Cleaning Electronic Medical Records Using Novel R Package MonoInc

Melyssa Minto, Michele Josey, Chantel I Nicolas. Catherine Hoyo, ClarLynda Williams-DeVane
Biomedical/Biotechnology Research Institute, North Carolina Central University

Abstract

Electronic medical records (EMR) have given researchers access to abundant data; however, EMR are infamous for their inaccuracies and inconsistencies. In the big data era of EMR, simple data cleaning has become a cumbersome effort requiring methods development to clean data in a manner that maintains its integrity. Using R, a free statistical computing software, MonoInc package was developed to impute monotonic data that is missing and outside of a given range, even data as noisy as EMR. This package also contains accessory functions; one to simply flag erroneous observations (mono.flag), one to calculate the proportion of entries inside of a given range (mono.range), and a function to check monotonicity (monotonic). The imputation methods offered in this package are Last and Next (LN), Nearest Neighbor (NN), Regression (Reg), Fractional Regression (FR), Last Observation Carried Forward (LOCF), and a weighted combination of those methods. Evaluations of all imputations in this package are based on each user's data rather than global data. Using a subset of data from the Newborn Epigenetic Study at Duke (n = 233), children's heights from months 0 to 120 were imputed using MonoInc. Each imputation method was analyzed to compare the imputation methods offered in MonoInc and to test its efficacy by examining the root mean sure error, mean absolute deviation, proportion variation, mean deviation and the residual from the mean curve, the proportion inside the pre-specified range, and the number of complete cases. Most weighted imputations that included LN performed well while weighted imputations with LOCF did not. Overall the weighted imputations outperformed single imputation methods. This analysis also shows which imputation method in MonoInc is best to use, based on the user's preference. MonoInc is a simple tool for cleaning erroneous values in data to ensure its monotonicity while maintaining the data's integrity.

Methods

Creating the Package:

1. A function was developed in R to flag erroneous values and impute them given an imputation method
2. From this function an R package was created with supporting functions
   - *mono.flag,*: check whether a single vector is monotonic or check whether a individuals data is monotonic in a dataset with multiple individuals
   - *mono.range*: computes the number of entries fall inside that a given range
   - *monotonic*: can flag where in the data is 'unusual' i.e. not monotone, outside of a range, or missing

The different imputation methods offered were:

1. Nearest Neighbor (NN): imputes based on the closet neighboring point (Paik et al., 2006)
2. Regression Imputation (REG): imputes the predicted value from linear regression coefficients of other points with the same ID (Paik et al., 2006)
3. Fractional Regression Imputation (FR): consists of Regression Imputation with an added error term (Paik et al., 2006)
4. Last Observation Carried Forward (LOCF): imputes the last recorded point (Powney et al. 2014)
5. Last & Next (LN): takes the average of the last and next recorded points (Engels et al. 2002))
6. Weighted average: compares any two imputation methods mentioned above

To compare the Imputation methods, they were ranked based off the following calculations:

- Root mean square error (RMS)
- Mean absolute deviation (MAD)
- Proportion variation (PV)
- Mean deviation (MD)
- Residual from the mean curve (Res)
- Proportion inside the range (RP)
- Number of complete cases (CC)

How to Use MonoInc

Checking Monotonicity of a vector

```
x<-c(1,2,3,4,5)
monotonic(x, direction='inc')
## [1] TRUE
x<-c(5,4,3,2,1)
monotonic(x,direction='dec')
## [1] TRUE
```

Checking Monotonicity of a dataset with multiple ids

```
#check of a data.matrix has monotonic  data
test <- monotonic(simulated_data, 1,3, direction = 'inc')
```

Seeing how many entries fall with in the normal ranges

```
mono.range(simulated_data, data.r, tol=4, xr.col=1 ,x.col=2, y.col=3)
[1] 0.6774194

test <- mono.flag(simulated_data, 1, 2, 3, 30, 175, data.r=data.r, direction=
'inc')
```

Impute missing monotonic data

```
locf <- MonoInc(simulated_data, 1,2,3, data.r,4,direction = 'inc', w1=0.3, mi
n=30, max=175, impType1='locf', impType2=NULL, sum=F)
```

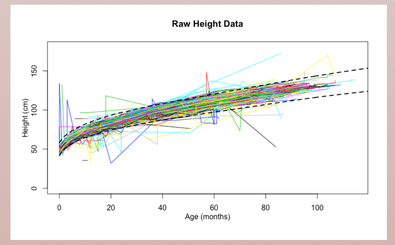

Figure 1: Shows the plot of each individuals growth from raw NEST data. The dotted lines are boundary lines that represent the upper and lower limits from the prespecified data with a tolerance of 4.

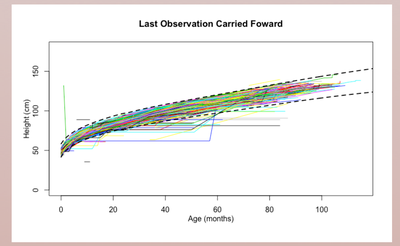

Figure 2: shows the plot of each individuals growth from NEST data after last observation carried forward imputation method. The dotted lines are boundary lines that represent the upper and lower limits from the prespecified data with a tolerance of 4.

