Missing Data in Large Scale Education Surveys

Developed by Paul Bailey, Michael Lee, Ting Zhang, and Howard Huade[*†]

May 30, 2019

The `EdSurvey` package gives users functions to efficiently analyze education survey data with complete cases. This document describes how to use the `mice` and `VIM` packages with `EdSurvey` to:

1) explore patterns of missingness,
2) impute data, and
3) run analyses on the imputed data.

These steps all must properly account for the survey weights and plausible values.

Characterizing Missing Values

For this section, the primary purpose of exploring missingness patterns is to see if data may be missing completely at random, and to discover the extent to which the data that is not missing supports imputation.

A variable that does not have any missing value does not need imputation, while a variable that is largely missing may not successfuly support imputation.

This vignette will use data from the Trends in International Mathematics and Science Study (TIMSS). To follow along with the examples to follow, consult the **Downloading Data** section of the *Using EdSurvey to Analyze TIMSS Data* vignette, or download and read the data as shown:

```
# library(EdSurvey)
# for Mac OS
# downloadTIMSS(years = 2015, root = '~\')
# fin4 <- readTIMSS("~\TIMSS\2015")
```

```
# library(EdSurvey)
# for Windows
# downloadTIMSS(years = 2015, root = 'C:/', cache=FALSE)
# fin4 <- readTIMSS("C:/TIMSS/2015")
```

```
library(EdSurvey)
fin4 <- readTIMSS(paste0(edsurveyHome,"/TIMSS/2015/"), countries="fin", gradeLvl=4)
```

Note that the data will be automatically stored in a folder in the directory that you specified. For example, the 2015 TIMSS data will be saved in the "TIMSS/2015" folder in the C drive. The R program assigns the folder name, but you can manually change it.

These series of explorations will focus the first plausible value for the mathematics subject scale `asmmat01` and the variable `atbg06f`, which characterizes parental involvement as shown via `searchSDF`:

```
searchSDF("atbg06f", fin4, levels = TRUE)
```

```
## Variable: atbg06f
## Label: GEN\CHARACTERIZE\PARENTAL INVOLVEMENT
## Levels (Lowest level first):
##      1. VERY HIGH
```

*This publication was prepared for NCES under Contract No. ED-IES-12-D-0002 with the American Institutes for Research. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

†The authors would like to thank (someone) for reviewing this document.

```
##      2. HIGH
##      3. MEDIUM
##      4. LOW
##      5. VERY LOW
##      9. OMITTED OR INVALID
```

The `EdSurvey` package allows us to explore the data connection without reading the full data set into our working memory. A subset of variables in the TIMSS data set that might relate to missingness in the `atbg06f` variable, such as:

- `"asmmat01"`, the first plausible vales for mathematics scores
- `"asbh20a"`, `"asbh22a"`, and `"asbh23a"` covers education level, employment status, and main job of fathers
- `"asbh20b"`, `"asbh22b"`, `"asbh23b"` cover the education level, employment status, and main job of mothers
- `"asbg04"`, `"asbg05d"`, `"asbg05e"` covers the number of books in the home, whether they have their own room, and whether their home has an internet connection
- `"asbg09"`, `"acbg13"` covers how often they have breakfast on school days and whether they have a school library

```
vars <- c("asmmat01", "asssci01",
          "asbh20a", "asbh22a", "asbh23a",
          "asbh20b", "asbh22b", "asbh23b",
          "asbg04", "asbg05d", "asbg05e",
          "asbg09", "acbg13", "atbg06f")
```

Missingness and the distribution of another variable

First, to explore the association between missingness and the distribution of another variable we use the `VIM` package's `marginplot` function.

```
library(VIM)
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

