Missing data analysis

Seminar in Criminology, Research and Analysis– Crim 7301
Week 9, 10/20/16
Andrew Wheeler

Class Overview

- Types of Missing data
- Missing at Random (MAR), Missing completely at random (MCAR)
- Techniques to account for missing data
  - Bad approaches: pairwise deletion, simple mean imputation
  - Recoding missing data
  - Full information maximum likelihood
  - Other selection models
  - Hot deck imputation
- Multiple Imputation through Chained Equations

Types of Missing Data

- Truncated (not in the sample)
  - Capture/Re-capture
- Censored (above/below a particular data value)
  - Time to Recidivism
  - Unknown time for burglary (interval censored)
- Missing data elements
  - Survey data non-response to particular questions

FIGURE 2

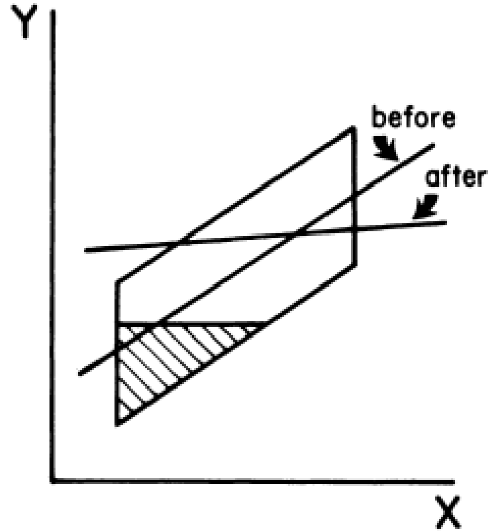


FIGURE 3

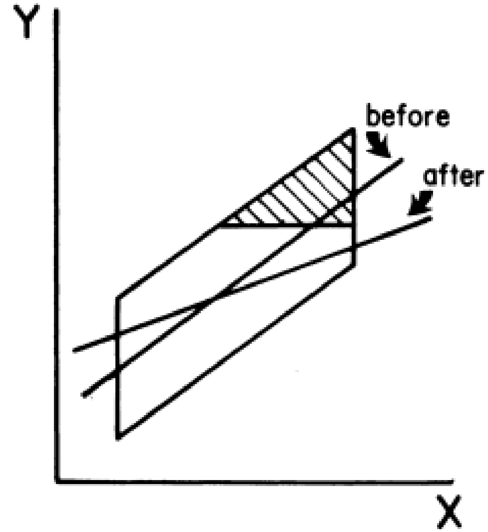


FIGURE 4

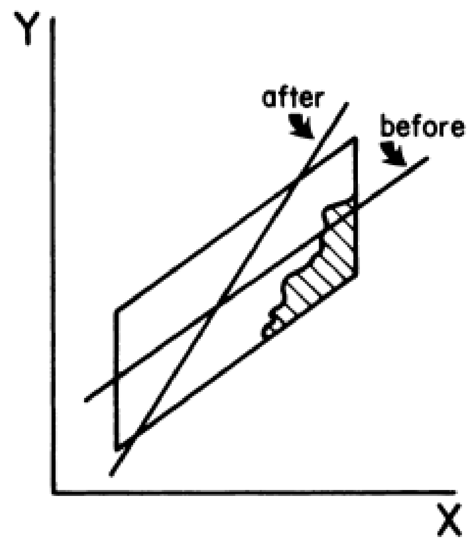
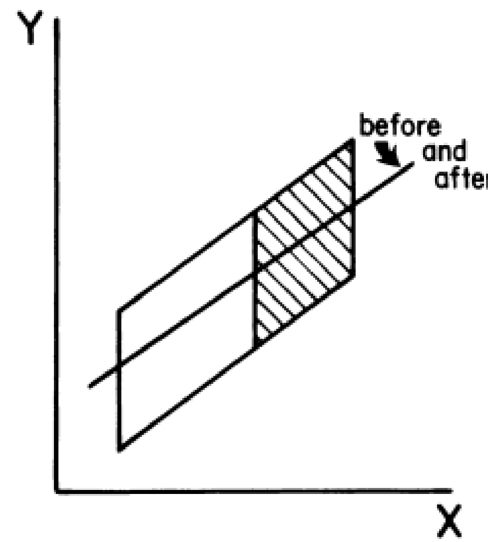


FIGURE 5



MAR & MCAR

- MCAR (missing completely at random)
  - Missing is not related to the actual data values
  - Missing on one item can be related to missing on another
  - Same as random sampling

- MAR (missing at random)
  - Missing can be related to data values, but other variables control for it
  - Example, $\Pr(Y \text{ Missing} \mid Y, X) = \Pr(Y \text{ Missing} \mid X)$
  - victimization, female and males.

