Part 6

Hierarchical modeling

ST740

North Carolina State University

Outline

- **Linear models**
- Generalized linear mixed models
- Hierarchical models
- Missing data and censoring

Bayesian one-sample (i.e., paired) t-test

- Say $Y_1, ..., Y_n \sim \text{Normal}(\mu, \sigma^2)$
- Typically Y_i is the difference of a pair of measurements, e.g., the post- minus pre-test for subject i
- Therefore the interest is to compare μ to zero
- We will consider two cases: σ^2 known and σ^2 unknown

Bayesian one-sample (i.e., paired) z-test

- Under the Jeffreys' prior $\pi(\mu) = 1$ with fixed σ,

$$\mu|\mathbf{Y}, \sigma \sim \text{Normal}\left(\bar{Y}, \frac{\sigma^2}{n}\right)$$

- Therefore the posterior mean is the sample mean,

$$\text{E}(\mu|\mathbf{Y}) = \bar{Y}$$

- The 95% credible set is the 95% confidence interval

$$\bar{Y} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

- For the test of $\mathcal{H}_0 : \mu \leq 0$ versus $\mathcal{H}_1 : \mu > 0$,

$$\text{Prob}(\mathcal{H}_0|\mathbf{Y}) = \text{Prob}(\mu \leq 0|\mathbf{Y}) = \Phi(\sqrt{n}\bar{Y}/\sigma)$$

  is the frequentist p-value

Bayesian one-sample (i.e., paired) t-test

- When σ^2 is unknown, the Jeffreys' prior is

$$\pi(\mu, \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{3/2}$$

- The marginal posterior integrating over uncertainty in σ^2 is

$$\mu|\mathbf{Y} \sim t_n\left(\bar{Y}, \frac{\hat{\sigma}^2}{n}\right)$$

  where $\hat{\sigma}^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2/n$

- This is very similar to the frequentist t-test, except that the degrees of freedom is n rather than $n-1$

- This is the effect of the prior

Bayesian two-sample z-test

- Say the n_1 observations from group 1 are

$$Y_i \sim \text{Normal}(\mu, \sigma^2)$$

  are the n_2 observations from group 2 are

$$Y_i \sim \text{Normal}(\mu + \delta, \sigma^2)$$

- The goal is to compare δ to zero
- With σ^2 known and Jeffrey's prior $\pi(\mu, \delta) = 1$,

$$\delta | \mathbf{Y}, \sigma^2 \sim \text{Normal}\left(\bar{Y}_2 - \bar{Y}_1, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$$

  and the results are identical to the two-sample z-test

Bayesian two-sample t-test

- When σ^2 is unknown, the Jeffreys' prior is

$$\pi(\mu, \delta, \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^2$$

- The marginal posterior integrating over uncertainty in σ^2 and μ is

$$\delta|\mathbf{Y} \sim t_n\left(\bar{Y}_2 - \bar{Y}_1, \frac{\hat{\sigma}^2}{n_1} + \frac{\hat{\sigma}^2}{n_2}\right)$$

  where the pooled variance estimator is

$$\hat{\sigma}^2 = \left[\sum_{i=1}^{n_1}(Y_i - \bar{Y}_1)^2 + \sum_{i=n_1+1}^{n_2}(Y_i - \bar{Y}_2)^2\right]/n$$

- This resembles the frequentist t-test, except that due to the prior the DOF is $n = n_1 + n_2$ rather than $n - 2$

Bayesian regression

- The likelihood remains

$$Y_i \sim \text{Normal}(\beta_0 + X_{i1}\beta_1 + ... + X_{ip}\beta_p, \sigma^2)$$

  independent for $i = 1, ..., n$ observations

- As with a least squares analysis, it is crucial to verify this is appropriate using qq-plots, added variable plots, etc.

- A Bayesian analysis also requires priors for β and σ

- We will focus on prior specification since this piece is uniquely Bayesian.

Priors

- For the purpose of setting priors, it is helpful to standardize both the response and each covariate to have mean zero and variance one.

- Many priors for β have been considered:
  1. Improper priors
  2. Gaussian priors
  3. Bayesian lasso
  4. Many, many more...

Improper priors

- With σ fixed, the Jeffreys' prior is flat $p(\beta) = 1$
- This is improper, but the posterior is proper under the same conditions required by least squares
- If σ is known then

$$\beta|\mathbf{Y} \sim \text{Normal}\left[\hat{\beta}_{OLS}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right]$$

- Therefore, the results should be similar to least squares
- How are they different?

Improper priors

- Of course we rarely know σ
- A conjugate uninformative prior is

$$\sigma^2 \sim \text{InvGamma}(a, b)$$

  with a and b set to be small, say $a = b = 0.01$.

- In this case the posterior of β follows a multivariate t centered on $\hat{\beta}_{OLS}$
- Again, the results are similar to OLS

Improper priors

- The objective Bayes Jeffreys prior is

$$p(\beta, \sigma^2) = \left(\frac{1}{\sigma^2}\right)^{p/2+1}$$

  which is the inverse gamma prior with $a = p/2$ and $b \to 0$

- This gives posterior (marginal over σ^2)

$$\beta|\mathbf{Y} \sim \mathrm{t}_n\left(\hat{\beta}_{OLS}, \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}\right)$$

  where $\hat{\sigma}^2 = (\mathbf{Y} - \mathbf{X}\hat{\beta}_{OLS})^T(\mathbf{Y} - \mathbf{X}\hat{\beta}_{OLS})/n$

- The posterior is proper in the same situations that the least squares solution exists

Multivariate normal prior

- Another common prior for is Zellner's g-prior

$$\beta \sim \text{Normal}\left[0, \frac{\sigma^2}{g}(\mathbf{X}^T\mathbf{X})^{-1}\right]$$

- This prior is proper assuming $\mathbf{X}$ is full rank
- The posterior mean is

$$\frac{1}{1+g}\hat{\beta}_{OLS}$$

- This shrinks the least estimate towards zero
- g controls the amount of shrinkage
- $g = 1/n$ is common, and called the unit information prior

