Lecture 3: More Missing Data Basics

Kyle M. Lang

Institute for Measurement, Methodology, Analysis & Policy
Texas Tech University
Lubbock, TX

September 8, 2015

Outline

- More on missing data mechanisms
- A few more missing data diagnostics
- Ad Hoc techniques and their problems
- More on MI and FIML

Missing Data Mechanisms

MCAR:

$$P(R|Y_{mis}, Y_{obs}) = P(R)$$

MAR:

$$P(R|Y_{mis}, Y_{obs}) = P(R|Y_{obs})$$

MNAR:

$$P(R|Y_{mis}, Y_{obs}) \neq P(R|Y_{obs})$$

Simulate Some Toy Data

```
nObs ← 1000 # Sample Size
pm ← 0.3 # Proportion Missing
sigma ← matrix(c(1.0, 0.5, 0.0,
                 0.5, 1.0, 0.3,
                 0.0, 0.3, 1.0),
               ncol = 3)
simDat ← as.data.frame(rmvnorm(nObs, c(0, 0, 0), sigma))
colnames(simDat) ← c("y", "x", "z")
x ← simDat$x
y ← simDat$y
z ← simDat$z
cor(y, x) # Check correlation between X and Y
```

```
[1] 0.4691526
```

MCAR Example I

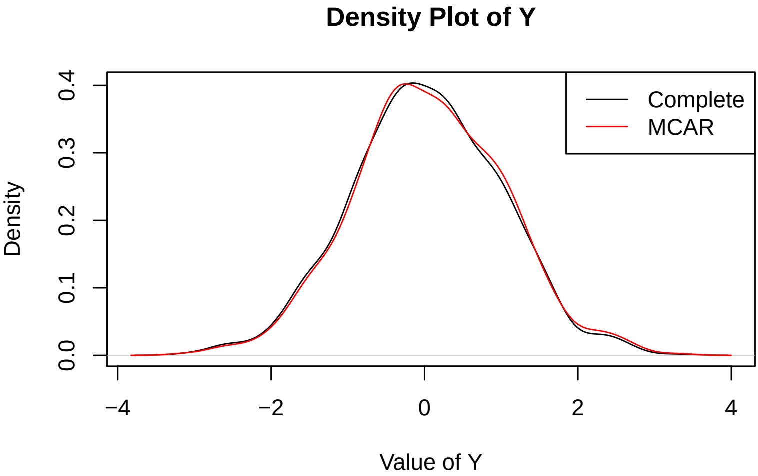
```
## Simulate MCAR Missingness:
rVec1 ← as.logical(rbinom(nObs, size = 1, prob = pm))
mean(rVec1) # Check the PM
```

```
[1] 0.301
```

```
y2 ← y
y2[rVec1] ← NA
cor(y2, x, use = "pairwise") # Look at correlation
```

```
[1] 0.4744126
```

MCAR Example II



MAR Example I

```
## Simulate MAR Missingness:
rVec2 ← pnorm(x, mean = mean(x), sd = sd(x)) < pm
mean(rVec2)
```

```
[1] 0.287
```

```
y3 ← y
y3[rVec2] ← NA
cor(y3, x, use = "pairwise") # Not looking so good :(
```

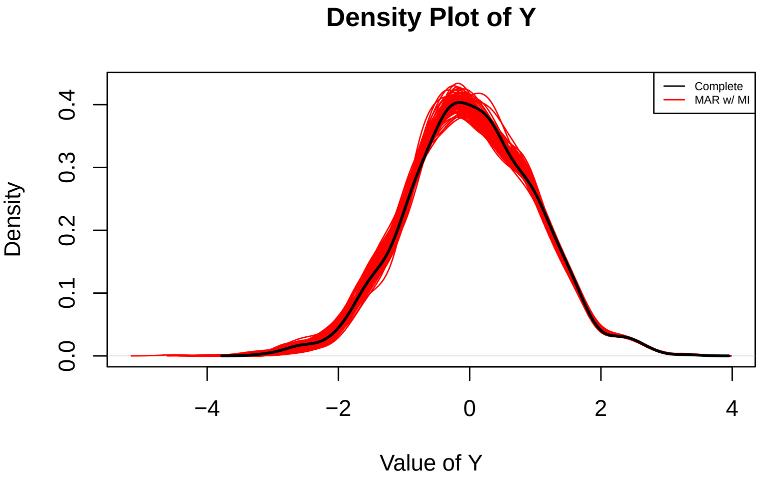
```
[1] 0.3715953
```

MAR Example II

```
## MI to the rescue:
miceOut1 ← mice(data.frame(y3, x),
                m = 100,
                maxit = 1,
                method = c("norm", ""),
                printFlag = FALSE)
impList1 ← list()
for(m in 1 : miceOut1$m) {
    impList1[[m]] ← complete(miceOut1, m)
}
corList ←lapply(impList1,
                FUN = function(impDat){
                    cor(impDat$x, impDat$y3)
                }
                )
mean(unlist(corList)) # Oh, much nicer :)
```

```
[1] 0.4835711
```

MAR Example III



MNAR Example I

```
## Simulate MNAR Missingness:
rVec3 ← pnorm(y, mean = mean(y), sd = sd(y)) < pm
mean(rVec3)
```

```
[1] 0.294
```

```
y4 ← y
y4[rVec3] ← NA
cor(y4, x, use = "pairwise") # Hmm...looks pretty bad.
```

