Missing Data Methods

Aaron Danielson

UCLA Department of Statistics

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick		17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83		77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			_
Carol		3	127	11	1	veggie	1	""""
Mandy	44	2	68	8	1	chicken		null

Overview

Missing Data Methods

Core Ideas

- Real data frequently contains missing values.
- Data can be missing in different ways. The mechanism of missingness determines whether it will effect statistical analysis.
- To avoid deletion of rows and columns of a matrix data, missing values can be imputed.
- The idea is to sample many complete datasets and average results across them.
- Imputation can help prediction if it preserves cases not represented in the complete data. (Ex: predict political party using income).

Let $\mathbf{X} \sim N \times K$ be a matrix of data and M be the missing data mechanism. Component ij is missing if $M_{ij} = 1$. We can write $\mathbf{X} = (\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}})$. We can define a probability distribution for the missing data mechanism

$$p(M|\psi)$$

where ψ is a vector of parameters.

Mechanisms of Missingness

Categorizing Missing Data

There are many ways that data can be missing.

- **Missing a priori.** By definition, the value does not exist:

$$p(M_{ij} = 1) = 1.$$

- **Missing Completely at Random (MCAR).** Missingness does not depend on the components that are missing:

$$p(M|\mathbf{X}, \theta) = p(M|\theta).$$

- **Missing at Random (MAR).** Missingness depends only on the observed data:

$$p(M|\mathbf{X}, \theta) = p(M|\mathbf{X}_{\text{obs}}, \theta).$$

- **Missing not at Random (NMAR).** Missingness depends on the unobserved components of the data.

The mechanism is said to be non-ignorable if it is NMAR since it cannot be estimated without knowledge of the missing values:

$$p(\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}, M|\theta_X, \psi) = p(M|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}, \psi)p(\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}|\theta_X).$$

Ignoring the presence of NMAR can result in biased statistical inference for the parameters of interest θ_X. And, if $\mathbf{X}$ is used to predict Y via parameter θ_Y, this parameter can also exhibit bias.

One can always use delete columns and rows of $\mathbf{X}$ to eliminate missing data. But, imputing the missing values may be useful to:

- Preserve degrees of freedom
- Remove bias due to the missingness mechanism.
- Maximize coverage of the covariate space (interpolation and extrapolation)
- Improve predictive performance.

Multiple Imputation with Parametric Models

The Big Idea

Sample from the posterior distribution $p(\theta_Y|Y, \mathbf{X}_{\text{obs}})$ by drawing $\mathbf{X}_{\text{mis}}^{(s)}$ from

$$p(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}}; \theta_X).$$

Then sample θ_Y with draws from

$$p(\theta_Y|Y, \mathbf{X}_{\text{mis}}^{(s)}, \mathbf{X}_{\text{obs}}).$$

Then

$$p(\theta_Y|Y, \mathbf{X}_{\text{obs}}) \approx \frac{1}{S}\sum_{s=1}^{S} p(\theta_Y|Y, \mathbf{X}_{\text{mis}}^{(s)}, \mathbf{X}_{\text{obs}}).$$

Moments and functions of the parameters can be estimated in this way.

There are many ways to impute missing values. Here are some common choices.

- Hot-deck.
- Mean/Median/Mode imputation.
- Ad hoc predictive models.
- Multivariate Approaches.
- RandomForest.

Multiple Imputation with Parametric Models

Averaging Across Imputations

Sample S complete data sets and compute the average parameter value

$$\hat{\theta} = \frac{1}{S}\sum_{s=1}^{S}\hat{\theta}^{(s)}$$

with variance

$$\mathrm{Var}(\hat{\theta}) = \hat{W} + \frac{S+1}{S}\hat{B}$$

where

$$\hat{W} = \frac{1}{S}\sum_{s=1}^{S}\hat{W}^{(s)} \text{ and } \hat{B} = \frac{1}{S-1}\sum_{s=1}^{S}(\hat{\theta}^{(s)} - \hat{\theta})^2$$

are the within-imputation and between-imputation variances, respectively.

- To make predictions about the variable Y with $\mathbf{X}$, we can use the averaged parameter values, $\hat{\theta}$.
- Or, for each $s = 1, \ldots, S$ average across predictions made by the separate models

$$\hat{Y} = \frac{1}{S}\sum_{s=1}^{S}\hat{Y}^{(s)}.$$

Suppose Y, a response, is completely observed and $\mathbf{X} = (\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}})$. If Y can be modeled in terms of $\mathbf{X}$, then

$$\begin{aligned} p(Y|\mathbf{X}_{\text{obs}}, \theta_Y) &= \int p(Y, \mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}}; \theta)\, d\mathbf{X}_{\text{mis}} \\ &= \int p(Y|\mathbf{X}_{\text{mis}}, \mathbf{X}_{\text{obs}}; \theta_Y) p(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}}; \theta_X)\, d\mathbf{X}_{\text{mis}}. \end{aligned}$$

If we ignore the missing data mechanism, then the likelihood of θ_Y satisfies

$$L_{\text{ignore}}(\theta_Y|Y, X_{\text{obs}}) \propto p(Y|X_{\text{obs}}, \theta_Y)$$

for all $\theta_Y \in \Theta$.

