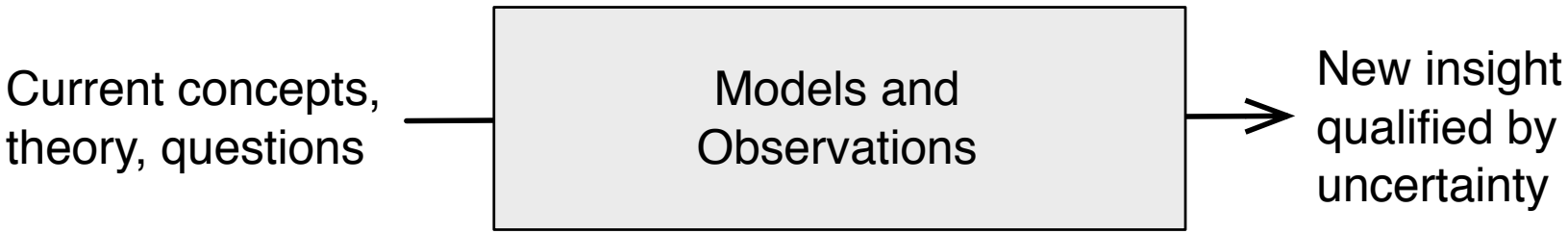
Models for Missing Data

ESS 575 Models for Ecological Data

N. Thompson Hobbs

April 6, 2019

Goal of this course



Missing data are random variables

Remember that Bayesian analysis treats all unobserved quantities as random variables. We see to understand the marginal posterior distributions that give rise to the unobserved quantities conditional on the observed ones. Missing data are treated in the same way as observed data (before they are observed), parameters, and latent variables.

Goal: A brief introduction to modeling missing observations

- Understand the concept of "ignorability."
- Understand the need to be thoughtful about missing data and how it arises.
- Understand the types of mechanisms giving rise to missing data.
- Be able to properly model missing observations of responses and covariates.

Roadmap

- The concept of "ignorability."
- Implications of ignorability for model specification
- Types of missing data
- Modeling missing data in covariates and responses
  - When "missingness" is ignorable
  - When "missingness" is not ignorable
- Best practices for researchers

Overarching concept: Ignorability

We start with a broad definition of "missing data".

Data are missing:

1. From a sample of a population because not all members of the population are included in the sample.
2. From a data set taken in a sample because of errors in recording, non-response in surveys, instrument failure, tag loss etc.

The question of ignorability asks "When do we need to include information about the data collection process in the model we use for analysis of the data?"

Overarching concept: Ignorability

Some notation: Let y be the *complete* set of data (a vector or matrix) of potential observations, the total number of potential observations in the population $= N$. The y can be divided into observed data y_{obs} and and unobserved data y_{miss}. Let I be a vector or matrix of indicator variables, $I = 1_1, ...I_N$, $I_j = 1$ if y_j is observed and $I = 0$ if y_j is missing. For a matrix, we index I as I_{ij}.

We define x as fully observed covariates and include them in distributions to reveal conditional dependence on x. It is critical to understand that the x used here can includes *indices* (i.e., indicators of blocks, strata, clusters, etc) as well as the usual values we expect for covariates.

We define ϕ as a parameter or parameters controlling the probability distribution of I.

Overarching concept: Ignorability

The data collection process can be ignored in the analysis if and only if

$$[\theta \mid x, y_{obs}] = [\theta \mid x, y_{obs}, I]$$

We can ignore the mechanisms that create missing observations when this is true.

Overarching concept: Ignorability

We can ignore the missing data mechanism if the posterior distribution of θ and the posterior predictive distribution of y_{miss} are entirely determined by the specification of a data model $[y|x, \theta]$ and the observed values of y_{obs}. This will be true if

- *Missing at random assumption* $[I \mid x, y, \phi] = [I \mid x, y_{obs}, \phi]$
- *Distinct parameters assumption.* The parameters of the missing data process are independent of the parameters in the data generating process: $[\phi \mid x, \theta] = [\phi \mid x]$. Example of non-distinct parameters: Assigning a larger number of samples to treatment that is suspected to have a small effect.

