Missing Data:
Theory and Practice

Marco Morales
mam2519@columbia.edu

GR5069
Topics in Applied Data Science
for Social Scientists

Spring 2018
Columbia University

Missing Data

The nature of the problem

unit	age	income	economic perceptions	education
1	33	25	3	14
2	22	?	-2	12
3	50	300	0	?
4	?	220	1	20
5	18	?	-1	11
6	45	180	2	13
7	76	50	-3	16
8	29	98	?	14

Missing Data

Consequences of the problem

- most algorithms **assume no missingness** in the data
  - typically not the case
- most common causes of data missingness:
  - **item non-response:** units provide information selectively (not everyone wants to reveal their income)
  - **unit non-response:** "units" provide no information (consequence of war)
  - **lost information:** miscoded information, lost records

Missing Data

Consequences of the problem

- potential **biases**:
  - projections outside of the support region
  - projections based on samples different from target population
  - incorrect - underestimated - variances (relevant on inferential problems)
- **Fundamental problem:** not using all available information
- **Consequence:** we may be generating **valid inferences/predictions for the wrong population**

Missing Data

Some theory and notation

$$D = \begin{bmatrix} 1 & 33 & 25 & 3 & 14 \\ 2 & 22 & \color{red}{20} & -2 & 12 \\ 3 & 50 & 300 & 0 & \color{red}{16} \\ 4 & \color{red}{30} & 220 & 1 & 20 \\ 5 & 18 & \color{red}{10} & -1 & 11 \\ 6 & 45 & 180 & 2 & 13 \\ 7 & 76 & 50 & -3 & 16 \\ 8 & 29 & 98 & \color{red}{2} & 14 \end{bmatrix} \qquad M = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

where

$D : \{D_{obs}, D_{miss}\}$
D_{miss} = **missing** data
D_{obs} = **observed** data
$M : \{1, 0\}$= missingness indicator matrix

Missing Data

Data Missingness mechanisms

- **Missing Completely at Random (MCAR)**: the probability of missingness is independent from the data (D)

$$P(M|D) = P(M)$$

- **Missing at Random (MAR)**: the probability of missingness only depends on observed data (D_{obs})

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

- **Non-Ignorable (NI)**: the probability of missingness depends both on observed (D_{obs}) and unobserved (D_{miss}) data

$$P(M|D_{obs}, D_{miss}) = P(M|D)$$

Missing Data

Data Missingness mechanisms

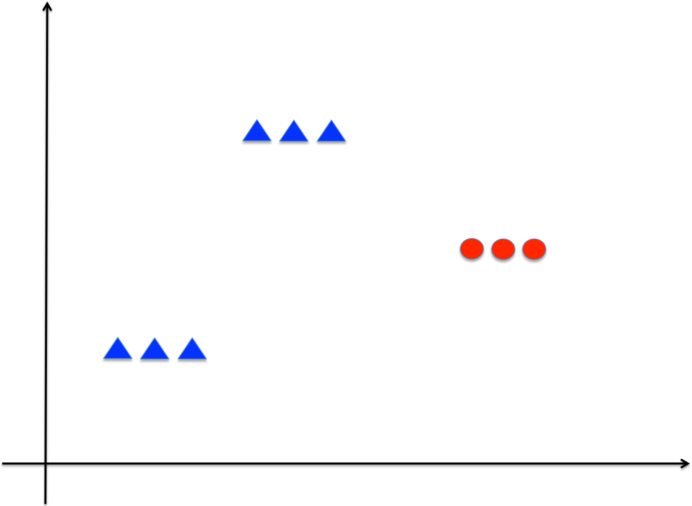
Mechanism	**Predict using**
Missing Completely at Random (MCAR)	–
Missing at Random (MAR)	D_{obs}
Non-ignorable (NI)	D_{obs} & D_{miss}

- data **imputation** can be used to address data missingness
  - imputation would **only work under MAR**
- MAR is an **assumption** (not directly verifiable)
  - ... but supported by some sort of theory about how missingness was generated

- **imputation methods** have been devised to handle missingness
  - **hot/cold deck imputation**: missing data is provided by a "nearest neighbor" donor unit
  - **mean imputation**: missing data is provided by the mean of observed data
  - **regression-based imputation**: missing data is generated by a regression model, conditional on observed data
  - **multiple imputation**: regression-based imputation that produces m vales for each missing value, conditional on observed data

Missing Data

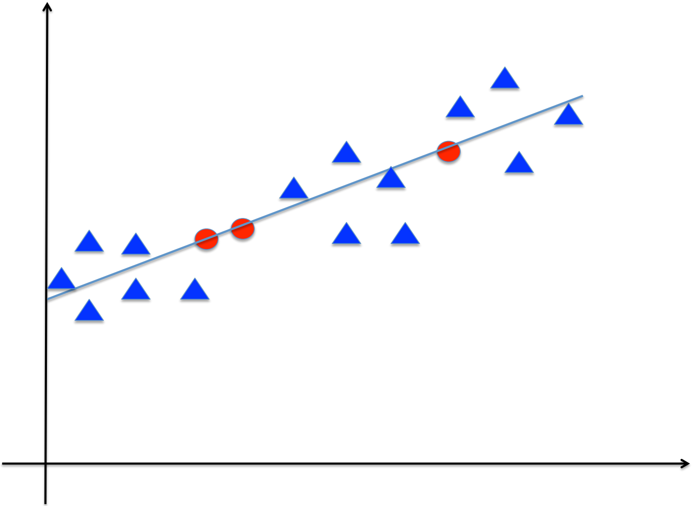
Mean Imputation



King, Honaker, Joseph & Scheve (1999)

