MISSING DATA AND MEASUREMENT ERROR

September 22, 2019

Jon Fintzi

Biostatistics Research Branch
National Institute of Allergy and Infectious Diseases
National Institutes of Health

Overview

- Bayesian inference *always* starts with a model for the **joint distribution** of θ and y:.

$$\pi(\theta, y) = f(y|\theta)\pi(\theta) = \pi(\theta|y)m(y).$$

- **Bayes rule** yields the **posterior distribution**

$$\pi(\theta|y) = \frac{f(y,\theta)}{m(y)} = \frac{f(y|\theta)\pi(\theta)}{m(y)} \propto Likelihood \times Prior.$$

.

Last week we talked about hierarchical models, all we did was iterate on this ideas:

- Model expressed that people are self-similar, but also are similar to one another.
- Individuals are *exchangeable* in the prior - reasonable to suppose that β_{Jon} and β_{Mike} come from the same distribution, but no prior information to differentiate Mike from Jon.
- We use the data to inform us about individuals *and* the population, individuals are no longer exchangeable in the posterior, i.e., $\pi(\beta_{Jon}|Y_{Jon}, Y_{Mike}) \neq \pi(\beta_{Mike}|Y_{Jon}, Y_{Mike})$.
- Different choices for model structure induce different features in the posterior, e.g., shrinkage with “mixed effects”, horseshoe for inducing posterior sparsity.

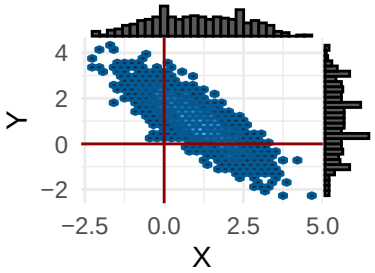
Example: divorce rate vs. state population size.

- Treat true divorce rate as an unknown parameter:

$$D_{obs,i} \sim N(D_{true,i}, D^2_{SE,i})$$
$$D_{true,i} \sim \pi(\theta).$$

- Effect is to shrink observed state divorce rates towards national average.
- If interested in divorce rate vs. population + marriage rate, can also model observed marriage as a noisy observation of the true marriage rate.
- Missing data is a form of measurement error.

- Common approaches to missing data:
  - Complete case analysis - (best case) introduce uncertainty, (worst case) introduce confounding.
  - Mean imputation, marginal imputation.



- Multiple imputation: simulate datasets from joint distribution, fit separately, and combine.
- Bayesian data augmentation: introduce missing data, Y_{miss} as latent variables. Target the joint posterior $\pi(Y_{miss}, \theta | Y_{obs})$.

Lecture 20 of Statistical Rethinking

- Different missingness mechanisms, MCAR, MAR, and MNAR, require different models.
- Imputation can improve precision for estimates of interest (shrinkage!).
- Bayesian inference always starts with a *joint* model for data, parameters, and covariates.

Plan for today

Two examples:

- Model BMI as a function of cholesterol and age.
  - Data augmentation with `brms` (Burkner, 2019).
  - Off-the-shelf, flexible, relatively straightforward syntax.
- Compartmental models for partially observed incidence data.
  - Introduce true incidence as a latent variable.
  - Ordinary differential equations describe the latent incidence.

Example: BMI vs. Cholesterol

Data (`nhanes` from the `mice` package)

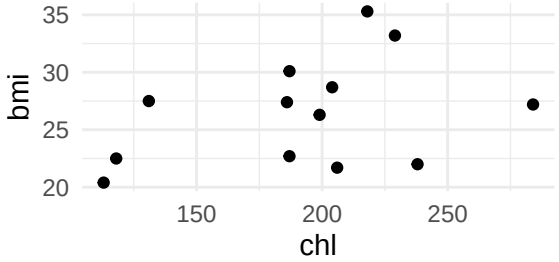
- 18 individuals, omit people missing both BMI and cholesterol.
- BMI (kg/m^2)
- Total serum cholesterol (mg/dL)

```
##   chl  bmi
## 2 187 22.7
## 3 187   NA
## 5 113 20.4
## 6 184   NA
## 7 118 22.5
## 8 187 30.1
```

Example: BMI vs. Cholesterol

Key features:

- Missingness in cholesterol and BMI, we'll assume MAR so need to impute but not model missingness (see Statistical Rethinking lecture 20 for the explanation of this).
- Looks like higher BMI associated with slightly higher cholesterol.