Univariate Gaussian priors

- If there are many covariates or the covariates are collinear, then $\hat{\beta}_{OLS}$ is unstable
- Independent priors can counteract collinearity

$$\beta_j \sim \text{Normal}(0, \sigma^2/g)$$

  independent over j

- The posterior mode is

$$\underset{\beta}{\text{argmin}} \sum_{i=1}^{n} (Y_i - \mu_i)^2 + g \sum_{j=1}^{p} \beta_j^2$$

- In classical statistics, this is known as the ridge regression solution and is used to stabilize the least squares solution

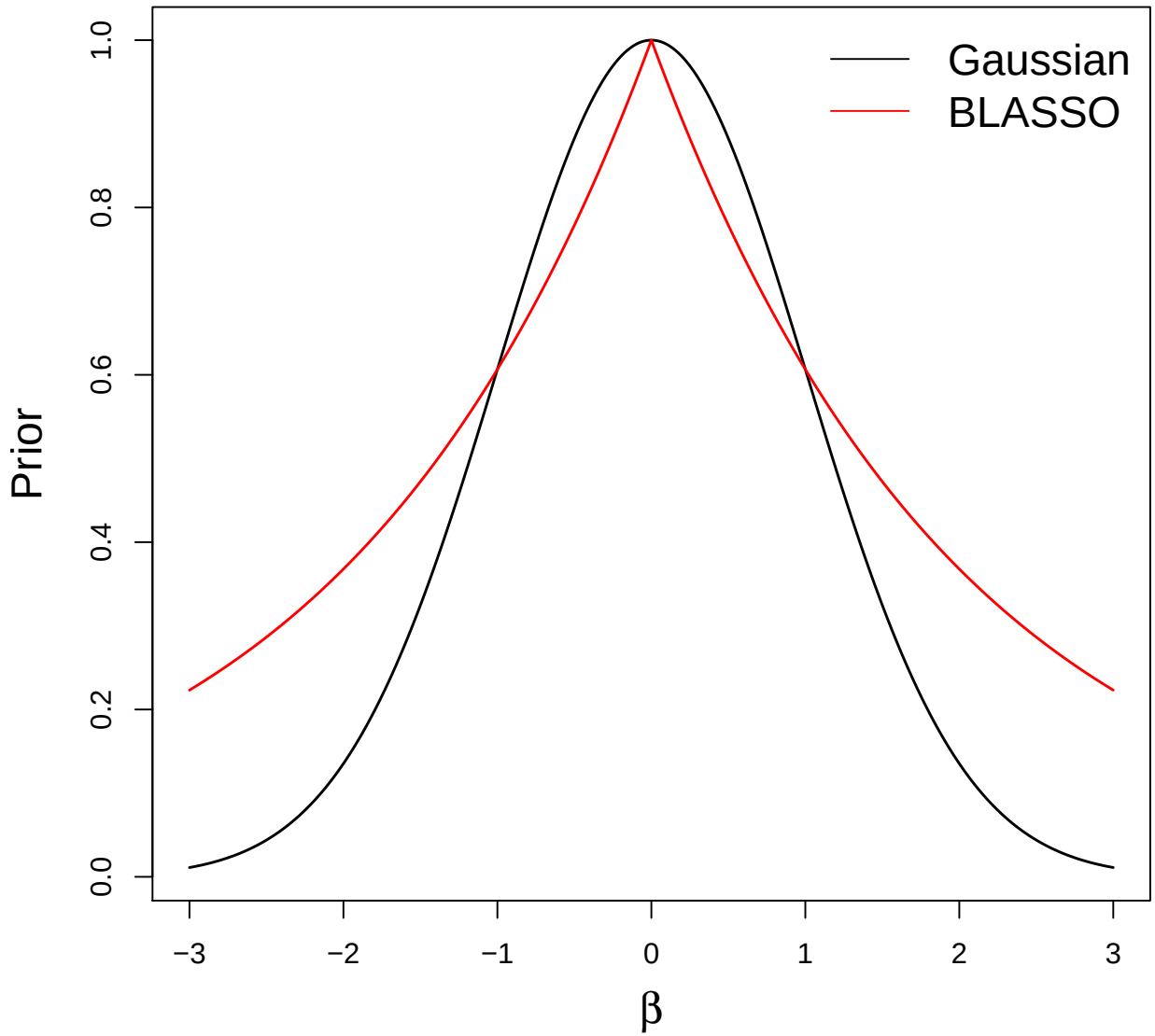
BLASSO

- An increasingly-popular prior is the double exponential or Bayesian LASSO prior
- The prior is $\beta_j \sim \text{DE}(\tau)$ which has PDF

$$f(\beta) \propto \exp\left(-\frac{|\beta|}{\tau}\right)$$

- The square in the Gaussian prior is replaced with an absolute value
- The shape of the PDF is thus more peaked at zero (next slide)
- The BLASSO prior favors settings where there are many β_j near zero and a few large β_j
- That is, p is large but most of the covariates are noise

BLASSO



BLASSO

- The posterior mode is the LASSO solution

$$\underset{\boldsymbol{\beta}}{\text{argmin}} \sum_{i=1}^{n} (Y_i - \mu_i)^2 + g \sum_{j=1}^{p} |\beta_j|$$

- It is popular because it adds stability by shrinking estimates towards zero, and also sets some coefficients to zero
- Covariates with coefficients set to zero can be removed
- Therefore, LASSO performs variables selection and estimation simultaneously
- BLASSO provides uncertainty about β_j and avoids picking a single g

BLASSO computation

- Bayesian LASSO can be fit using Gibbs sampling with the introduction of auxiliary variables
- Derivation:

Summarizing the results

- The standard summary is a table with marginal means and 95% intervals for each β_j
- This becomes unwieldy for large p
- Picking a subset of covariates is a crucial step in a linear regression analysis
- We will discuss this later in the course
- Common methods include cross-validation, information criteria, and stochastic search

Predictions

- Say we have a new covariate vector $\mathbf{X}_{new}$ and we would like to predict the corresponding response Y_{new}
- A plug-in approach would fix β and σ at their posterior means $\hat{\beta}$ and $\hat{\sigma}$ to make predictions

$$Y_{new}|\hat{\beta}, \hat{\sigma} \sim \text{Normal}(\mathbf{X}_{new}\hat{\beta}, \hat{\sigma}^2)$$