Ignorability and reserach design

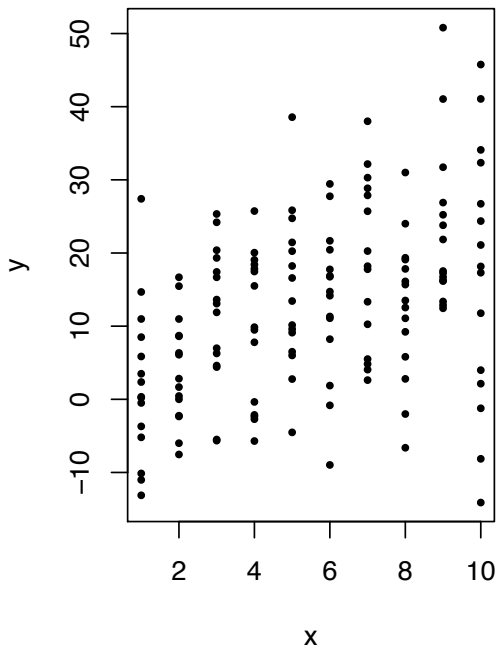
- Designs that are ignorable require no indices other that an index for the individual observations (i.e. y_i) and the covariates (x_i). Simple random sampling and completely randomized experiments have ignorable designs. In these cases $[I \mid x, y, \phi] = [I]$.
- Designs that are not completely random, for example, randomized complete block experiments, stratified random sampling, cluster, and others are not ignorable and must include information on the design in the analysis. Usually, proper indexing and information about the sample sizes specifies all of the needed information. In these cases, $[I \mid x, y, \phi] = [I|x]$.
- Samples from finite populations where the population of potential samples is not many times larger that the sample size are not ignorable. We must include information about the number of samples and the population size in the analysis.

Example: proper indexing

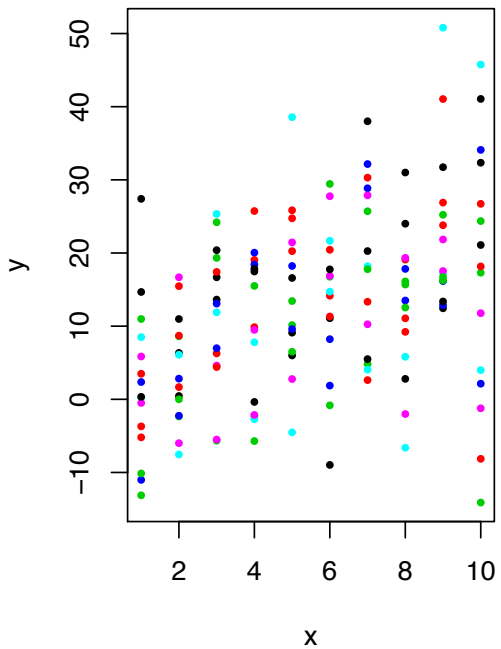
$$\mu_i = \beta_0 + x_i$$
$$y_i \sim \text{normal}(\mu_i, \sigma^2)$$
$$i = 1, ..., n$$

$$\mu_{ij} = \beta_{0_j} + x_{ij}$$
$$y_{ij} \sim \text{normal}(\mu_{ij}, \sigma^2)$$
$$\beta_{0_j} \sim \text{normal}(\mu_{\beta_0}, \sigma^2_{\beta_0})$$
$$j = 1, ..., J$$
$$i = 1, ..., n_j$$

Completely random design



Grouped design



Example: inference to population across strata

$$
\begin{aligned}
\mu_{ijk} &= \beta_{0_{jk}} + \beta_1 x_{ijk} \\
y_{ijk} &\sim \text{normal}(\mu_{ijk}, \sigma^2) \\
\beta_{0_{jk}} &\sim \text{normal}(\mu_{\beta_{0_k}}, \sigma^2_{\beta_{0_k}}) \\
j &= 1, ..., J_k \\
i &= 1, ..., n_j \\
k &= 1, ..., K
\end{aligned}
$$