Example: BMI vs. Cholesterol

Model:

$$
\begin{aligned}
BMI_{obs,i} &\sim LogNormal(\mu_{bmi,i}, \sigma^2_{bmi}) \\
BMI_{miss,i} &\sim LogNormal(\mu_{bmi,i}, \sigma^2_{bmi}) \\
\mu_{bmi,i} &= \beta_0 + \beta_1 CHL_i \\
CHL_{obs,i} &\sim Normal(\mu_{chl}, \sigma^2_{chl}) \\
CHL_{miss,i} &\sim Normal(\mu_{chl}, \sigma^2_{chl}) \\
&+ Priors \ldots
\end{aligned}
$$

- MAR $\implies$ observed and missing variables exchangeable in the prior.
- If MNAR, have to model probability of missing given latent value (Chapters 8 and 18 of Gelman et al., 2013).
- If we fit this in `Stan`, declare observed values as data and missing values as parameters, which we estimate just like any other parameters.

Interlude: Algorithmic Implementation

Example - normal means problem with missing values: $y_i \sim N(\mu, \sigma^2)$.

```
data {
  int<lower=0> N_obs; # number observed
  int<lower=0> N_mis; # number missing
  real y_obs[N_obs];  # vector of observed values
}
```

Interlude: Algorithmic Implementation

Example - normal means problem with missing values: $y_i \sim N(\mu, \sigma^2)$.

```
parameters {
  real mu;              # mean parameter
  real<lower=0> sigma;  # standard deviation
  real y_mis[N_mis];    # missing values are parameters
}
```

Interlude: Algorithmic Implementation

Example - normal means problem with missing values: $y_i \sim N(\mu, \sigma^2)$.

```
model {
  # joint distribution for observed and missing variables
  y_obs ~ normal(mu, sigma);
  y_mis ~ normal(mu, sigma);
}
```

Example: BMI vs. Cholesterol

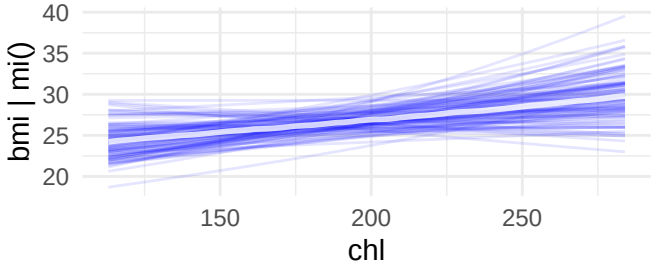
Trivial to fit using `brms`:

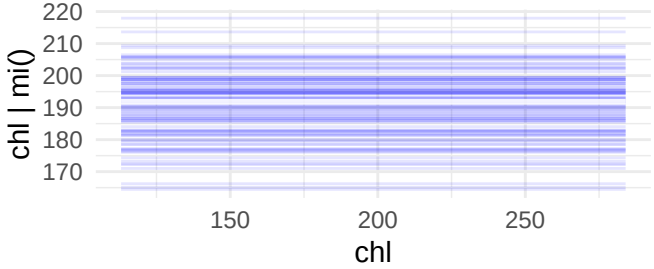
```
# model formula, mi() indicates that missing values should be estimated
bform <-
    bf(bmi | mi() ~ 1 + mi(chl), family = "lognormal") +
    bf(chl | mi() ~ 1) + set_rescor(FALSE)

# call to fit the model
nhanes_fit <- brm(formula = bform,
                  data = nhanes,
                  prior = prior(student_t(3,0,5), class = "b"), # change priors
                  refresh = 0) # silent compilation and fitting
```

Example: BMI vs. Cholesterol

Posterior is full of lines for BMI vs. cholesterol and values for cholesterol.





