Lec 11: Missing Data - Solutions

MATH 456 - Spring 2016

Navbar: [Home] [Schedule] [Data] [Week 15 Overview] [HW Info] [Google Group]

```
lib <- c("VIM", "xtable", "dplyr", "mice", "missForest", "ggplot2","scales", "lattice")
invisible(lapply(lib, library, character.only=T))
```

1. For each distribution listed below, draw a random sample of $n = 100$ and delete 20% of the data completely at random and calculate the bias. State if the observed mean over- or under-estimates the true mean.
   a. $X \sim \mathcal{N}(10, 2)$
   b. $Y \sim \mathcal{POISSON}(10)$
   c. $B \sim \mathcal{BINOMIAL}(10, .5)$
   d. $F \sim \mathcal{BINOMIAL}(10, .9)$
   Does the effect of MCAR on the bias differ for the different distributions?
2. Repeat #1 but set the missing data mechanism to b NMAR, where p is negatively correlated with the data.
3. **Using the Parental HIV data set, consider** only** the following variables: `Age`, `Gender`, `livwith`, `BSI_overall`, `Frnds`, and `Hookey`.**

```
hiv <- read.delim("C:/GitHub/MATH456/data/PARHIV_022216.txt")
names(hiv) <- tolower(names(hiv))
hiv <- hiv %>% select(age, gender, livwith, bsi_overall, frnds, hookey)
```

a. What percent of the data set overall is missing?

```
table(is.na(hiv))
```

```
##
## FALSE  TRUE
##  1506     6
```

```
round(mean(is.na(hiv))*100, 2)
```

```
## [1] 0.4
```

Only 6 pieces of data (0.4%) are missing.

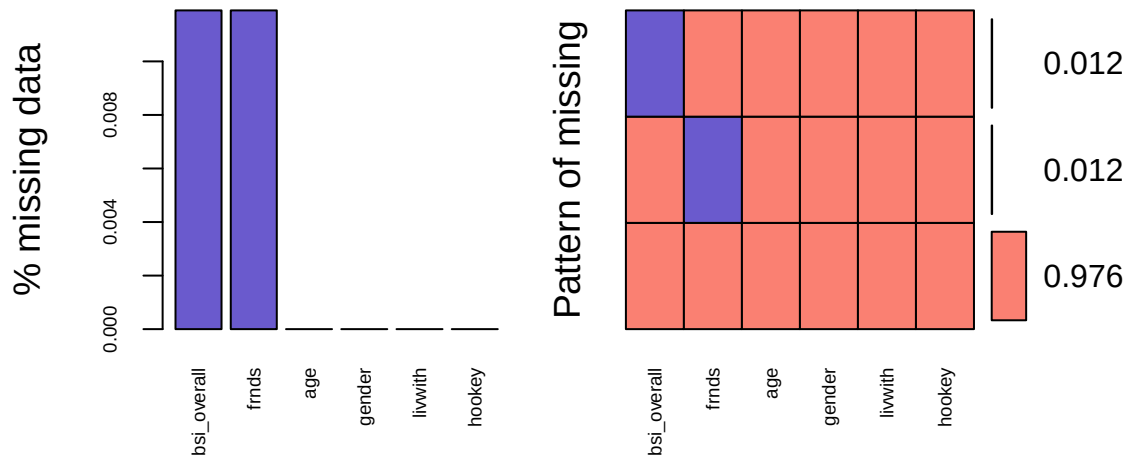
b. How much missing data is there per variable?

```
apply(hiv, 2, function(x) round(sum(is.na(x)),4))
```

```
##         age      gender     livwith bsi_overall       frnds      hookey
##           0           0           0           3           3           0
```

c. Describe the missing data pattern. (*Hint: Use `aggr` from the `VIM` package for part b and c*)

```
aggr(hiv, col=c("salmon", "slateblue"), numbers=TRUE, sortVars=TRUE, labels=names(hiv),
     cex.axis=.6, gap=3, ylab=c("% missing data", "Pattern of missing"))
```

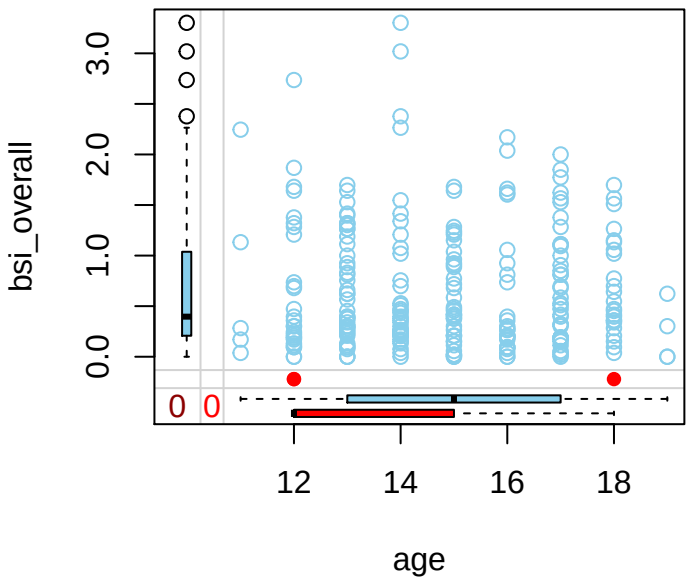



```
##
##  Variables sorted by number of missings:
##     Variable      Count
##  bsi_overall 0.01190476
##        frnds 0.01190476
##          age 0.00000000
##       gender 0.00000000
##      livwith 0.00000000
##       hookey 0.00000000
```

Only `BSI_overall` and `Frnds` have any missing data, with 3 records each, and they are never missing at the same time.

d. Describe the relationship of missing data between Age and BSI overall.

```
marginplot(hiv[,c(1,4)])
```



It appears that BSI overall is only missing for youth that are either very young (12), or very old (18).

4. Multiply impute this data set $m = 5$ times.
    a. State the imputation models used for each variable.
5. After controlling for age, who the student lives with, their overall friendships, and overall BSI score, what is the effect of gender on the likelihood a student will skip school? (I.e. Fit a logistic regression model using `hookey` as the outcome and all other covariates as predictors. Calculate the OR and 95% CI for the effect of gender)
    a. Fit this model on the complete cases (no imputation).
    b. Fit this model on the multiply imputed data sets and pool the results.
    c. Compare these two intervals. Which one is wider? Why?
    d. Discuss the amount of information lost (in terms of sample size and varance) using the complete case method.