Algorithm. At each MCMC interation indexed by t, draw:

$$\mathbf{p} = \left(\frac{N_1}{N}, ..., \frac{N_k}{N} \right)', \; N = \sum_{k=1}^{K} N_k \tag{1}$$

$$k^t \sim \text{categorical}(\mathbf{p}) \tag{2}$$

$$\tilde{\beta}_0^t \sim \text{normal}(\mu^t_{\beta_{0_{k^t}}}, \sigma^{2t}_{\beta_{0_{k^t}}}), \text{ New intercept} \tag{3}$$

$$\tilde{\mu}_i^t = \tilde{\beta}_0^t + \beta^t_{1k^t} \hat{x}_i, \text{ New prediction at } \hat{x}_i \tag{4}$$

$$\tilde{y}_i^t \sim \text{normal}(\tilde{\mu}_i^t, \sigma^{2t}_{k^t}), \text{ New observation at } \hat{x}_i \tag{5}$$

Strata level inference from converged MCMC

- Mean of distribution of intercepts across strata

$$\mu_{\beta_{all}} = \frac{1}{T} \sum_{i=1}^{T} \tilde{\beta_0}^t$$

- Mean of predictions of new observations across strata at specified value of covariate, $\hat{x}_i$

$$\mu_{all,i} = \frac{1}{T} \sum_{t=1}^{T} \tilde{\mu}_i^t$$

Can also compute credible intervals, standard deviations, etc.

Coding

Pseudo JAGS code:

```
#N is vector of number of potential samples for each of
#the K strata
p <- (N/sum(N))
k.new ~ dcat(p[])
beta_0.new ~ normal(mu_beta_0[k.new], sigma_beta_0^-2)
for(i in 1:x.hat){
  mu.hat[i] <- beta_0.new + beta_1[k.new]*x.hat[i]
  y.new[i] ~ dnorm(mu.hat[i], sigma[k.new])
}
```

Example: Finite population sampling

Our usual assumption is that we take a sample of n observations that is far smaller than the number of possible samples in the population we seek to understand $n << N$. However, in some cases this assumption does not hold. Proper inference requires including information on n and N in computation of means, medians, and their credible intervals but not on the "superpopulation" parameters in our model θ.

Recall the posterior predictive distribution

$$[y^{miss}|\mathbf{y}] = \int_{\boldsymbol{\theta}} [y^{miss} \mid \boldsymbol{\theta}][\boldsymbol{\theta} \mid \mathbf{y}] d\boldsymbol{\theta}$$

Can you derive this expression using rules of probability?

① $[y^{new}, \underline{\theta}, | \underline{y}] \propto [y^{new}, \underline{\theta}, \underline{y}]$

$[\underline{y}, y^{new}, \underline{\theta}] =$

$[y | y_{new}, \theta][y^{new}|\theta][\theta]$

y is conditionally independent of y^{new}

$[\underline{y}|\underline{\theta}][y^{new}|\underline{\theta}][\underline{\theta}]$

see that $[\underline{y}|\underline{\theta}][\underline{\theta}] \propto [\underline{\theta}|\underline{y}]$

& substitute

$[y^{new}|\underline{\theta}][\underline{\theta}|\underline{y}]$

from 1

$[y^{new}, \underline{\theta}, |\underline{y}] = [y^{new}|\underline{\theta}][\underline{\theta}|\underline{y}]$

Integrate out $\underline{\theta}$:

$$[y^{new}|\underline{y}] = \int_{\underline{\theta}} [y^{new}|\underline{\theta}][\underline{\theta}|\underline{y}]$$

Algorithm for Monte Carlo integration:

