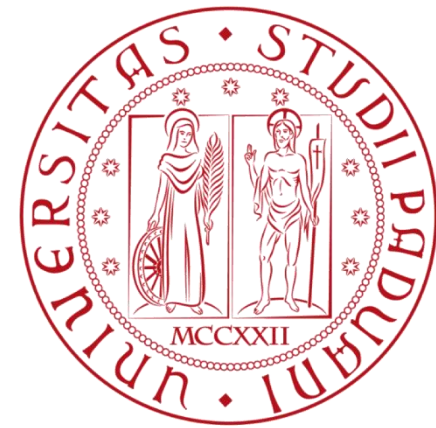




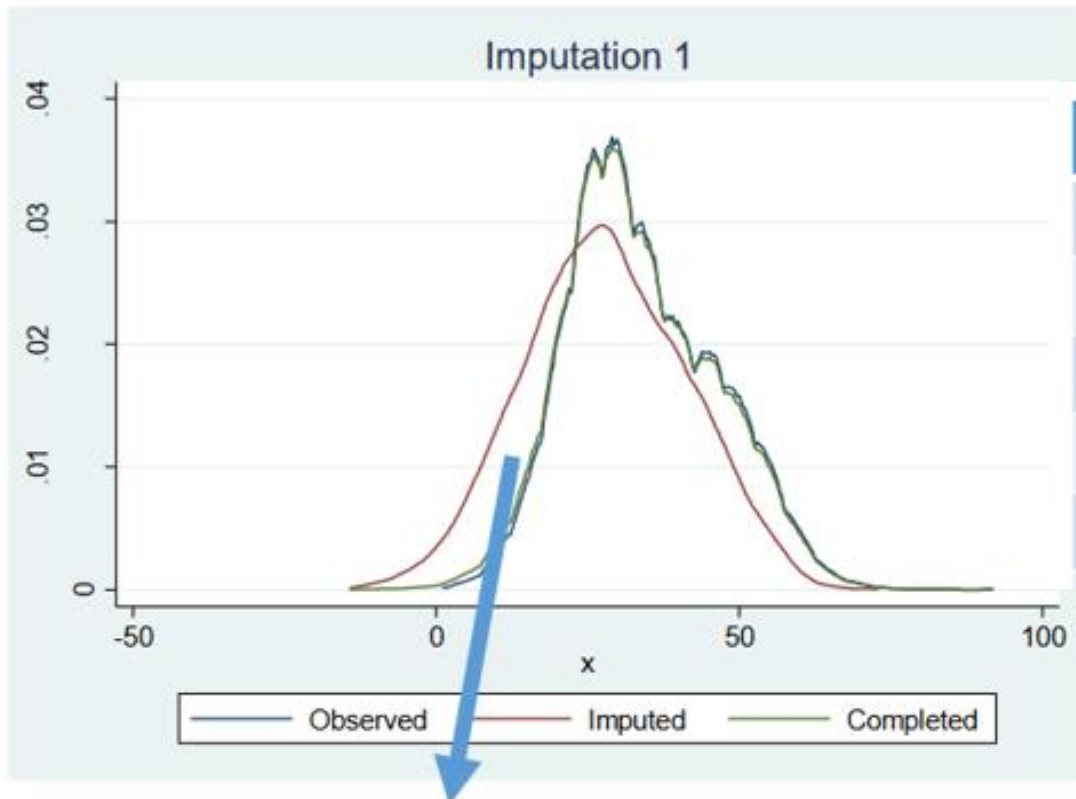
Dati multi-fonte e analisi territoriali

Marco Tosi, Irene Barbiera, e Federico Gianoli

Dipartimento di Scienze Statistiche

Modello di Imputazione

- La scelta del modello di imputazione ha conseguenze sia sulla distribuzione delle variabili sia sui risultati del modello analitico.