Example: BMI vs. Cholesterol

Interrogate the posterior predictive distribution to examine fit.

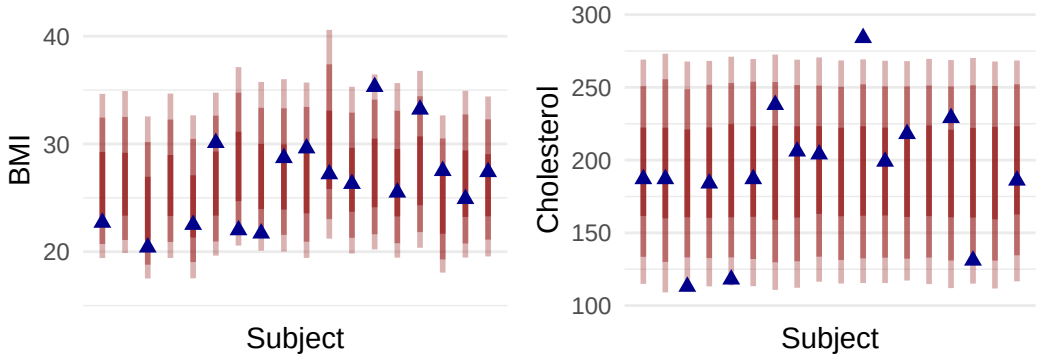
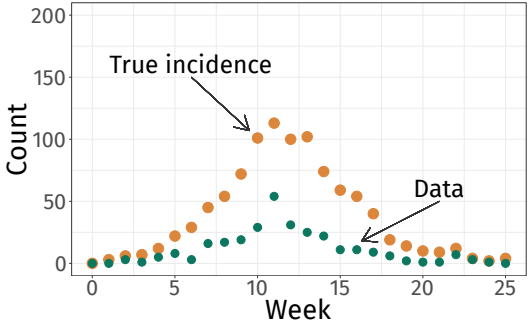


Figure 1: Posterior predicted BMI and cholesterol.

Example: Partially Observed Epidemic Count Data

Parially observed incidence data:

- $N_{SI}(t_\ell) =$ Cumulative infections up to t_ℓ,
- $Y_\ell =$ new cases seen in $(t_{\ell-1}, t_\ell]$,
- $Y_\ell \sim Neg.Binomial(\mu = \rho \times (N_{SI}(t_\ell) - N_{SI}(t_{\ell-1})),\ \sigma^2 = \mu(1 + \mu/\phi))$..



Example: Partially Observed Epidemic Count Data

Important:

- Only observe a *fraction of cases* at *discrete times*.
- Data come from an outbreak that evolves *continuously* in time.

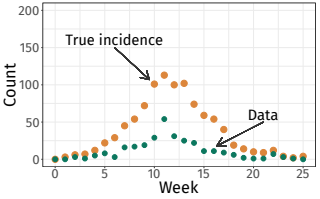
What do we want to learn?

- How many people were infected? How many people were infected?
- How to characterize the transmission dynamics of the outbreak?

Example: Partially Observed Epidemic Count Data

What makes this difficult?

1. *Under-reporting:* epidemic process, $\mathbf{X}$, only partially observed.



2. *Dependent happenings:* $\implies$ dependent data, $\mathbf{Y} = (\mathbf{Y_1}, \ldots, \mathbf{Y_L})$.
   - Observed data likelihood:

$$L(\mathbf{Y}|\boldsymbol{\theta}) = \prod_{\ell=\mathbf{1}}^{\mathbf{L}} \pi(\mathbf{Y}_\ell|\mathbf{Y_1}, \ldots, \mathbf{Y}_{\ell-\mathbf{1}}, \boldsymbol{\theta}) \neq \prod_{\ell=\mathbf{1}}^{\mathbf{L}} \pi(\mathbf{Y}_\ell|\boldsymbol{\theta}).$$

