Missing Data

EPSY 887: Data Science Institute

Jason Bryer

https://github.com/jbryer/EPSY887DataScience
jason@bryer.org

September 30, 2014

Agenda

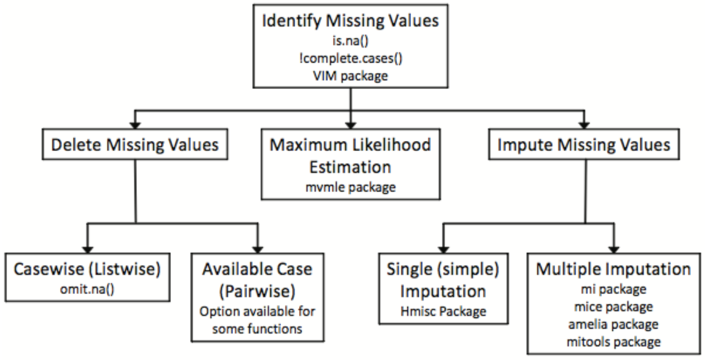
1. Evaluating the Nature of Missiness

2. Imputing Missing Values

Classifications of Missing Data

MCAR	*Missing completely at random*
MAR	*Missing at random*
NMAR	*Not missing at random*

Methods for Handling Missing Data[1]



Identify Missing Values
is.na()
!complete.cases()
VIM package

Delete Missing Values

Maximum Likelihood Estimation
mvmle package

Impute Missing Values

Casewise (Listwise)
omit.na()

Available Case (Pairwise)
Option available for some functions

Single (simple) Imputation
Hmisc Package

Multiple Imputation
mi package
mice package
amelia package
mitools package

[1] Kabacoff, 2011, p. 355

Mammal Sleep Data

```
> data(sleep)
> str(sleep)
```

```
'data.frame': 62 obs. of  10 variables:
 $ BodyWgt : num  6654 1 3.38 0.92 2547 ...
 $ BrainWgt: num  5712 6.6 44.5 5.7 4603 ...
 $ NonD    : num  NA 6.3 NA NA 2.1 9.1 15.8 5.2 10.9 8.3 ...
 $ Dream   : num  NA 2 NA NA 1.8 0.7 3.9 1 3.6 1.4 ...
 $ Sleep   : num  3.3 8.3 12.5 16.5 3.9 9.8 19.7 6.2 14.5 9.7 ...
 $ Span    : num  38.6 4.5 14 NA 69 27 19 30.4 28 50 ...
 $ Gest    : num  645 42 60 25 624 180 35 392 63 230 ...
 $ Pred    : int  3 3 1 5 3 4 1 4 1 1 ...
 $ Exp     : int  5 1 1 2 5 4 1 5 2 1 ...
 $ Danger  : int  3 3 1 3 4 4 1 4 1 1 ...
```

Complete Cases

```
> complete.cases(sleep)

 [1] FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TF
[12]  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE  TF
[23]  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TF
[34]  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TF
[45]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FAI
[56] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE

> head(sleep[complete.cases(sleep),])

  BodyWgt BrainWgt NonD Dream Sleep Span Gest Pred Exp Danger
2 1.0e+00      6.6  6.3   2.0   8.3  4.5   42    3   1      3
5 2.5e+03   4603.0  2.1   1.8   3.9 69.0  624    3   5      4
6 1.1e+01    179.5  9.1   0.7   9.8 27.0  180    4   4      4
7 2.3e-02      0.3 15.8   3.9  19.7 19.0   35    1   1      1
8 1.6e+02    169.0  5.2   1.0   6.2 30.4  392    4   5      4
9 3.3e+00     25.6 10.9   3.6  14.5 28.0   63    1   2      1
```

Incomplete Cases

```
> head(sleep[!complete.cases(sleep),])
   BodyWgt BrainWgt NonD Dream Sleep Span Gest Pred Exp Danger
1  6654.00   5712.0   NA    NA   3.3   39  645    3   5      3
3     3.38     44.5   NA    NA  12.5   14   60    1   1      1
4     0.92      5.7   NA    NA  16.5   NA   25    5   2      3
13    0.55      2.4  7.6   2.7  10.3   NA   NA    2   1      2
14  187.10    419.0   NA    NA   3.1   40  365    5   5      5
19    1.41     17.5  4.8   1.3   6.1   34   NA    1   2      1
```

How much is missing?