Assunzioni sulla distribuzione normale

Tipo di modello:	Coef. hhsiz	Coef. maxgrip
No imputation	0.005 (0.002)	-0.012 (0.000)
regress	0.026 (0.003)	-0.022 (0.000)
truncreg	0.026 (0.003)	-0.022 (0.000)
PMM (1)	0.026 (0.003)	-0.021 (0.000)
PMM (5)	0.026 (0.003)	-0.021 (0.000)

Modello di Imputazione e Assunzioni

- I modelli di imputazione che abbiamo visto (regressione lineare OLS) sono parametrici quindi imputano i valori mancanti estraendoli dalla distribuzione a posteriori predetta dal modello. Questo implica che stiamo facendo **forti assunzioni**:

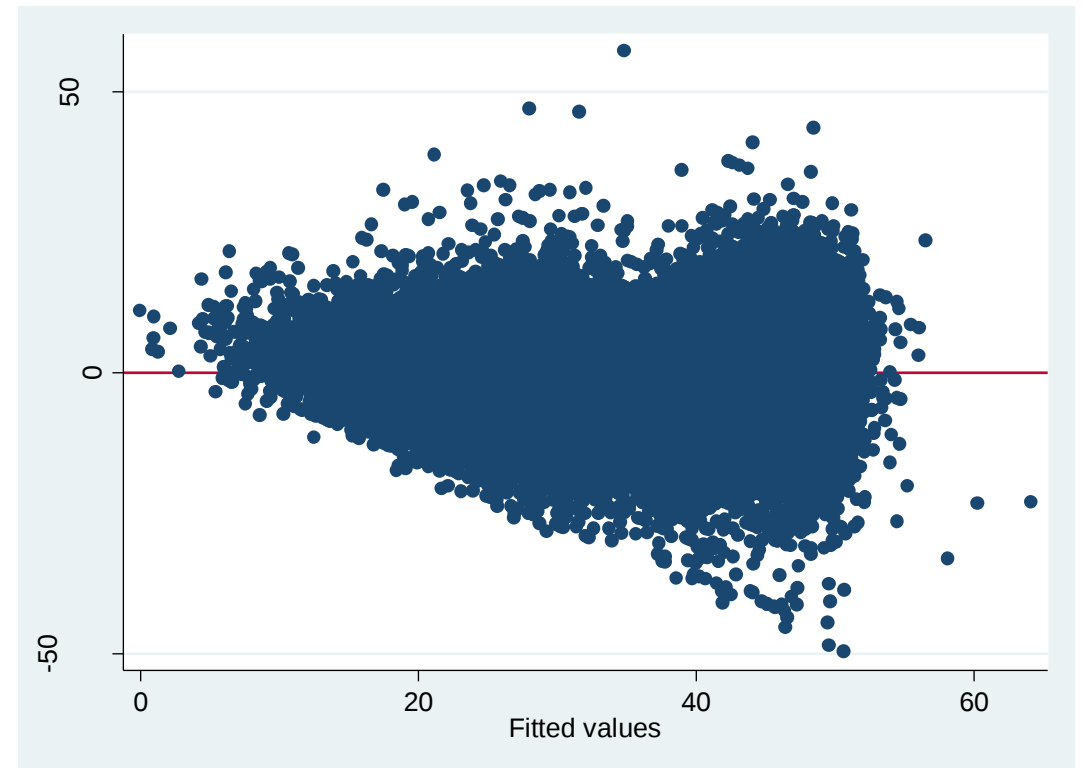
linearità,

normalità dei residui,

Indipendenza dei residui,

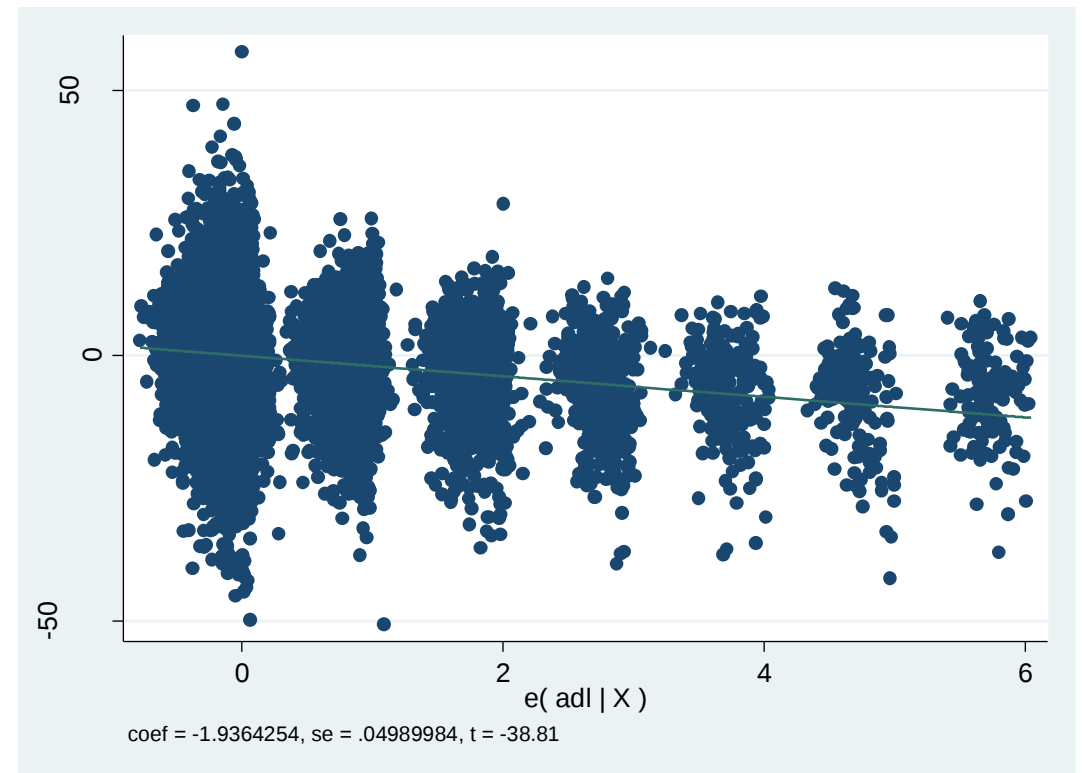
omoschedasticità dei residui,

non collinearità dei predittori



Trasformazioni dei predittori

- Alcune soluzioni possibili quando i dati sono «dispersi» per variabili categoriali: quando abbiamo troppe categorie in una variabile qualitativa conviene **aggregarle**. Oppure trattare una variabile ordinale come continua può dare soluzioni più robuste in alcuni casi. Oppure utilizzare variabili dicotomiche può essere preferibile.
- Tuttavia il problema di **specificazione del modello** (parametrico) persiste.



Distribuzioni continue distorte (skewed)

- Spesso ci troviamo davanti il problema delle distribuzioni di variabili continue che **non possono essere approssimate alla Normale**. Abbiamo **distribuzioni skewed** verso destra quando la distribuzione è spostata verso i valori positivi (es, soddisfazione verso il rapporto coi genitori), e skewed verso sinistra quando la distribuzione è spostata verso lo zero (es, redditi).
