# POLS/CS&SS 503: Advanced Quantitative Political Methodology

# MISSING DATA

May 19, 2015

Jeffrey B. Arnold

UNIVERSITY of WASHINGTON
DEPARTMENT OF POLITICAL SCIENCE

CENTER for STATISTICS and the SOCIAL SCIENCES

# Overview

What's the Problem?

Methods of Dealing with Missing Data

References

What's the Problem?

Methods of Dealing with Missing Data

References

# Types of Missingness

- **MCAR** Missingness completely at random
- **MAR** Missingness at random
- **MNAR** Missingness that depends on unobserved variables, or **NI** Non-ignorable missingness

## Funamental Problem with Missing Data

Cannot tell from data alone whether missingness is MAR or MNAR.

# What we will cover and not cover

- Covering: MCAR
  - Missing values in $X$
  - Methods: listwise-deletion, multiple imputation
- Not-covering: MNAR models
  - Selection models
  - Censoring, Truncation

What's the Problem?

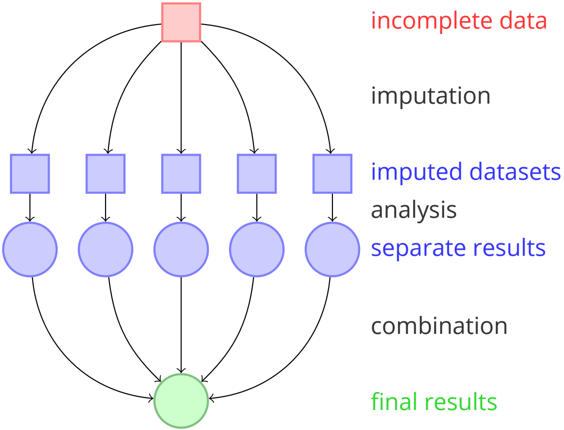
Methods of Dealing with Missing Data

References

# Methods

- Complete case (Listwise deletion)
  - Consistent and valid inferences when MCAR (or MAR but missingness does not depend on the dependent variable)
  - Even in MCAR, inefficient
- Available case (pairwise deletion):
  - E.g. Covariance matrix. Calculate $\sum_i (x_{i,j} - \bar{x}_j)(x_{i,k} - \bar{x}_k)$ for all obs in which $x_{i,j}, x_{i,k}$ are not-missing, regardless of missingness of other variables.
  - Does not work for all analyses
  - Can result in nonsensical results
- Unconditional Mean Imputation (Mean substitution)
  - preserves mean of variables; reduced variance
  - attenuates relationships between variables
  - overstates certainty—increases “effective” sample size and distorts inference

# Overview of Multiple Imputation



# When is Listwise Deletion Preferable to MI?

1. All of the following need to hold
   - Analysis model is conditional on $X$ and correctly specified
   - There is NI missingness in $X$
   - Missingness in $X$ is not a function of $Y$, and unobserved variable affecting $Y$ do not exist
   - Number of observations after deletion is large
2. Know $X$ well enough that we don't trust it to impute, but trust it enough to analyze $Y$
3. Rarely would you prefer listwise deletion to multiple imputation

# Multiple Imputation Estimator Combines Individual Estimates

Given $B_j^{(1)}, \ldots, B_j^{(g)}$, and $\text{SE}(B_j^{(1)}), \ldots, \text{SE}(B_j^{(g)})$ from $g$ imputations:
Estimate for **single coefficients is:**

$$\begin{aligned}
\text{Point Estimate} \qquad & \tilde{\beta}_j = \frac{\sum_{l=1}^{g} B_j^{(l)}}{g} \\
\text{Std. Error.} \qquad & \tilde{\text{SE}}\left(\tilde{\beta}_j\right) = \sqrt{V_j^{(W)} + \frac{g+1}{g} V_j^{(B)}} \\
\text{Within Imputation Variance} \qquad & V_j^{(W)} = \frac{1}{g} \sum_{l=1}^{g} \text{V}(B_j^{(l)}) \\
\text{Between Imputation Variance} \qquad & V_j^{(B)} = \frac{1}{g-1} \sum_{l=1}^{g} \left(B_j^{(l)} - \tilde{\beta}_j\right)^2
\end{aligned}$$

where $\tilde{\beta}_j$ distributed $t$ with complicated d.f. (see Fox, 564)

# Why we don't need to run many imputations

## Relative efficiency of multiple imputation

$$RE(\tilde{\beta}_j) = V(\tilde{\beta}_j^{MLE})/V(\tilde{\beta}_j^{MI}) = \frac{g}{g + \gamma_j}$$

where $\gamma_j$ is the relative rate of missing information

$$\gamma_j = \frac{R_j}{R_j + 1} \qquad R_j = \frac{g+1}{g} \times \frac{V_j^{(B)}}{V_j^{(W)}}$$

## Main point!

Suppose $R_j = \gamma$, then with $g = 5$ iterations, then efficiency is

$$\frac{5}{5 + 0.5} = 0.91$$

# Advice on Missing Data

- Include all relevant variables in the imputation; at least all used in the estimation, including the dependent variable.
- Even if data are not multivariate normal, multivariate normal works okay.
- Transform data to approximate normality (Amelia has options)
- See TSCS extensions in Amelia
- Post-hoc adjustments okay. Impute naively and adjust, e.g.round to integers, or 0/1.
- If need to save time, work with a single iteration until “final” analysis.
- Potential problems: complex interactions between variables
- Try default methods; they often work.
- If not ...
  - Multiple Chained Equations: **mice**, **mi** packages
  - Hot-deck imputation
  - Full Bayesian models

What's the Problem?

Methods of Dealing with Missing Data

References

# References


- Gelman and Hill, Ch. 25 "Missing Data Imputation"
- Fox, Ch 20 "Missing Data in Regression Models"
- Blackwell, Matthew, James Honaker, and Gary King. 10030. "A Unified Approach to Measurement Error and Missing Data: Overview." *Sociological Methods and Research*. http://j.mp/jqdj72.
- Honaker, James, Gary King, and Matthew Blackwell. 2011. "Amelia II:A Program for Missing Data." *Journal of Statistical Software* 45(7). http://www.jstatsoft.org/v45/i07/.
- Honaker, James, and Gary King. 2010. "What to Do about Missing Values in Time-Series Cross-Section Data." *American Journal of Political Science* 54(2): 561–81. http://onlinelibrary.wiley.com/doi/10.1111/j.1540-5907.2010.00447.x/abstract (May 19, 2015).
- King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* null (01): 49–69. http://journals.cambridge.org/article_S0003055401000235 (May 19, 2015).