Missing Data Analysis

William Murrah

Packages we will use:

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

```
doctype <-
   "latex"
   # "html"
txreg <- ifelse(doctype == "latex", texreg, htmlreg)
options(ztable.type = doctype)
```

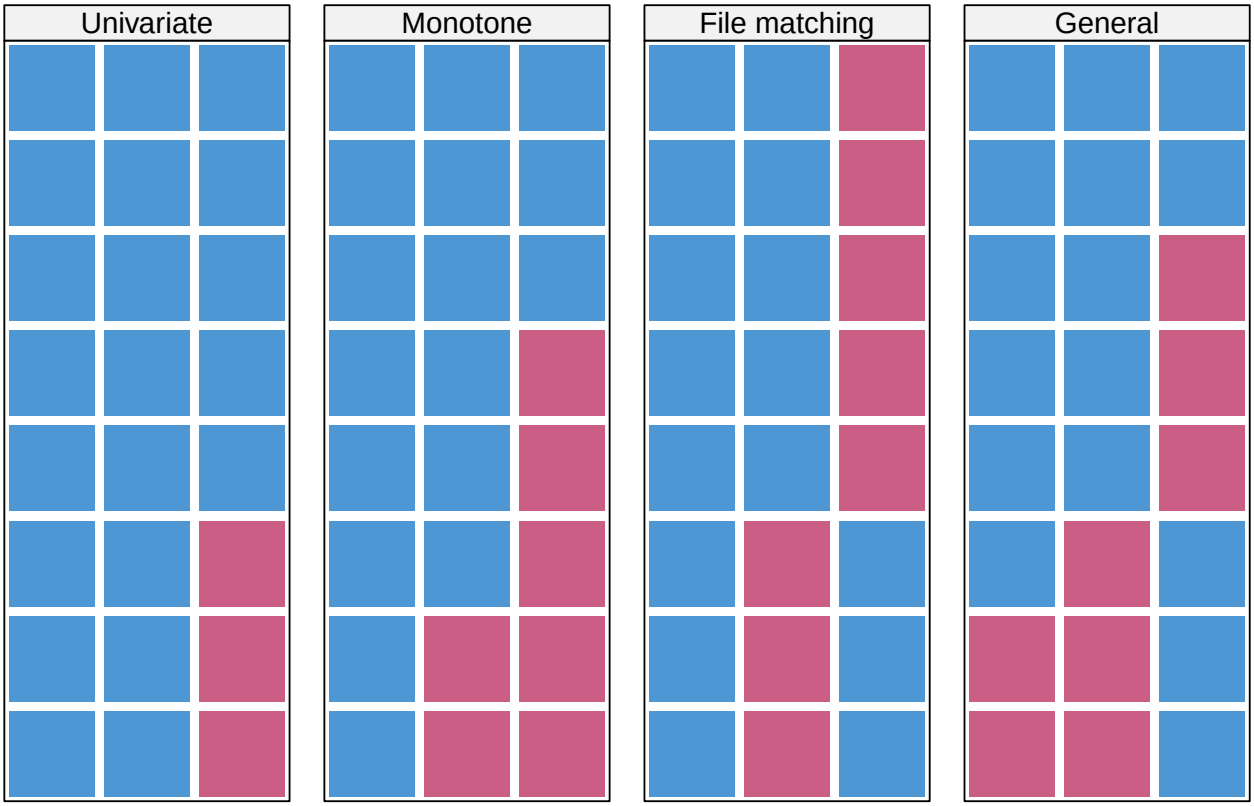
Notation we will use:

- 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 on n units for p variables in the sample.
- R = response indicator, a $n \times p$ matrix with each cell containing either a `0` or a `1`, where

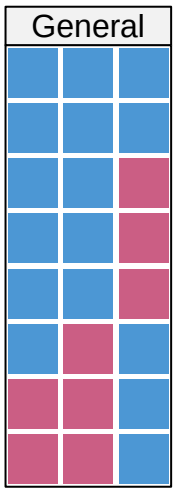
$$r_{ij} = \begin{cases} 1 \text{ if } y_{ij} \text{ is observed and,} \\ 0 \text{ if } y_{ij} \text{ is missing.} \end{cases} \tag{1}$$