```
## Loading required package: data.table
```

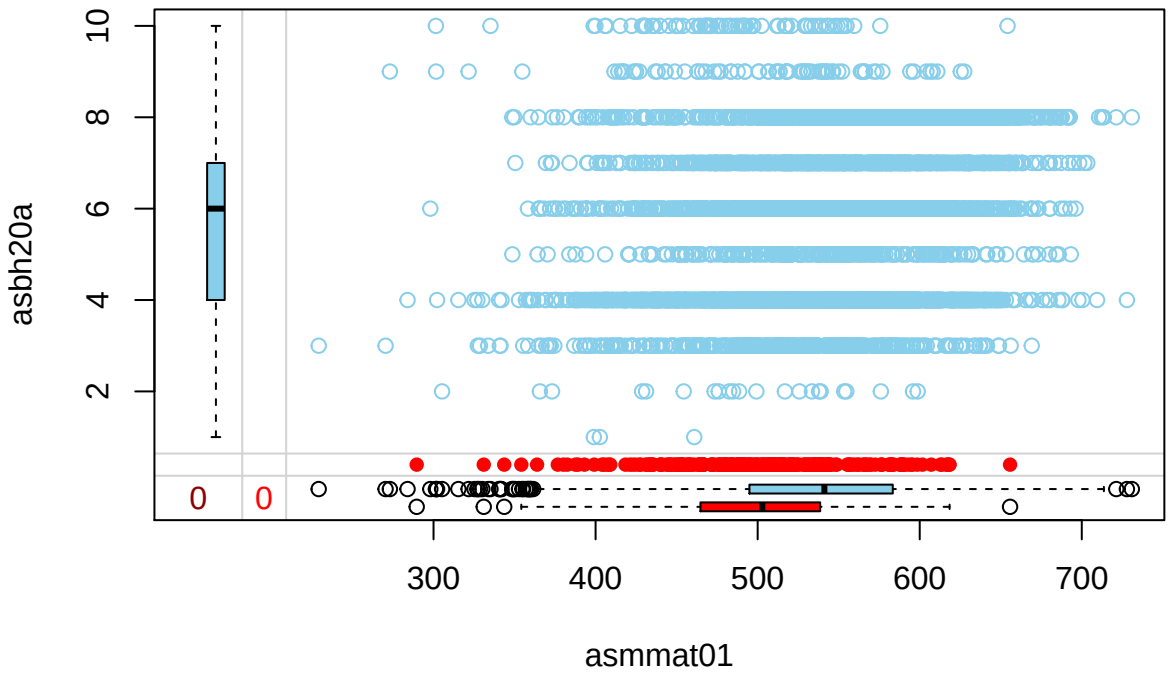
```
## VIM is ready to use.
##  Since version 4.0.0 the GUI is in its own package VIMGUI.
##
##           Please use the package to use the new (and old) GUI.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues
```

