Basic Concepts in Missing Data

William Murrah

A great book to learn about dealing missing data an how to deal with it is *Applied Missing Data Analysis* (Enders 2010), which covers full information maximum likelihood and multiple imputation. Much of the material covered in this repository and accompanying tutorials is based on the book *Flexible Imputation of Missing Data* (Van Buuren 2012), particularly chapters 4-6.

Packages we will use:

```
require(texreg)
require(mice)
require(VIM)
```

Notation we will use:

- n = number of units (number of cases or people) indexed by i.
- p = number of variables (including outcome and predictors), indexed by j.
- $Y = n \times p$ matrix containing the data values for p variables for n units in the sample.
- R = response indicator, a $n \times p$ matrix with each cell containing either a `0` or a `1`, where

$$r_{ij} = \begin{cases} 1 \text{ if } y_{ij} \text{ is observed and,} \\ 0 \text{ if } y_{ij} \text{ is missing.} \end{cases} \tag{1}$$

- Y_{obs} = the observed data, collectively (i.e. contains all elements y_{ij} where $r_{ij} = 1$).
- Y_{mis} = the missing data, collectively (i.e. contains all elements y_{ij} where $r_{ij} = 0$).

A simple data frame with some missing data:

```
pattern1
```

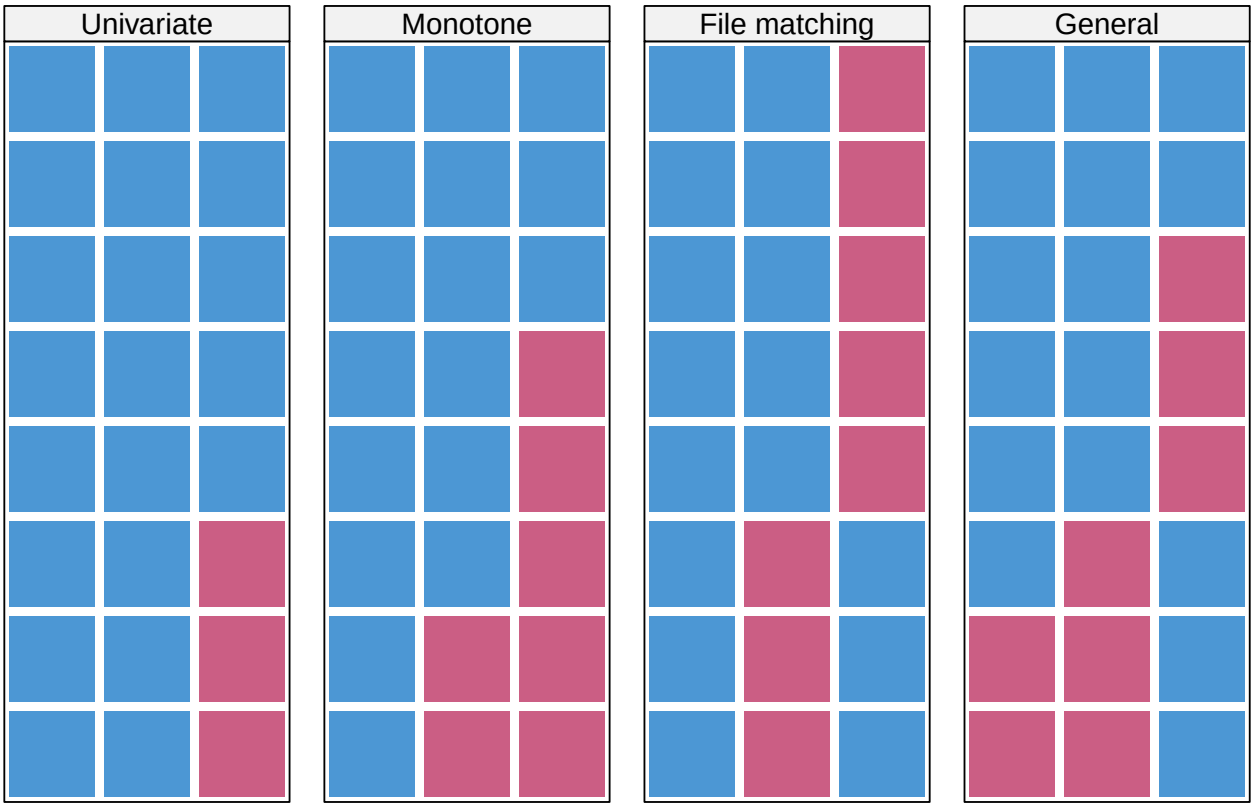
```
   A   B  C
1 12  31 24
2 63  51 71
3 61  19 31
4 63  76 51
5 87  21  6
6 65  26 NA
7  1 100 NA
8 24  81 NA
```