- 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

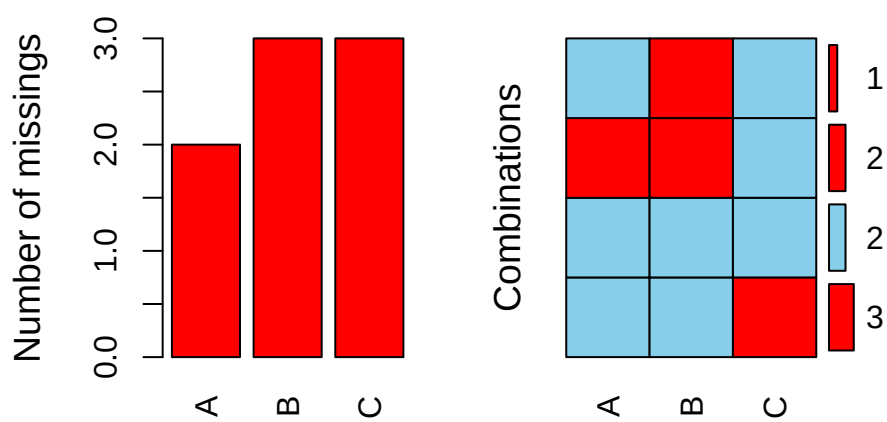
```r
R <- 1 - is.na(general)
R
```

```
  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)
```



Proportion of usable cases

Imputing Y_j from Y_k, the *proportion of usable cases* is the number of cases missing in Y_j that are observed in Y_k divided by the number of missing cases in Y_j.

```
p <- md.pairs(general)
p$mr/(p$mr + p$mm) # proportion of usable cases.
```

```
          A B C
A 0.0000000 0 1
B 0.3333333 0 1
C 1.0000000 1 0
```

Influx and Outflux

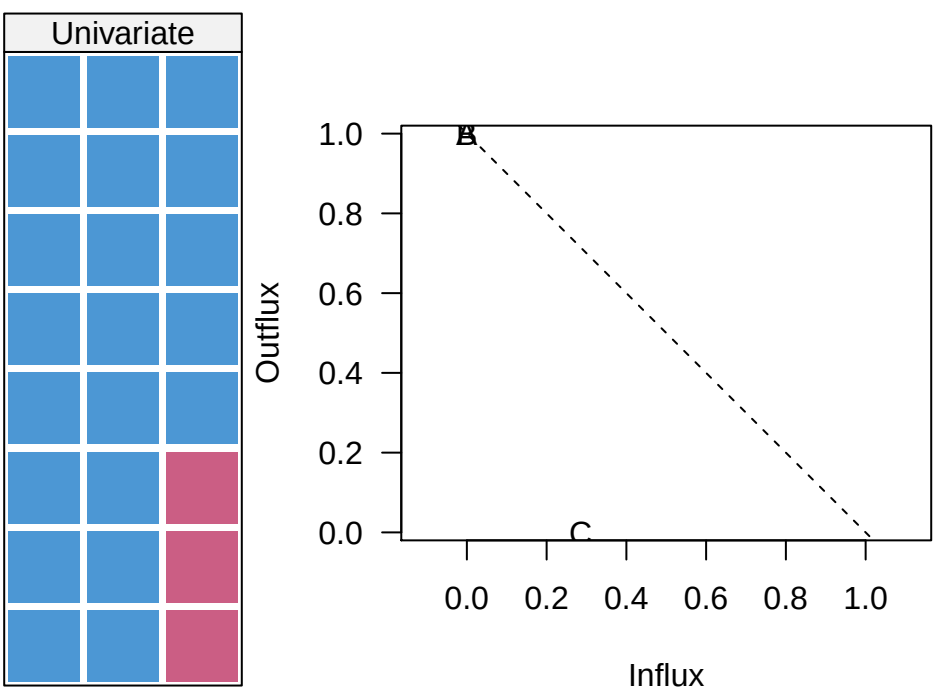
Influx - I_j the *influx coefficient* is how well the other variables connect to Y_j.