- However this plug-in approach suppresses uncertainty about β and σ
- Therefore these prediction intervals will be slightly too narrow leading to undercoverage

Posterior predictive distribution (PPD)

- We should really account for all uncertainty when making predictions, including our uncertainty about β and σ
- We really want the PPD

$$\begin{aligned} p(Y_{new}|\mathbf{Y}) &= \int f(Y_{new}, \beta, \sigma|\mathbf{Y}) d\beta d\sigma \\ &= \int f(Y_{new}|\beta, \sigma) f(\beta, \sigma|\mathbf{Y}) d\beta d\sigma \end{aligned}$$

- Marginalizing over the model parameters accounts for their uncertainty
- The concept of the PPD applies generally (e.g., logistic regression) and means the distribution of the predicted value marginally over model parameters

Posterior predictive distribution (PPD)

- MCMC naturally gives draws from Y_{new}'s PPD
  - For MCMC iteration t we have $\beta^{(t)}$ and $\sigma^{(t)}$
  - For MCMC iteration t we sample

$$Y_{new}^{(t)} \sim \text{Normal}(\mathbf{X}\beta^{(t)}, {\sigma^{(t)}}^2)$$

  - $Y_{new}^{(1)}, ..., Y_{new}^{(S)}$ are samples from the PPD
- This is an example of the claim that "Bayesian methods naturally quantify uncertainty"

Outline

- Linear models
- **Generalized linear mixed models**
- Hierarchical models
- Missing data and censoring

Generalized linear models (GLMs)

- GLMs extend linear models to non-Gaussian data
- A general formulation is

$$g[\mathrm{E}(Y_i|\beta)] = \eta_i = \mathbf{X}_i\boldsymbol{\beta}$$

- The linear predictor is η_i
- The link function g projects the mean from its support to $\mathcal{R}$ where modeling in unconstrained
- For example, logistic regression takes $g(x) = \log[x/(1-x)]$

Steps to selecting a Bayesian GLM

1. Identify the support of the response distribution
2. Select the likelihood by picking a parametric family of distributions with this support
3. Choose a link function g that transforms the range of parameters to the whole real line
4. Specify a linear model on the transformed parameters
5. Select priors for the regression coefficients

Logistic regression (LR)

- The model for binary responses is

$$\text{Prob}(Y_i = 1|\boldsymbol{\beta}) = \frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})}$$

- The coefficient β_j is interpreted as the increase in log odds of Y_i if X_j increases by one with all other covariates fixed
- Bayesian logistic regression requires a prior for $\boldsymbol{\beta}$
- All of the prior we have discussed for linear regression (Zellner, BLASSO, etc) apply
- The full conditional distributions are no longer conjugate but you can use Metropolis sampling (`MCMClogit`)

Logistic regression (LR)

- LR can be used to compare two proportions
- Say population $j \in \{1, 2\}$ has success probability π_j
- Then set $X_i = I(\text{observation } i \text{ is from population 2})$, and

$$\text{logit}\left[\text{Prob}(Y_i = 1|\boldsymbol{\beta})\right] = \beta_1 + X_i\beta_2$$

- The populations have the same probability if $\beta_2 = 0$
- How to pick priors for β_j that resemble the Jeffrey Beta(1/2,1/2) priors for the π_j?
- Prior predictive check are a simple/informal method

Probit regression

- The model for binary responses is

$$\text{Prob}(Y_i = 1|\beta) = \Phi(\mathbf{X}_i\beta) \tag{1}$$

  where Φ is the standard normal CDF

- The interpretation of β is not as clear as in LR
- The model is equivalent to a truncated normal model
- Say there are latent outcomes

$$Z_i \sim \text{Normal}(\mathbf{X}_i\beta, \sigma^2)$$

- Rather than observing Z_i, we observe only $Y_i = I(Z_i > 0)$
- For identifiability we set $\sigma = 1$, giving (1)

Probit regression computation

- Exact Gibbs sampling can be achieved[1]
- Derivation:

[1] https://www.jstor.org/stable/2290350

Logistic regression computation

- Exact Gibbs sampling can be achieved using the Polya-Gamma sampler [2]
- This also applies to negative binomial regression for count data
- Derivation:

[2] https://www.tandfonline.com/doi/abs/10.1080/01621459.2013.829001

Generalized linear mixed models (GLMMs)

- GLMs assume the observations are independent
- This is invalid if data are grouped
- For example, n classrooms each have m students
- It might be reasonable to assume the classrooms are independent, but the students within a class are likely dependent
- Random effects are a natural way to account for this dependence

Generalized linear mixed models (GLMMs)

- GLMMs extend GLMs to correlated data
- A general formulation is

$$g[\mathrm{E}(Y_i|\beta, \mathbf{u})] = \eta_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{u}$$

- The random effects distribution is $\mathbf{u} \sim \mathrm{Normal}(0, \Sigma)$
- Given $\mathbf{u}$, the observations are independent, but marginal over $\mathbf{u}$ (e.g., via MCMC) they are correlated
- The correlation between linear predictors is

$$\mathrm{Cor}(\mathbf{Z}_i\mathbf{u}, \mathbf{Z}_j\mathbf{u}) = \mathbf{Z}_i\Sigma\mathbf{Z}_j^T$$

- This induces correlation between observations, although expressions for $\mathrm{Cor}(Y_i, Y_j)$ are complicated

GLMMs example #1

- $Y_{ij} \in \{0, 1\}$ is the results of attempt j by kicker i
- The probability of success depends on distance, X_{ij}
- To account for dependence we add a random kicker effect, $u_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$
- The random effects logistic regression model is

$$\text{logit}\left[\text{Prob}(Y_{ij} = 1|\beta, u_i)\right] = \beta_0 + X_{ij}\beta_1 + u_i$$

- The vector $\mathbf{Z}_{ij}$ is zero everywhere except a one in element i
- The random effect distribution is $\mathbf{u} \sim \text{Normal}(0, \sigma^2\mathbf{I})$

