How do we handle missing data?

Friday, November 30, 2012
2:20 PM

- It is extremely common for observations to be partially or completely missing for important data sets, especially observational data sets

X

	Obs 1	Obs 2	Obs 3	Obs 4	Obs 5
y	1	4 ✓	2	?	?
x	4	?	?	2 ✓	?
z	2	1 ✓	?	1 ✓	?

$y \sim x + z$

- So what?
  - Models cannot be estimated on missing data without some kind of processing
  - Missing data contains potentially important information
  - Simply removing the missing cases can cause more variable (best case) or even biased (worst case) estimates

- Given that we have to use these flawed data sets… how do we use them in a way that minimizes the harm to inference

Missingness Patterns

Friday, November 30, 2012
2:38 PM

systematic missingness

- Usually, data is missing in three ways:

  1) Missing Completely at Random (MCAR): the occurrence of missing values for a variable is not related to the missing value, the values of any other variables, or the pattern of missingness in other variables

     → resources
     mistakes

  2) Missing at Random (MAR): the occurrence of missing values for a variable is random, contingent on the value or missingness of observable variables conditionally MCAR

     } modelable missingness

  3) Missing Not at Random (MNAR): the occurrence of missing values is systematically related to unknown or unmeasured covariate factors

     ☹

- Each form of missingness has different potential consequences

MCAR — efficiency

MAR (unmodeled) → bias & efficiency
sample selection

model { MAR →
MCAR maybe (less) efficiency problem.

  - Idea: we can fix MCAR and MAR data to look like non-missing data by filling in values based on what we DO know, reducing bias and inefficiency

NMAR/MNAR: ☹ bias & efficiency.

Early ideas

Friday, November 30, 2012
2:35 PM

- Scholars have traditionally used many ad hoc methods for handling missing data

1) Listwise deletion: drop any case with missing data on ANY observation
   a. Probably the most common naïve method for handling missing data
   b. Implemented by default in Stata and (usually) R
   c. Maximum loss of information

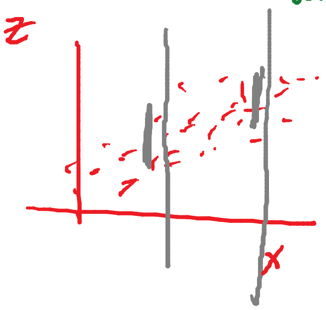
MCAR — efficiency
MAR — bias, efficiency.

y	x	z	w
6	8	9	- 8.5
-	-	-	-

$\mu_w = 8.5$

2) Mean (or multivariate distribution) imputation: replace missing values with the observed mean (or draws from the multivariate distribution) of the variable
   a. Understates variability in the imputed variable
   b. Makes no (or a limited) attempt to recover associations between the variables

underestimating SEs.



$\rho_{xz} = .6$

y	x	z
9	3	8
6	4	5

$\mu_x = 3$ $\mu_z = 5$

3) Regression-based imputation
   a. Implemented in the "impute" command in Stata
   b. Uncertainty about the quality of the estimate is not included in estimates using the imputed data

$y = \beta_0 + \beta_1 x + \beta_2 \omega + \beta_3 z$

$\hat{x} = f(W\hat{\alpha})$

efficiency →
SE too small

$x = f(\underline{W}\alpha)$ $\ddot{SE}$ too small → $\hat{\beta}$

Uncertainty in $\hat{\underline{\alpha}}$ →

Uncertainty in model of y

β

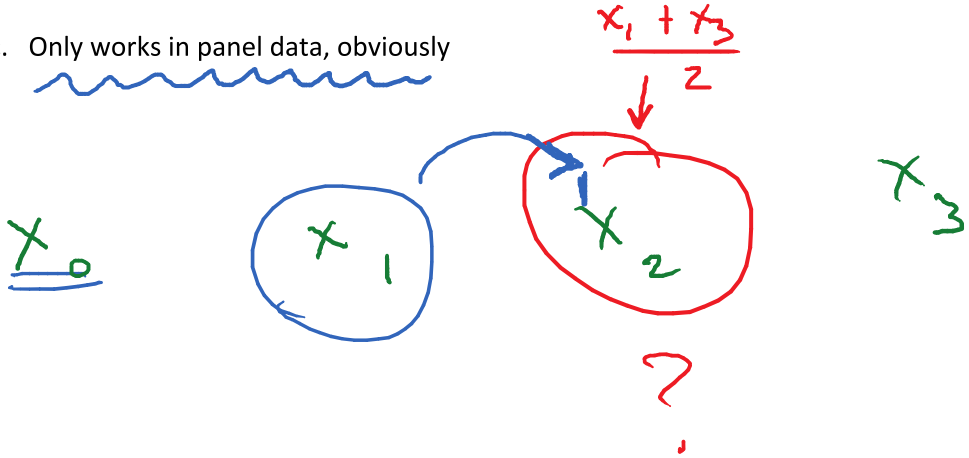
$$\underline{x} = \boxed{\hat{\alpha}_0} + \hat{\alpha}_1 W + \hat{\alpha}_2 Z + \hat{\alpha}_3 \underline{y}$$

