Missing data

Patrick Breheny

April 23

Introduction

- Our final topic for the semester is missing data
- Missing data is very common in practice, and can occur for a variety of reasons: patients skip hospital visits, investigators run out of time, investigators run out of money, experiments fail, subjects refuse to answer questions or permit tests, etc.
- There is an extensive literature on the topic of missing data and a wide variety of methods and approaches; although we cannot hope to do justice to the entire topic in a week, we can illustrate some of the issues and how they can be addressed in a Bayesian MCMC framework

Missing data in the Bayesian paradigm

- In Bayesian statistics, there is no fundamental difference between "data" and "parameters"; both are random, the only difference is that data are observable whereas parameters are not
- Missing data, then, is simply another unobserved quantity in the model – it requires a prior and will have a posterior distribution
- The "requires a prior" remark is more complicated than it sounds, though – depending on the type of missing data, we may need to give a lot of thought to modeling the missing data mechanism
- Furthermore, we do not typically model the distribution of covariates; this adds an additional layer of complexity when covariates are missing

Missing data models

- The complexity of these missing-data models can vary considerably depending on the mechanism that may have resulted in missing data
- In the simplest case, it may be reasonable to assume $y_i \overset{\text{iid}}{\sim} p(\boldsymbol{\theta})$
- Somewhat more complicated, it may be the case that the value of the missing data depends on other values recorded for that observation: $y_i \sim p(\boldsymbol{\theta}, \mathbf{x}_i)$
- Still more complicated, the probability that an observation is missing may depend on the missing value itself (*e.g.*, obese individuals may be less likely to report their weight, minorities may be less likely to report their ethnicity), or upon unobserved covariates

Imputation

- The most common approach to handling missing data is probably to throw out observations with missing values, but this has two large drawbacks:
  - Doing so throws away information and reduces efficiency
  - Doing so may introduce selection bias
- Another approach, also common and usually better, for handling missing data is *imputation*: filling in missing values with reasonable guesses as to what the value may have been
- Imputation methods range from simple (fill in missing values with the mean) to complex, model-based approaches, and may be deterministic or random

Multiple imputation

- A refinement of "single" imputation is *multiple imputation*, in which missing observations are repeatedly replaced with random values (typically from the predictive distribution of a model), creating several imputed data sets in which the observed values remain the same but the missing data varies from data set to data set
- In a sense, this is similar to what occurs in a Bayesian model fit using Gibbs sampling: with each iteration, missing data are drawn conditional on the current values of $\boldsymbol{\theta}$, then $\boldsymbol{\theta}$ is drawn from its full conditional given the current ("imputed") values of the missing data
- Multiple imputation, however, requires extra adjustments to simultaneously account for uncertainty in the parameter estimates and uncertainty in the values of the missing data (and does not entirely account for all sources of uncertainty)

Introduction

- We begin with the case of missing responses and illustrate using the CD4 data
- One can consider this data set to contain missing values in the sense that some subjects do not have recordings for all visits 1, 4, 7, 10, 13, 16, and 19
- We will analyze this data two ways: the first assuming that the missingness is essentially "ignorable", and the second assuming that missing values are more likely to have low CD4 percentages

New data structure

Thus, instead of:

Visit	ID	CD4Pct
1	2	1.00
4	2	0.30

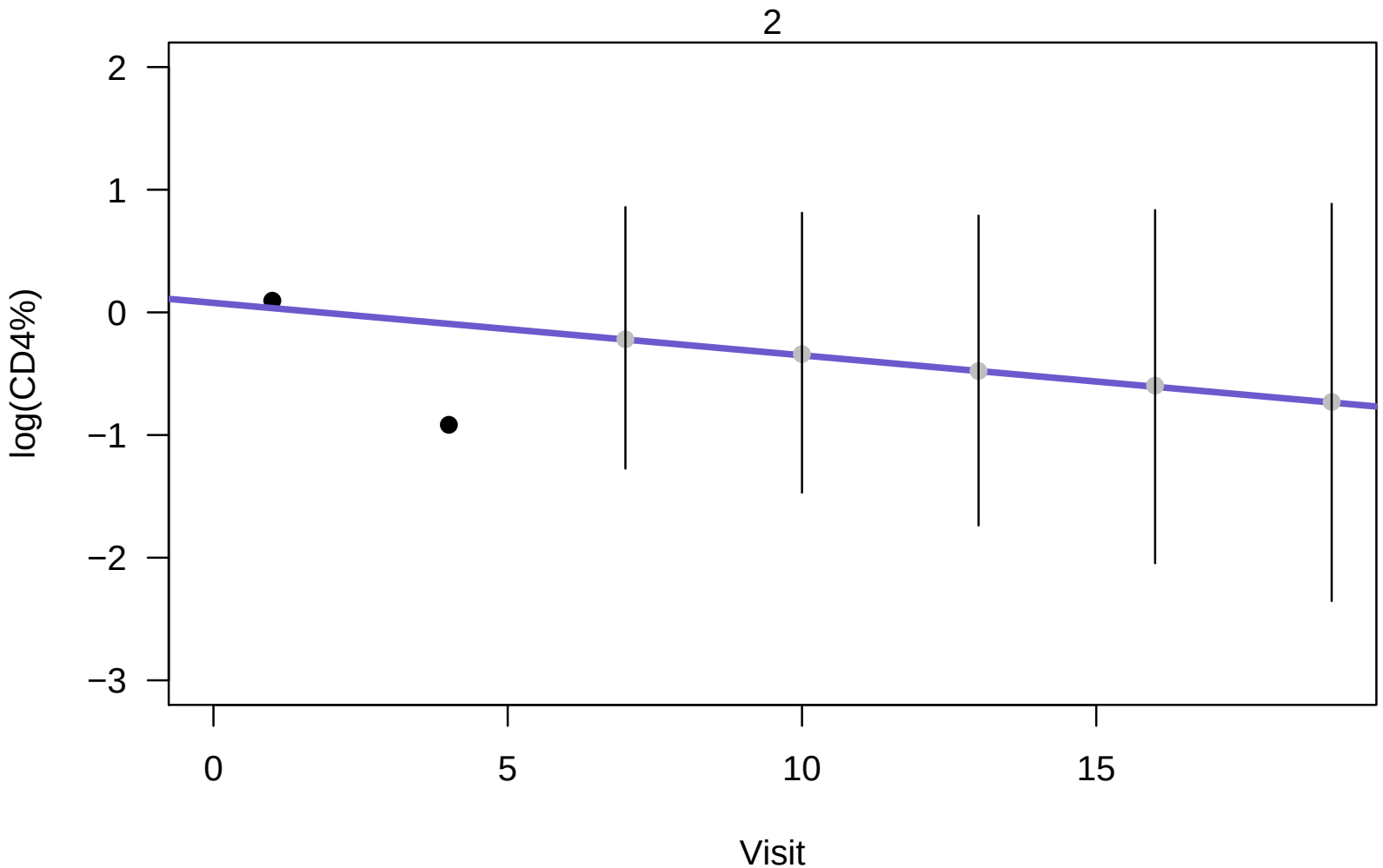
we now have

Visit	ID	CD4Pct
1	2	1.00
4	2	0.30
7	2	NA
10	2	NA
13	2	NA
16	2	NA
19	2	NA

Results

- First, we consider fitting the exact same model as in the assignment (model 3, specifically)
- Not surprisingly, we obtain the exact same posterior; the only difference is that (if we choose to monitor them), JAGS returns draws from the posterior predictive distribution of y for the missing observations
- Be careful, however, in monitoring all of y, as it can be a considerable computational burden to store an enormous array of y values, especially when the majority of values are identical from one iteration to the next

Subject 2



Posterior for γ_β

For future reference, here is the posterior for γ_β, the slope (in years)

	Median	2.5%	97.5%
$\gamma_{\beta 1}$	-0.23	-0.35	-0.11
$\gamma_{\beta 2}$	0.04	-0.13	0.21

In other words, the control group had average slope -0.23, while the treatment group had average slope -0.19

Informative missingness

- Alternatively, it is possible that patients drop out of the study because they are too sick to continue, or have died
- Let `Miss` indicate whether or not y is missing for a given visit; consider adding the following logistic regression component to our model:

$$\begin{aligned} \texttt{Miss}_i | \pi_i &\sim \text{Binom}(1, \pi_i) \\ \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \gamma_{m1} + \gamma_{m2} y_i \\ \gamma_{m1} &\sim \text{N}(0, 10000) \\ \gamma_{m2} &\sim \text{N}(0, 1/10) \end{aligned}$$