```
##
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
marginplot(fin4[,c("asmmat01", "asbh20a")])
```



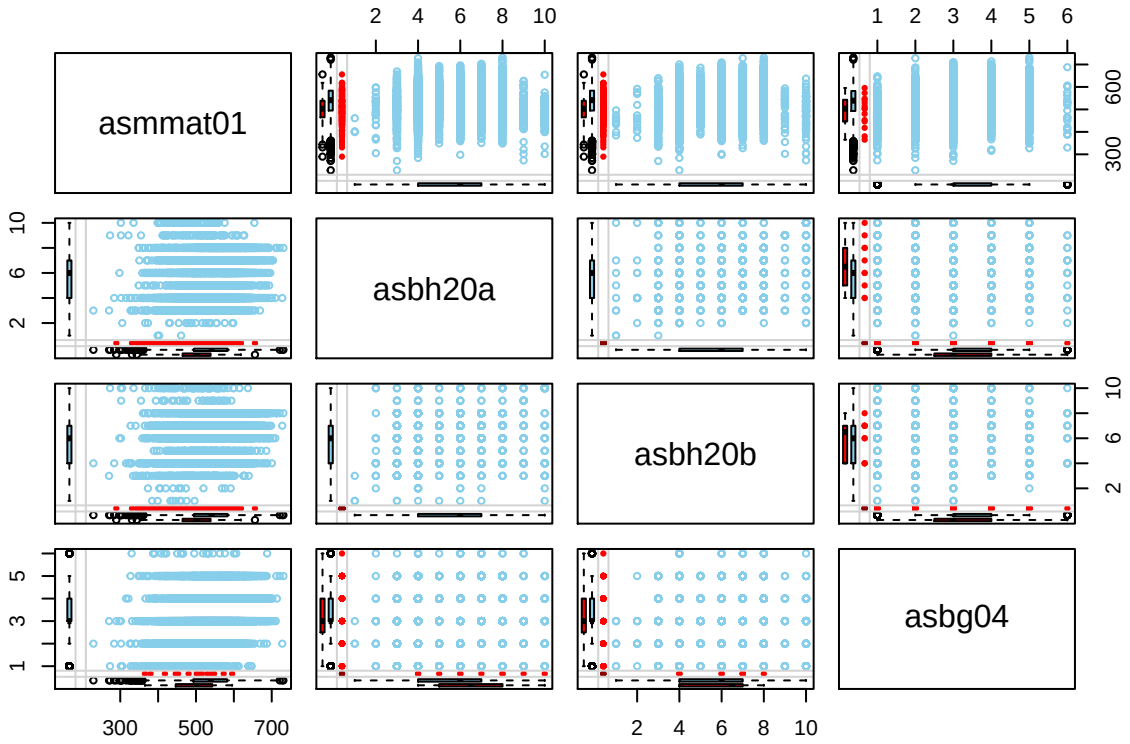
The scatterplot displayed in the call to `marginplot` shows the distribution of values containing non-missing values for each variable. The margins of the plot dislays several mesasure of missingness within and between the two selected variables. The bottom left corner of the figure includes a matrix of missingness. In this example, the left portion of the matrix shows 346 missing values for the variable `"asbh20a"`, while the right portion of the matrix shows 0 missing values for `"asmmat01"`, and 0 instances of missing data values that overlap.

The left and bottom margins of the figure display distributions of missingness for each variable as they overlap with the other. The bottom margin in the figure shows two box and whisker plots of the variable `"asmmat01"` to compare the distributions where the variable `"asbh20a"` is missing (the red box and whisker) and non-missing (the blue box and whisker)

Missingness and the distribution of multiple variables

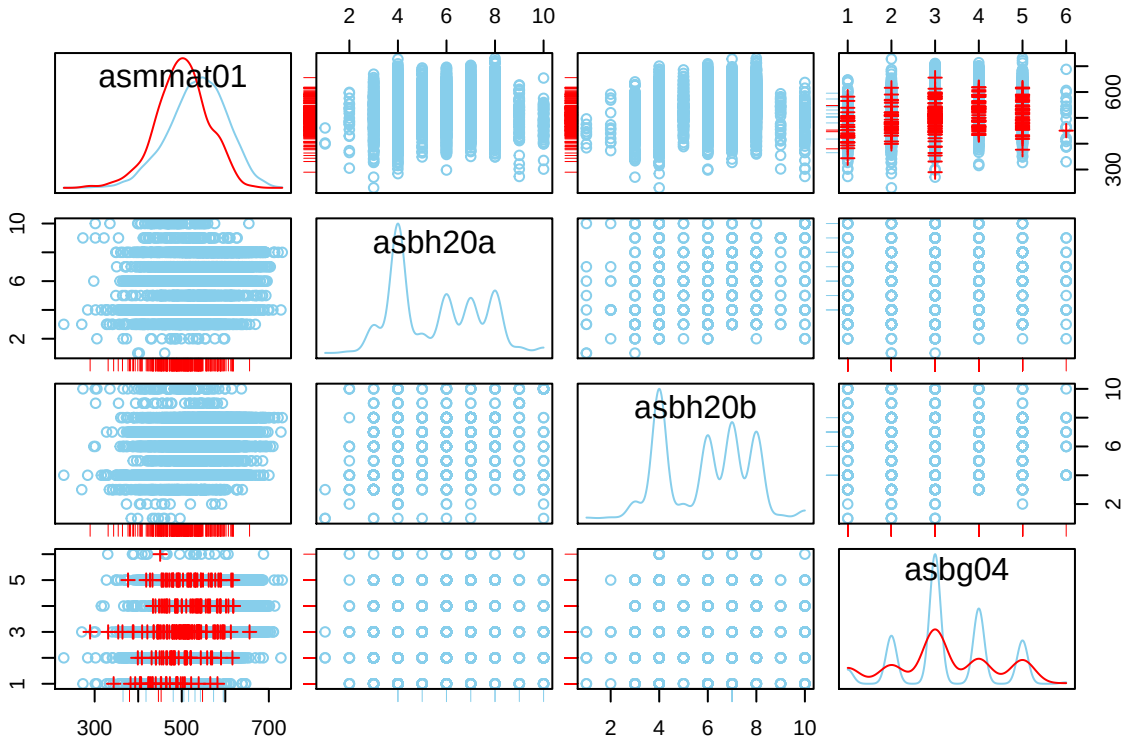
Building upon the comparisons in the `marginplot` function, the missingness distributions of multiple variables can be explored with `marginmatrix`.