- I_j is `0` for a completely observed variable
- I_j is `1` for a completely missing variable

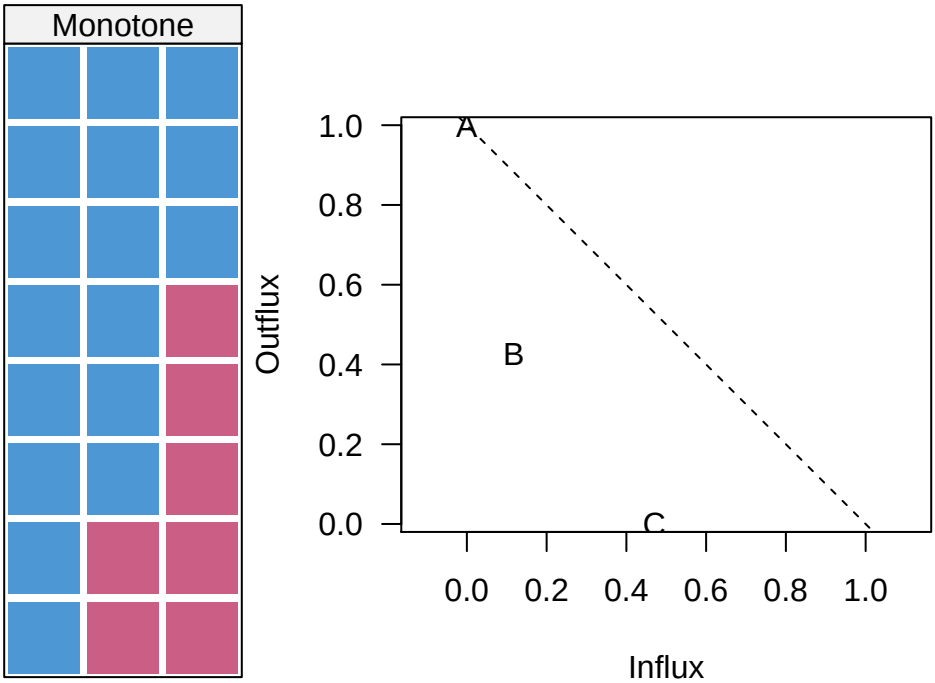
Outflux - O_j the *outflux coefficient* is how well Y_j is connected to other variables.

- O_j is `1` for a completely observed variable
- O_j is `0` for a completely missing variable

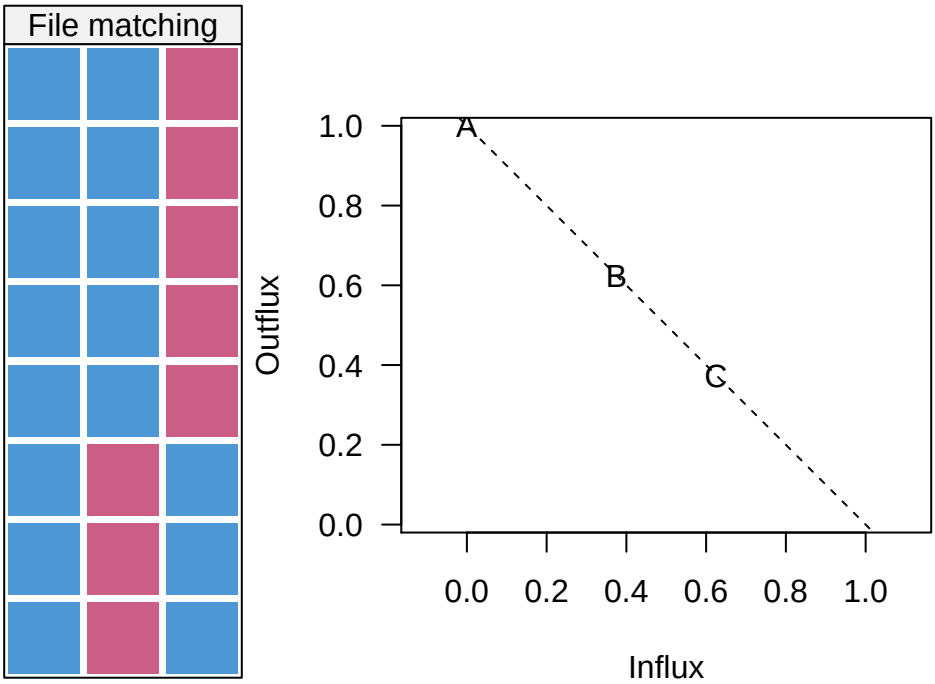
```
print(tp41[1])
fluxplot(pattern1, main = NULL)
```