Remarks

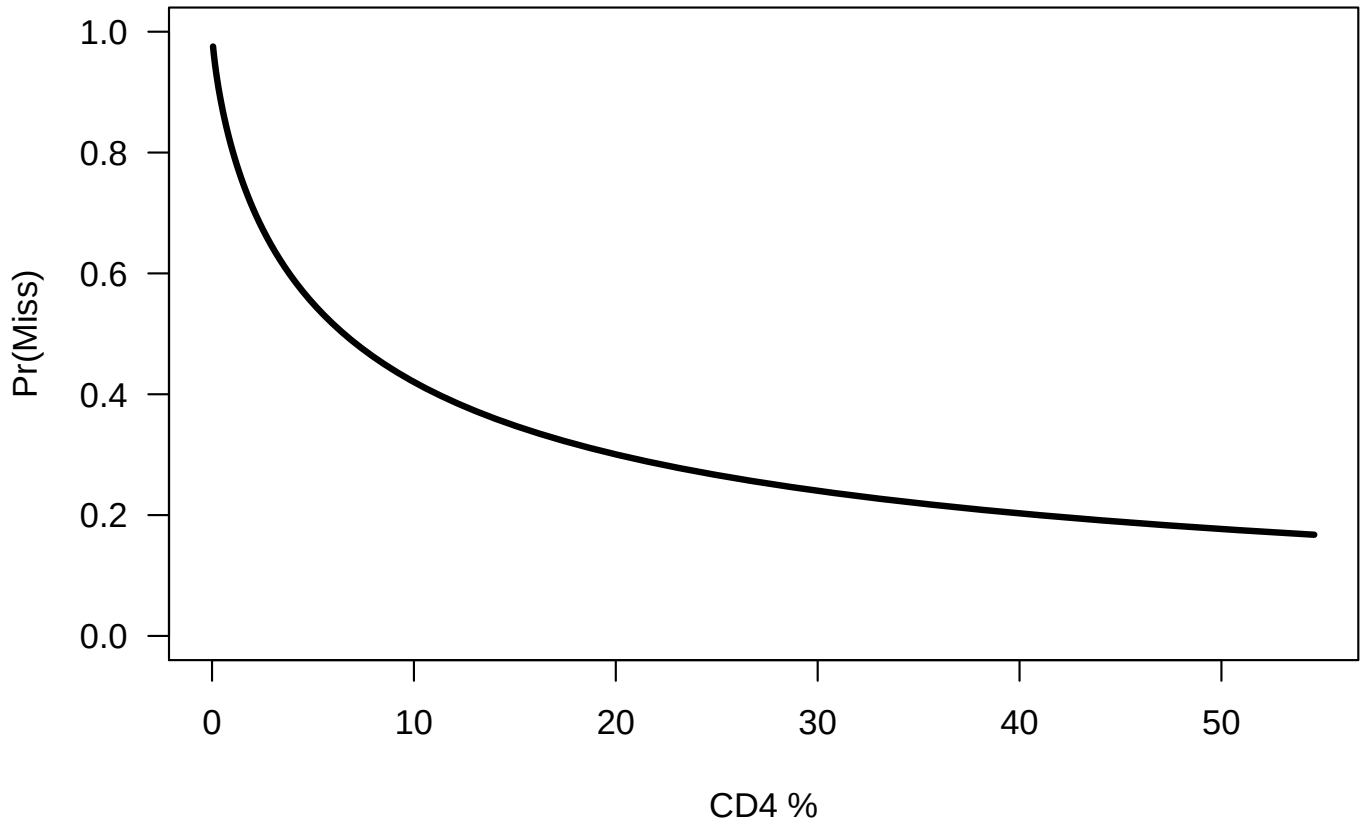
- Under the new model, the fact that an observation is missing may potentially inform us as to what its value may have been
- Note that we could not possibly fit this model "classically" without imputing y, as y would be missing (by definition) in all cases where the outcome equals 1
- The model depends on "imputing" values for y, although note that this is done dynamically here in the sense that the imputation informs us about the missingness parameters ($\boldsymbol{\gamma}_m$), which in turn informs us about the missing values of y, which affects $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, which affect the next imputation, etc.

Code

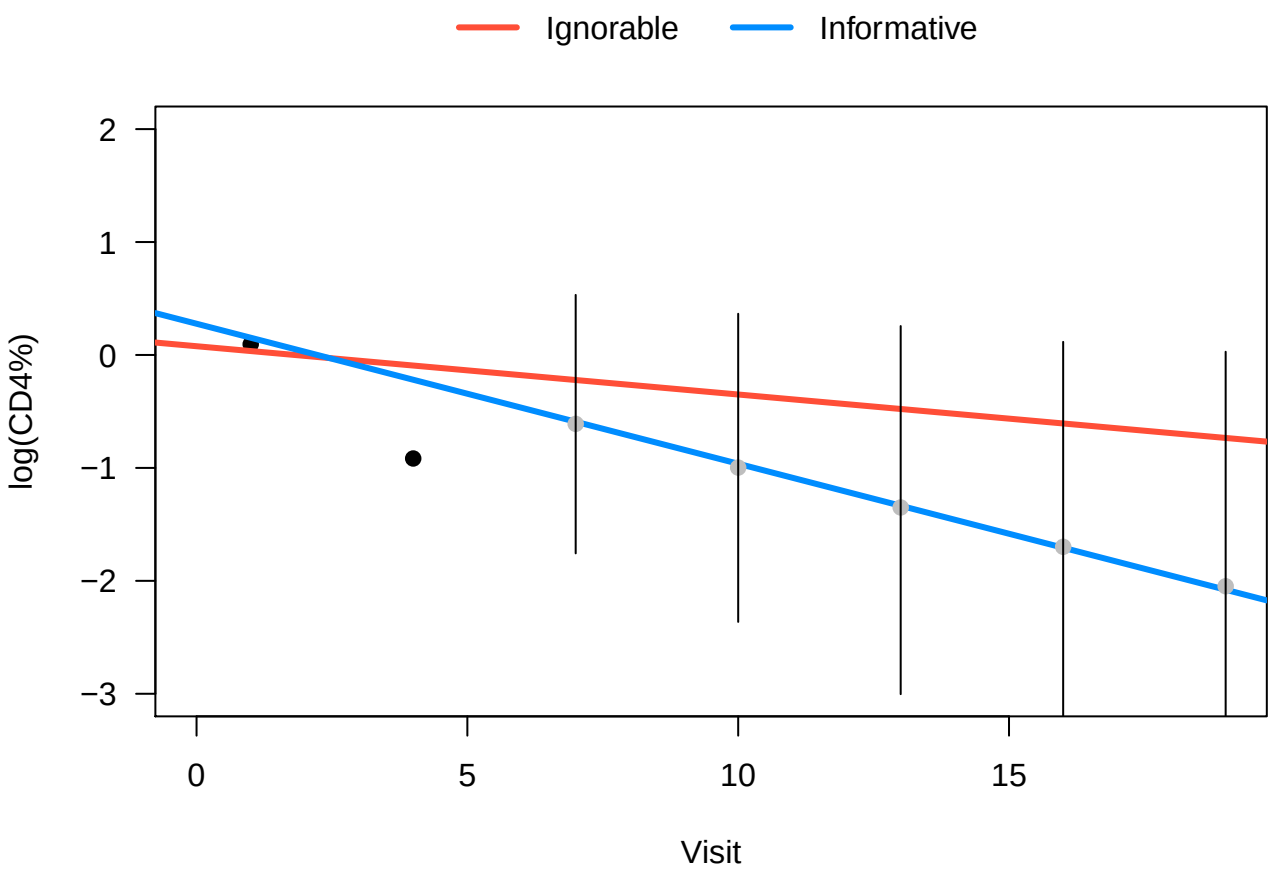
```
## Visit level
for (i in 1:n) {
  y[i] ~ dnorm(a[ID[i]] + b[ID[i]]*x[i], sigma[1]^(-2))
  Miss[i] ~ dbern(p[i])
  logit(p[i]) <- gm[1] + gm[2]*y[i]
}
gm[1] ~ dnorm(0, 0.0001)
gm[2] ~ dnorm(0, 10)
```