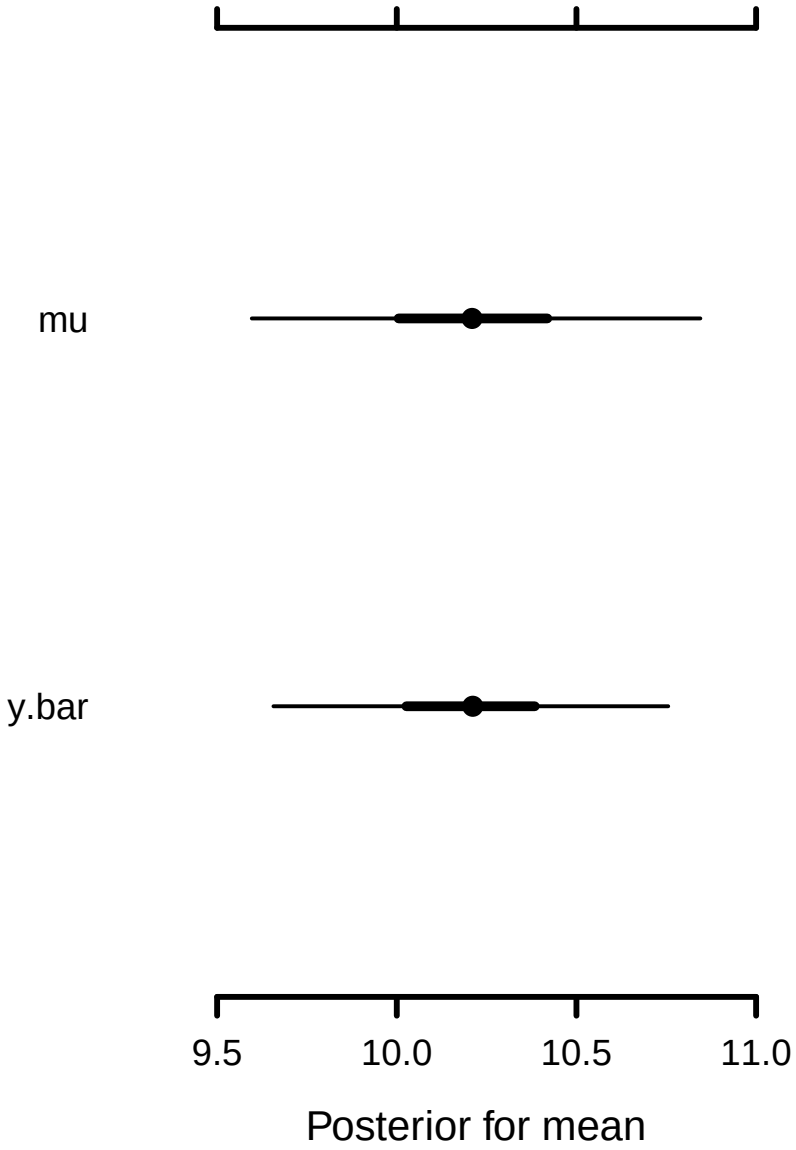
1. Fit the model to find values for θ.
2. At each MCMC iteration (indexed by t) make draws of $i = 1, ..., N - n$ *unobserved* values $y^t_{miss,i} \sim [y^t_{miss,i} \mid \theta^t, x_i]$
3. Compute $\bar{y}^t = \frac{n}{N}\text{mean}(\mathbf{y}_{obs}) + \frac{N-n}{N}\text{mean}(\mathbf{y}^t_{miss})$

