Hierarchical and Generalized Linear Models

Lin Zhang

Department of Biostatistics
School of Public Health
University of Minnesota

Hierarchical Regression Modeling

- Previously we discussed regular Bayesian linear regression with an independent vague prior for each coefficient, i.e.

$$\mathbf{y} \sim N(X\boldsymbol{\beta}, \sigma^2 I_n)$$
$$\pi(\boldsymbol{\beta}) = 1 \text{ or } \boldsymbol{\beta} \sim N(0, \tau^2 I_J)$$

  where τ^2 is a *large* constant.

- This is common setting for fixed effects regression models.

- Now we consider hierarchical linear models with varying coefficients, with the prior

$$\boldsymbol{\beta} \sim N(\mathbf{1}\alpha, \sigma_\beta^2 I_J)$$

  where α and σ_β^2 are unknown hyperpriors, and $\mathbf{1}$ is a $J \times 1$ vector of ones.

Simple varying-coefficient models

- Varying-coefficient models are hierarchical models in which groups of the regression coefficients are exchangeable and are modeled with normal population distribution.
- The simplest varying coefficient model is random effects models

$$\begin{aligned} Y_i &\sim N(\theta_i, \sigma^2)\ ,\ i = 1, \ldots, n \\ \theta_i &\sim N(\mu, \tau^2) \end{aligned}$$

- We can rewrite in the hierarchical form

$$\begin{aligned} \mathbf{Y} &\sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n) \\ \boldsymbol{\beta} &\sim N(\mathbf{1}\alpha, \sigma_\beta^2 I_n) \end{aligned}$$

with $(\boldsymbol{\beta}, \alpha, \sigma_\beta^2)$ in place of $(\boldsymbol{\theta}, \mu, \tau^2)$ and $\mathbf{X}$ is an $n \times n$ identity matrix.

Hyperpriors

- We usually place a flat or vague normal prior on the population mean α.
- Some common non-informative priors for σ^2_β
  - Flat prior on σ_β
  - Scaled-inverse chi-squared distribution on σ^2_β with small degrees of freedom
  - Flat prior on log($\sigma_b eta$) CANNOT be used as it will result in an improper posterior
- Cautious: Results may or may not be sensitive to prior on σ_β. Therefore, it is useful to conduct a *sensitivity analysis.*

Connection with intraclass correlation

- Assume data $y_1, \ldots, y_n$ fall into J batches/groups, that is
$$Y_i \sim N(\beta_j, \sigma^2) \ ,$$
where $j \in \{1, \ldots, J\}$ for each i.

- This is equivalent to
$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n)$$
where $\mathbf{X}$ is a $n \times J$ indicator matrix with $X_{ij} = 1$ if unit i in batch j and 0 otherwise.

- The correlations between two units in the same group is
$$\rho = \frac{\sigma_\beta^2}{\sigma^2 + \sigma_\beta^2}$$

$\Rightarrow$ Varying coefficient models are used for correlated/clustered data!

Mixed effects model

- The previous models assume all coefficients are random effects.
- A more common scenarios is that some coefficients are treated as random effects, which are modeled hierarchically, while others are treated as fixed effects.
- Mixed effects models take the form

$$\begin{aligned} \mathbf{Y} &\sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}, \sigma^2 I_n) \\ \pi(\boldsymbol{\beta}) &= 1 \\ \boldsymbol{\gamma} &\sim N(\boldsymbol{\alpha}, \Sigma_\gamma) \end{aligned}$$

  where $\mathbf{X}$ and $\mathbf{Z}$ are design matrices for fixed and random effects.
- A simple example

$$\begin{aligned} \mathbf{Y} &\sim N(\mathbf{1}\mu + I_n\boldsymbol{\theta}, \sigma^2 I_n) \\ \pi(\mu) &= 1 \\ \boldsymbol{\theta} &\sim N(\mathbf{0}, \tau^2 I_n) \end{aligned}$$

  $\Rightarrow$ a variation of the random effects model!

Clusters of varying coefficients

- To generalize, γ can be divided into K clustered, each cluster having different population mean and variance.

$$\gamma = \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_K \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{1}\alpha_1 \\ \vdots \\ \mathbf{1}\alpha_K \end{bmatrix}, \begin{bmatrix} \sigma^2_{\gamma_1} I & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2_{\gamma_k} I \end{bmatrix} \right)$$

- Examples:
  - A county as a cluster, with a random effect for within-cluster coefficients
  - A categorical variable as a cluster, with indicator matrix for within-cluster coefficients
- Priors on α_k
  - Uniform or vague prior
  - Normal prior with common variance $\Rightarrow$ Nested models!

