PoliSci 4782 Political Analysis II

Missing Data

Seamus Wagner

The Ohio State University

Why Worry about Missing Data

- Many political science data sets have:
  - Some amount of missing data
  - Non-random occurrence of missing observations
- Casewise deletion (CWD) can bias results (incorrect inference)
- Single imputation techniques have many problems, the greatest of which is an inability to account for imputation uncertainty
- Multiple Imputation (MI) fills in missing observations and produces better inferences
- Yes, this is "making stuff up," but it is provably less wrong than CWD... sometimes.

Missing Data Mechanisms

- Define:

$$\mathbf{D}_{\text{mis}} = (\mathbf{Y}_{\text{mis}}, \mathbf{X}_{\text{mis}})$$

$$\mathbf{D}_{\text{obs}} = (\mathbf{Y}_{\text{obs}}, \mathbf{X}_{\text{obs}})$$

- We stipulate a $n \times k$ matrix, $\mathbf{M}$, corresponding to $\mathbf{D}$ that contains 0 when the $\mathbf{D}$ matrix data value is *not* missing, and 1 when it is missing.

Missing Completely At Random (MCAR)

$$\Pr(\mathbf{M}|\mathbf{D}) = \Pr(\mathbf{M})$$

- This is the best-case scenario, though not common.
- Means that there is no underlying associative process that causes the absence of data
- When missingness is independent to the value of the data, missing or observed, the missing data are said to be MCAR
- Intuition: both the observed and missing data, independently, have the properties of a random sample of the entire data set
- More specifically, the data for which we have responses ($\mathbf{D}_{\text{obs}}$) and the data for which we do not have responses ($\mathbf{D}_{\text{mis}}$) have the same distribution, whatever that distribution may be: $\mathbf{D}, \mathbf{D}_{\text{obs}}, \mathbf{D}_{\text{mis}} \sim f(\mu, \sigma^2)$.

Missing At Random (MAR)

$$\Pr(\mathbf{M}|\mathbf{D}) = \Pr(\mathbf{M}|\mathbf{D}_{\text{obs}})$$

- More commonly, missing data can be assumed to be *missing at random* (MAR)
- If MAR, then the missingness can be related to the observed data $\mathbf{D}_{\text{obs}}$, but not to the unobserved data $\mathbf{D}_{\text{mis}}$
- Intuition: missingness in one variable can be related to other variables but those other variables have to be recorded in the data set
- Example: missingness in income could be related to education, occupation, neighborhood, etc... and we observe those values
- This is the condition which produces most of the bias in social science (kind of; see next slide)

Non-Ignorable (NI)

$$\Pr(\mathbf{M}|\mathbf{D}) = \Pr(\mathbf{M}|\mathbf{D})$$

- Can be thought of as the “you’re totally screwed” condition
- Nonignorable missing data occur when missingness is related to unknown and unobserved parameters
- Example: missingness in income, but we *do not* observe the education, occupation or neighborhood
- When NI holds, the expression $\Pr(\mathbf{M}|\mathbf{D})$ cannot be simplified
- If missingness is NI, <u>no method</u> can reliably produce unbiased results
- There is simply not enough information in the data set to make valid imputations or produce unbiased results
- Problem: impossible to test for NI

Some of the Ways to Deal with Missing Data

- Casewise deletion (CWD)
- Imputation
  - best guess imputation
  - hot desk imputation
  - mean substitution
  - $\hat{y}$ regression imputation
  - $\hat{y} + \epsilon$ regression imputation
    - single imputation $\rightsquigarrow$ multiple imputation (MI)

Casewise Deletion Is *Risky*

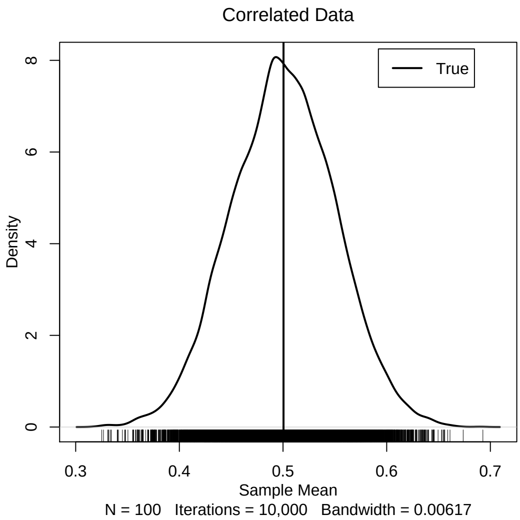
- What is casewise (listwise) deletion?
  - Deleting observations with any missingness
- Best case: inefficient
- Worst case: biased
- If MCAR = TRUE: unbiased but inefficient
  - The deletion will not pull results in a particular direction
  - But we throw way lots of information unnecessarily
- If MCAR = FALSE: biased results!
  - People with high incomes are less likely to report them
- Take-home point: Be aware of possible bias with casewise deletion
- DANGER: all software packages (that I know of) have casewise deletion as their default

