = 67572
```

```
Average RVI = 0.0358
```

```
Largest FMI = 0.3915
```

```
Complete DF = 67552
```

```
DF adjustment: Small sample
```

```
DF: min = 129.64
```

```
avg = 39446.66
```

```
max = 63933.35
```

```
Model F test: Equal FMI
```

```
F( 19,54211.9) = 450.82
```

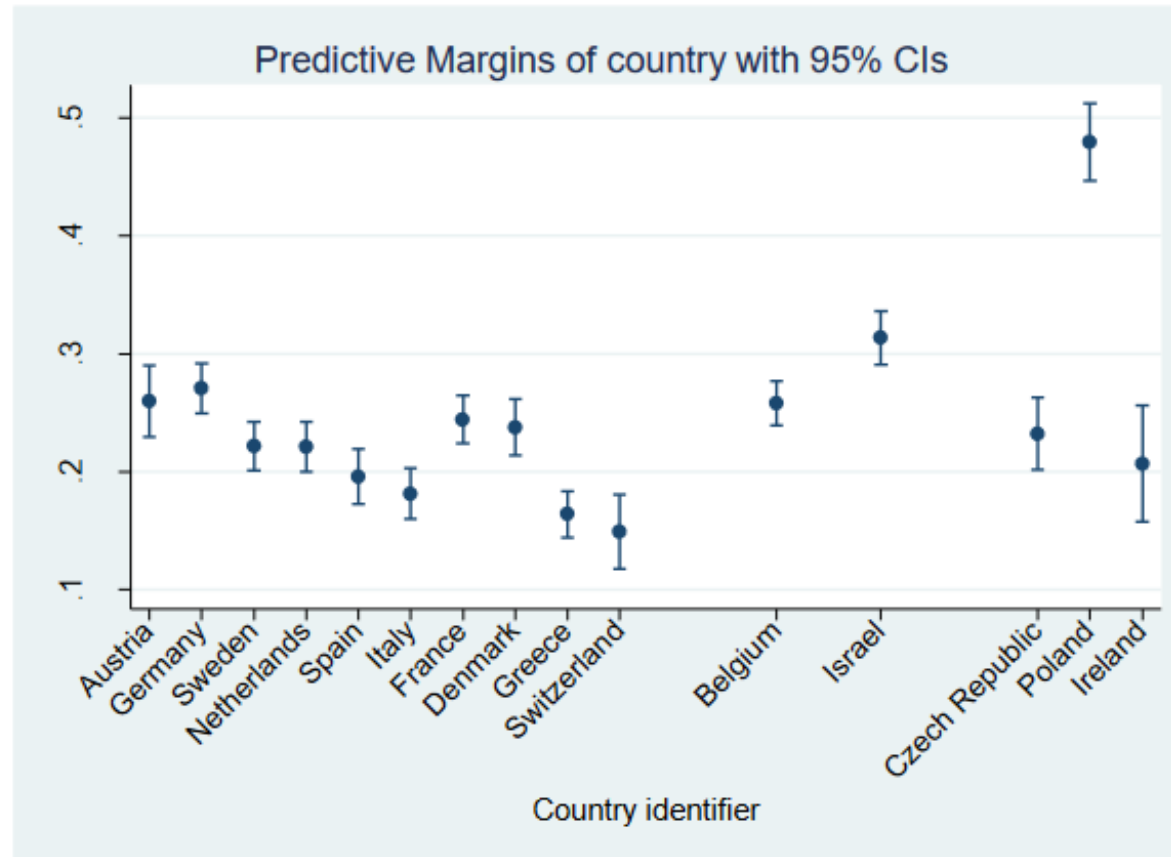
```
Within VCE type: OLS
```

```
Prob > F = 0.0000
```

adl	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
maxgrip	-.0229264	.0005198	-44.11	0.000	-.0239547	-.0218981
country						
12	.0105547	.0186557	0.57	0.572	-.0260106	.0471199
13	-.0385966	.0185377	-2.08	0.037	-.0749306	-.0022625

Modello analitico

- Dipende da cosa vogliamo studiare/ le finalità dell'analisi



Modelli analitici a seconda del tipo di imputazione

Tipo di modello:	Coef. hhsize	Coef. maxgrip	Coef. gender
No imputation	0.005 (0.002)	-0.012 (0.000)	-0.168 (0.007)
regress	0.026 (0.003)	-0.022 (0.000)	-0.315 (0.011)
truncreg	0.026 (0.003)	-0.022 (0.000)	-0.299 (0.011)
PMM (1)	0.026 (0.003)	-0.021 (0.000)	-0.272 (0.011)
PMM (5)	0.026 (0.003)	-0.021 (0.000)	-0.265 (0.011)

Take-home message

- La procedura di Imputazione Multipla:
 - È finalizzata a «sostituire» i dati mancanti con un set di **valori simulati per avere dataset completi**.
 - Le stime dei parametri ottenute dal dataset completo **tengono in considerazione dell'incertezza generata dalle non-risposte**.
 - Lo scopo non è quello di «sostituire» i valori mancanti con quelli più vicini a quelli reali ma ottenere un dataset completo per **l'inferenza statistica** (Rubin 1996).
- La fase di **imputazione è separata da quella analitica**. Quindi gli studiosi (a prescindere dal loro scopo analitico) possono ripetere l'imputazione e coloro che raccolgono i dati, che di solito hanno variabili ausiliarie aggiuntive, possono fornire dati imputati migliori.

Tipo di modello di imputazione

- Imputazione Univariata tiene in considerazione della natura delle variabili imputate:

regress	linear regression for a continuous variable
pmm	predictive mean matching for a continuous variable
truncreg	truncated regression for a continuous variable with a restricted range
intreg	interval regression for a partially observed (censored) continuous variable
logit	logistic regression for a binary variable
ologit	ordered logistic regression for an ordinal variable
mlogit	multinomial logistic regression for a nominal variable
poisson	Poisson regression for a count variable
nbreg	negative binomial regression for an overdispersed count variable