GLMMs example #2

- $Y_i \in \{0, 1, ...\}$ is the number of cancer cases in county i
- The model is

$$Y_i|\lambda_i \sim \text{Poisson}(N_i\lambda_i)$$

where N_i is the population of county i

- The relative risks are modeled as

$$\log(\lambda_i) = \beta_0 + u_i$$

where $\mathbf{u} \sim \text{Normal}(0, \Sigma)$ ($\mathbf{Z}$ is the identify matrix)

- The spatial covariance matrix Σ has (i, j) element

$$\Sigma_{ij} = \sigma^2 \exp(-d_{ij}\phi)$$

where d_{ij} is the distance between counties i and j

Confusion about random effects

- MCMC does not distinguish between random effects and other parameters
- For example, β, $\mathbf{u}$ and σ^2 are all treated as random in a Bayesian analysis
- However, u_i is called a "random" effect because it represents one random draw from a population distribution
- Often for GLMMs, we are less interested in particular u_i and more interested in the population distribution via Σ

Bayesian analysis of GLMMs

- There are not really any special techniques needed to implement a Bayesian GLMM
- Gibbs and Metropolis can be used
- As in all analyses, we require priors
- The main advantages of Bayesian implementation is the ability to incorporate prior information and account for uncertainty in the variance components
- For example, MLE analyses of GLMMs use plug-in estimators of the variance components and rely on normal approximations for the fixed and random effects

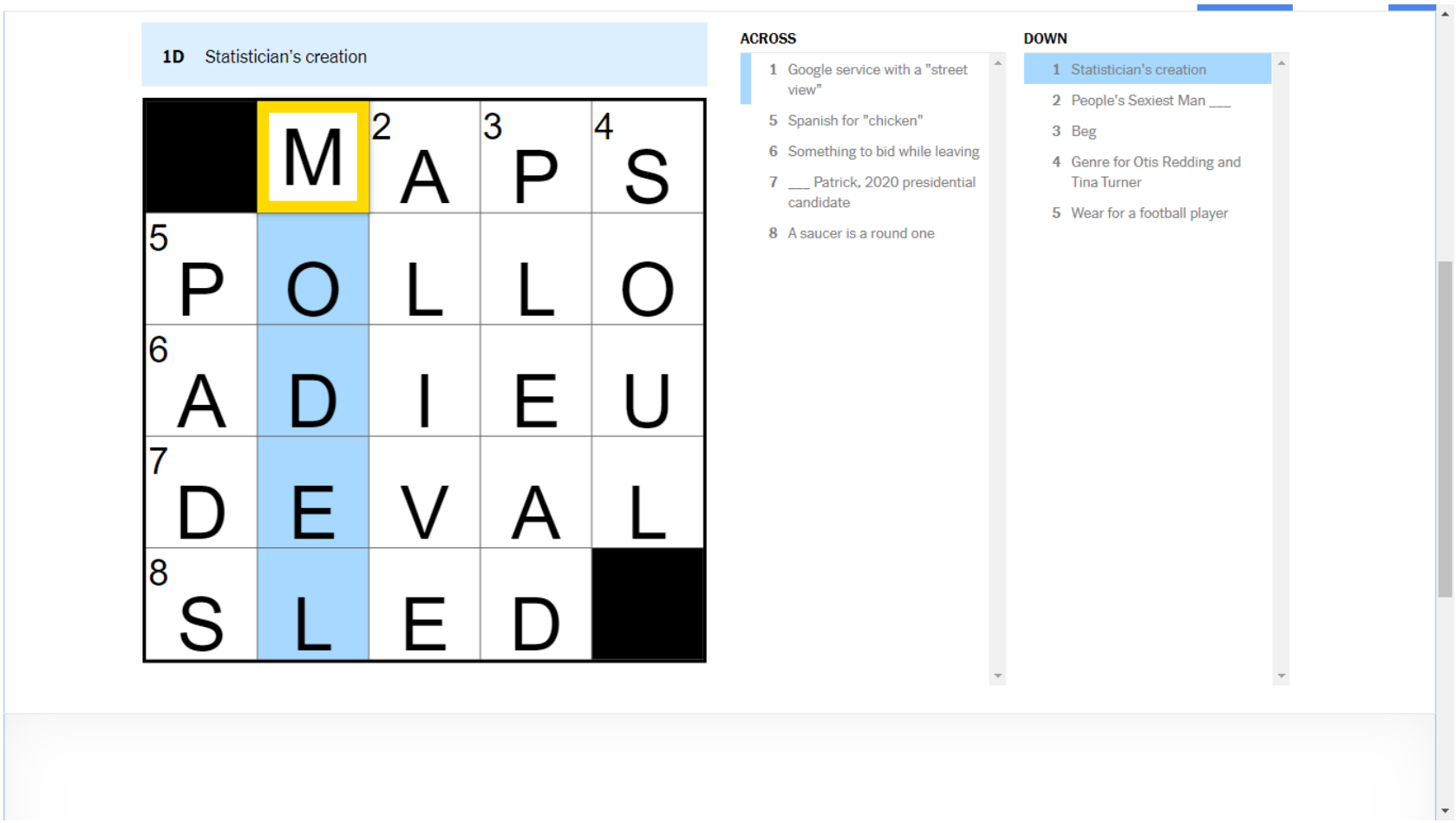
Outline

- Linear models
- Generalized linear mixed models
- **Hierarchical models**
- Missing data and censoring

Hierarchical models

- Hierarchical modeling provides a framework for building complex and high-dimensional models from simple and low-dimensional building blocks
- Of course, it is possible to analyze these models using non-Bayesian methods
- However, this modeling framework is popular in the Bayesian literature because MCMC is conducive to hierarchical models
- Both “divide and conquer” big problems by splitting them into a series of smaller problems in the same way

We build models!