Don’t Do These!

- Pairwise deletion – uses different subsets of data to calculate correlations
  - Correlation matrix not guaranteed to be positive-definite
  - No one knows the correct degrees of freedom for subsequent models
- Simple mean imputation – recode missing data to the mean of the observed data values (or mode for categories)
  - Inappropriately reduces variance, by a lot!
- Listwise deletion is not very bad compared to either of these options

Recoding Missing Data

- Not a problem if missing is intentionally design missing, examples:
  - if you have a survey that ask if person took drugs, then a second question asks how often they took the drugs
  - Age of menarche for females and sexual activity
  - Zero values for logged independent variable

Other Selection Models

- Full information maximum likelihood
  - For some models, can partition likelihood between missing and non-missing data
  - Available in many sem fitting software/functions

- Other selection models
  - Heckmen selection [missing on the dependent variable] – depends on having instruments to predict missing
  - Step 1: $\Pr(\text{Missing}) = \Phi(\beta_0 + \beta_1 Z) = \hat{p}$
  - Step 2: $\hat{I} = \phi(\widehat{p}) / \Phi(\widehat{p})$ [Inverse Mill's Ratio, ϕ is pdf and Φ is cdf of normal distribution]
  - Step 3: $Y = \beta_0 + \beta_1 X + \gamma \hat{I}$

Hot Deck Imputation

- Draws at random from observed cases to use as imputation, so is always in the data
- Can further condition on other categorical covariates
- Useful for categories with many levels

Multiple Imputation through Chained Equations

- Step 1: Predict missing data from other variables

- Fits sequential models to predict missing data values
  - E.g. a linear model to predict continuous variables
  - Logistic to predict 0/1
  - Multinomial to predict more than two categories
  - Ordinal Logistic to predict ordinal categories

- Step 2: Once models have converged, generate *M* imputed complete datasets
  - Predicted values versus predictive mean matching

- Step 3: Estimate models for each subset, then combine coefficients into pooled estimate

Multiple Imputation through Chained Equations

- Equation for pooling coefficients:
  - $x_1, x_2, \dots x_k$ coefficients with $s_1, s_2, \dots s_k$ standard errors
  - $m = \#$ of imputed datasets

$$\bar{x} = \mathrm{m}^{-1} \sum_{k}^{m} x_i$$

$$\mathbb{V}(\bar{x}) = m^{-1} \left(\Sigma_k \, s_k^2 \right) + (1 + m^{-1}) \cdot (m - 1)^{-1} \cdot \Sigma_k (x_k - \bar{x})^2$$

Or more simply:

$$\mathbb{V}(\bar{x}) = \mathbb{E}(s^2) + [1 + 1/m] \cdot \mathbb{V}(x)$$

Multiple Imputation through Chained Equations

- Example (R Code)

```
######################################
library(mice)

x <- c(0.5,0.6,0.7,0.6)
s <- c(0.2,0.4,0.3,0.1)
m <- length(x)

#by hand results
v <- mean(s^2) + (1 + 1/m)*var(x)
c(mean(x),sqrt(v))

#via function in the mice package
res <- pool.scalar(x,s^2,method="rubin")
c(res$qbar,sqrt(res$t))
#######################################
```

Multiple Imputation through Chained Equations

- Need to make sure each individual equation is consistent with the subsequent model
  - Passive transformations
  - Do the values need to be rounded?

- Can be difficult to converge with many variables or a lot of missing data

- Can include other auxiliary variables to predict missing

- Tends to not need many imputations

Homework & Next Weeks Class

Lab Assignment

Conduct multiple imputation for a survey of citizen perceptions of public safety in Dallas. Property versus Violent predicted by income. Code snippets in R, Stata and SPSS

For Next Week – Social Network Statistics

- McGloin, J. M. (2005). Policy and intervention considerations of a network analysis of street gangs. *Criminology & Public Policy*, 4(3):607-635.
- McGloin, J. M. and Kirk, D. S. (2010). An overview of social network analysis. *Journal of Criminal Justice Education*, 21(2):169-181.
- Papachristos, A. V. (2011). The coming of a networked criminology. *Measuring Crime & Criminality: Advances in Criminological Theory*, 17:101-140.
- Papachristos, A. V., Hureau, D. M., and Braga, A. A. (2013). The corner and the crew: The influence of geography and social networks on gang violence. *American Sociological Review*, 78(3):417-447.
- Papachristos, A. V. and Kirk, D. S. (2015). Changing the street dynamic: Evaluating Chicago's group violence reduction strategy. *Criminology & Public Policy*, 14(3):525-558.