```
marginmatrix(fin4[,c("asmmat01", "asbh20a", "asbh20b","asbg04")])
```



We can also look at this another way with `scattmatrixMiss`. Whereas `marginmatrix` focuses patterns of missingness in the margins, `scattmatrixMiss` highlights missing data directly in the plots.

```
scattmatrixMiss(fin4[,c("asmmat01", "asbh20a", "asbh20b","asbg04")],
                highlight=c("asbh20a", "asbh20b"))
```

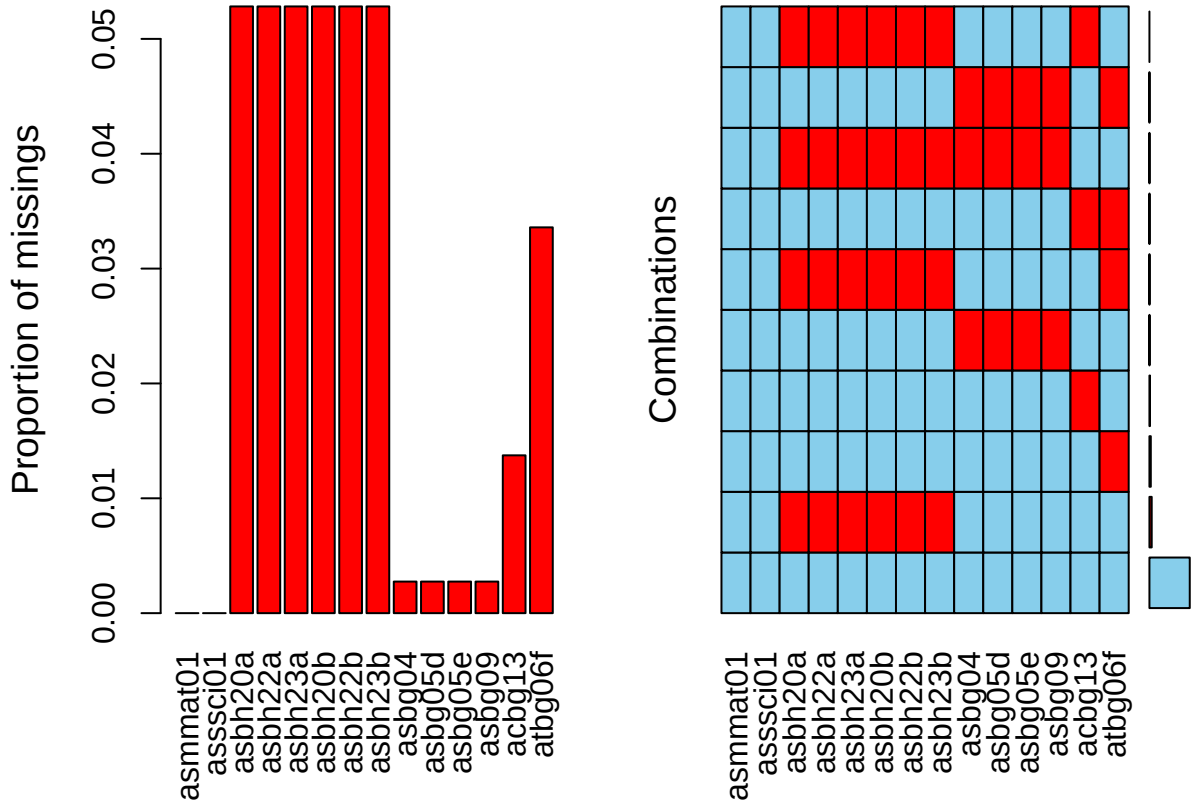


The argument `highlight` is used to displays the missing values for a selected vector of variables, in this example `"asbh20a"` and `"asbh20b"`. Blue lines and dashes through each panel display data where data is not missing and red lines and dashes denote missing data for the highlighted variables. Note that the top left panel indicates that mathematics scores are lower for respondents where data is missing with respect to the mothers and fathers education level variables.

Finding Variables to Impute

To impute data it is necessary to find variables with a sufficient number of cases with non-missing data related to the variables of interest. After exploring the patterns of missingness with `marginplot`, `marginmatrix`, and `scattmatrixMiss`, the function `aggr` assists in finding variables that can be imputed.