4) Interpolation of panel data

a. If value is missing, either use the observation from the last time period OR a linear interpolation of the previous and next observation

b. Some evidence that this technique creates bias and overconfidence in estimates

c. Only works in panel data, obviously



Better idea: multiple imputation

$y \sim x + z$

Friday, November 30, 2012
3:31 PM

imp. → $x \sim \alpha_0 + \alpha_1 z + \alpha_2 y$

- We want to account for the uncertainty in our imputed variable
- use a model of some kind (e.g., regression?) to predict the missing observations using non-missing observations...
- Instead of simply picking one value for a missing value, we pick many... and our uncertainty is represented in the VCV matrix of the β coefficients we use to predict the missing values!

draw m (α) out of asymptotic dist, normal.

1) We pick m many values of $\hat{\alpha}$ out of its asymptotic distribution, the multivariate normal, using our estimates of $\hat{\alpha}$ and the VCV $\hat{\Sigma}$ to fill in the mean and VCV of this distribution $\Phi(\hat{\alpha}, \hat{\Sigma})$

2) predict m many values of the missing value, creating m many data sets

3) re-calculate m new estimates of $\tilde{\beta} = \sum_{m=1}^{M} \tilde{\beta}_m$ using each of the imputed datasets, and then calculate the standard error of our final $\tilde{\beta}$ using the following formula developed by Donald Rubin: → inflation in the SEs of $\tilde{\beta}$ that we do to correct for imputation.

$$V_\beta = W + \left(1 + \frac{1}{m}\right)B$$

OLS estimate of σ^2

where $W = \frac{1}{m}\sum_{m=1}^{M} s_m^2$ and $B = \frac{1}{m-1}\sum_{m=1}^{M}(\tilde{\beta}_m - \tilde{\beta})^2$ ← variance due to uncertainty in imputation of x.

- W and B are estimates of the within-imputation and between-imputation variation.

$y \sim \beta_0 + \beta_1 x + \beta_2 z$

$[\alpha_0 + \alpha_1 z + \alpha_2 y]$

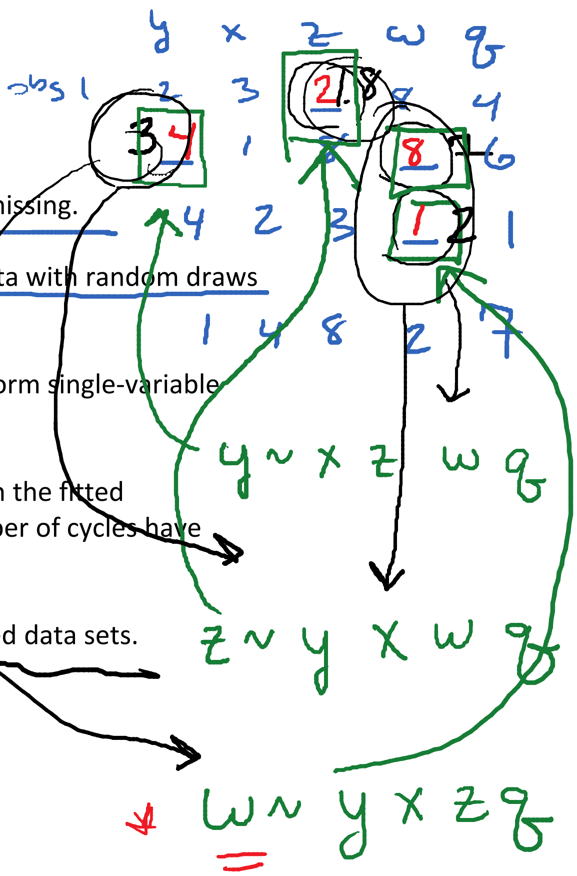
drew m many copies of $\hat{\alpha}$ from the asy. dist of α

