KU Leuven Summer School
Segment 2B
More Missing Data

Paul Gustafson

September 15, 2022

Stylized context (very similar to Segment 2A)

X_1: Binary indicator of high blood pressure (1=Yes)

X_2: Binary indicator of regular exercise (1=No)

Y: Binary indicator of heart disease (1=Yes)

Statistical model:

$\text{logit}\{Pr(Y = 1|X_1, X_2)\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

But there is a "disturbance in the force:"

- Obtain (X_1, Y) for **all** n study subjects from electronic health records.
- Obtain X_2 for **only some** study subjects from survey, turns out to have a low response rate.

First (mystery) dataset

n=1000

```
summary(dat1)
```

```
##        x1              x2              y        
##  Min.   :0.000   Min.   :0      Min.   :0.000  
##  1st Qu.:0.000   1st Qu.:0      1st Qu.:0.000  
##  Median :0.000   Median :0      Median :0.000  
##  Mean   :0.406   Mean   :0      Mean   :0.353  
##  3rd Qu.:1.000   3rd Qu.:1      3rd Qu.:1.000  
##  Max.   :1.000   Max.   :1      Max.   :1.000  
##                  NA's   :603                   
```

First dataset, continued

```
table(dat1, exclude=NULL)
```

```
## , , y = 0
##
##    x2
## x1    0   1 <NA>
##   0 141  27  244
##   1  32  65  138
##
## , , y = 1
##
##    x2
## x1    0   1 <NA>
##   0  53  15  114
##   1  13  51  107
```

2×2×3 table

"sufficient"

Second (mystery) dataset

```r
table(dat2, exclude=NULL)
```

```
## , , y = 0
##
##    x2
## x1    0   1 <NA>
##   0 188  28  196
##   1  43  84  108
##
## , , y = 1
##
##    x2
## x1    0   1 <NA>
##   0  75  29   78
##   1   5  32  134
```

Third (mystery) dataset

```r
table(dat3, exclude=NULL)
```

```
## , , y = 0
##
##     x2
## x1     0   1 <NA>
##    0 351  61    0
##    1  81 154    0
##
## , , y = 1
##
##     x2
## x1     0   1 <NA>
##    0 136  13   33
##    1  30  32  109
```

weird

Answer by package - JAGS

```
genmod.string <- "model{

  ### prior distribution
  alpha0 ~ dnorm(0, 0.1)
  alpha1 ~ dnorm(0, 0.1)
   beta0 ~ dnorm(0, 0.1)
   beta1 ~ dnorm(0, 0.1)
   beta2 ~ dnorm(0, 0.1)

  ### statistical model
  for (i in 1:n) {
    x2[i] ~ dbern(pr.x2[i])
    logit(pr.x2[i]) <- alpha0 + alpha1*x1[i]

    y[i] ~ dbern(pr.y[i])
    logit(pr.y[i]) <- beta0 + beta1*x1[i] + beta2*x2[i]
  }

}"
```

– prec.
weakly informative
for log odds or
log OR (as
earlier)

Pause to comment on this prior and stat model specification

X_1 unmodelled

$$\text{logit}\ Pr(X_2 = 1 \mid X_1) = \alpha_0 + \alpha_1 X_1$$ "saturated"

$$\text{logit}\ Pr(Y = 1 \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

assume
no
interaction

JAGS, continued

```
### generative model, data go in
mod <- jags.model(textConnection(genmod.string),
         data=list(x1=dat1$x1, x2=dat1$x2, y=dat1$y,
         n.chains=4)
```

includes NAs

```
update(mod, 2000) # burn-in

###  MC output comes out
opt1.JAGS <- coda.samples(mod, n.iter=10000,
  variable.names=c("alpha0","alpha1","beta0","beta1",
    "beta2","x2[7]","x2[8]"))
```

JAGS, continued

```
summary(opt1.JAGS)
```

E(B1|data)