```
a <- aggr(fin4[,vars])
```



The histogram on the left panel the proportion of missineness by variable. The right panel displays a histogram of the count of each unique combination of variable missingness. As indicated bt the tall blue bar, the overwhelmingly most common combination of responses is all non-missing values, while the second most common combination of responses are all non-missing values except for the variables describing a mother and father's education level, employment status, and main job.

In addition to the visualizations, using `summary` on the saved object returned via `aggr` returns a brief statement on the missingness of the selected variables:

```
summary(a)
```

```
##
##  Missings per variable:
##  Variable Count
##  asmmat01     0
##  asssci01     0
##   asbh20a   346
##   asbh22a   346
##   asbh23a   346
##   asbh20b   346
##   asbh22b   346
##   asbh23b   346
##    asbg04    18
##   asbg05d    18
##   asbg05e    18
##    asbg09    18
##    acbg13    90
```

```
##   atbg06f   220
##
##  Missings in combinations of variables:
##                 Combinations Count     Percent
##  0:0:0:0:0:0:0:0:0:0:0:0:0:0  5901 90.09160305
##  0:0:0:0:0:0:0:0:0:0:0:0:0:1   200  3.05343511
##  0:0:0:0:0:0:0:0:0:0:0:0:1:0    81  1.23664122
##  0:0:0:0:0:0:0:0:0:0:0:0:1:1     8  0.12213740
##  0:0:0:0:0:0:0:0:1:1:1:1:0:0    11  0.16793893
##  0:0:0:0:0:0:0:0:1:1:1:1:0:1     3  0.04580153
##  0:0:1:1:1:1:1:1:0:0:0:0:0:0   332  5.06870229
##  0:0:1:1:1:1:1:1:0:0:0:0:0:1     9  0.13740458
##  0:0:1:1:1:1:1:1:0:0:0:0:1:0     1  0.01526718
##  0:0:1:1:1:1:1:1:1:1:1:1:0:0     4  0.06106870
```

Multiple Imputation

To impute data you need the relevant variables (show example implied by final example from previous section, read mice documentation for how to do this).

methods of imputation

There are many imputation schemes, we highlight the mechanism behind two of them in this section: regression with residual, and predicted mean matching.

regression

show equation, regression picture, explain why you need residual

predicted mean matching

show figure of PMM, explain benefit in terms of support.

imputation with plausible values

We will follow LSS book and impute M' values per plausible value.

methods of analysis

I don't know how analysis works for MI data, so we'll have to fill that in.