Missing Data

Regression-based imputation

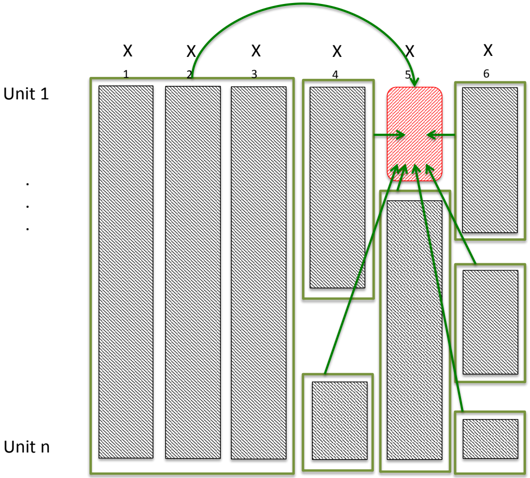


King, Honaker, Joseph & Scheve (1999)

- **single-value imputations** may have important shortcomings:
  - potential bias in point estimates
  - understate uncertainty surrounding imputed values (underestimate variances)
- **multiple imputation** overcomes some of these shortcomings
  - assigns m plausible values from a conditional distribution
  - provide variance estimates that converge to the true variance
- particularly important when trying to generate **valid inferences**

Missing Data

Multiple Imputation



Missing Data

Multiple Imputation - Rubin (1977)

1) **impute** m values for each missing data
   - employ an algorithm to impute missing data m times
   - existing data remain unchanged
   - a stochastic value is assigned for missing values
2) **analyze** each one of the m data bases
   - use each of m data bases *as if* it had full information
   - perform analyses on each data base: compute descriptive statistics, regression, etc
3) **combine** m estimates to compute point estimates and variances of quantities of interest(q)

- **Point estimates** of quantities of interest

$$\tilde{q} = \frac{1}{m} \sum_{j=1}^{m} q_j \tag{1}$$

where

$\tilde{q} =$ point estimate of quantities of interest
$q_j =$ quantity of interest for imputation j
$m =$ number of imputations

Missing Data

Multiple Imputation - quantities of interest (q)

- **Variance of the point estimate** of the quantity of interest
- sum of *within* and *between* imputation variance

$$
\begin{aligned}
SE(q)^2 &= \bar{w} + b \\
&= \frac{1}{m}\sum_{j=1}^{m} SE(q_j)^2 + \left(1 + \frac{1}{m}\right)\frac{\sum_{j=1}^{m}(q_j - \bar{q})^2}{m-1}
\end{aligned}
\tag{2}
$$

where

$\bar{w}$ = *within* imputation variance
b = *between* imputation variance
$\tilde{q}$ = point estimate of the quantity of interest
q_j = quantity of interest on imputation j
m = number of imputations

Missing Data

Multiple Imputation - quantities of interest (q)

- q is distributed t with degrees of freedom defined by

$$d.f. = (m-1)\left[1 + \frac{1}{m+1}\frac{\bar{w}}{b}\right]^2 \qquad (3)$$

where

$\bar{w}$ = *within* imputation variance
b = *between* imputation variance
m = number of imputations

Missing Data

Multiple Imputation - advantages

- Accurately reflects **imputation uncertainty**
  - imputation with useful information have low variances
  - includes *between* imputation variance to avoid underestimating the general variance

Missing Data

Imputation Software

- packages
  - `mi`
  - `mice`
  - `Amelia`

Missing Data

Missing Data in Big Data environments

- there seems to be a belief that as size tends towards **big data**, missingness becomes less relevant
  - belief that **asymptotics** kick in and solve everything
  - belief that **large samples** are, by definition, unbiased
  - a number of **implementations of common classifiers** handle missing data natively
- **problem:** missingness may be generating a **biased sample** of observed data... regardless of size
- **consequence:** biased training and testing sets $\neq$ general population

Missing Data

Missing Data in Big Data environments

- **question:** how good are these algorithms at recovering the original data distribution if using a biased sample (Zadrozny 2004)?
  - **local learners:** output depends asymptotically on $P(y|x)$
    - logistic regression, hard margin SVM
  - **global learners:** output depends asymptotically on $P(y|x)$ and $P(x)$
    - Bayesian classifiers, decision trees, soft margin SVM
- **results:** local learners not affected by sample selection bias, but global learners are

Missing Data:
Theory and Practice

Marco Morales
mam2519@columbia.edu

GR5069
Topics in Applied Data Science
for Social Scientists

Spring 2018
Columbia University