   - *Intractable observed data likelihood* State space of $\mathbf{N}$ is *huge*, even in small populations!

$$L(\mathbf{Y}|\boldsymbol{\theta}) = \int \prod_{\ell=\mathbf{1}}^{\mathbf{L}} \pi(\mathbf{Y}_\ell|\mathbf{Y_1}, \ldots, \mathbf{Y}_{\ell-\mathbf{1}}, \mathbf{N}, \boldsymbol{\theta}) \pi(\mathbf{N}|\boldsymbol{\theta}) \mathbf{dN}$$

Example: Partially Observed Epidemic Count Data

Strategy:

- *Bayesian data augmentation* - introduce incident event processes, $\mathbf{N} = (\mathbf{N_{SI}}, \mathbf{N_{IR}})$, as latent variables in the model.
- Target the joint posterior, $\pi(\mathbf{N}, \boldsymbol{\theta} | \mathbf{Y})$.

Challenge: Need a tractable representation for the transition density of $\mathbf{N}(\mathbf{t}_\ell) | \mathbf{N}(\mathbf{t}_{\ell-\mathbf{1}}), \boldsymbol{\theta}$.

- In large populations, not unreasonable to represent $\mathbf{N}$ with a deterministic system of ODEs.
- Classical tools in the disease modeling literature, see Allen (2008) and Blackwood (2018) for an overview.

Example: Partially Observed Epidemic Count Data

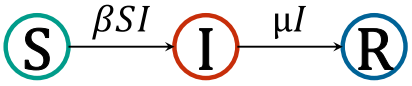
Deterministic SIR model:

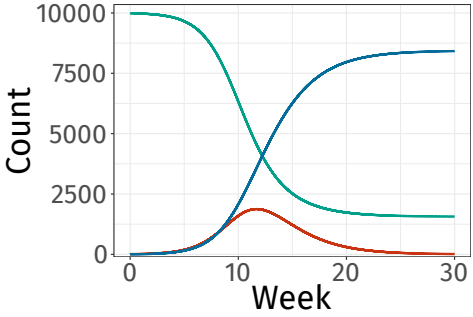
Incidence paths are solutions to systems of differential equations,

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} N_{SI} \\ N_{IR} \end{pmatrix} = \begin{pmatrix} \beta SI \\ \mu I \end{pmatrix},$$
$$= \begin{pmatrix} \beta(S_0 - N_{SI})(I_0 + N_{SI} - N_{IR}) \\ \mu(I_0 + N_{SI} - N_{IR}) \end{pmatrix},$$

subject to $\mathbf{X_0} = (\mathbf{S_0}, \mathbf{I_0}, \mathbf{R_0})$, $\mathbf{N_0} = \mathbf{0}$.

- β = per-contact infection rate.
- μ = recovery rate.
- Priors on $1/\mu$ = mean infectious period and $\mathcal{R}_0 = \beta N/\mu =$ basic reproduction number.





Example: Partially Observed Epidemic Count Data

Joint model, $\pi(\mathbf{Y}, \mathbf{N}, \boldsymbol{\theta})$, where $\mathbf{N}$ has the *Markov* property.

- Data, $\mathbf{Y}$, are conditionally independent given $\mathbf{N}$.
- *Simplified complete data likelihood:*

$$L(\mathbf{Y}, \mathbf{N}|\boldsymbol{\theta}) = \pi(\mathbf{N}(\mathbf{t_0})|\boldsymbol{\theta}) \prod_{\ell=1}^{\mathbf{L}} \pi(\mathbf{Y}_\ell|\mathbf{N}(\mathbf{t}_\ell), \boldsymbol{\theta}) \pi(\mathbf{N}(\mathbf{t}_\ell)|\mathbf{N}(\mathbf{t}_{\ell-1}), \boldsymbol{\theta}).$$