Missing-data model

Posterior mean for γ_{m2} is -0.76, with 95% interval (-0.88, -0.61)



Subject 2



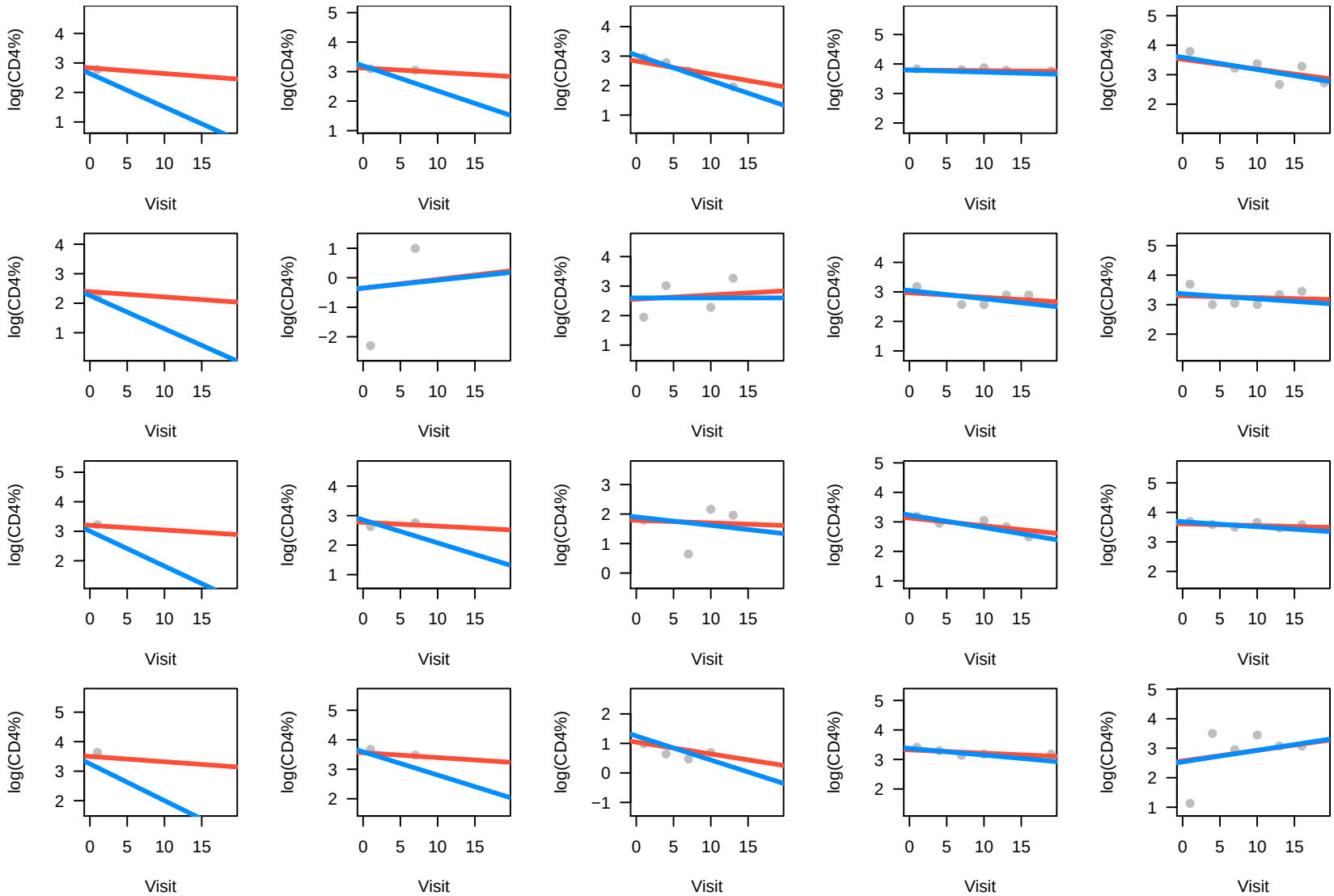
Posterior for γ_β

Under the informative missingness model, we have the following posterior for the population-level slope parameters:

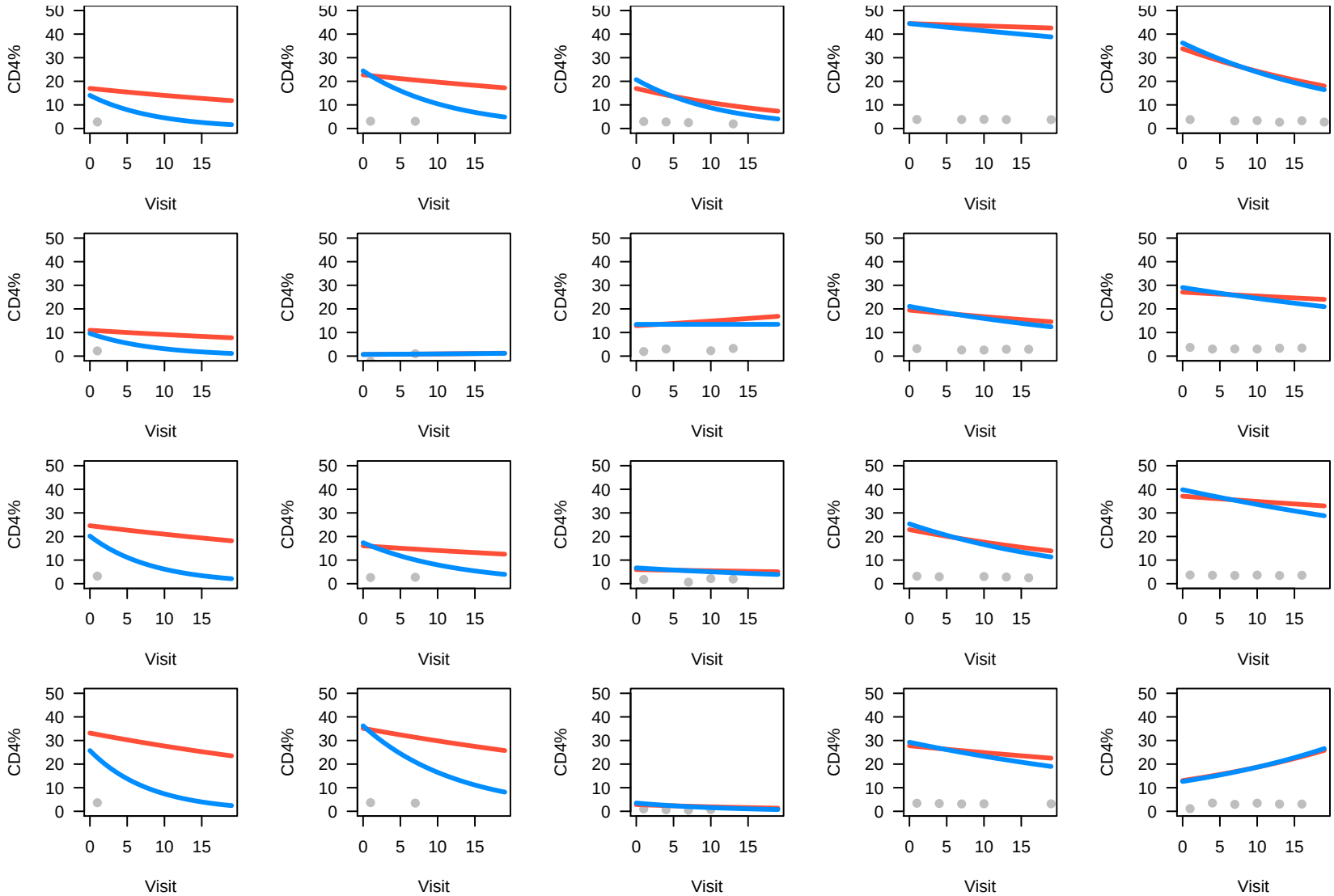
	Median	2.5%	97.5%
$\gamma_{\beta 1}$	-0.59	-0.77	-0.42
$\gamma_{\beta 2}$	-0.06	-0.29	0.18

In other words, the average progression (in both groups) is considerably more rapid if we assume an informative missingness than if we ignore it; there's still no solid evidence that treatment has any effect on disease progression

Representative patients



Representative patients



THM study: Introduction

- We now turn to the case where we have missing covariate data
- To illustrate, we will analyze data from a study of trihalomethanes (THM) in domestic tap water
- THMs are a chemical byproduct of the treatment process used to disinfect the public water supply
- As we mentioned earlier in the course, THMs are thought to be carcinogenic in humans at high concentrations, but may also have other deleterious consequences; here we investigate the relationship between THM levels and the probability that a child is born with a low birthweight

Data

- The study used data from the UK (United Kingdom, not University of Kentucky) National Births Register, which contains data on the birth weight (here categorized as `lbw`), where the mother lives (from which the `THM` levels in her water supply could be estimated, dichotomized into $> 60\mu g/L$ or not) and the sex of the baby (`Male`)
- Socioeconomic status may play a role; as a rough measure of that, we include `Dep`, an indicator for whether the mother lived in a deprived local area