Number of missing values

```
> sum(is.na(sleep$Dream))
```

```
[1] 12
```

Percent missing

```
> mean(is.na(sleep$Dream))
```

```
[1] 0.19
```

Percent of rows with missing data

```
> mean(!complete.cases(sleep))
```

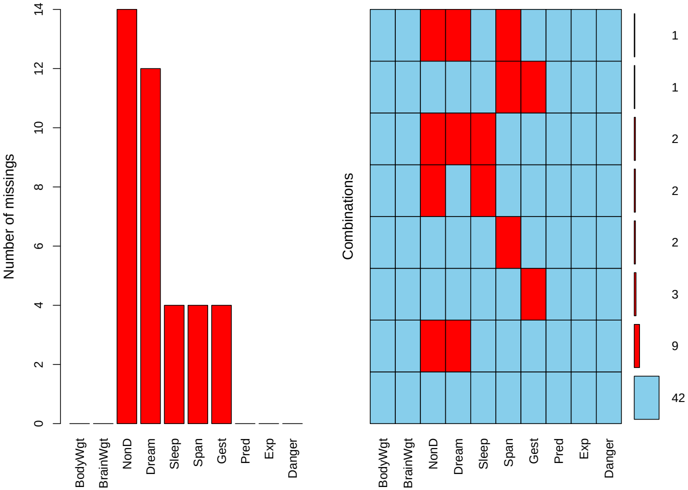
```
[1] 0.32
```

Pattern of Missingness

```
> md.pattern(sleep)
   BodyWgt BrainWgt Pred Exp Danger Sleep Span Gest Dream NonD   
42       1        1    1   1      1     1    1    1     1    1  0
 2       1        1    1   1      1     1    0    1     1    1  1
 3       1        1    1   1      1     1    1    0     1    1  1
 9       1        1    1   1      1     1    1    1     0    0  2
 2       1        1    1   1      1     0    1    1     1    0  2
 1       1        1    1   1      1     1    0    0     1    1  2
 2       1        1    1   1      1     0    1    1     0    0  3
 1       1        1    1   1      1     1    0    1     0    0  3
         0        0    0   0      0     4    4    4    12   14 38
```

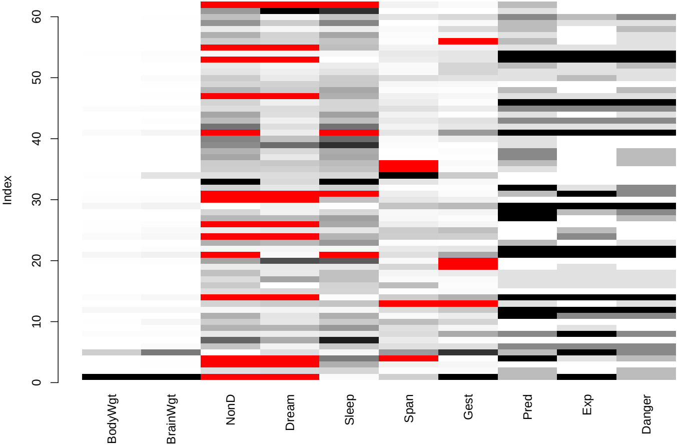
Visualizing Missingness

```
> aggr(sleep, prop=FALSE, numbers=TRUE)
```