Create an R matrix

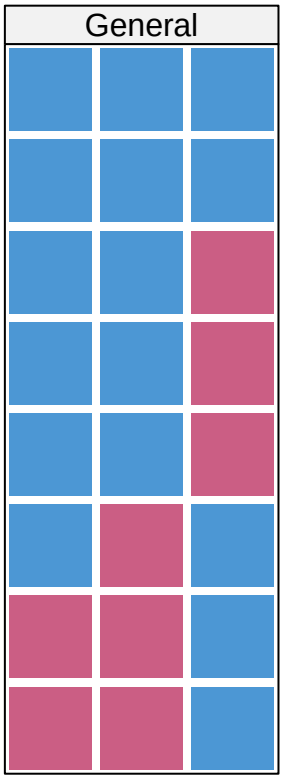
```
1 - is.na(pattern1)
```

```
  A B C
1 1 1 1
2 1 1 1
3 1 1 1
4 1 1 1
5 1 1 1
6 1 1 0
7 1 1 0
8 1 1 0
```

Missing Data Patterns




```
print(tp41[4])
```




```r
1 - is.na(pattern4)
```

```
   A B C
25 1 1 1
26 1 1 1
27 1 1 0
28 1 1 0
29 1 1 0
30 1 0 1
31 0 0 1
32 0 0 1
```

```r
md.pattern(pattern4)
```

```
  A B C
2 1 1 1 0
1 1 0 1 1
3 1 1 0 1
2 0 0 1 2
  2 3 3 8
```

Session Information

```
R version 3.2.2 (2015-08-14)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 14.04.3 LTS

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] grid      stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] VIM_4.4.1        data.table_1.9.6 colorspace_1.2-6 mice_2.22
 [5] Rcpp_0.12.1      texreg_1.35      knitr_1.11       mosaic_0.11
 [9] mosaicData_0.9.1 car_2.1-0        ggplot2_1.0.1    lattice_0.20-33
[13] dplyr_0.4.3

loaded via a namespace (and not attached):
 [1] zoo_1.7-12          reshape2_1.4.1      splines_3.2.2
 [4] htmltools_0.2.6     yaml_2.1.13         mgcv_1.8-7
 [7] chron_2.3-47        e1071_1.6-7         nloptr_1.0.4
[10] DBI_0.3.1           sp_1.2-0            plyr_1.8.3
[13] robustbase_0.92-5   stringr_1.0.0       MatrixModels_0.4-1
[16] munsell_0.4.2       gtable_0.1.2        evaluate_0.8
[19] SparseM_1.7         lmtest_0.9-34       quantreg_5.19
[22] pbkrtest_0.4-2      parallel_3.2.2      class_7.3-14
[25] vcd_1.4-1           DEoptimR_1.0-3      proto_0.3-10
[28] scales_0.3.0        formatR_1.2.1       lme4_1.1-10
[31] gridExtra_2.0.0     digest_0.6.8        stringi_0.5-5
[34] tools_3.2.2         magrittr_1.5        randomForest_4.6-12
[37] ggdendro_0.1-17     MASS_7.3-44         Matrix_1.2-2
[40] assertthat_0.1      minqa_1.2.4         rmarkdown_0.8.1
[43] R6_2.1.1            rpart_4.1-10        nnet_7.3-11
[46] nlme_3.1-122
```

References

Enders, Craig K. 2010. *Applied Missing Data Analysis*. Guilford Publications.

Van Buuren, Stef. 2012. *Flexible Imputation of Missing Data*. CRC press.