Missing data

- Smoking and ethnicity are known risk factors for low birthweight, as well as potential confounders with THM levels due to their spatial patterns
- It would certainly be desirable to control for smoking and ethnicity, but these variables are not recorded in the Births Register
- They are, however, available for a separate group of mothers who participated in a study known as the national birth cohort study
- The full sample, therefore, consists of 1,000 births from the cohort study with complete data and 3,000 births from the registry with missing values for smoking and ethnicity

Simple model

- We begin with a simple model to illustrate the basic idea of modeling the distribution of covariates
- No matter how ignorable the missing data may be, we still must specify a distribution for the covariates
- In this particular case, this is straightforward – both `Smk` and `Eth` are indicator variables (for maternal smoking during pregnancy and non-white ethnicity, respectively), so a Bernoulli distribution is a natural choice

Implementation

- This is straightforward to implement:

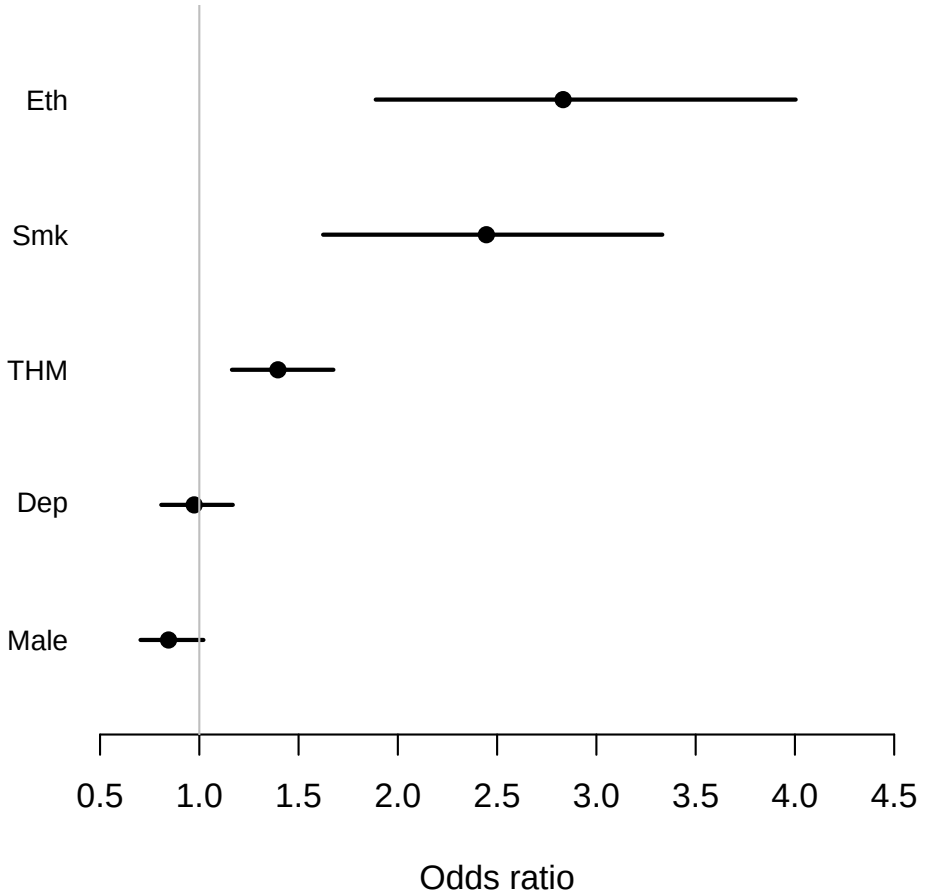
```
# Imputation
for (i in 1:n) {
  Smk[i] ~ dbern(theta[1])
  Eth[i] ~ dbern(theta[2])
}
for (j in 1:2) {
  theta[j] ~ dunif(0,1)
}
```

- Note that in doing this, we're assuming that smoking and ethnicity are independent of the other explanatory variables

Comparison with external imputation

- Note that we are not, however, assuming that smoking and ethnicity are independent of the outcome (low birth weight); the fact that smoking and ethnicity appear in the logistic regression model allows their imputation to be influenced by the response
- This is in stark contrast to "external" imputation, in which the imputation model must include the response, otherwise effect estimates will be biased towards 0

Results



Imputation illustration

	lbw	THM	Dep	Male	Smk	Eth	SmkImp	EthImp
14	1	1	0	0			0.49	0.37
1476	0	0	1	0			0.32	0.19

The sample means (among the complete cases) for `Smk` and `Eth` were 0.355 and 0.228, respectively

Comparison: Imputation

To illustrate the bias towards zero phenomenon, let's compare the results of our Bayesian model with an "external" imputation, in which we impute `Smk` and `Eth` and then fit a logistic regression model to the imputed data set:

	OR estimates	
	Smoking	Ethnicity
Two-step	1.4	1.5
Bayes	2.4	2.8

Comparison: Complete-case

- How does this affect the estimate of the effect of THM?
- Let's compare our odds ratio estimates with that of the unadjusted logistic regression and the complete-case analysis:

		95% Interval	
	OR	Lower	Upper
Bayes	1.4	1.2	1.7
Unadjusted	1.4	1.2	1.7
Complete-case	1.1	0.8	1.6

Ignorable missingness?