Multiple Imputation through Chained Equations

Friday, November 30, 2012
3:35 PM

- Developed by van Buuren [and collaborators]

1. Discard all observations for which everything is missing.
2. For all missing observations, fill in the missing data with random draws from the observed values.
3. Move through the columns of variables and perform single-variable imputation using some method.
1. Replace the original (random) replacements with the fitted replacements. Repeat step 3 until a certain number of cycles have completed.
5. Do stages 1-4 m many times to create m imputed data sets.



- There are many ways to impute variables using the MICE algorithm
  - Regression (linear, or logistic , or multinomial) $\hat{w}$ or $f(\hat{w})$ sample.
  - Predictive Mean Matching (PMM) -- the default in MICE for continuous variables
    - Create predicted value for missing variable from regression model
    - Pick the three cases that have the closest predicted values (in terms of Euclidean distance)
    - Randomly choose one of the three values to impute

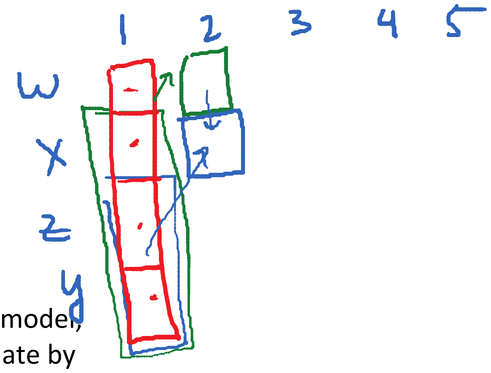
x~ y z w

y	x	z	w	$\hat{x}$
4	2	1	6	1
6	4	3	1	2

6	4	3	1	2
2	4	8	7	1.6 ✓
9	4	1	3	8
2	8	9	6	6

Bayesian Data Augmentation

Friday, November 30, 2012
3:52 PM

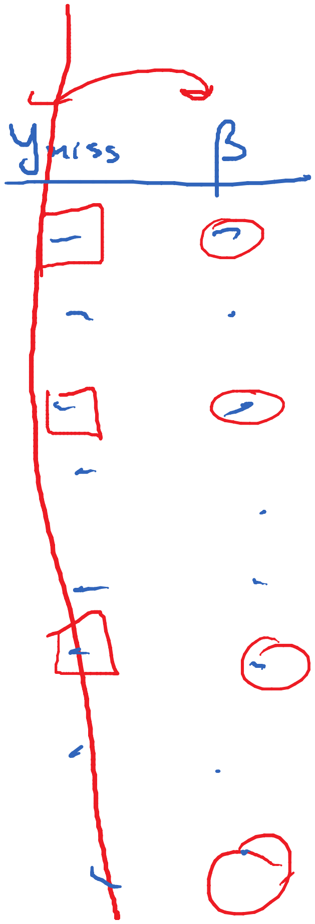


- MICE imputations are a form of Markov Chain

- You know where else we build Markov Chains?
- Build a missing data model right into a Bayesian (hierarchical) model, treating the missing values as just another parameter to estimate by drawing out of its posterior distribution

$$f(\beta, y_{miss}|y_{obs}) \propto f(y_{obs}|\beta, y_{miss})f(\beta, y_{miss})$$

$f(\beta)$ $\quad$ $f(\beta = \beta_0)$



- What we want to do is integrate out the missing values by sampling from the total distribution, then averaging out the beta distributions over the space of missing data points

$$f(\beta|y_{obs}) \propto \int_{Y_{miss}} f(y_{all}|\beta)f(\beta)\left[\int_B f(y_{miss}|\beta, y_{obs})f(\beta)\, d\beta\right] dy_{miss}$$

- As long as the data are MAR and that the likelihood of missingness is not related to β (the "ignorability") assumption, this works fine

$\beta_0 \quad \beta_1 \quad \beta_2$

$$f(\beta_0, \beta_1, \beta_2)$$
$$f(\beta_0 | \beta_1, \beta_2)$$
$$f(\beta_0)$$