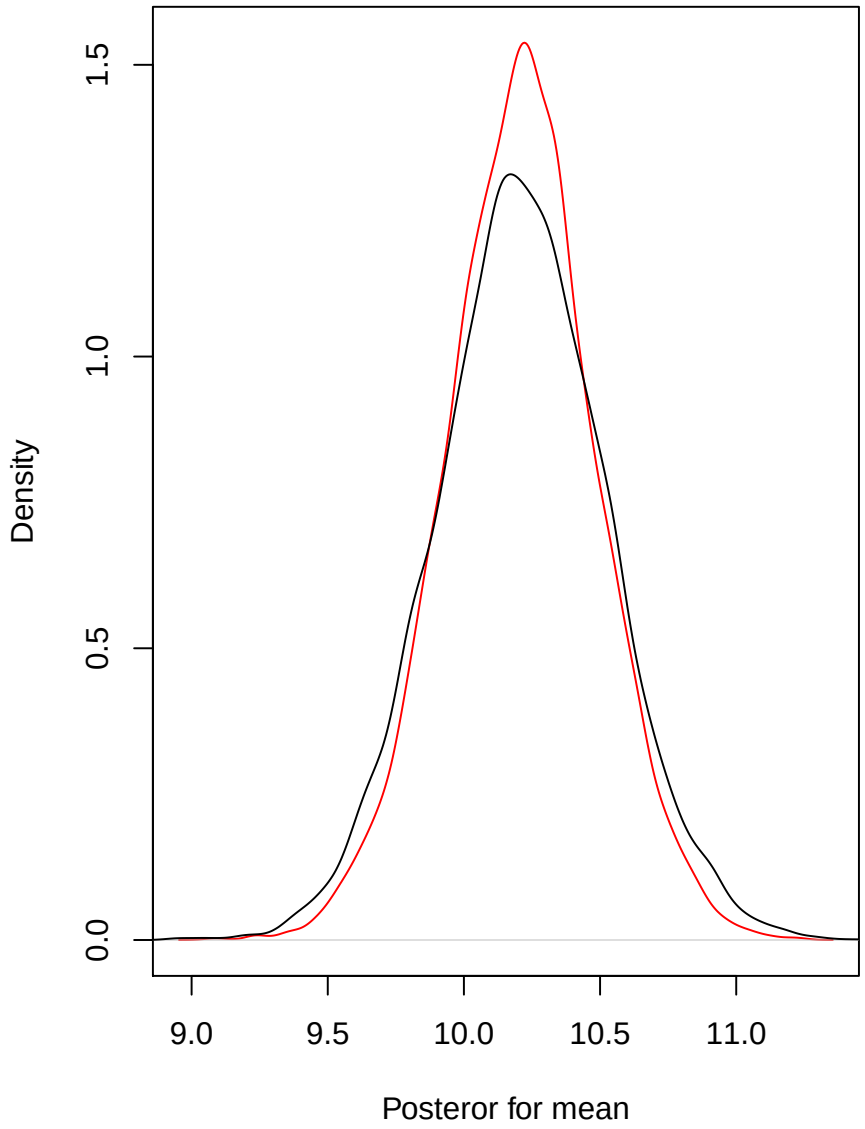
Could compute any function of the simulated and obserevd vectors of observations: median of $\mathbf{y}$, mean of $\log(\mathbf{y})$, etc.

Code for algorithm

```
{
sink("FiniteMean.R")
cat("
model{
mu ~ dnorm(0,.00001)
sigma ~ dunif(0,20)
for(i in 1:length(y.obs)){
  y.obs[i] ~ dnorm(mu, sigma^-2)
}
for(i in 1:(N-n)){
  y.miss[i] ~ dnorm(mu, sigma^-2) #simulate missing data
}
y.bar = n/N * mean(y.obs) + (N-n)/N * mean(y.miss)
} #end of model
",fill=TRUE)
sink()
}
```

Output: Note shrikage of finite mean (red line)





When do you need to consider finite samples?

If n is small and N is large, $\bar{y}$ converges on μ.

$$\bar{y}|y_{obs} \approx \text{normal}\left(\bar{y}, \frac{1}{n} - \left(\frac{1}{N}\right)\sigma^2\right)$$

You can see that N as small as 1000 with a relatively small n (50) will cause less than a 2% change in the variance.

Overarching concept: Ignorability

We start with a broad definition of "missing data".

Data are missing:

1. From a sample of a population because not all members of the population are included in the sample.
2. From a data set taken in a sample because of errors in recording, non-response in surveys, instrument failure, tag loss, etc.

