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)
```

```
doctype <-
   "pdf"
  # "html"
oreg <- ifelse(doctype == "pdf", texreg, screenreg)
```

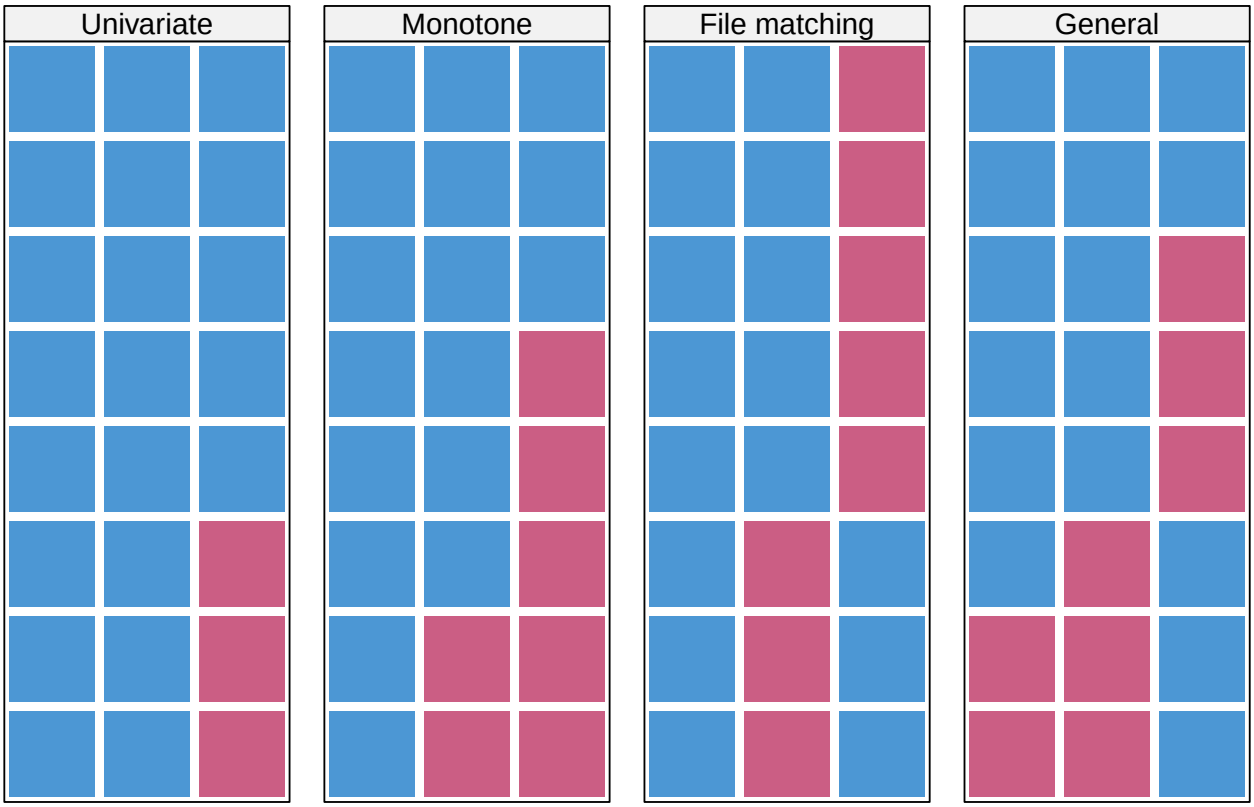
Notation we will use:

- 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

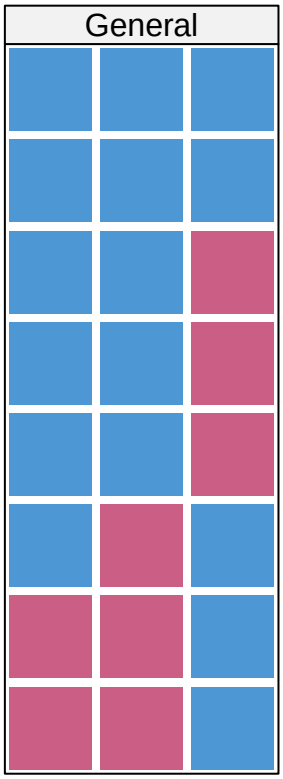
$$r_{ij} = \begin{cases} 1 \text{ if } y_{ij} \text{ is observed and,} \\ 0 \text{ if } y_{ij} \text{ is missing.} \end{cases} \tag{1}$$

- 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 $r_{ij} = 0$).

Missing Data Patterns




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



A simple data frame with some missing data:

```
general
```

```
   A  B  C
1 22 50 17
2 82 76 90
3 53 18 NA
4 92 85 NA
5 84 87 NA
6  5 NA 14
7 NA NA 11
8 NA NA 52
```

Create an R matrix

```
R <- 1 - is.na(general)
R
```

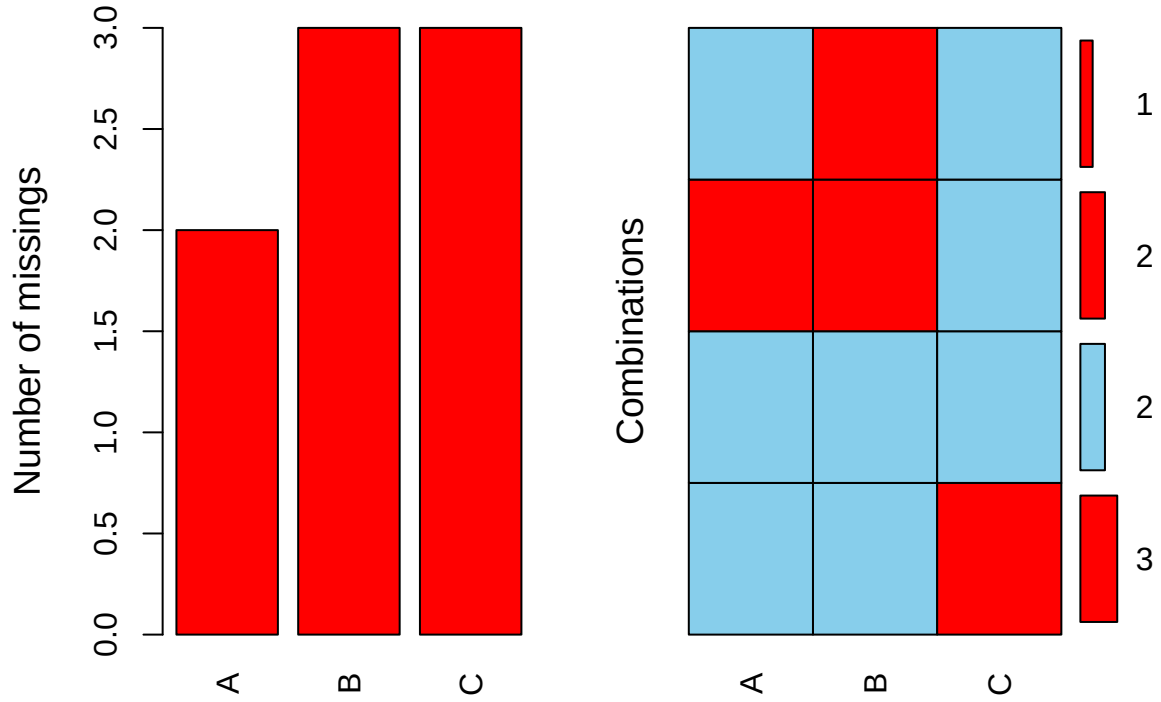
```
  A B C
1 1 1 1
2 1 1 1
3 1 1 0
4 1 1 0
```

```
5 1 1 0
6 1 0 1
7 0 0 1
8 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
```

```r
aggr(general, numbers = TRUE, prop = FALSE)
```




```r
# matrixplot(general, sortby = "A")
```

	pmA	pmB	pmC
(Intercept)	−23.57	2.47	−3.25
	(278033.04)	(70269.47)	(8.63)
B	0.00		−0.08
	(4322.16)		(0.11)
A		−3.24	0.13
		(7895.50)	(0.18)
C		2.67	
		(7420.88)	
AIC	4.00	6.00	10.34
BIC	1.39	3.30	9.17
Log Likelihood	-0.00	-0.00	-2.17
Deviance	0.00	0.00	4.34
Num. obs.	2	3	5

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 1: Statistical models

Predictors of missingness

```
pmA <- glm(is.na(A) ~ B + C, family = binomial, data = general)
pmB <- glm(is.na(B) ~ A + C, family = binomial, data = general)
pmC <- glm(is.na(C) ~ A + B, family = binomial, data = general)
oreg(list(pmA, pmB, pmC), custom.model.names = c("pmA", "pmB", "pmC"))
```

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_0.12
 [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.