SD(B1|data)

```
##
## Iterations = 3001:13000
## Thinning interval = 1
## Number of chains = 4
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##          Mean    SD Naive SE Time-series SE
## alpha0 -1.519 0.170 0.000851        0.00300
## alpha1  2.477 0.244 0.001222        0.00434
## beta0  -0.918 0.104 0.000518        0.00130
## beta1   0.227 0.195 0.000976        0.00322
## beta2   0.508 0.256 0.001279        0.00509
## x2[7]   0.231 0.421 0.002106        0.00218
## x2[8]   0.000 0.000 0.000000        0.00000
##
## 2. Quantiles for each variable:
##
```

latent

obs

And for comparison: complete-case analysis

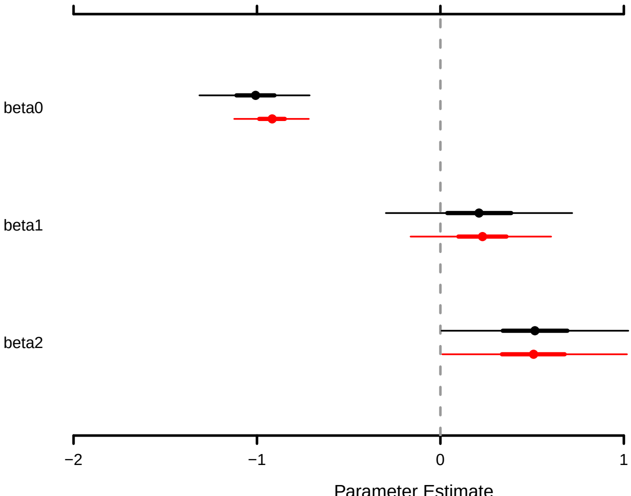
```
cmplt <- !is.na(dat1$x2)
cmplt[1:8]
```

```
## [1]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
```

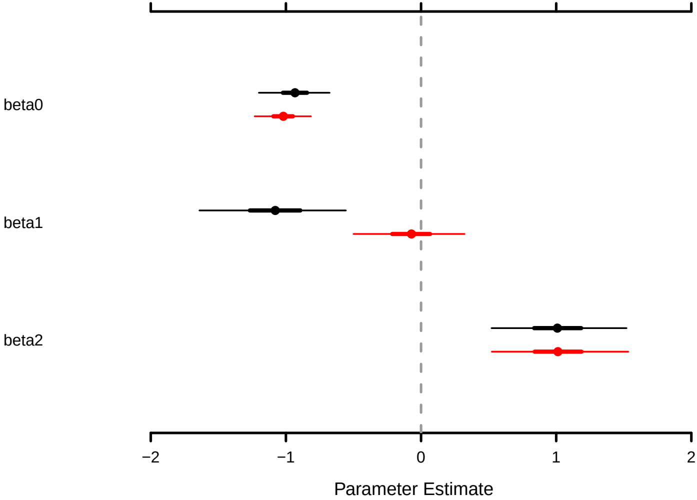
```
mod <- jags.model(textConnection(genmod.string),
  data=list(x1=dat1$x1[cmplt],
            x2=dat1$x2[cmplt],
            y=dat1$y[cmplt],

opt1.cc.JAGS <- coda.samples(mod,
                  variable.names=c("beta0","beta1","beta2"),
                  n.iter=10000)
```

Comparison: Dataset 1, complete-case versus latent

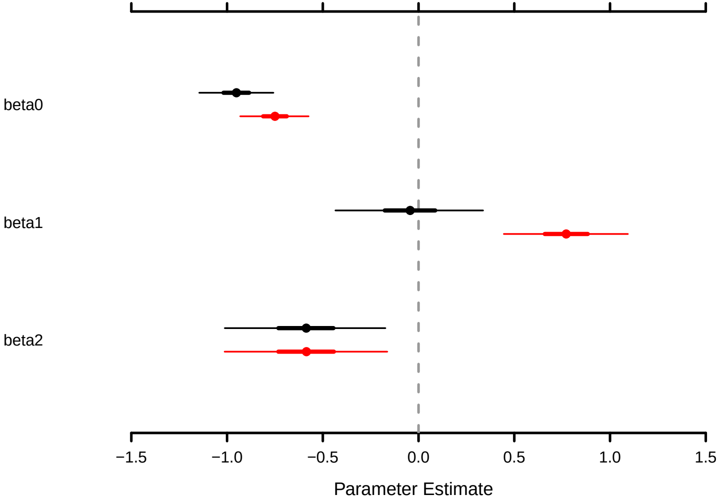
```r
MCMCplot(opt1.cc.JAGS, opt1.JAGS,
  params=c("beta0","beta1","beta2"))
```



