Missing Data and Imputation

Hoff Chapter 7, GH Chapter 25

April 21, 2017

Bednets and Malaria

- `Y`:presence or absence of parasites in a blood smear
- `AGE`: age of child
- `BEDNET`: bed net use (exposure)
- `GREEN`:greenness of the surrounding vegetation based on satellite photography
- `PHC`: whether a village is part of a primary health-care system

Bednets and Malaria

```
malaria = read.csv("gambia.dat", header=TRUE)
summary(malaria)

       Y               AGE              BEDNET           GREEN
 Min.    :0.0000   Min.    :1.000   Min.    :0.0000   Min.    :28.85   Min.
 1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:40.85   1st
 Median :0.0000   Median :2.000   Median :1.0000   Median :40.85   Medi
 Mean    :0.3093   Mean    :2.399   Mean    :0.7049   Mean    :39.84   Mean
 3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:40.85   3rd
 Max.    :1.0000   Max.    :4.000   Max.    :1.0000   Max.    :47.65   Max.
                                    NA's    :317
```

39% missing

More about missingness

- Consider Probability of missingness - are certain groups more likely to have missing data?

More about missingness

- Consider Probability of missingness - are certain groups more likely to have missing data?
- Are certain responses more likely to be missing? (i.e. individuals with high income are more likely to not report it) probability of missing depends on value of outcome.

More about missingness

- Consider Probability of missingness - are certain groups more likely to have missing data?
- Are certain responses more likely to be missing? (i.e. individuals with high income are more likely to not report it) probability of missing depends on value of outcome.
- Analysis depends on assumptions about missingness

Mechanisms for Missingness

- Missing Completely at random (MCAR): missingness does not depend on outcome or other variables

Mechanisms for Missingness

- Missing Completely at random (MCAR): missingness does not depend on outcome or other variables
- Missing at Random: missing does not depend on value of variable, but may depend on other variables.

Mechanisms for Missingness

- Missing Completely at random (MCAR): missingness does not depend on outcome or other variables
- Missing at Random: missing does not depend on value of variable, but may depend on other variables.
- Missing Not at Random: missingness depends on the variable that is missing

Mechanisms for Missingness

- Missing Completely at random (MCAR): missingness does not depend on outcome or other variables
- Missing at Random: missing does not depend on value of variable, but may depend on other variables.
- Missing Not at Random: missingness depends on the variable that is missing

Cannot tell from data

Modeling

- Delete subjects with any missing observations. This would remove 39 % of the data and reduces power. Induces Bias if data are not missing completely at random!

Modeling

- Delete subjects with any missing observations. This would remove 39 % of the data and reduces power. Induces Bias if data are not missing completely at random!
- Replace each missing value with an estimated mean (plug-in approach). This implies that we are certain about the values of the missing cases, so any measures of uncertainty in parameter estimates are overly optimistic (too narrow). Distorts correlation structure in data

Modeling

- Delete subjects with any missing observations. This would remove 39 % of the data and reduces power. Induces Bias if data are not missing completely at random!
- Replace each missing value with an estimated mean (plug-in approach). This implies that we are certain about the values of the missing cases, so any measures of uncertainty in parameter estimates are overly optimistic (too narrow). Distorts correlation structure in data
- Work with likelihoods based on observed data; this will be a product of marginal distributions, difficult to work with

Modeling

- Delete subjects with any missing observations. This would remove 39 % of the data and reduces power. Induces Bias if data are not missing completely at random!
- Replace each missing value with an estimated mean (plug-in approach). This implies that we are certain about the values of the missing cases, so any measures of uncertainty in parameter estimates are overly optimistic (too narrow). Distorts correlation structure in data
- Work with likelihoods based on observed data; this will be a product of marginal distributions, difficult to work with
- Model Based Methods

Observed Data

- $(Y_{i,1}, Y_{i,2}, Y_{i,3}, Y_{i,4}, Y_{i,5})$
- $(O_{i,1}, O_{i,2}, O_{i,3}, O_{i,4}, O_{i,5})$

Observed Data

- $(Y_{i,1}, Y_{i,2}, Y_{i,3}, Y_{i,4}, Y_{i,5})$
- $(O_{i,1}, O_{i,2}, O_{i,3}, O_{i,4}, O_{i,5})$ where $O_{i,j}$ is 1 if $Y_{i,j}$ is observed and $O_{i,j}$ is 0 if $Y_{i,j}$ is missing

Observed Data

- $(Y_{i,1}, Y_{i,2}, Y_{i,3}, Y_{i,4}, Y_{i,5})$
- $(O_{i,1}, O_{i,2}, O_{i,3}, O_{i,4}, O_{i,5})$ where $O_{i,j}$ is 1 if $Y_{i,j}$ is observed and $O_{i,j}$ is 0 if $Y_{i,j}$ is missing