  - $\pi(Y_\ell|\mathbf{N}(\mathbf{t}_\ell), \boldsymbol{\theta})$ – sampling model, negative binomial.
  - $\pi(\mathbf{N}(\mathbf{t}_\ell)|\mathbf{N}(\mathbf{t}_{\ell-1}), \boldsymbol{\theta})$ – transition density for latent epidemic, SIR

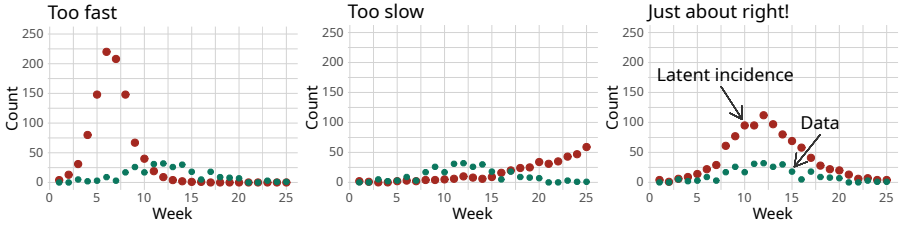
- Here, $\boldsymbol{\theta}$ maps 1:1 onto $\mathbf{N}$ so no need to sample $\mathbf{N}$ explicitly.
- Stochastic representations of $\mathbf{N}$ require sampling latent paths. Tradeoff realism and compuational tractability.

Key point: true incidence is missing data. In the Bayesian paradigm we estimate it like any other parameter by including it in our joint model and targeting the posterior!

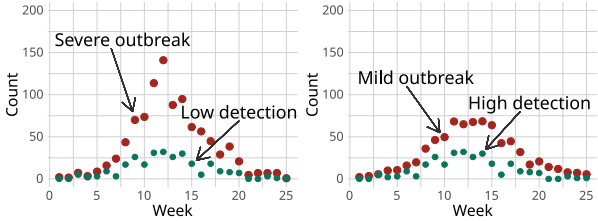
Example: Partially Observed Epidemic Count Data

Goal: Infer $\pi(\boldsymbol{\theta}, \mathbf{N}|\mathbf{Y}) \propto \mathbf{L}(\mathbf{Y}|\mathbf{N}, \boldsymbol{\theta})\pi(\mathbf{N}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

- **Outbreak dynamics:** $\pi(\mathbf{N}|\boldsymbol{\theta})$



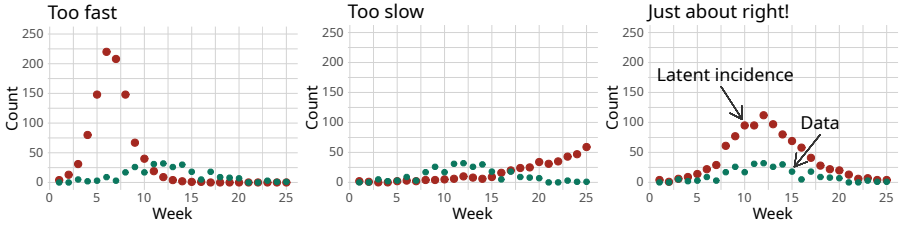
- **Observation model:** $L(\mathbf{Y}|\mathbf{N}, \boldsymbol{\theta})$



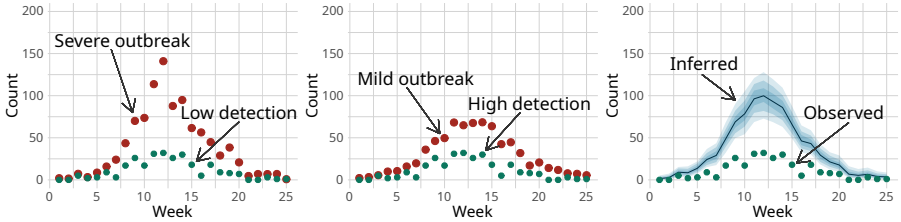
Example: Partially Observed Epidemic Count Data

