Saturday, June 3, 2023

Seeing the leaves through the forest

A primer on using random forest models for missing data problems

Anonymized for review

ABSTRACT Political scientists often struggle with decisions about what to do with incomplete cases for their regression analyses, and one can often make several decisions that influence one's ultimate substantive conclusions. In many areas of research outside of political science and the social sciences, scholars take advantage of an extension of multiple imputation, which offers the choice to leverage machine learning models for predicting values in missing data. This manuscript provides a summary of missing data and its consequences for our regression models along with providing an explanation of how to implement random forest models with an expanded form of the multiple imputation procedure, called multiple imputation with chained equation to handle complex causes for non-random missingness in our data. After providing a primer on standard missing data procedures in political science and random forest with multiple imputation with chained equations, I examine its performance on simulated data. I conclude by providing recommendations for dealing with missing data in practice.

KEYWORDS Missing data; Multiple imputation; Machine learning

```
# Install libraries

install.packages(

  c(

      "AMELIA"

      , "mice"

      , "miceRanger"

  )

)

# Load libraries

library(Amelia) # for MI
```

```r
library(mice) # for many MICE and interpolation procedures

library(miceRanger) # for RF-Mice procedure

# Dataset


df


# Listwise Deletion

dfImputed <- df[complete.cases(df), ] # exclude rows that have missing values in any column


# Interpolation

dfImputed <- mice(

  df # dataframe

  , m = 10 # number of imputations

  , method = "mean" # mean interpolation

)


# Amelia

dfImputed <- amelia(

  df # dataframe

  , m = 10 # number of imputations

)


# Linear Bayesian MICE

dfImputed <- mice(

  df # dataframe

  , m = 10 # number of imputations

  , method = "linear" # Bayesian linear MICE
```

```
)


# RF-MICE with mice package
dfImputed <- mice(
  df # dataframe
  , m = 10 # number of imputations
  , method = "rf" # RF-MICE
)


# RF-MICE with miceRanger package
dfImputed <- miceRanger(
  df # dataframe
  , m = 10 # number of imputations
)
```