The overarching question of ignorability asks "When do we need to include information about the data collection process in the model we use for analysis of the data?"

Missing data: mechanisms creating missing data are ignorable.

We now think of I as a 0 or 1 indicator of "missingness"[1] of covariates or responses in a datset.

- Missing completely at random: The probability of missingness is the same for all units (e.g. responses or covariates). Ignorable.

$$[I \mid y, x, \phi] = [I|\phi]$$

- Missing at random: The probability of missingness depends only the observed covariates. Ignorable conditional on the covariates.

$$[I \mid y_{obs}, x_{obs}, \phi] = [I|\phi, x_{obs}]$$

[1] My wife, Saran, asked "Wouldn't if be better usage to say "absence" instead of "missingness"?"

Missing data: mechanisms creating missing data are *not* ignorable.

Missing not at random: The probability of missingness depends on unobserved covariates or the responses themselves.

$$[y, x, I \mid \theta, \phi] = [y \mid \theta][I \mid x_{unobs}, \phi]$$

Examples

Imagine a series of plots with automated gas-flux analyzers. We are interested in carbon dioxide emissions in response to a suite of covariates including plant biomass, soil water, and the other usual suspects. There is a nearby fire. Some of the plots burn because windblown embers land on them. We loose a week of data in those plots. Can we ignore the lost data?

- *Missing compeletly at random*: The only thing that determines if an analyzer was lost is whether a spark landed on the plot. There is no spatial pattern in burned plots. These can be viewed as ignorable, demonic intrusions.
- *Missing at random*: Grass biomass and soil water influenced whether a plot burned. Low biomass, high moisture plots did not lose analyzers. Ignorable because covariate account for probability of missingness.
- *Missing not at random* Plots near the fire burn. Others don't. There is a spatial gradient in the lost analyzers.

Exercise

What could we do to make the third case ignorable?

Exercise

- Some analyzers fail because their connections to batteries are chewed by field mice (really). Are the missing data ignorable?
- Some analyzers have intake vents blocked by grass, causing data to be missed. Are the missing data ignorable?
- Lighting strikes a single plot with an analyzer, sizzling the instrument. (This is probably not ignorable in at least some sense.)
- Lighting strikes the grid of plots, rendering them all ineffective for a week. Inorable or not?

How do you know which type of missinginess?

- You probably don't care if there are a small number of missing values relative to total number of observations,
- How do you know if missing data not ignorable (i.e, missing not at random)?
  - Remember, you have a data vector I for all responses and covariates.
  - Prepare bivariate plots of I against covariates, Box plots of frequencies of missingess in different sites, strata, or experimental treatments, plots of pattern of missingness against spatial coordinates.
  - Knowledge of your subject matter. Classification example.

Modeling ignorable missing data

- Responses: Including NA's in the data will automatically model them because the likelihood is a model of the responses.
- Covariates: Cannot be NA. However, when ignorabale they can be modeled as

$$x_i \sim [x_i \mid \mu_x, \sigma_x^2] \text{ or:, better for multiple covariates}$$

$$\mathbf{x}_i \sim \text{multivariate normal}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma})$$

. Remember, the parameters of these distributions must have priors.The multivariate case can be complicated when covariates have different support. You will need vectors of transformed (log or logit) values of the covariates.

We must explicitly model non-ignorable missing data when:

A "substantial" fraction of the observations ($> 5\%$?) *and* when observations of responses or covariates are missing not at random.

Example: Modeling response data missing not at random

Example 1: Modeling response data missing not at random.

We are investigating the height of willows adjacent to streams in 2 x 2 factorial experiment with four blocks. Treatments are simulated beaver dams and fences. An graph of missing data ($I = 0$ if missing, 1 otherwise) shows that missing values tend to increase with height. We hypothesize that missingness is positively related to height because tall willows exist in dense clumps making it difficult for field technicians to find tags on stems.