Hierarchical models

Often Bayesian models can we written in the following layers of the hierarchy

1. **Data layer**: $[\mathbf{Y}|\theta, \alpha]$ is the likelihood for the observed data **Y** given the model parameters

2. **Process layer**: $[\theta|\alpha]$ is the model for the parameters θ that define the latent data generating process

3. **Prior layer**: $[\alpha]$ prior for hyperparameters

Epidemiology example - Data layer

- Let S_t and I_t be the number of susceptible and infected individuals in a population, respectively, at time t
- The data Y_t is the number of observed cases at time t
- The data layer models our ability to measure the process I_t
- **Data layer**: $Y_t | I_t \sim \text{Binomial}(I_t, p)$
- This assumes no false positives and false negative probability p

Epidemiology example - Process layer

- Scientific understanding of the disease is used to model disease propagation
- We might select the simple Reed-Frost model

$$\textbf{Process layer}: \quad I_{t+1} \sim \text{Binomial}\left[S_t, 1-(1-q)^{I_t}\right]$$
$$S_{t+1} = S_t - I_{t+1}$$

- This assumes all infected individuals are removed from the population before the next time step
- Also that q is the probability of a non-infected person coming into contact with and contracting the disease from an infected individual

Epidemiology example - Prior layer

- The epidemiological process-layer model expresses the disease dynamics up to a few unknown parameters
- The Bayesian model is completed using priors, say,
- **Prior layer:**

$$\begin{aligned} I_1 &\sim \text{Poisson}(\lambda_1) \\ S_1 &\sim \text{Poisson}(\lambda_2) \\ p, q &\sim \text{beta}(a, b) \end{aligned}$$

When to stop adding layers?

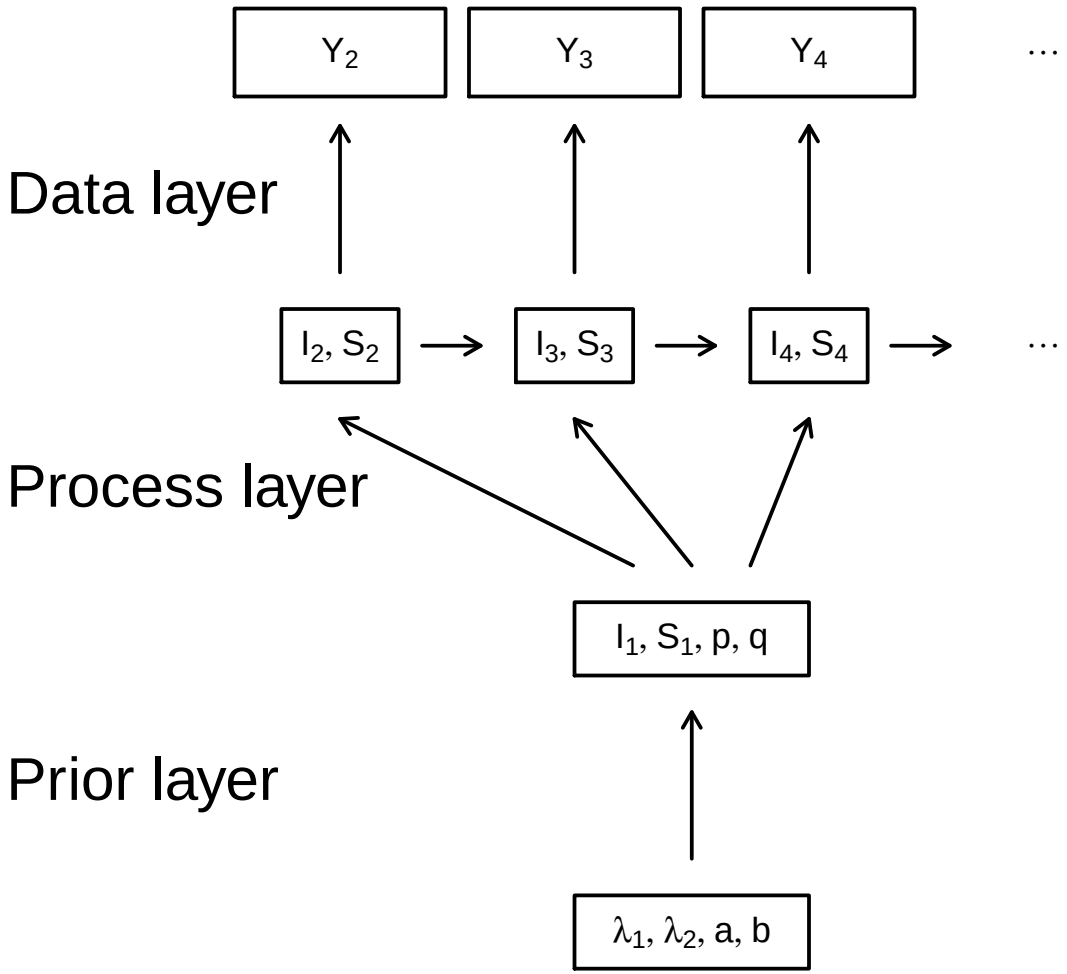
- In the previous example a, b, λ_1 and λ_2 are fixed
- But we will have uncertainty about the correct value
- Maybe replace a fixed value with another layer, say $a \sim \text{Uniform}(0, \theta)$?
- Then maybe $\theta \sim \text{Exponential}(\xi)$, $\xi \sim \text{Uniform}(0, \eta)$, etc.
- Rule of thumb: Be careful assigning priors to parameters in layers without replication.
- For example, even if we knew p exactly this would be just one value and we couldn't hope to estimate the parameters of its beta distribution.

Directed acyclic graphs (DAGs)

- A DAG is a graphical representation of a hierarchical model
- DAGS sometimes go by the name Bayesian networks
- Each observation and parameter is a node
- An arrow for X to Y means that the conditional distribution of Y depends on X
- "Directed" means that arrows only go one way
- Acyclic means there are no cycles, e.g.,

$$X \to Y \to Z \to X$$

Epidemiology example - DAG



Directed acyclic graphs (DAGs)