- Una soluzione è la trasformazione logaritmica.
- Per l'eccesso di Zeri si possono imputare separatamente due variabili, una dicotomica per zero o non-zero, e un'altra continua per i valori positivi.

Modelli Semi-Parametrici

- Un approccio alternativo è quello di utilizzare i **modelli semi-parametrici** come il **Predictive Mean Matching (PMM)** per rilassare le assunzioni di Normalità (e in generale le assunzioni dei modelli parametrici per variabili continue).
- Questi modelli imputano i valori mancanti (non partendo dalla distribuzione dei parametri stimata dal modello di regressione) ma **utilizzando i valori dei dati completi**.
- La specificazione del modello risulta meno problematica.

Predictive Mean Matching (PMM)

- **Predictive Mean Matching (PMM)** è un metodo che imputa i valori mancanti estraendoli **dai dati osservati** della variabile stessa. Quindi le unità statistiche complete servono ad imputare direttamente i dati mancanti.
- **Vantaggi:** la **distribuzione dei valori imputati è simile (se non uguale) a quella dei dati osservati** ed evita che le imputazioni cadano fuori dal range dei dati osservati. Questa è una caratteristica desiderabile quando le assunzioni di Normalità sono violate o quando la relazione tra variabili non è lineare.
- **Svantaggi:** necessità **di un campione abbastanza ampio** di dati osservati che coprano il range di valori della popolazione reale, e non possiamo imputare dati di sotto-gruppi particolari della popolazione che hanno valori fuori range rispetto ai dati osservati.

