

# Multiple Imputation (MI) of Missing Data

- **Why do multiple imputation?**

To get an unbiased precision of the estimate:

- Suppose that we are estimating  $Y$  on the basis of  $X$ . For every situation with  $X = 5$ , for example, we will impute the same value of  $Y$ .
- This leads to an underestimate of the standard error of our final estimate (e.g., regression coefficient), because we have less variability in our imputed data than we would have had if those values had not been missing.

# How does MI works?

- Takes the predicted values of  $Y$  and then add, or subtract, an error component drawn randomly from the residual distribution of  $Y - \hat{Y}$ .
- It will still underestimate the variance
- Solve this problem by repeating the imputation problem several times, generating multiple sets of new data whose coefficients varying from set to set.

# Obtaining final estimate from MI

- If we are estimating  $Q$ , such as a variable mean or regression coefficient, the overall point estimate  $q^*$  of  $Q$  is the average of the  $m$  separate estimates  $q_j$ .
- That is,  $q^* = (q_1 + q_2 + \dots + q_m) / m$
- The estimated variance  $T$  of  $Q$  is obtained by components-of-variance argument (Rubin's rule):  
$$T = W + (1 + 1/m) B$$
where  $B$  = the within data set variance, and  
 $W$  = variance across the  $m$  data sets

According to King, et al, about 5 or 10 imputed data sets is often satisfactory.

- **Implemented in SAS or STATA, this standard error calculation won't be needed.**

# Is it appropriate to SWC data?

Imputation is appropriate when data are:

**Either** missing completely at random (MCAR),

Formally,  $M$  is independent of  $D$ , so  $P(M | D) = P(M)$ .

**Or** Missing at random (MAR), Formally,

$M$  is dependent on  $D_{\text{obs}}$  but not on  $D_{\text{mis}}$ :

$$P(M | D) = P(M | D_{\text{obs}})$$

*(Not necessary to be causal)*

What if Non-ignorable, where  $P(M | D)$  cannot be simplified?

# Available options for MI

- **proc mi** in SAS
  - Assumes multivariate normal distribution
- **ice** in STATA
  - Based on conditional density of target variable given all other variables
- MI implementation in WinBUGS
  - Assumes multivariate normal distribution

*All of the above can work with missing covariate*

# Why chose STATA instead of SAS?

- no multivariate joint distribution assumption; this reason alone makes it appealing since it allows different types of variables to be imputed together
- allowing different kinds of weights, as long as the regression models allow them
- Allows boot option to relax the multivariate normality assumption of the regression coefficients
- easy to understand;
- easy to use.

# How does ice work?

Two steps:

Step1: imputes a single variable given a set of predictor variables using **uvis** program

Step2: "Regression switching", cycles through all the variables to be imputed using **uvis**

**micombine** does analysis separately for each dataset and reports summary estimate

# Comparison

- Have a complete data set with no missing THM or covariate
- Randomly deleted 70% of the THM values
- Used **ice** to impute 5 data sets
- Implemented WinBUGS for data with missing dependent variable

# WinBugs implementation for MI

- For this method, we assume that all of the data are jointly distributed multivariate normal with some unknown mean (denoted  $\mu$ ) and covariance (denoted  $\Sigma$ ).
- We then estimate the mean and covariance parameters, and simulate missing values from the multivariate normal distribution per King's method, or we can just use the draws from the posterior means and covariances for our final analyses directly.

# Working with missing data in WinBUGS

- “Best-Guess” imputation method
- Imputing missing covariates and analyzing the data at the same time
- Requires:
  - Assumption of MCAR
  - Information priors for missing values of the covariates

Suppose that  $y_i \sim N(\mu_i, \tau)$  and there is no missingness in the vector  $Y$ ,

$$\text{where } \mu_i = b_1 + b_2 X_{2i} + \dots + b_p X_{pi}$$

$$b_j \sim N(0, .001) \quad \text{and} \quad \tau \sim \text{Gamma}(.001, .001)$$

**\* Needs to be more clever than WinBUGS to implement**

# References

- Royston, P. 2005. Multiple imputation of missing values: update. *Stata Journal* 5(2): 188–201.
- Van Buuren, S., H. C. Boshuizen, and D. L. Knook. 1999. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 18: 681-694
- Rubin, D. B. 1987. Multiple imputation for non response in surveys. New York: John Wiley & Sons
- King, G, Honaker, J, Joseph, A, and Scheve, K. 2001. Analysing incomplete political data: an alternative algorithm for multiple imputation. *American Political Science Review* 18 (1): 49-69

THANK YOU