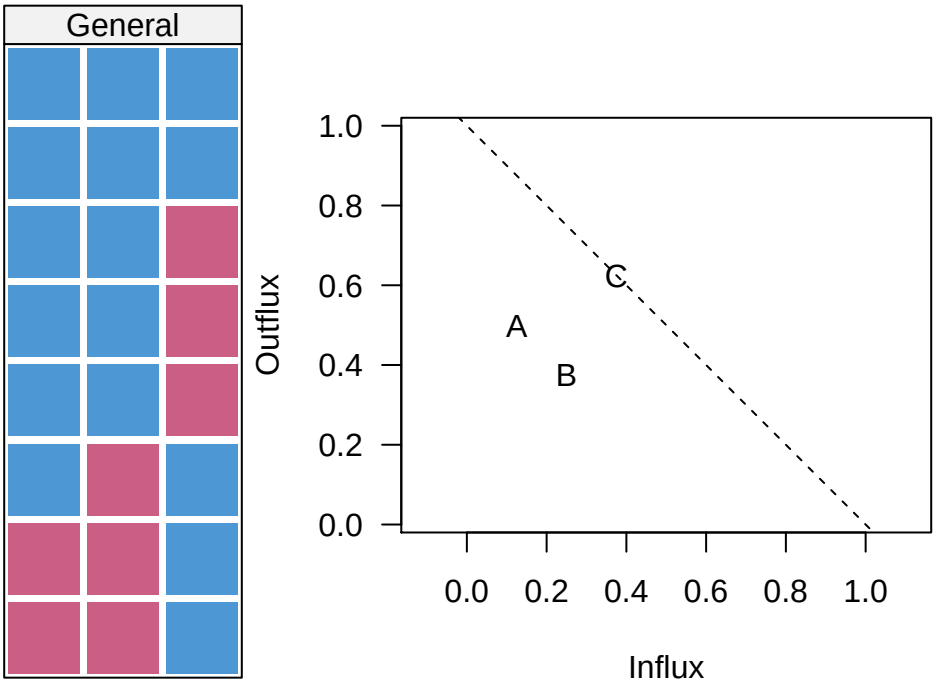
```r
print(tp41[2])
fluxplot(pattern2, main = NULL)
```




```r
print(tp41[3])
fluxplot(pattern3, main = NULL)
```




```r
print(tp41[4])
fluxplot(general, main = NULL)
```


General
Outflux
Influx
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C

Missing Data Analysis

Data we will use:

```r
#  ?nhanes
data("nhanes2")
ztable(nhanes)
```

	age	bmi	hyp	chl
1	1.00			
2	2.00	22.70	1.00	187.00
3	1.00		1.00	187.00
4	3.00			
5	1.00	20.40	1.00	113.00
6	3.00			184.00
7	1.00	22.50	1.00	118.00
8	1.00	30.10	1.00	187.00
9	2.00	22.00	1.00	238.00
10	2.00			
11	1.00			
12	2.00			
13	3.00	21.70	1.00	206.00
14	2.00	28.70	2.00	204.00
15	1.00	29.60	1.00	
16	1.00			
17	3.00	27.20	2.00	284.00
18	2.00	26.30	2.00	199.00
19	1.00	35.30	1.00	218.00
20	3.00	25.50	2.00	
21	1.00			
22	1.00	33.20	1.00	229.00
23	1.00	27.50	1.00	131.00
24	3.00	24.90	1.00	
25	2.00	27.40	1.00	186.00