Predictive Mean Matching (PMM)

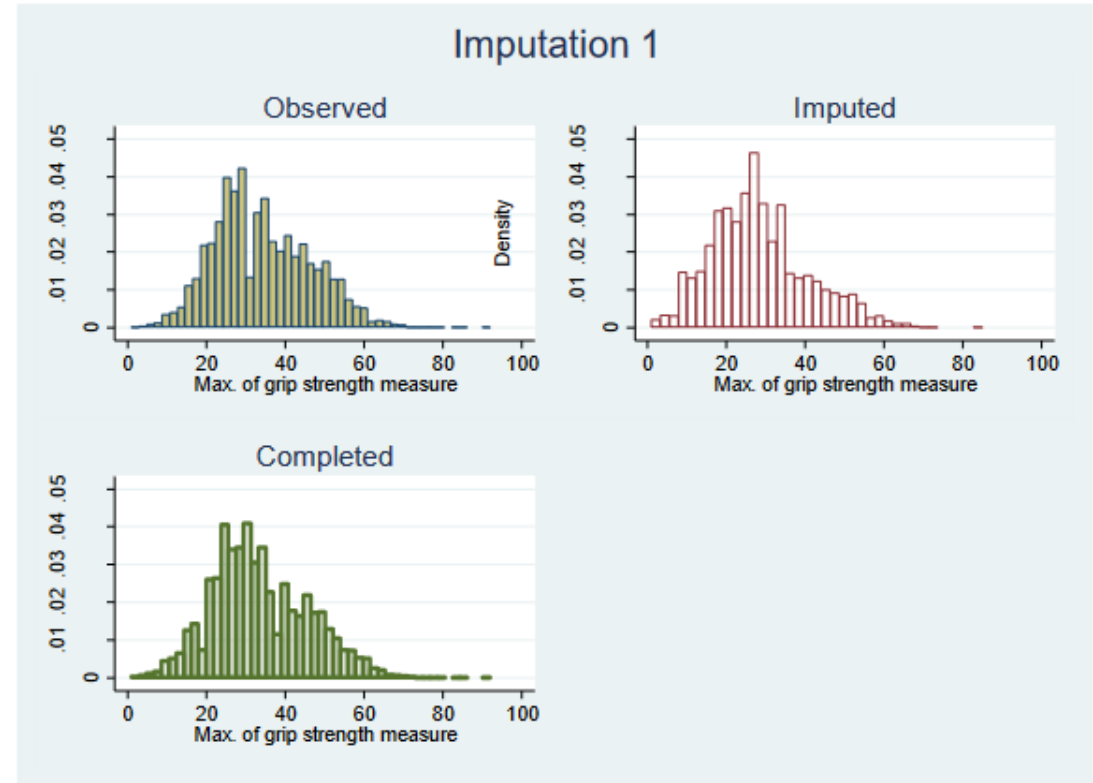
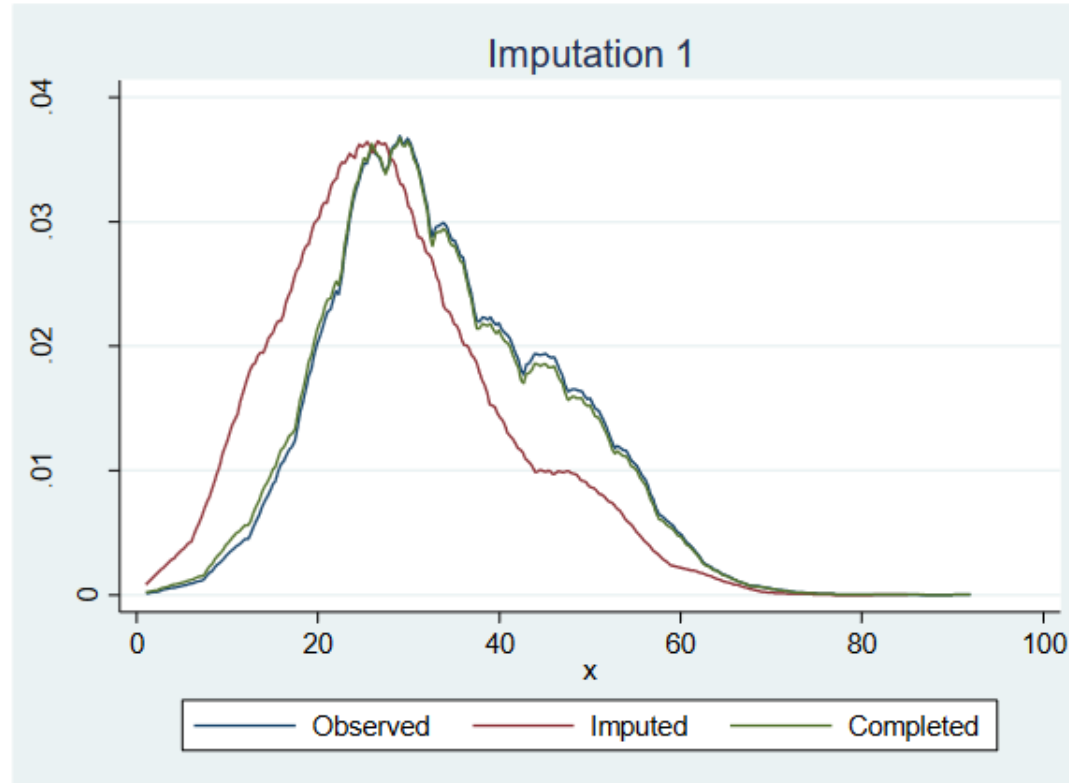
1. **Definire la media predittiva** («predictive mean») di una unità statistica in base alle variabili osservate incluse nel modello. Ossia è il **valore predetto da un modello di regressione lineare** (parte parametrica).
2. Ogni unità incompleta (con valori mancanti) viene **abbinata («matching»)** ad una osservazione estratta casualmente dal gruppo di unità complete (le «unità donatrici») che hanno medie predittive vicine all'unità incompleta.
3. Le unità incomplete vengono **imputate** nei loro valori mancanti assegnando il **valore *reale* delle unità donatrici più vicine** («nearest neighbor»).

