KU Leuven Summer School
Segment 2C
Missing (Perhaps?) Not at Random

Paul Gustafson

September 15, 2022

Minimal Working Example

```
expit <- function(z) {1/(1+exp(-z))}
logit <- function(p) {log(p)-log(1-p)}

n <- 2000
x <- rbinom(n, size=1, prob=.4)
y <-rbinom(n, size=1,
            prob=expit(cbind(1,x) %*%
                       c(-1, 0.2)))
```

$$\text{logit } Pr(Y=1|X) = \beta_0 + \beta_1 X$$

Creating the problem

```
r <- rbinom(n, size=1,
  prob=expit(0.1 + 0.3*x - 0.6*x*y))

x.obs <- rep(NA, n); x.obs[r==1] <- x[r==1]
dat <- data.frame(x=x.obs, y=y)
```

X subject to nonignorable missingness

MAR analysis (i)

```
genmod.mar.string <- "
model {
  alpha ~ dnorm(0, 0.1)
  beta0 ~ dnorm(0, 0.1)
  beta1 ~ dnorm(0, 0.1)

  for (i in 1:n) {
    x[i] ~ dbern(pr.x)

    y[i] ~ dbern(pr.y[i])
    logit(pr.y[i]) <- beta0 + beta1*x[i]
  }
  logit(pr.x) <- alpha
}"
```

along the lines of what we've seen

$f(x,y) = f(x)f(y|x)$

MAR analysis (ii)

```
### generative model, data go in
mod <- jags.model(
  textConnection(genmod.mar.string),
  data=list(x=dat$x, y=dat$y,
            n=dim(dat)[1]),
         n.chains=3)

update(mod, 2000) # burn-in

###  MC output comes out
opt.mar.JAGS <- coda.samples(mod, n.iter=15000,
  variable.names=c("alpha","beta0","beta1"))
```

MAR Analysis (iii)

```
MCMCsummary(opt.mar.JAGS)
```

```
##          mean     sd   2.5%    50%   97.5% Rhat n.eff
## alpha -0.382 0.0626 -0.505 -0.382 -0.2590    1 11722
## beta0 -0.843 0.0744 -0.991 -0.843 -0.6988    1  6559
## beta1 -0.217 0.1446 -0.500 -0.216  0.0629    1  5586
```

95% credible interval for β_1

but truth is $\beta_1 = 0.20$

Now explictly model the missingness

```
statmod.mnar.string <- "
for (i in 1:n) {
  x[i] ~ dbern(pr.x)

  y[i] ~ dbern(pr.y[i])
  logit(pr.y[i]) <- beta0 + beta1*x[i]

  r[i] ~ dbern(pr.r[i])
  logit(pr.r[i]) <- gamma0 + gamma1*y[i] +
    gamma2*x[i] + (gamma3-gamma2)*x[i]*y[i]
}
logit(pr.x) <- alpha"
```

same as before

MAR $\Leftrightarrow (\gamma_2, \gamma_3) = (0, 0)$

4 df to represent $R|X,Y$

parameterized such that $\log OR(R, X|Y) = \begin{cases} \gamma_2 & \text{if } Y=0 \\ \gamma_3 & \text{if } Y=1 \end{cases}$

With this prior specification

allowing that their could be small departures from MAR

```
prior.mnarA.string <- "
   alpha ~ dnorm(0, 0.1)
   beta0 ~ dnorm(0, 0.1)
   beta1 ~ dnorm(0, 0.1)
  gamma0 ~ dnorm(0, 0.1)
  gamma1 ~ dnorm(0, 0.1)
  gamma2 ~ dnorm(0, 100)
  gamma3 ~ dnorm(0, 100)"
```

$N(0, 0.1^2)$

```
genmod.mnarA.string <-  paste(
  "model {",prior.mnarA.string,
  statmod.mnar.string, "}")
```

Any pertinent remarks about our generative model?

$$f(\alpha) f(\beta) f(\gamma)$$

$$\prod_{i=1}^{n} f(x_i \mid \alpha)\, f(y_i \mid x_i, \beta)\, f(r_i \mid y_i, x_i, \gamma)$$

$$\Rightarrow$$

$$f(\alpha, \beta, \gamma, x^{(mis)} \mid x^{(obs)}, y, r)$$

Turn the crank

```
### generative model, data go in
mod <-  jags.model(
  textConnection(genmod.mnarA.string),
  data=list(x=dat$x, y=dat$y,
            r=as.vector(!is.na(dat$x)),
            n=dim(dat)[1]),
           n.chains=3)

update(mod,2000) # burn-in

###  MC output comes out
opt.mnarA.JAGS <- coda.samples(
  mod,
  n.iter=15000,
  variable.names=c("alpha","beta0","beta1",
  "gamma0","gamma1","gamma2","gamma3"))
```

