Longitudinal data analysis in R: Imputation of missing data

Nelson Nazzicari

(Bioinformatician & data scientist) **CREA**, Lodi (Italy)

Ideally: how things “should” be

	Time 1	Time 2	Time 3	Time 4	Time 5
Variable 1	2.37	9.16	3.83	2.28	1.22
Variable 2	5.14	9.12	22.47	17.04	7.23
Variable 3	-1.74	-0.77	-0.13	0.74	-0.71

Realistically: how things often are

	Time 1	Time 2	Time 3	Time 4	Time 5
Variable 1	2.37	9.16	3.83	missing	1.22
Variable 2	5.14	missing	22.47	missing	7.23
Variable 3	-1.74	-0.77	-0.13	0.74	-0.71

Reasons for missing data points

- Experimental errors
  - Machine failure, contamination, external causes
- Intrinsic
  - Some techs work on a statistical basis (e.g. GBS)
- Data merging
  - Different sources, labs, sensors
  - Improvement in tech
- Bad luck ¯_(ツ)_/¯

Why is it a problem?

$$y = \beta_0 + \beta_1 x + e$$

- $y = \beta_0 + \beta_1 x + e$
- $\hat{y} = \beta_0 + \beta_1 x$
- $\beta_0 = 5, \beta_1 = 3$

x	ŷ
1	8
-3	-4
missing	?

Coping via filtering

	Time 1	Time 2	Time 3	Time 4	Time 5
Variable 1	2.37	9.16	3.83	**missing**	1.22
Variable 2	5.14	**missing**	22.47	**missing**	7.23
Variable 3	-1.74	-0.77	-0.13	0.74	-0.71

- Removing Time 2 and Time 4
- Removing Variable 2 and Time 4
- What to sacrifice?
- You can only go so far…

Real cases are worse

	Time 1	Time 2	Time 3	Time 4	Time 5
Variable 1	2.37	9.16	3.83	missing	missing
Variable 2	5.14	9.12	22.47	missing	7.23
Variable 3	missing	missing	-0.13	0.74	-0.71

- Common for have a variable (subject) to disappear…
- …or to enter the study after some point
- Measurement can change, especially on long studies (different technology, different teams)

Imputation: what is it?

- Filling the blanks (via some algorithm)
- It’s a kind of prediction
- Don’t expect miracles
  - But results may be very good (redundancy in data)
- NAIVE: use the arithmetic mean of the variable
- LESS NAIVE: linear regression between before and after, interpolation, locf/nocb (Last Observation Carried Forward / Next Observation Carried Backward)
- EVEN LESS NAIVE:

Random Forest imputation

- Ensemble learning (thus: forest)
  - Regression (numerical)
  - classification (categorical)

Random Forest imputation

Iterative approach:

1. Fill the blanks (median/average, or most common)
2. For each variable X (in ascending order of missing rate):
    - X becomes the dependent variable to be predicted (thus: y)
    - Predict ŷ training RF on all data except X
    - ŷ becomes the new X
3. Repeat point 2 for either a max number of loops or until updates under threshold

Random Forest imputation

- Complex relationship, mixed types, multivariate
- Slow, problems with NMAR (not missing at random)
- In R: library(missForest), function “missForest”
- In Python: fancyImpute package

Random Forest imputation

https://doi.org/10.1186/s12874-020-01080-1

Hong and Lynn *BMC Medical Research Methodology* (2020) 20:199
https://doi.org/10.1186/s12874-020-01080-1

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction

Shangzhi Hong and Henry S. Lynn*

Abstract

Background: Missing data are common in statistical analyses, and imputation methods based on random forests (RF) are becoming popular for handling missing data especially in biomedical research. Unlike standard imputation approaches, RF-based imputation methods do not assume normality or require specification of parametric models. However, it is still inconclusive how they perform for non-normally distributed data or when there are non-linear relationships or interactions.

Methods: To examine the effects of these three factors, a variety of datasets were simulated with outcome-dependent missing at random (MAR) covariates, and the performances of the RF-based imputation methods missForest and CALIBERrfimpute were evaluated in comparison with predictive mean matching (PMM).

Results: Both missForest and CALIBERrfimpute have high predictive accuracy but missForest can produce severely biased regression coefficient estimates and downward biased confidence interval coverages, especially for highly skewed variables in nonlinear models. CALIBERrfimpute typically outperforms missForest when estimating regression coefficients, although its biases are still substantial and can be worse than PMM for logistic regression relationships with interaction.

Conclusions: RF-based imputation, in particular missForest, should not be indiscriminately recommended as a panacea for imputing missing data, especially when data are highly skewed and/or outcome-dependent MAR. A correct analysis requires a careful critique of the missing data mechanism and the inter-relationships between the variables in the data.

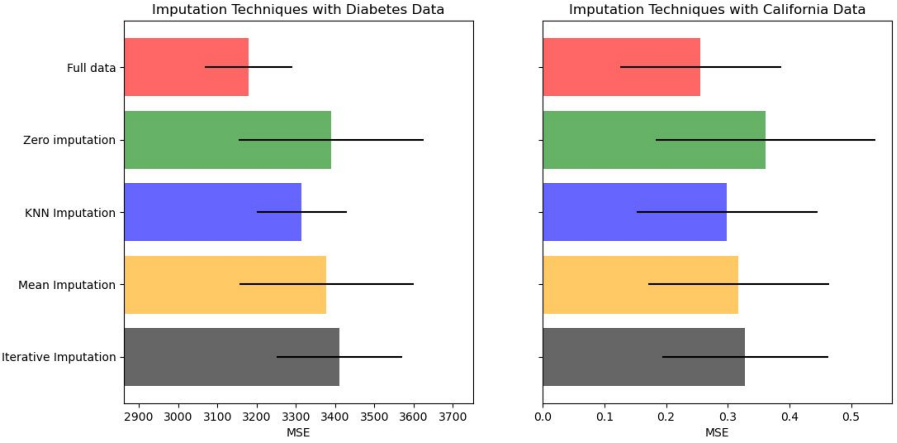
Keywords: Missing data imputation, Imputation accuracy, Random forest

K-Nearest Neighbors imputation

- Well known clustering algorithm
- Algorithmic approach:
  - Find the K closest examples
  - Fill the blank using the mode
- How to compute the distance?
  - Euclidean on complete set
- How to choose K?
  - Small Ks (1-3): high variance, risk of overfitting
  - Big Ks (4+): high bias, risk of losing details
- Fast, but sensitive to outliers
- In R: library(scrime), function “knncatimpute” (categorical only)
- In Python: packages fancyImpute or sklearn.impute

Notebook on imputation

Take home message



From https://scikit-learn.org/stable/auto_examples/impute/plot_missing_values.html

- Imputation can be relevant
  - How many missing points?
  - MAR?
- Part of EDA

Is imputation really needed?

XGboost (Extreme Gradient Boosting)

- ensemble of decision trees
- https://doi.org/10.1145/2939672.29397
- https://arxiv.org/abs/1603.02754
- Still boosting (ensemble of weak trees)
- Gradient boosting → updates to trees for loss minimization
- Newton–Raphson method for optimization
- Highly parallel (CPUs, GPUs, cache management)
- Many supported frameworks (python, R, scala, julia, ruby…)
- Missing values?

XGboost and missing values

- accommodates sparse feature formats
- “Sparsity Aware Split Finding” algorithm
  - assesses potential splits
  - look for maximum information gain
  - a missing value is encountered → informed decision about whether to go left or right in the tree structure
- default branches during inference

Notebook on XGboost