- Building models this way ensures we will always have a valid joint distribution
- For example, say we need to specify the joint distribution of (X, Y, Z)
- Any joint distribution can be written as

$$f(X, Y, Z) = f(X)f(Y|X)f(Z|X, Y)$$

- This is a fully-connected DAG
- Ad-hoc constructions like

$$f(X, Y, Z) = f(X|Z)f(Y|X)f(Z|X, Y)$$

may or may not give a valid joint PDF

Hierarchical models and MCMC

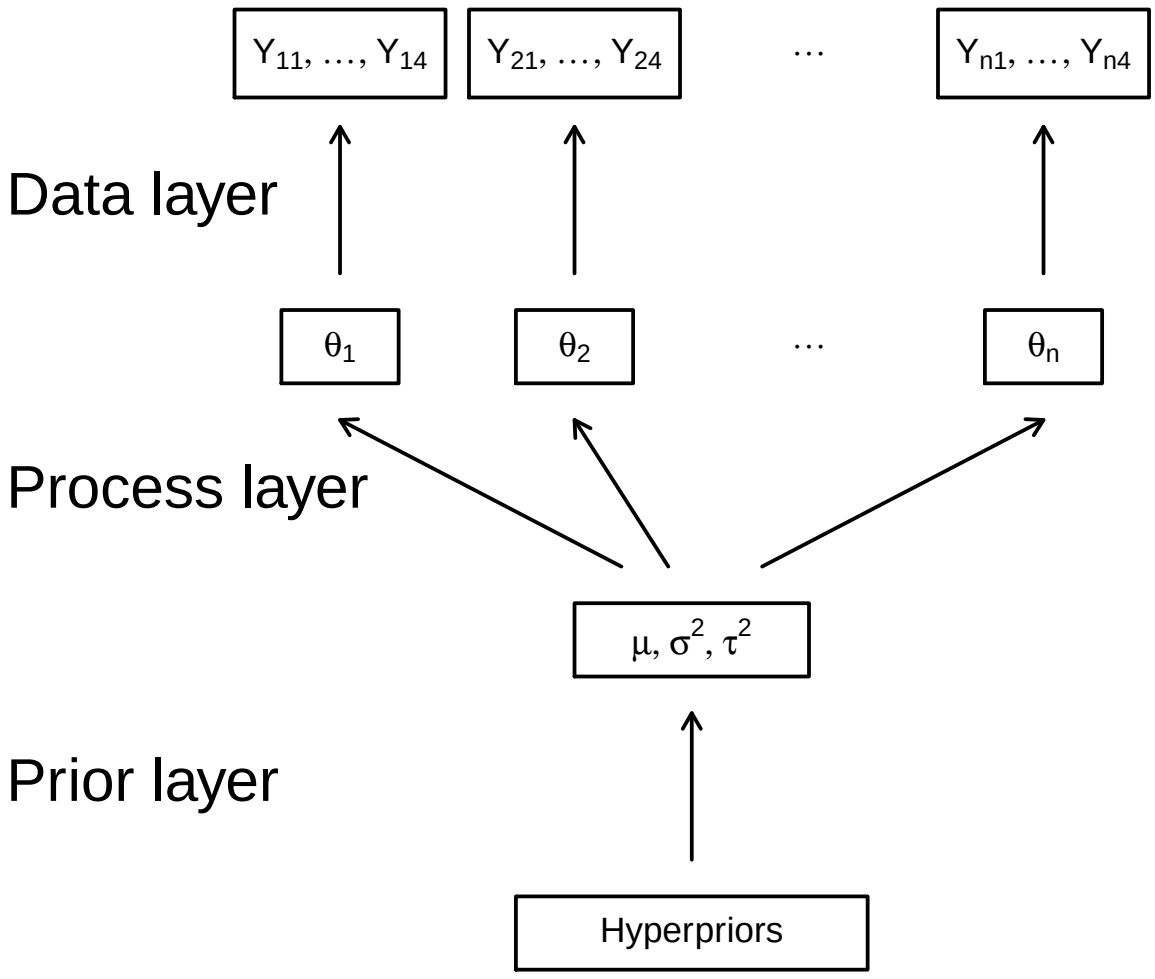
- Consider the classic one-way random effects model:

$$Y_{ij} \sim \mathrm{N}(\theta_i, \sigma^2) \quad \text{and} \quad \theta_i \sim N(\mu, \tau^2)$$

  where Y_{ij} is the j^{th} replicate for unit i and $\boldsymbol{\alpha} = (\mu, \sigma^2, \tau^2)$ has an uninformative prior

- This hierarchy can be written using a directed acyclic graph

Random effects example - DAG



Hierarchical models and MCMC

- MCMC is efficient in this case even if the number of parameter or levels of the hierarchy is large
- You only need to consider "connected nodes" when you update each parameter
- For example, consider the random effect θ_1

$$
\begin{aligned}
p(\theta_1|\cdot) &\propto \left[\prod_{i,j} f(Y_{ij}|\theta_i,\tau^2)\right]\left[\prod_{i=1}^{n}\pi(\theta_i|\boldsymbol{\alpha})\right]\pi(\boldsymbol{\alpha}) \\
&\propto \left[\prod_{j} f(Y_{1j}|\theta_1,\tau^2)\right]\pi(\theta_1|\boldsymbol{\alpha})
\end{aligned}
$$

- This only includes data for subject 1 and the prior for θ_1, so our old normal/normal conjugacy rules apply
- Each of these updates is a draw from a standard one-dimensional normal or inverse gamma

Classes of hierarchical models

- Most hierarchical models we fit could be classified as **multi-level** statistical models
- Here we have different parameters for different levels/groups
- The distribution of parameters across groups follows a random-effects model
- The GEV/random slopes model on the first exam is a good example
- Another class is the **mathematical/statistical model**
- Here we quantify bias and uncertainty in a mathematical, often differential equation, model

