Missing data models

• We will deal with missing data in the linear regression context, but the ideas apply to all models.

• The model is

\[ Y_i \sim \text{Normal}(\beta_0 + \beta_1 X_i + \ldots + \beta_p X_p, \sigma^2). \]

• Often either \( Y_i \) or elements \( X_{ij} \) are missing.

• We will study separately the case of missing responses and missing covariates.
Missing responses

• If the response is missing this is essentially a prediction problem.

• We have seen how to handle this in JAGS.

• We obtain samples from the posterior predictive distribution:

  • This distribution accounts for random error as well as uncertainty in the model parameters.

  • If only responses are missing, can we delete them for the purpose of estimating $\beta$?
Missing responses

- Now say all responses are observed, but a some covariates are missing.

- The simplest approach is imputation:
  
  - Bayesian methods handle this well using MCMC.

  - The main idea is to treat the missing values as unknown parameters in the Bayesian model.

  - Unknown parameters need priors, so missing $X_i = (X_{i1}, \ldots, X_{ip})^T$ must have priors such as $X_i \sim \text{MVN}(\mu_X, \Sigma_X)$.

  - Assumptions about missing data:

    - Of course if the prior is way off, the results will be invalid. For example, if in reality the data are not missing at random the Bayesian model will likely give bad results.

    - Example of non-random missingness:

      - If specified correctly, the model will lead to inference for $\beta$ that properly accounts for uncertainty about the missing data.
Hierarchical linear regression model with missing data

- $Y_i|X_i, \beta, \sigma^2 \sim N(X_i^T \beta, \sigma^2)$
- $X_i|\mu, \Sigma \sim MVN(\mu, \Sigma)$
- $p(\beta) \propto 1$
- $\sigma^2 \sim \text{InvG}(0.01, 0.01)$.
- $\mu \sim MVN(0, 100^2 I_p)$
- $\Sigma \sim \text{InvWishart}(0.01, 0.01 I_p)$

If some observations have missing $Y$ and some have missing $X$, can we delete those with missing $Y$? Can we delete those with missing $X$?
Missing data

- Overview of the Gibbs sampling algorithm:

- The full conditional of missing $Y_i$ is:

- The full conditional of missing $X_i$ is:

- JAGS does not handle missing $X$ well. Let’s use this as an opportunity to explore OpenBUGS in the stacks handout.