Missing at Random Data:

- O_i and Y_i are independent given $\boldsymbol{\theta}$

Observed Data

- $(Y_{i,1}, Y_{i,2}, Y_{i,3}, Y_{i,4}, Y_{i,5})$
- $(O_{i,1}, O_{i,2}, O_{i,3}, O_{i,4}, O_{i,5})$ where $O_{i,j}$ is 1 if $Y_{i,j}$ is observed and $O_{i,j}$ is 0 if $Y_{i,j}$ is missing

Missing at Random Data:

- O_i and Y_i are independent given $\boldsymbol{\theta}$
- distribution for O_i does not depend on $\boldsymbol{\theta}$

Observed Data

- $(Y_{i,1}, Y_{i,2}, Y_{i,3}, Y_{i,4}, Y_{i,5})$
- $(O_{i,1}, O_{i,2}, O_{i,3}, O_{i,4}, O_{i,5})$ where $O_{i,j}$ is 1 if $Y_{i,j}$ is observed and $O_{i,j}$ is 0 if $Y_{i,j}$ is missing

Missing at Random Data:

- O_i and Y_i are independent given $\boldsymbol{\theta}$
- distribution for O_i does not depend on $\boldsymbol{\theta}$

Marginal Model for observed data

$$\begin{aligned} p(o_i, y[o_i = 1] \mid \boldsymbol{\theta}) &= p(o_i) p(y[o_i = 1] \mid \boldsymbol{\theta}) \\ &= p(o_i) \int \left\{ p(y_{i,1}, y_{i,2}, y_{i,3}, y_{i,4}, y_{i,5} \mid \boldsymbol{\theta}) \prod_{y_{i,j} \ni o_{i,j}=0} dy_{i,j} \right. \end{aligned}$$

Observed Data

- $(Y_{i,1}, Y_{i,2}, Y_{i,3}, Y_{i,4}, Y_{i,5})$
- $(O_{i,1}, O_{i,2}, O_{i,3}, O_{i,4}, O_{i,5})$ where $O_{i,j}$ is 1 if $Y_{i,j}$ is observed and $O_{i,j}$ is 0 if $Y_{i,j}$ is missing

Missing at Random Data:

- O_i and Y_i are independent given $\boldsymbol{\theta}$
- distribution for O_i does not depend on $\boldsymbol{\theta}$

Marginal Model for observed data

$$p(o_i, y[o_i = 1] \mid \boldsymbol{\theta}) = p(o_i)p(y[o_i = 1] \mid \boldsymbol{\theta})$$
$$= p(o_i) \int \left\{ p(y_{i,1}, y_{i,2}, y_{i,3}, y_{i,4}, y_{i,5} \mid \boldsymbol{\theta}) \prod_{y_{i,j} \ni o_{i,j}=0} dy_{i,j} \right.$$

Integrate over the missing variables to obtain the likelihood

Use the Gibbs Sampler to Integrate

If we had "complete data" then we would draw $\boldsymbol{\theta}$ from the condition distribution of $\boldsymbol{\theta} \mid \mathbf{Y}$ class for sampling $\boldsymbol{\mu}$ and Σ. Add a step at each iteration to generate the missing data:

Use the Gibbs Sampler to Integrate

If we had "complete data" then we would draw $\boldsymbol{\theta}$ from the condition distribution of $\boldsymbol{\theta} \mid \mathbf{Y}$ class for sampling $\boldsymbol{\mu}$ and Σ. Add a step at each iteration to generate the missing data:

- Generate $Y_{\text{miss}}^{(t+1)}$ from $p(Y_{\text{miss}} \mid Y_{\text{obs}}, \boldsymbol{\theta}^{(t)})$ and fill in the missing data to obtain a "complete" matrix Y from Y_{obs} and Y_{miss}

Use the Gibbs Sampler to Integrate

If we had "complete data" then we would draw $\boldsymbol{\theta}$ from the condition distribution of $\boldsymbol{\theta} \mid \mathbf{Y}$ class for sampling $\boldsymbol{\mu}$ and Σ. Add a step at each iteration to generate the missing data:

- Generate $Y_{\text{miss}}^{(t+1)}$ from $p(Y_{\text{miss}} \mid Y_{\text{obs}}, \boldsymbol{\theta}^{(t)})$ and fill in the missing data to obtain a "complete" matrix Y from Y_{obs} and Y_{miss}
- Generate $\boldsymbol{\theta}^{(t+1)}$ from $p(\boldsymbol{\theta} \mid Y_{\text{obs}}, Y_{\text{miss}}^{(t+1)},)$