Model

$$\text{chl}_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{bmi}_i + \beta_3 \text{hyp}_i + \epsilon_i \tag{2}$$

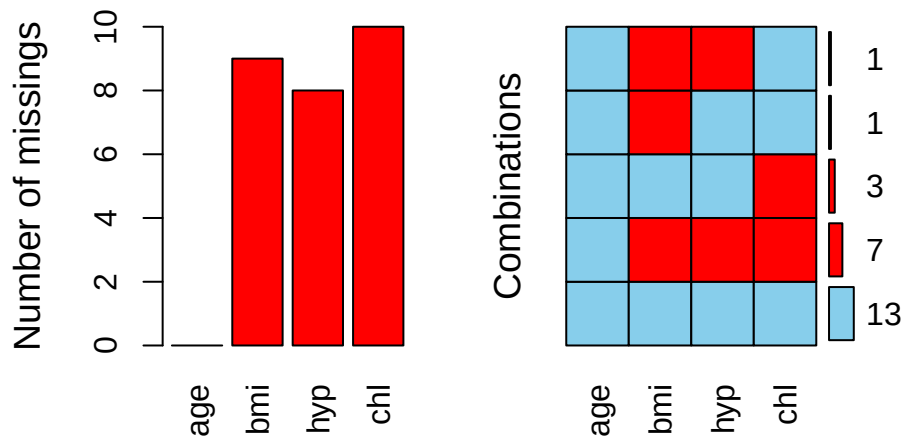
Missing Data Patterns

```r
md.pattern(nhanes)
```

```
   age hyp bmi chl
13   1   1   1   1  0
 1   1   1   0   1  1
 3   1   1   1   0  1
```

```
 1    1    0    0    1   2
 7    1    0    0    0   3
      0    8    9   10  27
```