Model

$$
\begin{aligned}
g(\boldsymbol{\beta}, \mathbf{x}_i, t) &= \beta_{0j} + \beta_1 t + \beta_1 x_1 + \beta_3 x_2 + \beta_4 x_1 x_2 \\
y_{ijt} &\sim \text{lognormal}\left(\log(g(\boldsymbol{\beta}, \mathbf{x}_i)), \sigma_j^2\right) \\
\beta_j &\sim \text{normal}(\mu_{\beta_0} \sigma_{\beta_0}^2) \\
h(\boldsymbol{\gamma}, y_{ijt}) &= \text{inverse logit}(\gamma_0 + \gamma_1 y_{ijt}) \\
I_{ijt} &\sim \text{Bernoulli}(h(\boldsymbol{\gamma}, y_{ijt})) \\
[I, \boldsymbol{\beta}, \boldsymbol{\gamma} \mu_{\beta_0}, \sigma_{\beta_0}^2, \boldsymbol{\sigma}^2 \mid \mathbf{y}] &\propto \prod_{i=1}^{n_j} \prod_{j=1}^{J} \prod_{t=1}^{T} [I_i \mid \boldsymbol{\gamma}, y_{ijt}][y_{ijt} \mid \log(g(\boldsymbol{\beta}, \mathbf{x}_i, t)), \sigma_j^2] \\
&\times \text{priors}
\end{aligned}
$$

Modeling response data missing not at random

We often need informed priors (at least weakly informred) on the parameters in the missingness model.

$$h(\boldsymbol{\gamma}, y_{ij}) = \text{inverse logit}(\gamma_0 + \gamma_1 y_{ij})$$
$$I_{ij} \sim \text{Bernoulli}(h(\boldsymbol{\gamma}, y_{ij}))$$

The $\gamma's$ can be informed by calibration studies of missingness.

Example 2: Modeling response data missing not at random

We have data on herbaceous canopy gap size (cm) on transects of fixed length. The data at the ends of the transect are *censored*. We know these gaps are at least as large as the distance from the transect end to the next nearest edge of vegetation. They are missing not at random because the response influences the missingness.

```
#likelihood for canopy gaps
for(i in 1:length(y)){
  mu[i] <- exp(B0[y.site.index[i]] + B1*x[i])
  y.isCensored[i] ~ dinterval( y[i]] , y.gapLimit[i] )
  y[i] ~ dlnorm(log(mu[i]), tau)
}
```

See http://doingbayesiandataanalysis.blogspot.com/2012/01/complete-example-of-right-censoring-in.html for a more detailed treatment of modeling censored data. See the .pdf "Math_for_ censored_data.pdf" in the repository for writing the posterior and joint distributions.

Best practices

- Always explore missing values if you have many of them. Do the plots described above and think about how missing values might arise.
- It is good to model missing $x's$, even if their missingness is ignorable, to avoid throwing away data when there is a single missing value in a vector of values.
- Describe how you treated missing data in your paper. Include missing data models in your posterior and joint distributions if you used them for inference. Justify your assumption of missing completely at random of missing at random if necessary.
- You may need to conduct calibration studies to inform priors in missing data models.

Further study of ignorability and missing data

- A. Gelman, J. B. Carlin, H. S. Stern, D. Dunson, A. Vehhtari, and D. B. Rubin. Bayesian data analysis. Chapman and Hall / CRC, London, UK, 2013. Chapter 8.
- A. Gelman and J. Hill. Data analysis using regression and multilievel / hierarchical modeling. Cambridge University Press, Cambridge, UK, 2009. Chapter 25
- http://www.columbia.edu/~cjd11/charles_dimaggio/DIRE/styled-4/styled-11/code-10/
- http://www.bias-project.org.uk/Missing2012/Lectures.pdfhttp://www.bias-project.org.uk/Missing2012/MissingIndex.htm
- http://www.bias-project.org.uk/Missing2012/Lectures.pdf
- https://web.as.uky.edu/statistics/users/pbreheny/701/S13/notes/4-23.pdf
- W. A. Link and R. J. Barker. Bayesian inference with ecological applications. Academic Press, 2010. Section 8.5