Evidence from Simulation

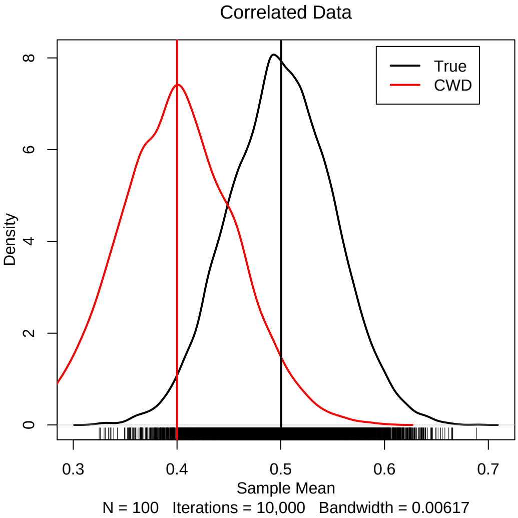
- $N = 100$; Monte Carlo iterations: $10,000$
- $Pr(y_i = 1) = 0.50$
- Missingness was introduced to $\mathbf{y}$ with the following properties:

$$y_i = \begin{cases} Pr(y_i = \text{NA}) = 0.40, & \text{if} \quad y_i = 1 \\ Pr(y_i = \text{NA}) = 0.10, & \text{if} \quad y_i = 0. \end{cases}$$

Evidence from Simulation



Evidence from Simulation



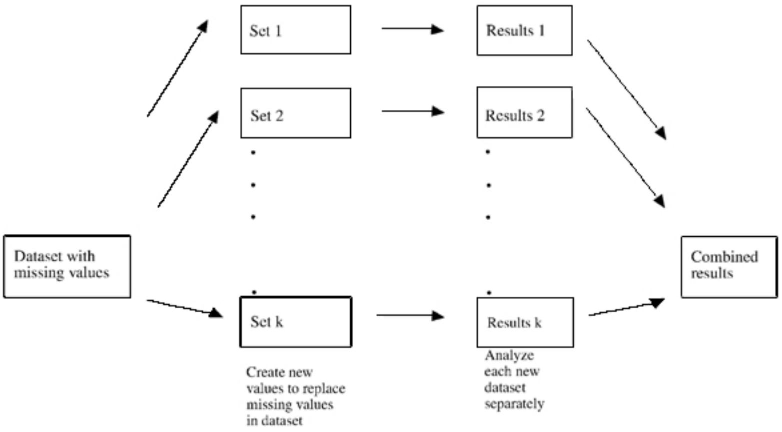
Single Imputation

- An improvement if methods are valid, but still problematic
- If we fill in each missing value with one imputed value, we get unbiased results (assuming a good imputation mechanism)
- The problem is that we would not capture our *additional uncertainty* in the imputed values
- We have more faith in the values we actually observe than in those we impute
- So, single imputation may result in good estimates but artificially small standard errors
- Multiple imputation fixes this problem so we can have unbiased coefficients and standard errors

Multiple Imputation in a Nutshell

- 3 steps:
  - Generate reasonable imputations of the missing the data m times to get m replicate datasets,
  - Analyze/regress each dataset separately,
  - Combine results to a single summary process.
- Imputation step assumes a conditional distribution for the missing data conditioning on observed values (no NI)
- Oddly enough $m = 5$ to 10 is sufficient
- Combination process uses means for coefficients and a stylized mean for standard errors.

Multiple Imputation in a Nutshell



Two Broad Ways of Doing MI

1. Joint Modeling ("Amelia")
    - The original MI idea (Rubin, 1976)
    - Specify an appropriate parametric density $P(Y|\theta)$ and an appropriate prior $P(\theta)$
    - Imputations are draws from the posterior predictive distribution $P(y^{\text{mis}}|y^{\text{obs}}, \theta)$
    - Loosely akin to predicted values from OLS
2. Fully Conditional Specification (MICE)
    - No need for an explicit density $P(\mathbf{Y}|\theta)$
    - Instead, specify a separate conditional density $P(Y_j|Y_{-j}, \theta_j)$ for each Y_{-j}
    - This density is used to impute y_j^{mis} given y_{-j}
    - i.e. LS or logit applied to cases in y_j^{obs}
    - Iterates over all conditionally specified imputation models, each iteration consisting of one cycle through all Y_j

- Both ways require a combination rule to aggregate the result from each imputed dataset

Rubin's Combination Rules

- Run your model on each of the M imputed data sets
- Let $\hat{\theta}_m, m = 1, \ldots, M$ be the estimates computed individually from the M imputed data sets
- Let $\Sigma_m, m = 1, \ldots, M$ be the associated variances for $\hat{\theta}_m$
- A single estimate of θ, $\bar{\theta}_M$, can be produced by taking the mean of $\hat{\theta}_m$ over all m:

$$\bar{\theta}_M = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_n.$$

- Not only is $\bar{\theta}_M$ a valid estimate of θ, but it is more statistically efficient than a single imputation technique could produce

Rubin's Combination Rules

- Calculating the variability of our estimate of θ is somewhat more complicated
- Have variance *within* datasets and *between* datasets
- The within imputation variance is simply the mean of individual variances:

$$\bar{W}_M = \frac{1}{M} \sum_{m=1}^{M} \Sigma_m.$$

- Between imputation variance is the mean of the squared differences between individual estimates $\hat{\theta}_m$ and the total estimate $\bar{\theta}_M$:

$$B_M = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \bar{\theta}_M)^2.$$

Rubin's Combination Rules

- The total variance associated with $\bar{\theta}_M$ is computed:

$$T_M = \bar{W}_M + \left(1 + \frac{1}{M}\right) B_M$$
$$= \frac{1}{M} \sum_{m=1}^{M} \Sigma_m + \left(1 + \frac{1}{M}\right) \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \bar{\theta}_M)^2$$

(note that $[1 + 1/M]$ is an adjustment for finite M)

Coming Up

- Lab tutorial on multiple imputation
- Lectures on causal inference and research design next week