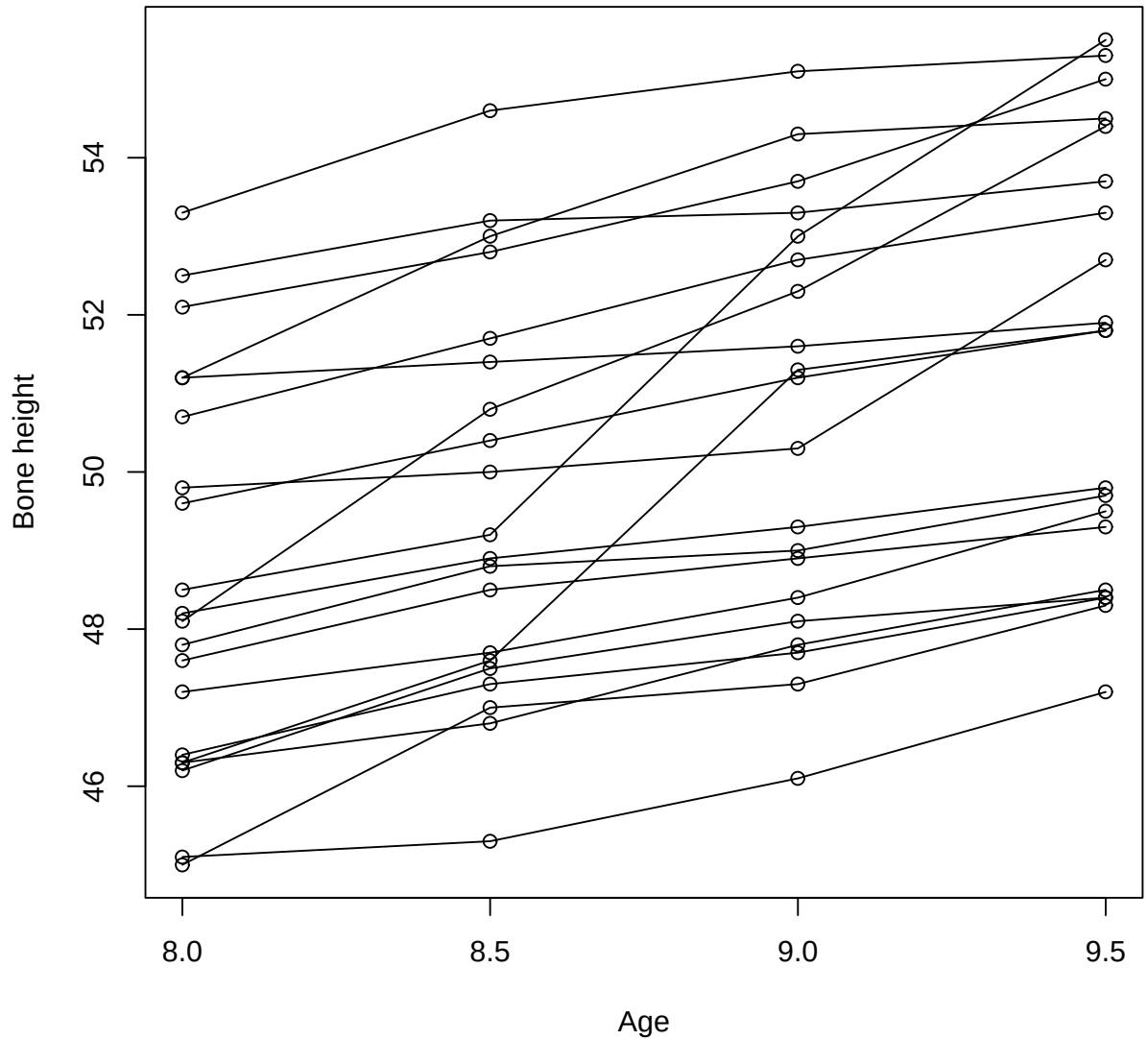
Multi-level Random slopes model

- Let Y_{ij} be the j^{th} observation for subject i
- As an example, consider the data plotted on the next slide were Y_{ij} is the bone density for child i at age X_j.
- Here we might specify a different regression for each child to capture variability over the population of children:

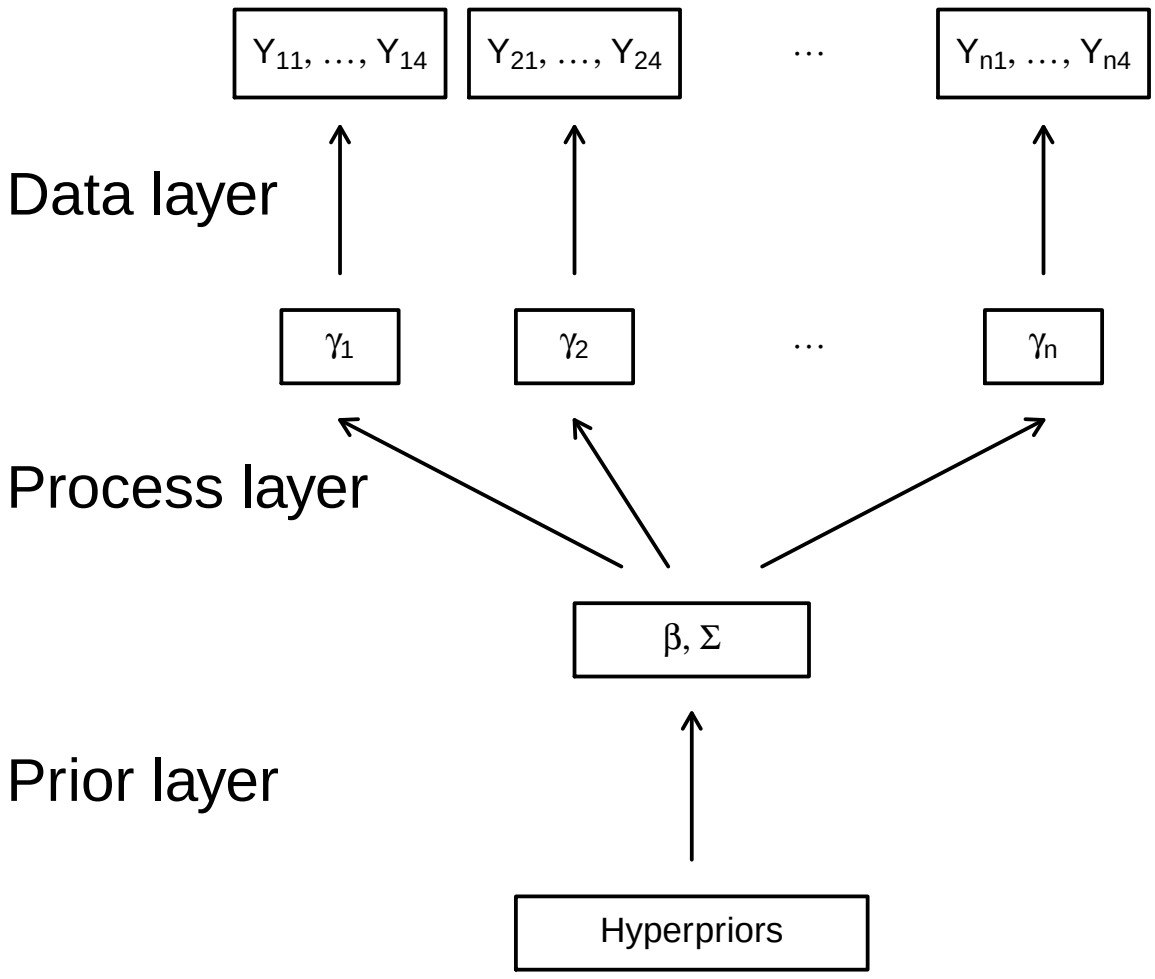
$$Y_{ij} \sim \text{Normal}(\gamma_{0i} + X_i\gamma_{1i}, \sigma^2)$$