- The assumption of ignorable missingness is perhaps a bit questionable in this case
- For example, 51% of smokers were exposed to high THM levels, compared with 40% of nonsmokers
- Likewise, 52% of non-white mothers were exposed to high THM levels, compared with 41% of white mothers
- Furthermore, 24% of non-white mothers smoked, compared with 39% of white mothers
- In other words, `Smk` and `Eth` do not seem to be independent of each other or the exposure of interest

Area variables

- Furthermore, the investigators have access to the area of residence for the mothers, which may contain useful clues about the smoking status and ethnicity of the mother
- The data set online also contains `SmkArea`, the proportion of mothers in a given mother's residenial area who smoke, and similarly for `EthArea`
- We can use these variables, along with `THM` and maybe others, to build an imputation model

Correlation

- If we only had one variable to impute, or if the variables were independent, we could consider a logistic regression model (or models)
- However, it seems plausible that the two variables are correlated, which complicates things a bit as we now need a model for a multivariate discrete outcome
- Correlations among normally distributed variables are relatively straightforward, but things are much more complicated for other distributions

Multivariate probit models

- For this reason, it is often convenient to work with a latent variable approach that treats the discrete variable as a coarse observation of an underlying continuous quantity
- Specifically, we consider the following model:

$$\begin{aligned}
\mathbf{z}_i &\sim \mathrm{N}(\mu_i, \boldsymbol{\Sigma}) \\
\mu_{i1} &= \delta_{11} + \delta_{12}\texttt{THM}_i \cdots + \delta_{15}\texttt{SmkArea}_i + \delta_{16}\texttt{EthArea}_i \\
\mu_{i2} &= \delta_{21} + \delta_{22}\texttt{THM}_i \cdots + \delta_{25}\texttt{SmkArea}_i + \delta_{26}\texttt{EthArea}_i
\end{aligned}$$

$$\texttt{Smk}_i = \begin{cases} 1 & z_{i1} > 0 \\ 0 & z_{i1} \leq 0 \end{cases}$$

$$\texttt{Eth}_i = \begin{cases} 1 & z_{i2} > 0 \\ 0 & z_{i2} \leq 0 \end{cases}$$

Implementation

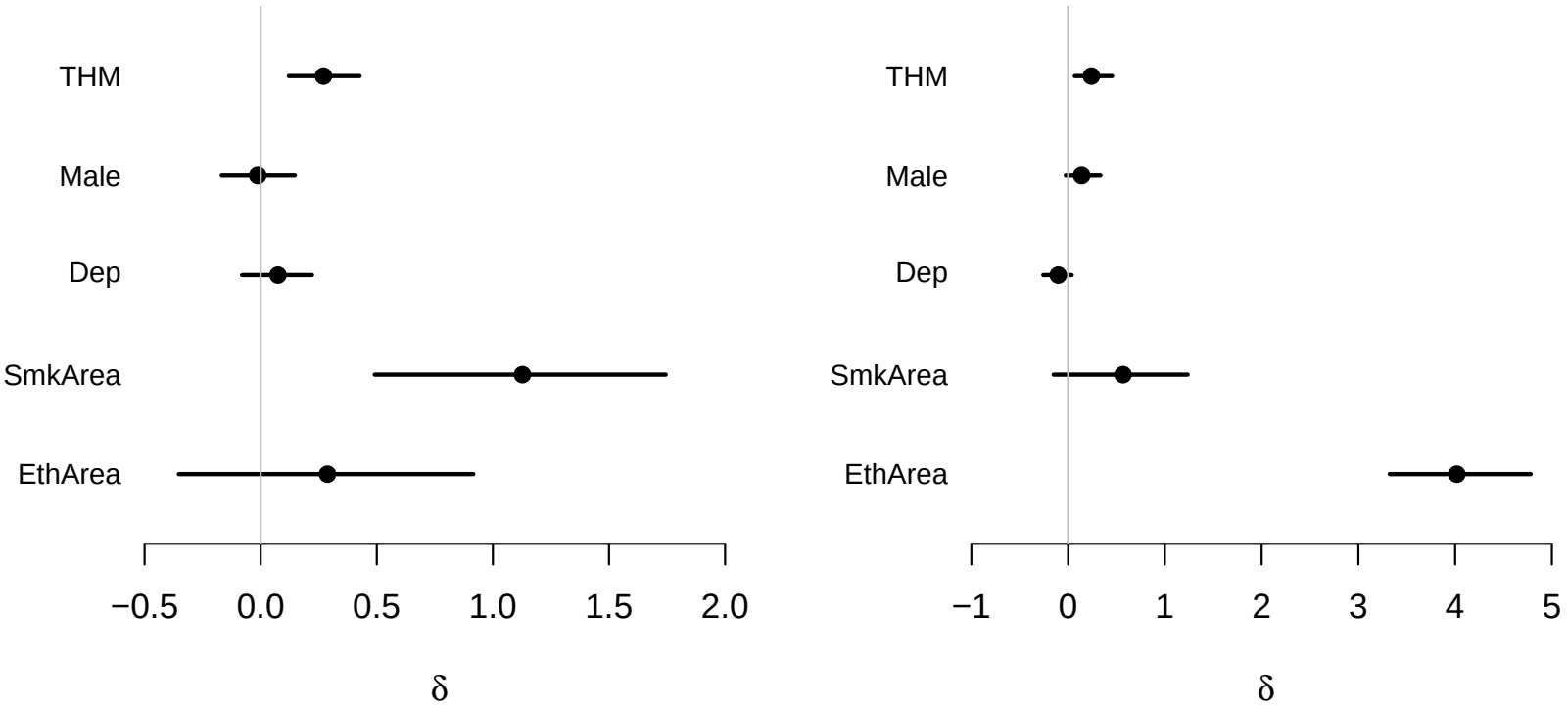
The model can be implemented in JAGS as follows:

```
for (i in 1:n) {
  Z[i,1:2] ~ dmnorm(mu[i,1:2], Omega[1:2,1:2])
  mu[i,1] <- d[1,1] + d[2,1]*THM[i] + d[3,1]*Male[i]
    + d[4,1]*Dep[i] + d[5,1]*SmkArea[i] + d[6,1]*EthArea[i]
  mu[i,2] <- d[1,2] + d[2,2]*THM[i] + d[3,2]*Male[i]
    + d[4,2]*Dep[i] + d[5,2]*SmkArea[i] + d[6,2]*EthArea[i]
  Smk[i] ~ dinterval(Z[i,1],0)
  Eth[i] ~ dinterval(Z[i,2],0)
}
```