Varying intercepts and slopes

- So far we have focused on hierarchical models for scalar parameters. We could have multiple parameters that vary by group.

- Consider the longitudinal model for rate weight data

$$Y_{ij} \overset{ind}{\sim} N\left(\alpha_i + \beta_i x_{ij}\ ,\ \sigma^2\right)\ ,$$

  Y_{ij}: the weight of the i^{th} rat at measurement point j,
  x_{ij}: rat's age in days,
  $i = 1, \ldots, k = 30$, and $j = 1, \ldots, 5$.

- Adopt the random effects model for joint distribution of (α_i, β_i)

$$\boldsymbol{\theta}_i \equiv \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \overset{\text{iid}}{\sim} N\left(\boldsymbol{\theta}_0 \equiv \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}, \Sigma\right),\ i = 1, \ldots, k\ .$$

  Nonzero Σ_{12} brings correlation between α_i and β_i.

Varying intercepts and slopes

- HyperPriors: Conjugate forms are available, namely

$$\begin{aligned} \sigma^2 &\sim IG(a,\, b)\ , \\ \boldsymbol{\theta}_0 &\sim N(\boldsymbol{\eta},\, C)\ , \quad \text{and} \\ \Sigma &\sim Inv - Wish\,((\rho R),\, \rho)\ . \end{aligned}$$

  Inverse-Wishart strongly constraints the variance parameters.

- We assume the hyperparameters $(a, b, \boldsymbol{\eta}, C, \rho,$ and $R)$ are all known, so there are $30(2) + 3 + 3 =$ 66 unknown parameters in the model.

- Yet the Gibbs sampler is relatively straightforward to implement here, thanks to the conjugacy at each stage in the hierarchy.

Posterior Sampling

- Full conditional of $\boldsymbol{\theta}_i$ is

$$\boldsymbol{\theta}_i | \mathbf{y}, \boldsymbol{\theta}_0, \Sigma^{-1}, \sigma^2 \sim N\left(D_i \mathbf{d}_i\ ,\ D_i\right) \quad \text{where}$$

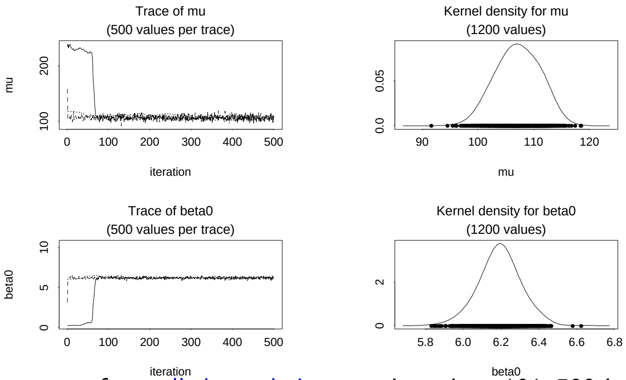
$$D_i^{-1} = \sigma^{-2} X_i^T X_i + \Sigma^{-1} \quad \text{and} \quad \mathbf{d}_i = \sigma^{-2} X_i^T \mathbf{y}_i + \Sigma^{-1} \boldsymbol{\theta}_0,$$

$$\text{for} \quad \mathbf{y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{pmatrix}, \quad \text{and} \quad X_i = \begin{pmatrix} 1 & x_{i1} \\ \vdots & \vdots \\ 1 & x_{in_i} \end{pmatrix}.$$

- Similarly, the full conditionals for σ^2, $\boldsymbol{\theta}_0$, and Σ emerge in closed form as inverse gamma, normal, and Inverse Wishart, respectively!

Posterior Sampling

- Using vague hyperpriors, run 3 initially overdispersed parallel sampling chains for 500 iterations each:



- The output from all three chains over iterations 101–500 is used in the posterior kernel density estimates (col 2)

- The average rat weighs about 106 grams at birth, and gains about 6.2 grams per day.