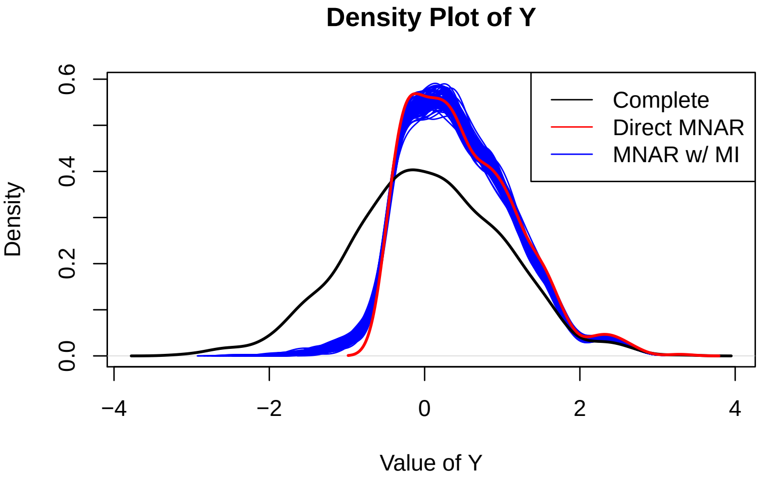
```
[1] 0.3708081
```

MNAR Example II

```
## Can MI help?
miceOut2 ← mice(data.frame(y4, x),
                 m = 100,
                 maxit = 1,
                 method = c("norm", ""),
                 printFlag = FALSE)
impList2 ← list()
for(m in 1 : miceOut2$m) {
    impList2[[m]] ← complete(miceOut2, m)
}
corList2 ←lapply(impList2,
                 FUN = function(impDat){
                     cor(impDat$x, impDat$y4)
                 }
                 )
mean(unlist(corList2)) # Not really
```

```
[1] 0.3914519
```

MNAR Example III



Indirect MNAR Example I

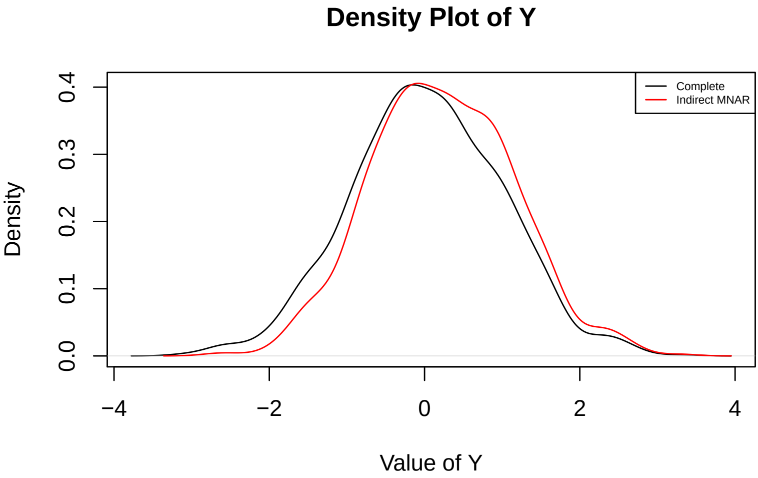
QUESTION: In our previous MAR example, what happens if we don't account for the predictor of the MAR missingness?

```
cor(y3, x, use = "pairwise") # Hmm...that's a problem.
```

```
[1] 0.3715953
```

ANSWER: We get *Indirect MNAR.*

Indirect MNAR Example II



Tricky Example I

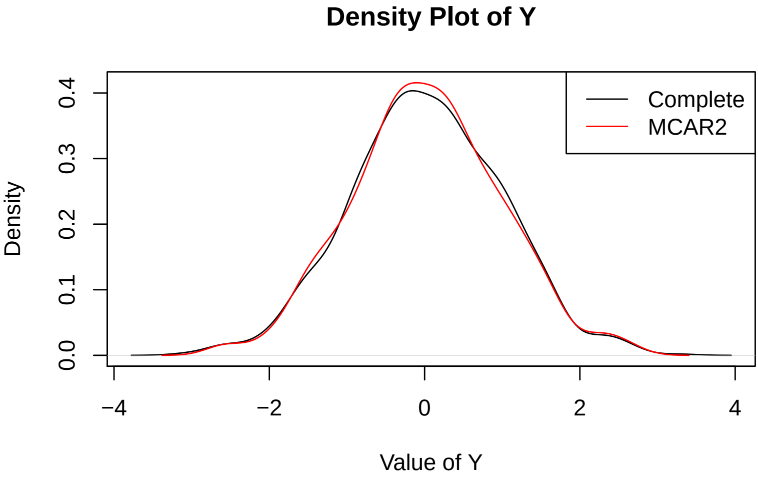
Question: What happens if we ignore the predictor of missingness, but that predictor is independent of our study variables?

```
rVec3 ← pnorm(z, mean = mean(z), sd = sd(z)) < pm
y5 ← y
y5[rVec3] ← NA
cor(y5, x, use = "pairwise")
```

```
[1] 0.472859
```

Tricky Example II

ANSWER: We get back to MCAR :)


Density Plot of Y
Complete
MCAR2
Density
Value of Y