Same comparison, but for Dataset 2



Same comparison, but for Dataset 3



So what have we actually done here? (thinking space 1)

Monte Carlo representation

of

$$f\left(\underset{\sim}{\alpha}, \underset{\sim}{\beta}, \underset{\sim}{X}_2^{(mis)} \,\middle|\, \underset{\sim}{X}_1, \underset{\sim}{X}_2^{(obs)}, \underset{\sim}{y}\right)$$

Let's think harder about how missing values became that way

Let R by a binary indicator, taking the value 1 if X_2 is observed, 0 if it's missing.

Need to think about the distribution of (X_1, X_2, Y, R).

And concede that in fact for a given patient we will observe an event which has one of these two forms:

- $(X_1 = x_1, X_2 = x_2, Y = y, R = 1)$
- $(X_1 = x_1, Y = y, R = 0)$

Aside to think about: Sometimes this would be written as $(X_1, Y, R, X_2 R)$ are the observable variables.

In generality, think of this generative model

$$\begin{aligned} f(\alpha, \beta, x_1, x_2, y, r) &= f(\alpha, \beta) f(x_1) f(x_2 | x_1, \alpha) \times \\ & f(y | x_1, x_2, \beta) f(r | x_1, x_2, y). \end{aligned}$$

- Have made *conditional independence* assumptions here.
- With terms in red, think if our answer depends on them, then we will have to know their forms.

Applying Bayes theorem to this generative model gets us to

$$f\left(\alpha, \beta, x_2^{(mis)} | x_1, x_2^{(obs)}, y, r\right) \propto f(\alpha,\beta) f(x_2|x_1,\alpha) f(y|x_1,x_2,\beta) \times$$
$$f(r|x_1, x_2, y)$$

(and again, think hard about the meaning of $\propto$ here)

So we have a hall-pass to stick with the analysis above so long as. . .

$f(r|x_1, x_2, y)$ doesn't actually depend on x_2

as a function of $\alpha, \beta, x_2^{(mis)}$ given everything else

i.e. $R \perp\!\!\!\perp X_2 \mid X_1, Y$

- missing at random (MAR)
- ignorable missingness

Ignorable missingess

In words, chance of missingess

DOES NOT DEPEND

on the underlying value that may/may not be obscured.

Two related things to ponder. In situations where we *aren't* comfortable making this assumption:

- Could we include a further unknown parameter (say λ) in the generative model and keep/augment the $f(r|x_1, x_2, y, \lambda)$ term?
- Can the data empirically provide evidence for/against the assumption?

Now for a grand reveal concerning the three mystery datasets

$$\text{logit}\{Pr(Y=1|X_1,X_2)\} \quad = \quad -1 \quad + \quad 0.2\ X_1 \quad + 0.6 X_2$$

Dataset 1

$Pr(R = 1|X_1, X_2, Y) =$ 0.4

missing completely at random

no dependence of R on (X_1, X_2, Y)

Intuition : complete case analysis valid but inefficient ✓ (slide 13)

Dataset 2

$Pr(R = 1|X_1, X_2, Y) =$ $0.55 - 0.3 X_1 Y$

complete cases no longer
a random subset of all cases

complete case analysis dubious (slide 14 ~)

but still ignorable missingness, no X_2 here
latent variable analysis should work

Dataset 3

$Pr(R = 1|X_1, X_2, Y) =$ $1 - .75 X_2 Y$

X_2 here

no reason to expect our latent var. analysis to work ✓ (slide 15)

And then a final thought to come back to

If we don't feel comfortable assuming ignorable missingness, why not just work with

$f(\alpha, \beta, \lambda) f(x_1) f(x_2|x_1, \alpha) f(y|x_1, x_2, \beta) f(r|x_1, x_2, y, \lambda)$

see 2C

big caveat

learning the association between X_2 and R from data on (R, RX_2) is impossible

Thought, continued