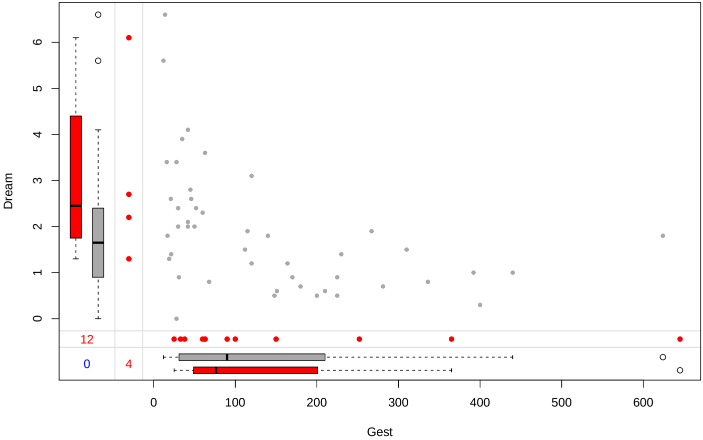
Visualizing Missingness

```
> matrixplot(sleep)
```



Visualizing Missingness

```
> marginplot(sleep[,c('Gest','Dream')], pch=c(20), col=c('darkgray',
```



Shadow Matrix

```
> sm <- as.data.frame(abs(is.na(sleep)))
> head(sleep)
```

```
  BodyWgt BrainWgt NonD Dream Sleep Span Gest Pred Exp Danger
1 6654.00   5712.0   NA    NA   3.3 38.6  645    3   5      3
2    1.00      6.6  6.3   2.0   8.3  4.5   42    3   1      3
3    3.38     44.5   NA    NA  12.5 14.0   60    1   1      1
4    0.92      5.7   NA    NA  16.5   NA   25    5   2      3
5 2547.00   4603.0  2.1   1.8   3.9 69.0  624    3   5      4
6   10.55    179.5  9.1   0.7   9.8 27.0  180    4   4      4
```

```
> head(sm)
```

```
  BodyWgt BrainWgt NonD Dream Sleep Span Gest Pred Exp Danger
1       0        0    1     1     0    0    0    0   0      0
2       0        0    0     0     0    0    0    0   0      0
3       0        0    1     1     0    0    0    0   0      0
4       0        0    1     1     0    1    0    0   0      0
5       0        0    0     0     0    0    0    0   0      0
6       0        0    0     0     0    0    0    0   0      0
```

Correlation of Missingness

Examine the correlation of missingness between variables

```
> #Extract varabibles that have some missingness
> y <- sm[which(sapply(sm, sd) > 0)]
> cor(y)

        NonD  Dream  Sleep   Span   Gest
NonD   1.000  0.907  0.486  0.015 -0.142
Dream  0.907  1.000  0.204  0.038 -0.129
Sleep  0.486  0.204  1.000 -0.069 -0.069
Span   0.015  0.038 -0.069  1.000  0.198
Gest  -0.142 -0.129 -0.069  0.198  1.000
```

Relationship between missingness and observed variables

```
> cor(sleep, y, use='pairwise.complete.obs')
          NonD  Dream  Sleep   Span   Gest
BodyWgt  0.227  0.223 0.0017 -0.058 -0.054
BrainWgt 0.179  0.163 0.0079 -0.079 -0.073
NonD        NA     NA     NA -0.043 -0.046
Dream   -0.189     NA -0.1890  0.117  0.228
Sleep   -0.080 -0.080     NA  0.096  0.040
Span     0.083  0.060 0.0052     NA -0.065
Gest     0.202  0.051 0.1597 -0.175     NA
Pred     0.048 -0.068 0.2025  0.023 -0.201
Exp      0.245  0.127 0.2608 -0.193 -0.193
Danger   0.065 -0.067 0.2089 -0.067 -0.204
```

Rows are observed variables, columns missing indicators. Nondreaming (`NonD`) sleep scores are more likely to be missing with larger body weights (`BodyWgt`) with r=0.227. Since the correlations are not very larger this suggests the nature of the missingness deviates minimally from the MCAR and MAR assumptions.

Understanding missingness

Kabacoff (2011, p. 362) suggests the following questions to address:

- What percentage of the data is missing?
- Is it concentrated in a few variables, or widely distributed?
- Does it appear to be random?
- Does the covariation of missing data with each other or with the observed data suggest a possible mechanism that's producing the missing values.