And report posterior quantities

. wider now than under MAR

. upper endpoint still not reaching truth of 0.2

```
MCMCsummary(opt.mnarA.JAGS)
```

```
##              mean     sd    2.5%      50%   97.5% Rhat n.eff
## alpha  -0.381089 0.0719 -0.5211 -0.38133 -0.2394    1  6076
## beta0  -0.844381 0.0793 -1.0003 -0.84397 -0.6912    1  4669
## beta1  -0.214351 0.1606 -0.5323 -0.21375  0.0977    1  4050
## gamma0  0.172698 0.0681  0.0415  0.17162  0.3076    1  6141
## gamma1 -0.236507 0.1144 -0.4625 -0.23676 -0.0136    1  8084
## gamma2  0.000243 0.0993 -0.1935  0.00073  0.1917    1  5126
## gamma3 -0.001862 0.0995 -0.1971 -0.00177  0.1926    1 10451
```

come back to

And now the same with this prior specification

now allowing slightly larger departures from MAR

```
prior.mnarB.string <- "
   alpha ~ dnorm(0, 0.1)
   beta0 ~ dnorm(0, 0.1)
   beta1 ~ dnorm(0, 0.1)
  gamma0 ~ dnorm(0, 0.1)
  gamma1 ~ dnorm(0, 0.1)
  gamma2 ~ dnorm(0, 25)
  gamma3 ~ dnorm(0, 25)"
```

$\sim N(0, 0.2^2)$

```
genmod.mnarB.string <-  paste(
  "model {",prior.mnarB.string,
  statmod.mnar.string, "}")
```

Crank

```r
### generative model, data go in
mod <-  jags.model(
  textConnection(genmod.mnarB.string),
  data=list(x=dat$x, y=dat$y,
    r=as.vector(!is.na(dat$x)),
    n=dim(dat)[1]),
  n.chains=3)

update(mod, 2000) # burn-in

###  MC output comes out
opt.mnarB.JAGS <- coda.samples(
  mod,
  n.iter=15000,
  variable.names=c("alpha","beta0","beta1","gamma0",
    "gamma1","gamma2","gamma3"))
```

Answer

wider still
- almost reaches truth of 0.2

```
MCMCsummary(opt.mnarB.JAGS)
```

```
##              mean     sd     2.5%      50%   97.5% Rhat n.eff
## alpha   -0.37978 0.0951 -0.56618 -0.37993 -0.1932    1  2213
## beta0   -0.84465 0.0917 -1.02539 -0.84368 -0.6687    1  2496
## beta1   -0.21456 0.1987 -0.60383 -0.21329  0.1720    1  2152
## gamma0   0.17851 0.1002 -0.00844  0.17576  0.3863    1  1835
## gamma1  -0.23858 0.1501 -0.53573 -0.23850  0.0576    1  2548
## gamma2  -0.00209 0.2014 -0.39610 -0.00159  0.3953    1  1579
## gamma3  -0.00483 0.2019 -0.40203 -0.00533  0.3888    1  3619
```

Take a bit of stock

Recall γ_2, γ_3 describe $\log OR(R, X|Y = y)$

	prior SDs	posterior SDs	quality of numerical approximation
MAR	(0,0)	(0,0)	
	(0.1, 0.1)	(0.1, 0.1)	degrading
	(0.2, 0.2)	(0.2, 0.2)	

no free lunch

can't learn the missingness mechanism from the data

Adjective soup

MAR	MNAR
$\gamma_2 = \gamma_3 = 0$ known (and don't need to model R)	γ_2, γ_3 unknown
• identified • regular • textbook	• partially identified • irregular • inconsistent
• Fisher info theory applies	"low info."
• nice	

Folk theorem

In these kinds of irregular problems, as you turn down the information knob,

the numerical approximation of posterior quantities worsens (when using off-the-shelf MCMC)

increasing prior variances on γ_1, γ_3

Important distinction related to folk theorem

The posterior dist of β when you don't know much about the missing data mechanism could be the scientifically pertinent thing to report - we just can't compute it easily

- a problem of Bayesian computation, not Bayesian analysis per se