- $\gamma_i = (\gamma_{i0}, \gamma_{i1})^T$ controls the growth curve for child i
- These separate regression are tied together in the prior, $\gamma_i \sim \text{Normal}(\beta, \Sigma)$, which borrows strength across children
- This is a linear mixed model: γ_i are random effects specific to one child and β are fixed effects common to all children

Bone height data



Random slopes example - DAG



Mathematical/statistical models

- Mathematicians and engineers often build models using differential equations
- Examples: weather/climate models, resilience of a airplane wing to strain, strength of a bridge
- These models often have parameters that are known well, e.g., response of steel to temperature
- But some parameters are known with less precision: response of hurricane intensity to increased SST
- Also, all models have bias, most observations have bias and/or error
- Embedding the mathematical model in a statistical model gives uncertainty quantification (UQ)

Mathematical/statistical models

- Let $g(\mathbf{X}, \boldsymbol{\theta})$ be a mathematical model, e.g., the solution to differential questions
- The design variables are $\mathbf{X}$ and the unknown parameters are $\boldsymbol{\theta}$
- Example: $g(\mathbf{X}, \theta)$ the true air pollution at a sensor
- Example: $\mathbf{X}$ is the power plant structure and the wind field
- Example: $\boldsymbol{\theta}$ is the true emission from the power plant

Mathematical/statistical models

Now say we observed n observations Y_i under conditions $\mathbf{X}_i$

- Can we estimate θ?
- Can we estimate model bias?
- Can we predict (with uncertainty) Y for a new $\mathbf{X}$?
- Can we find the optimal $\mathbf{X}$?

Mathematical/statistical models

- A common model[3] is

$$Y_i = g(\mathbf{X}_i, \boldsymbol{\theta}) + \delta(\mathbf{X}_i) + \varepsilon_i$$

- The parameters $\boldsymbol{\theta}$ often have informative priors
- The discrepancy term δ captures systematic bias, and can be modeling with splines or a Gaussian processes
- The measurement error term is $\varepsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$
- This would be straightforward, except that often evaluation g takes hours or days

[3] https://rss.onlinelibrary.wiley.com/doi/10.1111/1467-9868.00294

Mathematical/statistical models

Examples

Outline

- Linear models
- Generalized linear mixed models
- Hierarchical models
- **Missing data and censoring**

Missing data models

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

- The model is

$$Y_i \sim \text{Normal}(\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}, \sigma^2)$$

- Either Y_i or elements of X_{ij} can be 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 obtain samples from the PPD of Y_i
- At each MCMC iteration we simply draw

$$Y_i \sim \text{Normal}(\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}, \sigma^2)$$

- This distribution accounts for random error as well as uncertainty in the model parameters
- For the other updates the data are essentially complete
- If only responses are missing, can we delete them for the purpose of estimating β?

Censored data

- Censored data often arise in survival analysis
- For example, Y_i is the time until an event for subject i
- If subjects are only monitored until time T, patients that do not have an event at the end of the study are censored and you know only that $Y_i > T$
- Another example is a detection limit so that all observations between zero and detection limit T are only know to be in the interval $(0, T)$

Censored data

- Handling censored data is really similar to missing data
- For example, if Y_i is censored and known be at least T, you make a draw from its PPD but restricted to (T, ∞)
- Given the imputed censored observation the remaining analysis proceeds as if the data are complete
- These ideas can also be used in modeling such as tobit and probit regression (see examples)

Missing covariates

- Now say all responses are observed, but a some covariates are missing
- The simplest approach is imputation, e.g., just plug in the sample mean of the covariate for the missing values
- This doesn't account for uncertainty in the imputations
- Bayesian methods handle this well using MCMC

Missing covariates

- The main idea is to treat the missing values as unknown parameters in the Bayesian model
- Unknown parameters need priors, so missing $\mathbf{X}_i = (X_{i1}, ..., X_{ip})^T$ must have priors such as

$$\mathbf{X}_i \sim \text{Normal}(\mu_X, \Sigma_X)$$

- Assumptions about missing data:
  - Missing status is independent of Y and $\mathbf{X}$
  - Covariates are Gaussian
- There are ways to relax both assumptions, but it becomes complicated

Missing covariates

- 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 β that properly accounts for uncertainty about the missing data

Hierarchical linear regression model with missing data

- $Y_i|\mathbf{X}_i, \beta, \sigma^2 \sim \text{Normal}(\mathbf{X}_i^T\beta, \sigma^2)$
- $\mathbf{X}_i|\mu, \Sigma \sim \text{Normal}(\mu, \Sigma)$
- $p(\beta) \propto 1$
- $\sigma^2 \sim \text{InvG}(0.01, 0.01)$
- $\mu \sim \text{Normal}(0, 100^2 I_p)$
- $\Sigma \sim \text{InvWishart}(0.01, 0.01 I_p)$

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?

Overview of the Gibbs sampling algorithm

- The full conditional of missing Y_i is:

$$Y_i|\mathbf{X}_i, \beta, \sigma^2 \sim \text{Normal}(\mathbf{X}_i^T \beta, \sigma^2)$$

- The full conditional of missing X_i is:

  The algebra is involved, but it has the same full conditional form as β

- In fact, all the full conditionals are conjugate

Derivation