Now suppose we want to estimate θ_Y (the parameters relating $\mathbf{X}$ and Y) and θ_X (the parameters describing $\mathbf{X}$). Then

$$\begin{aligned} p(Y, \mathbf{X}_{\text{obs}}; \theta_Y, \theta_X) &= \int p(Y, \mathbf{X}_{\text{mis}}, \mathbf{X}_{\text{obs}}; \theta)\, d\mathbf{X}_{\text{mis}} \\ &= \int p(Y|\mathbf{X}_{\text{mis}}, \mathbf{X}_{\text{obs}}; \theta_Y) p(\mathbf{X}_{\text{mis}}, \mathbf{X}_{\text{obs}}; \theta_X)\, d\mathbf{X}_{\text{mis}}. \end{aligned}$$

If we ignore the missing data mechanism, then the likelihood of (θ_Y, θ_X) satisfies

$$L_{\text{ignore}}(\theta_Y, \theta_X | Y, \mathbf{X}_{\text{obs}}) \propto p(Y, \mathbf{X}_{\text{obs}}; \theta_Y, \theta_X)$$

for all $(\theta_Y, \theta_X) \in \Theta_y \times \Theta_X$.

Likelihood Models with Missing Data

Observed Likelihood (Non-Ignorable)

The joint probability of Y and the missingness mechanism M conditional on $\mathbf{X}_{\text{obs}}$ is

$$\begin{aligned} p(Y, M|\mathbf{X}_{\text{obs}}, \theta_Y, \theta_X, \psi) &= \int p(Y, M, \mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}}; \theta)\, d\mathbf{X}_{\text{mis}} \\ &= \int p(M|Y, \mathbf{X}_{\text{mis}}, \mathbf{X}_{\text{obs}}; \psi) p(Y|\mathbf{X}_{\text{mis}}, \mathbf{X}_{\text{obs}}; \theta_Y) p(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}}; \theta_X)\, d\mathbf{X}_{\text{mis}}. \end{aligned}$$

Then the observed data likelihood of the full model parameters satisfies

$$L_{\text{full}}(\theta_Y, \psi|Y, X_{\text{obs}}, M) \propto p(Y, M|\mathbf{X}_{\text{obs}}, \theta_Y, \theta_X, \psi)$$

for all $\theta_Y \in \Theta$ and $\psi \in \Psi$. If components of $\mathbf{X}$ are MAR, then

$$p(Y, M|\mathbf{X}_{\text{obs}}, \theta_Y, \theta_X, \psi) = p(M|Y, \mathbf{X}_{\text{obs}}, \psi) \int p(Y|\mathbf{X}_{\text{mis}}, \mathbf{X}_{\text{obs}}; \theta_Y) p(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}}; \theta_X)\, d\mathbf{X}_{\text{mis}}.$$

Likelihood Models with Missing Data

Implications for Estimation: What's the point?

- If data is MCAR or MAR, then the missing data mechanism need't be modeled and

$$\theta_Y^* = \text{argmax}_{\theta_Y \in \Theta} L_{\text{ignore}}(\theta_Y | Y, \mathbf{X}_{\text{obs}}).$$

- For joint estimation of θ_Y and θ_X,

$$(\theta_Y^*, \theta_X^*) = \text{argmax}_{\theta_Y \in \Theta} L_{\text{ignore}}(\theta_Y, \theta_X | Y, \mathbf{X}_{\text{obs}}).$$

- If the data is NMAR, then the missing data mechanism and the data must be modeled jointly. Then maximum likelihood estimates are

$$(\theta_Y^*, \psi^*) = \text{argmax}_{(\theta_Y, \psi) \in \Theta \times \Psi} L_{\text{full}}(\theta_Y | Y, \mathbf{X}_{\text{obs}}).$$

Estimation methods vary by application, but here are a few ways.

- The EM algorithm. Fill in missing values with their conditional expectations. Then maximize the conditional expectation.
- Data Augmentation. At iteration s, draw $\mathbf{X}_{\text{mis}}^{(s+1)}$ from $p(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}})$ and then draw $\theta_Y^{(s)}$ from $p(\theta|Y, \mathbf{X}_{\text{mis}}^{(s+1)}, \mathbf{X}_{\text{obs}})$.

Due to time constraints, we won't discuss these in detail.

Random Forest and Missing Values

Missing Predictor Values for Trees

- Discard?
- Impute with another mechanism.
- Create NA category for categorical variables.
- Use proximities.

There are two methods implemented by the RandomForest algorithm:

(1) Impute the median (mode) if data is continuous (categorical).
(2) Proximity-based Method
    a. Use (1) to get initial imputations.
    b. Compute proximities.
    c. Replace missing values in unit *i* by a weighted average of non-missing values, with weights proportional to the proximity between case *i* and the cases with the non-missing values.

    Repeat steps [a.] and [b.]

Software for Imputation

R Packages

Check out the following R packages to perform imputation.

1. Amelia
2. mice
3. missForest
4. missMDA
5. VIM

- S. van Buuren and Karin Groothuis-Oudshoorn. *mice: Multivariate imputation by chained equations in R*. Journal of statistical software, 1-68, 2010.
- Bradley Efron. *Missing Data, Imputation, and the Bootstrap*. Journal of the American Statistical Association, 89(426):463-475, 1994.
- James Honaker, Gary King, and Matthew Blackwell. *Amelia II: A program for missing data.* Journal of statistical software 45.7, 1-47, 2011.
- Alexander Kowarik and Matthias Templ. *Imputation with R package VIM.* Journal of Statistical Software 74.7 (2016): 1-16.
- Roderick JA Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Vol. 333., 2014.
- Benjamin Marlin. *Missing data problems in machine learning*. PhD dissertation, 2008.

D Stekhoven and P Buhlmann. *MissForest ? Non-Parametric Missing Value Imputation for Mixed-Type Data*. Bioinformatics, 28(1), 112?118. doi:10.1093/bioinformatics/ btr597.