Seeing all imputation options

```
sum <- MonoInc(simulated_data, 1,2,3, data.r,4,direction = 'inc', w1=0.3, min
=30, max=175, impType1=NULL, impType2=NULL, sum=T)
head(sum)
```

Table 1: Shows the output of the MonoInc function when sum=TRUE

ID	X	Y	Nearest.Neighbor	Regression	LOCF	Last.Next	Fractional.Reg	Decreasing	Outside.Range
30	108	NA	134.13068	139.05812	125.67741	129.90405	142.10143	TRUE	TRUE
30	117	134.13068	134.13068	134.13068	134.13068	134.13068	134.13068	FALSE	FALSE
31	22	82.41794	82.41794	82.41794	82.41794	82.41794	82.41794	FALSE	FALSE
31	40	NA	93.33662	95.72835	93.33662	98.59616	100.58994	TRUE	FALSE
31	48	NA	103.85569	99.46063	103.85569	103.89058	104.32223	TRUE	FALSE
31	56	103.92547	103.92547	103.92547	103.92547	103.92547	103.92547	FALSE	FALSE

Results

Table 2: Ranks of the imputation methods

	Res	PV	MD	MAD	RMS	RP	CC	Total
lnfr	5	6	14	3	3	4	3	38
nnreg	1	3	5	8	8	11	2	38
nnlocf	10	7	4	1	1	13	2	38
lnlocf	3	11	15	2	2	3	3	39
lnreg	2	8	13	4	4	5	4	40
lnnn	6	9	10	5	5	2	4	41
reglocf	13	1	1	6	7	14	2	44
frlocf	14	2	3	7	10	10	1	47
nnfr	11	5	7	9	9	7	2	50
ln	4	12	6	12	11	1	5	51
nn	15	4	2	11	12	9	5	58
regfr	8	14	8	13	13	8	2	66
locf	12	10	9	10	6	15	5	67
reg	7	13	11	14	14	12	5	76
fr	9	15	12	15	15	6	5	77

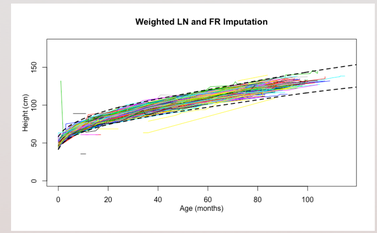

Figure 3: shows the plot of each individuals growth from simulated data after weighted last and next and fractional regression imputation The dotted lines are boundary lines that represent the upper and lower limits from the prespecified data with a tolerance of 4.

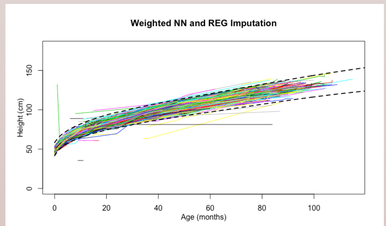

Figure 4: shows the plot of each individuals growth from simulated data after weighted nearest neighbor and regression imputation The dotted lines are boundary lines that represent the upper and lower limits from the prespecified data with a tolerance of 4.

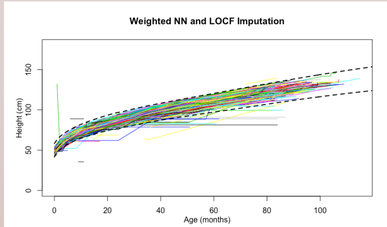

Figure 5: shows the plot of each individuals growth from simulated data after weighted nearest neighbor and last observation carried forward imputation The dotted lines are boundary lines that represent the upper and lower limits from the prespecified data with a tolerance of 4.

Conclusion

Based on the results of almost any other imputation ethos paired with LN did relatively better than the other weighted imputations. Regression by itself had poor results. If you want your data to maintain the same integrity after imputation, LN, NN, and FR are strongly recommended, however of you don't mind a little shift weighted imputations like LNNN and LNREG are strongly recommended. LOCF is the the worst imputation method, not only did were the results poor but the integrity of the data after this imputation is very damaged. Overall the weighted imputations helps reduce the non-monotonicity, the data outside of a prespecified range, and missing. The more inaccurate the original data is, the worst the imputation will be.

References

Engles, J.,Diehr,P. (2002) Imputation of missing longitudinal data: a comparison of methods. Journal of Clinical Epidemiology **46**, 968-976.
Minto, M., Josey, M., and Williams-DeVane, C. (2016) MonoInc: Monotonic Increasing. R package version 1.1. https://CRAN.R-project.org/package=MonoInc.
Paik, M. and Larsen, M. (2006) Fractional Regression Nearest Neighbor Imputation. *ASA Section on Survey Research Methods*, 3500-3507.
Powney, M, Williamson, P, Kirkham, J, Kolamunnage-Dona R. (2014) A review of the handling of missing longitudinal outcome data in clinical trials. *Trials*, **15**, 237.