No solution? Evaluating Multiple Imputation

Anthony S. Chapman, Dr Steven Turner, Dr. Wei Pang

Abstract

Blah

1. Introduction

Data collection has been increasing Missing data is inevitable (human and computing reasons, i.e. people not putting it in or computer corrupting it,) Non-computing people either imputate willy-neely or ignore missing data - need to use as much data as you can. (Ask Graham about theory about using as much data as possible for better analysis) Many imputation algorithms out there with many parameters, which is best? Need

2. Background

Talk about something [1] [7]

3. The Problems

3.1. Incompleteness

With so many new data being collected daily [7], it was inevitable that some of the data would have missing values [4], whether they be through human error or computational inefficiency. Although there are ways to combat missing data such as mean-value imputation or multiple imputation [4, 6, 2], many researchers whom are not very computational or statistically confident would rather disregard any records with missing values [1, **?**, **?**, **?**, **?***]. As an example, in [1], the authors decided to use 2,758 records for analysis out of the possible 44,261 mainly due to missing data, this is a mere 6.2% out of the records available. There must be a way for even non-computing or non-statistical researchers to benefit from the tools available.*

3.2. Will it work on my data?

This next problem arises when a researchers does decide to use the data with missing values but does not have sufficient knowledge to apply the available methods, like Muliple Imputation by Chained Equations (MICE [9]) using the computational language R [5] or the Impute Missing Values function in the statistical software SPSS [3]. The problem is, how do you know if the imputed values are representative to the truth, how do you know whether record 2,754 column 5 is male or not after you apply the imputation method.

Even if the imputation method has been proven to work on someone else's dataset such as [8], this no indication it will work for yours, unless you have the exact dataset as them, which is unrealistic. This is due to the many reasons and ways that missing data is creates, for example there might be a relationship between one missing value and another one.

In order to test whether an imputation method works on your dataset, you need something to compare the results to, a benchmark, like this one would be able to analyse what effect of the methods. Unfortunately, it is very difficult to find a complete dataset which contains the same characteristics of your own dataset, there will always be differences.

3.3. Which imputation is best for me

The following problem applies to researchers, even those computationally competent, who wish to find out whether one imputation method is better than another. There is nothing to easily compare results from different imputation methods or same imputation methods with slightly different parameters. The main problem arises when one tries to compare the outcomes from one method to another, here an adequate analogy would be that compare imputation method A to method B would be like comparing chocolate with a bicycle; the outcomes might not be comparable.

There should a way to compare different methods without having to create your own computer software in the process. Although

4. Possible Solutions

4.1. Incompleteness

Imputations solves incompleteness, you just have to be careful how you use it.

Can't blindly impute something as it might result in bias and unreliable results.

4.2. Testing your own data

Use your own level or missingness as a bench-mark and create mini-me's as bench-mark. You are the closest thing to yourself. Group theory stuff, multidimensional-mixed data distance measurements, Gower, medoids, widths and dissimilarities.

Just because it worked on someone else, doesn't mean it works for you, cite papers who test specific datasets.

4.3. Comparing Imputations

Will now be able to compare different imputations on your own dataset with "normalised " results for comparison.

5. Conclusion

It's better to use all the data you can but can't blindly imputation. This framework indicates whether your data

6. Discussion

Working on implementing this, ClEMI, any re-searcher regardless the computing ability will be able to use it.

References

[1] Amy M. Branum, Jennifer D. Parker, Keim Sarah A., and Schempf Ashley H. Prepregnancy body mass index and gestational weight gain in relation to child body mass index among siblings. *American Journal of Epidemiology*, 174(10):1159–1165, 2011.

[2] ALAN C. COCK. Working with missing values. *Journal of Marriage and Family*, 174(67):10121028, 2005.

[3] SPSS Inc. *SPSS Statistics for Windows, Version 17.0.* Chicago: SPSS Inc, 2008. `http://www-01.ibm.com/software/uk/analytics/spss`.

[4] Therese D. Pigott. A review of methods for missing data. *Educational Research and Evaluation*, 7(4):. 353–383, 2001.

[5] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0 `http://www.R-project.org`.

[6] Donald B. Rubin. An overview of multiple imputation.

[7] ScienceDaily. Big data, for better or worse. 2013 (accessed: January 18, 2016). `http://www.sciencedaily.com/releases/2013/05/130522085217.htm`.

[8] Anoop D. Shah and Jonathan W. Bartlett. Comparison of parametric and random forest mice in imputation of missing data in survival analysis. 2014.

[9] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. `http://www.jstatsoft.org/v45/i03/`.