with standard uninformative priors on δ and a $(1, \rho, \rho, 1)$ prior on $\mathbf{\Sigma}$

Imputation model results

95% interval for ρ: (-0.41, -0.21)

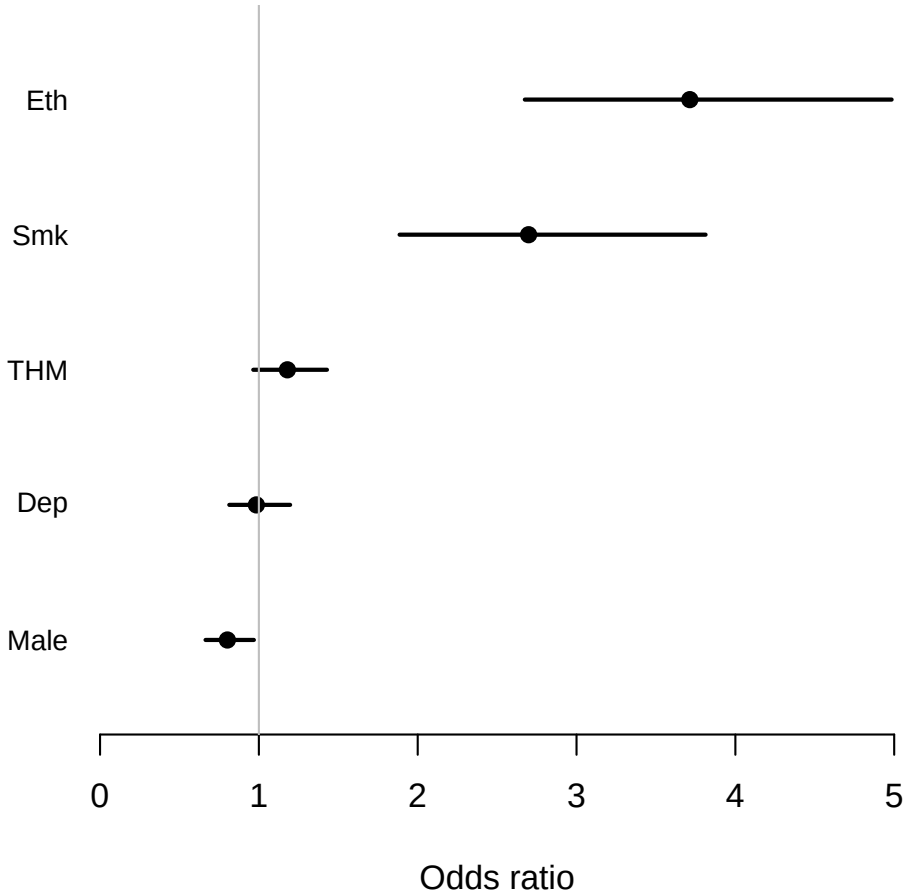


Imputation examples

For the same two patients as before:

	lbw	THM	SmkArea	EthArea	SmkImp	EthImp
14	1	1	0.06	0.18	0.46	0.47
1476	0	0	0.45	0.00	0.36	0.06

Results: Birthweight model



Comparison

		95% Interval	
	OR	Lower	Upper
Bayes (model)	1.2	1.0	1.4
Bayes (ignorable)	1.4	1.2	1.7
Unadjusted	1.4	1.2	1.7
Complete-case	1.1	0.8	1.6

Final remarks

- THM levels are correlated with smoking and ethnicity; failing to adjust for this effect leads to an overestimation of the risk posed by THM
- Treating the missing data as ignorable/missing completely at random also fails to properly adjust for these confounding variables
- The complete-case analysis seems reasonable in this case, but we have to throw away 75% of our data
- The multivariate model for the missing covariates properly adjusts for confounding, with a substantial reduction in uncertainty compared to the complete-case analysis
- THM seems to pose a borderline significant risk for low birth weight, although the effect is rather small