Le Unità donatrici

1. Utilizzando il metodo PMM ci troviamo davanti al problema di **definire un numero di potenziali unità donatrici**. Più piccolo è il numero di unità donatrici per ogni unità incompleta, e più sarà alta sia la correlazione tra le imputazioni e sia la variabilità dei valori imputati nelle ripetizioni dell'Imputazione Multipla (ossia nell'estrazione casuale). D'altra parte aumentare il numero delle unità donatrici **risulta in distorsioni maggiori**, allontanandoci dal valore «vero». Trade-off tra bias e variabilità dell'imputazione.
2. Schenker et al. (1996): simulazioni mostrano che 3 unità donatrici minimizzano il bias.

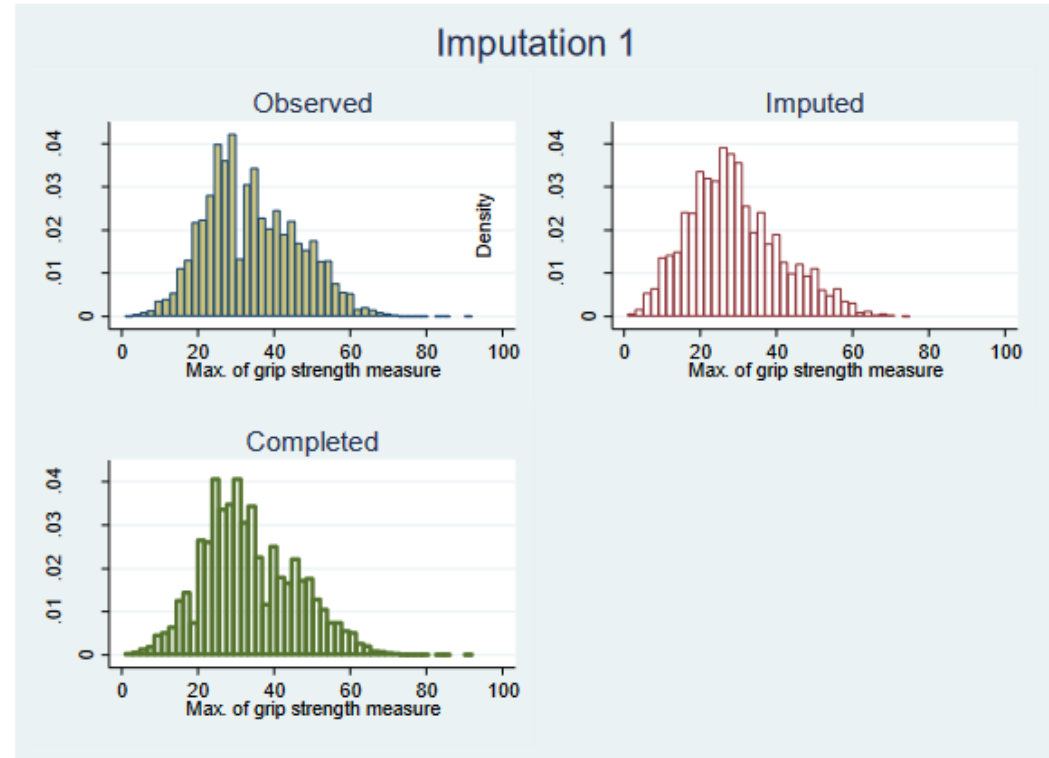
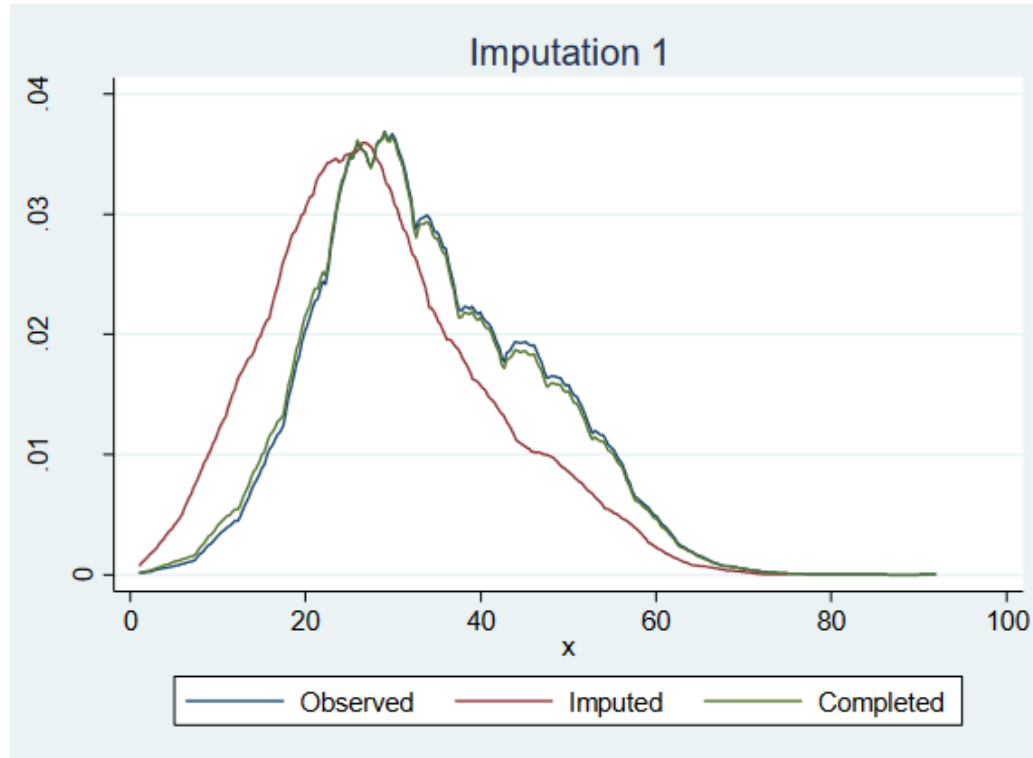
PMM

- Consideriamo solo il caso più vicino a seconda delle variabili incluse nel modello



PMM

- Consideriamo i 5 casi più vicini



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