Goal: Infer $\pi(\boldsymbol{\theta}, \mathbf{N}|\mathbf{Y}) \propto \mathbf{L}(\mathbf{Y}|\mathbf{N}, \boldsymbol{\theta})\pi(\mathbf{N}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

- **Outbreak dynamics:** $\pi(\mathbf{N}|\boldsymbol{\theta})$



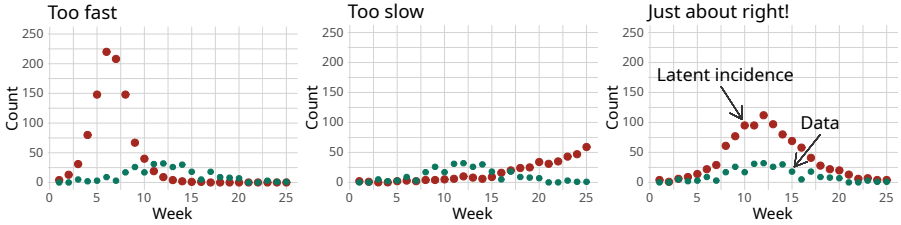
- **Observation model:** $L(\mathbf{Y}|\mathbf{N}, \boldsymbol{\theta})$



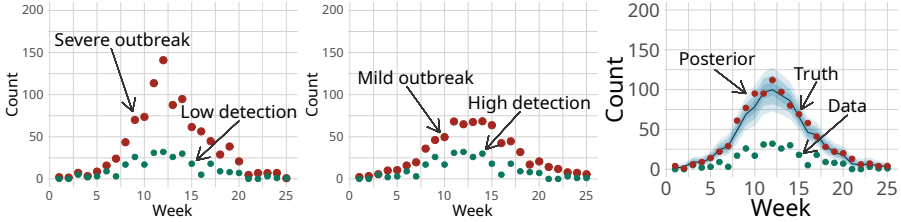
Example: Partially Observed Epidemic Count Data

Goal: Infer $\pi(\boldsymbol{\theta}, \mathbf{N}|\mathbf{Y}) \propto \mathbf{L}(\mathbf{Y}|\mathbf{N}, \boldsymbol{\theta})\pi(\mathbf{N}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

- **Outbreak dynamics:** $\pi(\mathbf{N}|\boldsymbol{\theta})$

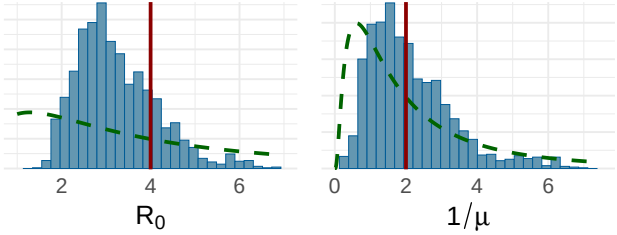


- **Observation model:** $L(\mathbf{Y}|\mathbf{N}, \boldsymbol{\theta})$



Example: Partially Observed Epidemic Count Data

Posterior distributions of model parameters:



Wrapping up

- Quantify two kinds of uncertainty, epistemic, which reflects subjective ignorance, and aleatory, which is uncertainty due to chance.
- A Bayesian model **always** defines a joint distribution for data and parameters.
- Some simple examples, PREVAIL II and linear regression; some complex hierarchical models and missing data.
- Failure modes of misspecified priors under poorly chosen scales, weakly informative priors as a reasonable strategy.
- Good workflow is like going to the dentist.
- Various computational tools.

References

Allen, Linda JS. "An introduction to stochastic epidemic models." Mathematical epidemiology. Springer, Berlin, Heidelberg, 2008. 81-130.

Blackwood, Julie C., and Lauren M. Childs. "An introduction to compartmental modeling for the budding infectious disease modeler." *Letters in Biomathematics* 5.1 (2018): 195-221.

P. Burkner. "Handle Missing Values with brms." `https://cran.r-project.org/web/packages/brms/vignettes/brms_missings.html` (2019).

Gelman, Andrew, et al. Bayesian data analysis. Chapman and Hall/CRC, 2013.