Use the Gibbs Sampler to Integrate

If we had "complete data" then we would draw $\boldsymbol{\theta}$ from the condition distribution of $\boldsymbol{\theta} \mid \mathbf{Y}$ class for sampling $\boldsymbol{\mu}$ and Σ. Add a step at each iteration to generate the missing data:

- Generate $Y_{\text{miss}}^{(t+1)}$ from $p(Y_{\text{miss}} \mid Y_{\text{obs}}, \boldsymbol{\theta}^{(t)})$ and fill in the missing data to obtain a "complete" matrix Y from Y_{obs} and Y_{miss}
- Generate $\boldsymbol{\theta}^{(t+1)}$ from $p(\boldsymbol{\theta} \mid Y_{\text{obs}}, Y_{\text{miss}}^{(t+1)},)$

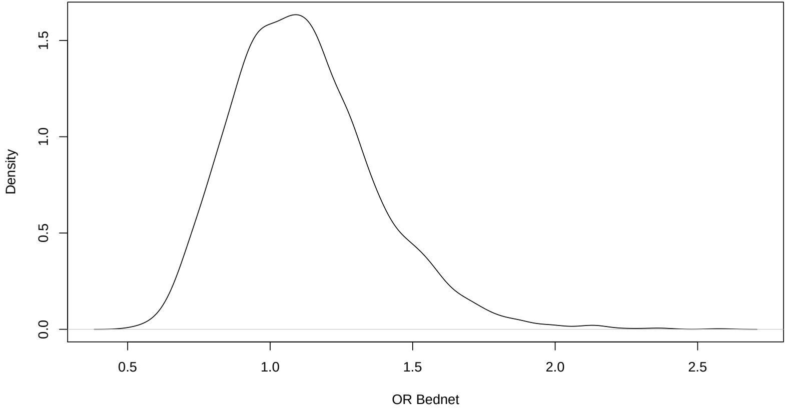
Averaging over the draws of Y_{miss} "integrates" marginalizes over the missing dimensions

JAGS Model

```
model = function() {
  for (i in 1:N) {
  Y[i] ~ dbern(p[i])
  logit(p[i]) <- alpha + beta.age*AGE[i] + beta.bednet*BEDNET[i]
                 +beta.green*GREEN[i] + beta.phc*PHC[i]
}
   # model for missing exposure variable
for (i in 1:N) {
  BEDNET[i] ~ dbern(q) #prior model for whether or not child
                       # sleeps under treated bednet
}
  #uniform prior (uniform) on prob of sleeping under bednet
  q ~ dbeta (1,1)
 #vague priors on regression coefficients
   alpha ~ dnorm(0,0.00000001)
   beta.age ~ dnorm(0,0.00000001)
   beta.bednet ~ dnorm(0,0.00000001)
   beta.green ~ dnorm(0,0.00000001)
   beta.phc ~ dnorm(0,0.00000001)
  # calculate odds ratios of interest
  OR.bednet <- exp(beta.bednet) #OR of malaria for children using bedne
}
```

Posterior Density

```
theta = as.data.frame(sim$BUGSoutput$sims.matrix)
plot(density(theta[,1]), xlab="OR Bednet", main="")
```



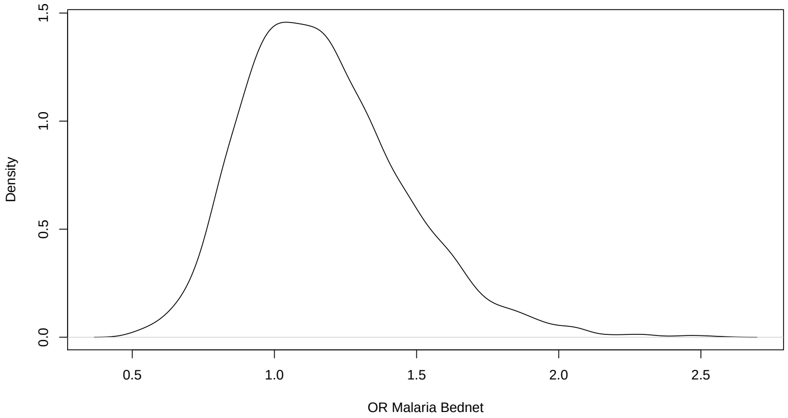
JAGS Model

```
model2 = function() {
  for (i in 1:N) {
  Y[i] ~ dbern(p[i])
  logit(p[i]) <- alpha + beta.age*AGE[i] + beta.bednet*BEDNET[i]
                 +beta.green*GREEN[i] + beta.phc*PHC[i]
}
   # model for missing exposure variable
for (i in 1:N) {
  BEDNET[i] ~ dbern(q[i]) #prior model for  bednet use
  logit(q[i]) <- gamma[1] + gamma[2]*PHC[i] #allow prob depend on PHC
}

 #vague priors on regression coefficients
   gamma[1] ~ dnorm(0,0.00000001)
   gamma[2] ~ dnorm(0,0.00000001)
   alpha ~ dnorm(0,0.00000001)
   beta.age ~ dnorm(0,0.00000001)
   beta.bednet ~ dnorm(0,0.00000001)
   beta.green ~ dnorm(0,0.00000001)
   beta.phc ~ dnorm(0,0.00000001)
  # calculate odds ratios of interest
  OR.bednet <- exp(beta.bednet) #OR of malaria for children using bedne
```