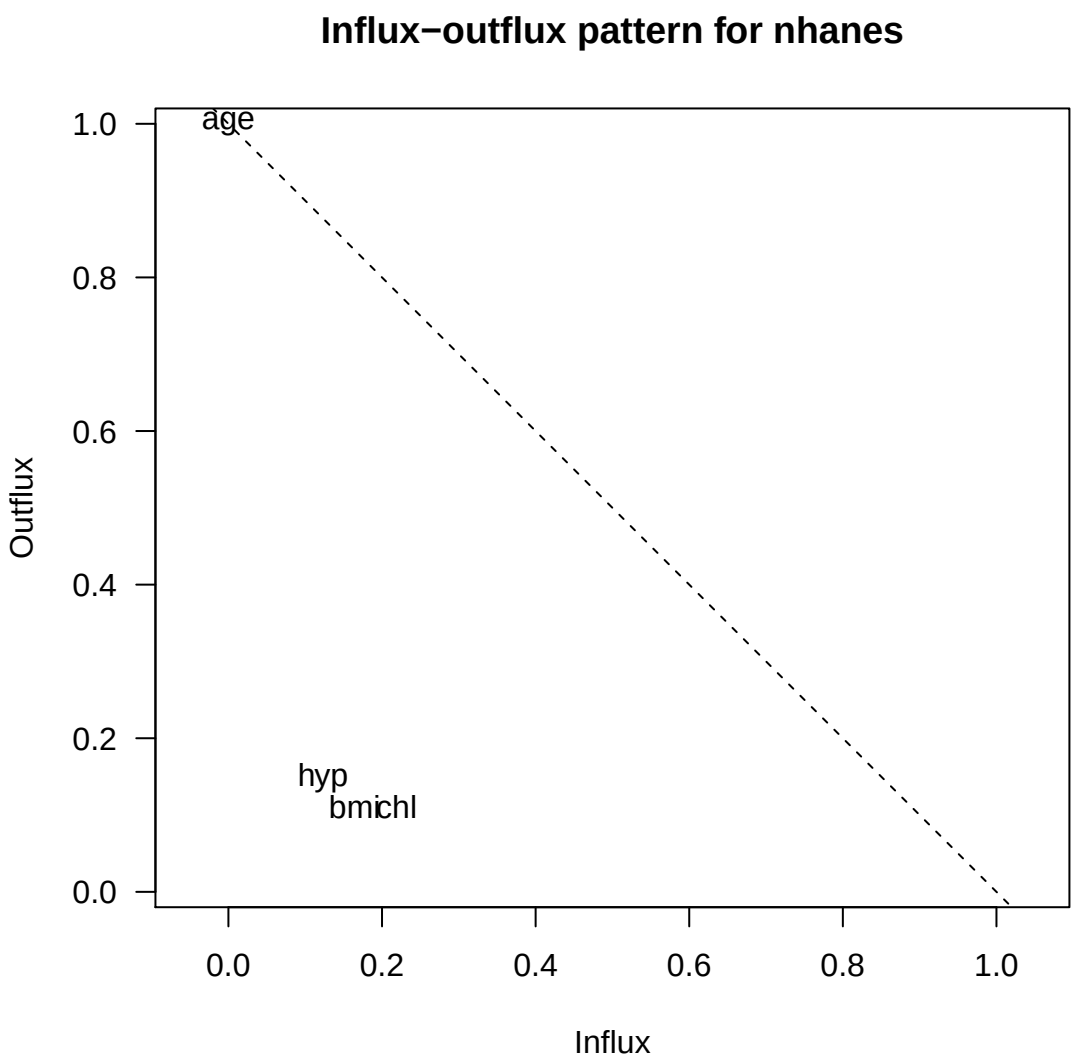
```r
aggr(nhanes, numbers = TRUE, prop = FALSE)
```




```r
p <- md.pairs(nhanes)
p$mr/(p$mr + p$mm)
```

```
    age bmi       hyp       chl
age NaN NaN       NaN       NaN
bmi   1 0.0 0.1111111 0.2222222
hyp   1 0.0 0.0000000 0.1250000
chl   1 0.3 0.3000000 0.0000000
```

```r
fluxplot(nhanes)
```



Predictors of missingness

```
round(x = cor(x = nhanes,
              use = "pairwise.complete.obs"),
      digits = 2)
```

```
      age   bmi  hyp  chl
age  1.00 -0.37 0.51 0.51
bmi -0.37  1.00 0.05 0.37
hyp  0.51  0.05 1.00 0.43
chl  0.51  0.37 0.43 1.00
```

```
pmA <- glm(is.na(chl) ~ age + bmi + hyp, family = binomial, data = nhanes)
pmB <- glm(is.na(bmi) ~ age + chl + hyp, family = binomial, data = nhanes)
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

	chl	bmi	hyp
(Intercept)	−9.18	14.75	−24.57
	(8.61)	(9598.75)	(303112.73)
age	1.72	−18.21	0.00
	(1.36)	(6513.98)	(95787.74)
bmi	0.20		0.00
	(0.26)		(13515.52)
hyp	−0.90	−0.31	
	(1.76)	(11600.34)	
chl		0.01	−0.00
		(0.03)	(1497.48)
AIC	21.19	13.55	8.00
BIC	24.28	16.10	10.26
Log Likelihood	-6.60	-2.77	-0.00
Deviance	13.19	5.55	0.00
Num. obs.	16	14	13

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

Table 1: Statistical models

```r
pmC <- glm(is.na(hyp) ~ age + bmi + chl, family = binomial, data = nhanes)
txreg(list(pmA, pmB, pmC), custom.model.names = c("chl", "bmi", "hyp"))
```

```
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] ztable_0.1.5     VIM_4.4.1        data.table_1.9.6 colorspace_1.2-6
 [5] mice_2.22        Rcpp_0.12.1      texreg_1.35      knitr_1.11
 [9] mosaic_0.12      mosaicData_0.9.1 car_2.1-0        ggplot2_1.0.1
[13] lattice_0.20-33  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
```