Imputations: Benefits, Risks and a Method for Missing Data

Nikolas Mittag
Harris School of Public Policy, University of Chicago

Introduction

Problems
- frequently missing variables or observations
- conditions for either analyzing complete cases or imputation often do not hold

Contributions
- examine consequences of invalid assumptions
- is omitting or imputing missing data better?
- evaluate common imputation methods
- discuss method to address these problems

When can Imputation Methods be Useful?

Missing Variables
- can reduce or avoid omitted variable bias
- problem: quality & comparability of source data

Missing Observations
- data often includes imputed observations
- under which conditions should they be used?

Missing data is . . .	Imputations use information from "outside the model": No	Imputations use information from "outside the model": Yes
ignorable	Potential efficiency gains from larger sample	Potential efficiency gains from larger sample and new information
not ignorable	Trade-off selection vs. imputation bias	Additional information may remove bias

Implied Desirable Features of Imputations

- conditioning on a lot of information
- reproducing relation of missing data to covariates and error term
- incorporate differences between source and outcome data
- transparency - researchers need to know which information was used

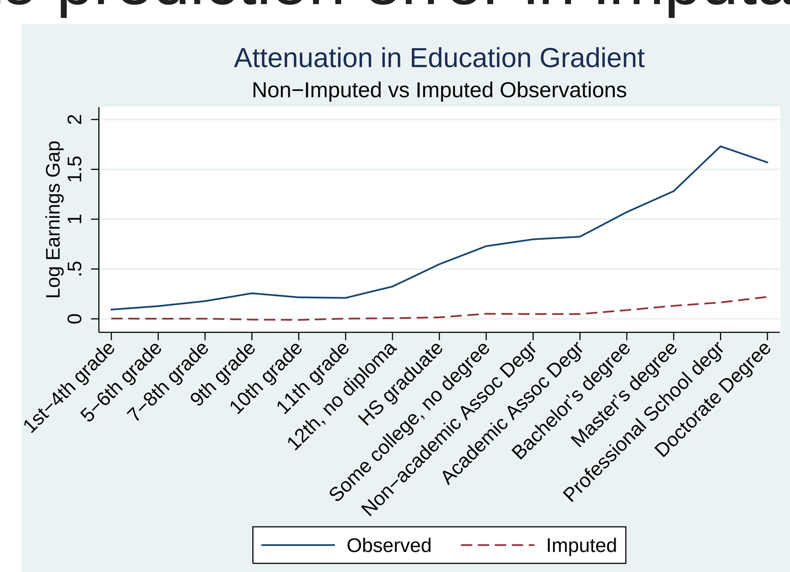
Common Sources of Bias

Conditioning Set
- imputed values are (conditionally) independent of any variable that is not in the conditioning set
- Literature has shown significant bias in OLS for an imputed dependent variable under MAR
- In addition, I discuss the bias
  - on coefficients of other included variables
  - when used as an independent variable
  - from including endogenous variables

$\Longrightarrow$ Ideal conditioning set is as large as possible, but excludes endogenous variables. Yet, it is often not even published in practice.

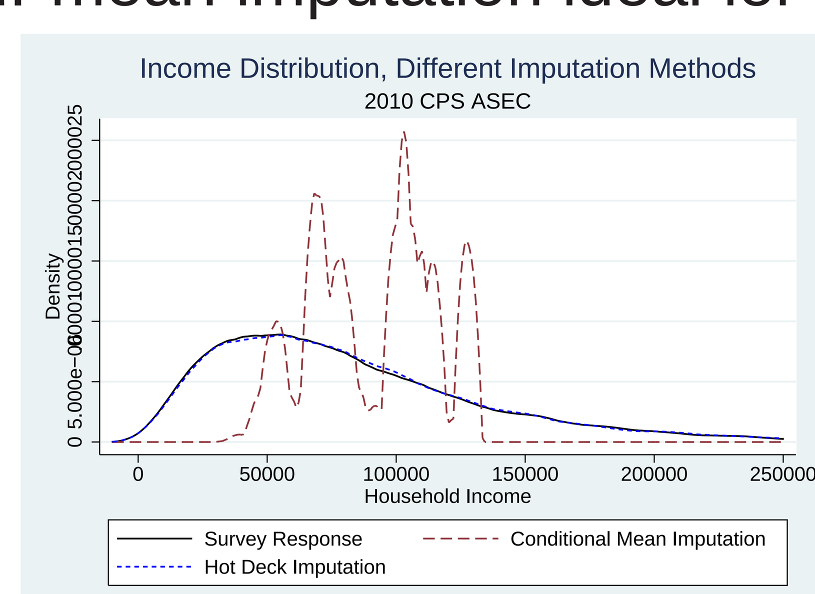
Using Imputed Values

1. Bias due to prediction error in imputations, e.g.:



2. Obtaining correct SEs is often not feasible

Ideal Imputation hinges on Outcome Model
E.g. cond. mean imputation ideal for OLS, but:



Implied Desirable Features of Imputations

- allow many conditioning variables to reduce bias from conditioning set and prediction error
- more information to avoid these biases
- flexibility to choose or adjust conditioning set
- replicability and known theoretical properties to correct bias and obtain correct SEs

Common Imputation Methods

Method	key advantages	main problems
Hot Decks	univariate stat., idiosyncratic data features	conditioning set, estimating SEs, multivar. models
Re-weighting	conditioning set, implementation	require MAR, limited scope
Parametric Models	implementation, theoretical properties	parametric restrictions
Semi-Parametric	relax parametric restrictions	implementation

Conditional Density Method

Basic Idea
- obtain a flexible parametric estimate of conditional density of missing data
- estimate model of interest by integrating over the estimated density using simulation

Main Advantages
- combines key advantages of parametric and semi-parametric methods
- easy to implement and obtain correct SEs
- no bias from prediction error
- conditioning set can be large and is adjustable
- works well with many outcome models
- makes "division of labor" simple

Performance

- Compare methods in two applications:
  - Imputing SNAP amounts from CPS in the ACS
  - Imputing hours worked in ACS under MAR
- hot decks reproduce marginal densities well, but fare poorly in multivariate applications
- conditional mean imputation is ideal for regressions, but poor in other cases
- conditional density method performs similar to ideal method in all applications

International Total Survey Error Workshop 2013, June 3, Ames, IA
email: `mittag@uchicago.edu` www: `http://home.uchicago.edu/~mittag/`