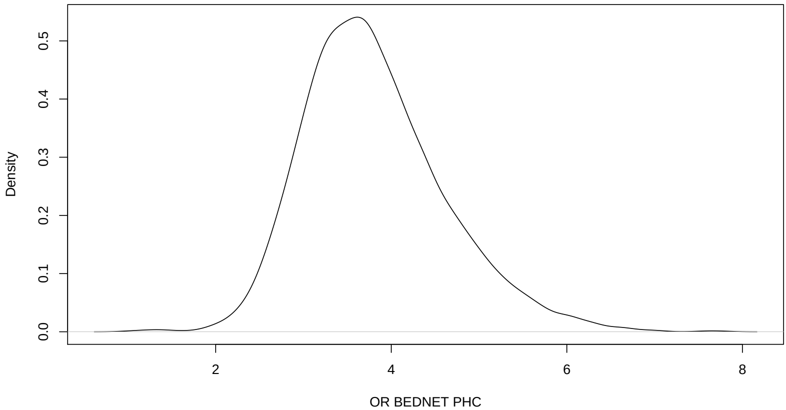
Posterior Density

```
theta.phc = as.data.frame(sim.phc$BUGSoutput$sims.matrix)
plot(density(theta.phc[,1]), xlab="OR Malaria Bednet", main="")
```



Posterior Density

```
plot(density(theta.phc[,"OR.bednet.PHC"]), xlab="OR BEDNET PHC", main="
```



intervals

```
exp(confint(glm(Y ~ . , data=malaria, family=binomial), parm="BEDNET"))

    2.5 %    97.5 %
0.7104643 1.7646674

HPDinterval(as.mcmc(theta))

                   lower        upper
OR.bednet       0.6730938    1.6168561
beta.bednet    -0.3414251    0.5189801
deviance     1564.7933630 1579.4668447
attr(,"Probability")
[1] 0.95

HPDinterval(as.mcmc(theta.phc))

                     lower       upper
OR.bednet         0.6752977    1.742158
OR.bednet.PHC     2.4186453    5.499151
deviance       1524.1955374 1539.457613
attr(,"Probability")
[1] 0.95
```

More than one variable with missing data

- Model each predictor (joint distribution)

More than one variable with missing data

- Model each predictor (joint distribution)
- Coherent sequential model of conditional distributions

More than one variable with missing data

- Model each predictor (joint distribution)
- Coherent sequential model of conditional distributions
- Handle Mix of Discrete and Continuous

More than one variable with missing data

- Model each predictor (joint distribution)
- Coherent sequential model of conditional distributions
- Handle Mix of Discrete and Continuous
- Categorical: Continuation Ratios easiest

More than one variable with missing data

- Model each predictor (joint distribution)
- Coherent sequential model of conditional distributions
- Handle Mix of Discrete and Continuous
- Categorical: Continuation Ratios easiest

Missing Not at Random

- probability of missing depends on predictor

Missing Not at Random

- probability of missing depends on predictor
- need to model joint missingness indicator and outcomes

Missing Not at Random

- probability of missing depends on predictor
- need to model joint missingness indicator and outcomes
- model missingness given variables

Missing Not at Random

- probability of missing depends on predictor
- need to model joint missingness indicator and outcomes
- model missingness given variables
- need more information !

Summary

- Make sure you know how missing data are coded!

Summary

- Make sure you know how missing data are coded!
- Think about why they are missing; i.e if there is no garage then there can be no garage condition.

Summary

- Make sure you know how missing data are coded!
- Think about why they are missing; i.e if there is no garage then there can be no garage condition.
- Joint Models require understanding more about the data and reasons for missingness and more sophisticated modelling

Summary

- Make sure you know how missing data are coded!
- Think about why they are missing; i.e if there is no garage then there can be no garage condition.
- Joint Models require understanding more about the data and reasons for missingness and more sophisticated modelling