Options for analyzing data with missing values

- Complete case analysis (listwise deletiong) - Use the `na.omit` function to remove any rows with missing values.
- Pairwise deletion
- Multiple imputation
- Simple imputation - replace values with a single value (e.g. mean, median, mode)

Agenda

1. Evaluating the Nature of Missiness

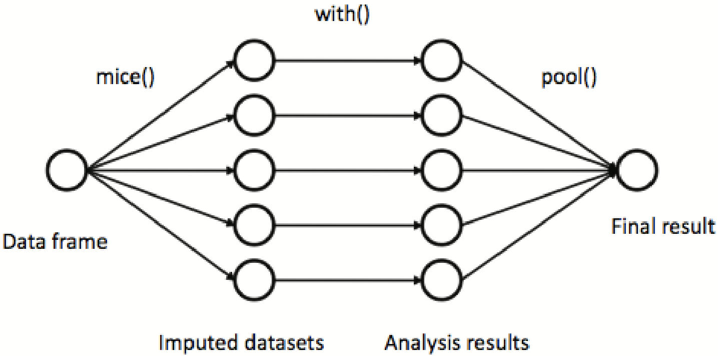
2. Imputing Missing Values

Multiple Imputation

- Imputes (fills in) missing values using repeated simulations.
- Utilizes Gibbs sampling.
- Each variable with missing values is predicted from other variables.
- Multiple complete datasets are created using differing distributions.
- As few as three, but typically five or more imputations are necessary.
- Analysis is conducted separately for each complete dataset and results are pooled.

See volume 45 of the *Journal of Statistical Software* which is a special volume on multiple imputation: `http://www.jstatsoft.org/v45/`.

Steps for Multiple Imputation[2]



[2] Kabacoff, 2011, p. 366

mice

Using the mice package to impute missing values.

```
> imp <- mice(sleep, printFlag=FALSE, seed=1234)
> imp
Multiply imputed data set
Call:
mice(data = sleep, printFlag = FALSE, seed = 1234)
Number of multiple imputations:  5
Missing cells per column:
 BodyWgt BrainWgt     NonD    Dream    Sleep     Span     Gest
       0        0       14       12        4        4        4
    Pred      Exp   Danger
       0        0        0
Imputation methods:
 BodyWgt BrainWgt     NonD    Dream    Sleep     Span     Gest
      ""       ""    "pmm"    "pmm"    "pmm"    "pmm"    "pmm"
    Pred      Exp   Danger
      ""       ""       ""
VisitSequence:
 NonD Dream Sleep  Span  Gest
```

mice

```
> dataset5 <- complete(imp, 5)
> head(dataset5)
```

```
  BodyWgt BrainWgt NonD Dream Sleep Span Gest Pred Exp Danger
1 6654.00   5712.0  3.2   0.3   3.3 38.6  645    3   5      3
2    1.00      6.6  6.3   2.0   8.3  4.5   42    3   1      3
3    3.38     44.5 11.0   1.3  12.5 14.0   60    1   1      1
4    0.92      5.7 12.8   3.4  16.5  4.5   25    5   2      3
5 2547.00   4603.0  2.1   1.8   3.9 69.0  624    3   5      4
6   10.55    179.5  9.1   0.7   9.8 27.0  180    4   4      4
```

mice

```
> fit <- with(imp, lm(Dream ~ Span + Gest))
> pooled <- pool(fit)
> summary(pooled)

                 est     se     t df Pr(>|t|)   lo 95    hi 95 nmis
(Intercept)  2.5462 0.2547 10.00 52  1.0e-13  2.0352  3.05724   NA
Span        -0.0045 0.0120 -0.38 52  7.1e-01 -0.0287  0.01961    4
Gest        -0.0039 0.0015 -2.67 56  1.0e-02 -0.0069 -0.00097    4
              fmi lambda
(Intercept) 0.087  0.053
Span        0.089  0.054
Gest        0.054  0.021
```