Generalized linear models

- Previously we discuss the linear regression models where the outcomes are normally distributed.
- Now we extend to the generalized linear models, which allows for regression for general non-normal outcomes.
- A generalized linear model is specified in three stages:
  - The linear predictor: $\eta = \mathbf{X}\boldsymbol{\beta}$
  - The link function $g(\cdot)$ that relates the linear predictor to the mean of the outcome: $\mu = g^{-1}(\eta) = g^{-1}(\mathbf{X}\boldsymbol{\beta})$
  - The distribution of the outcome variable with mean $E(y|\mathbf{X}) = \mu$, which may depend on one or more nuisance parameters.

Common Generalized Linear Models

- Logistic Regression for binary/binomial data:
  - Distribution of Y: $y_i \sim Bin(n_i, p_i)$
  - Link function: $\log\left(\frac{\mu_i}{1-\mu_i}\right) = \log\left(\frac{p_i}{1-p_i}\right) = \eta_i = \mathbf{X}_i\boldsymbol{\beta}$
  - Likelihood:

$$p(\mathbf{y}|\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\begin{array}{c} n_i \\ y_i \end{array} \right) \left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right)^{y_i} \left(\frac{1}{1+e^{\eta_i}}\right)^{n_i - y_i}$$

- Probit Regression for binary data:
  - Distribution of Y: $y_i \sim Bin(1, p_i)$
  - Link function: $\Phi^{-1}(\mu_i) = \Phi^{-1}(p_i) = \eta_i = \mathbf{X}_i\boldsymbol{\beta}$, where Φ is the standard normal cdf
  - Likelihood:

$$p(\mathbf{y}|\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right)^{y_i} \left(\frac{1}{1+e^{\eta_i}}\right)^{1-y_i}$$

Probit versus Logistic

- The probit and logit models will be similar, in practice, differing only in the extremes of the tails
- The probit model is attracting in Bayesian analysis, as $Pr(y_i = 1) = \Phi(\mathbf{X}_i\beta)$ is equivalent to the hierarchy

$$\begin{aligned} u_i &\sim N(\mathbf{X}_i\beta, 1) \\ y_i &= \begin{cases} 1, & \text{if } u_i > 0 \\ 0, & \text{if } u_i < 0 \end{cases} \end{aligned}$$

  where u_i is a latent continuous variable. $\Rightarrow$ closed-form full conditional for Bayesian computation and easy to include random effects.

- However, the logit model is often preferred as it has easier interpretation in terms of log-odds.

Common Generalized Linear Models

- Poisson Regression for count data:
  - Distribution of Y: $y_i \sim Poisson(\mu_i)$
  - Link function: $\log(\mu_i) = \eta_i = \mathbf{X}_i\boldsymbol{\beta}$
  - Likelihood:

$$p(\mathbf{y}|\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{1}{y_i!} e^{-e^{\eta_i}} \left(e^{\eta_i}\right)^{y_i}$$

- Overdispersed Poisson Model:
  - Link function: $\log(\mu_i) = \mathbf{X}_i\boldsymbol{\beta} \ + \ \epsilon_i \ , \ \epsilon_i \sim N(0, \sigma^2)$
  - Account for overdispersion in count data.
- Hurdle Poisson Model:
  - Distribution of Y: $y_i \sim p_i\delta_0 + (1 - p_i)Poisson(\mu_i)$
  - Link function: $\log\left(\frac{p_i}{1-p_i}\right) = \mathbf{X}_i\boldsymbol{\beta}_1, \ \log(\mu_i) = \mathbf{X}_i\boldsymbol{\beta}_2$
  - Account for excessive zeros in count data.

Prior distributions for GLMs

- Bayesian analysis of GLMs can be completed using flat or non-informative priors on the regression parameters $\Rightarrow$ this will be similar to MLEs
- More often, though, normal priors (either non-informative or hierarchical priors) are used for regression parameters.
- Informative priors are useful in situations where identifiability is challenging

Example: beetles under CS_2 exposure

- Recall the bettles data, which record the number of beetles killed after exposure to CS_2.

Dosage X_i	# killed y_i	# exposed n_i
1.6907	6	59
1.7242	13	60
1.7552	18	62
1.7842	28	56
1.8113	52	63
1.8369	52	59
1.8610	61	62
1.8639	60	60

- The outcome, the number killed, follows a binomial distribution

$$y_i \sim bin(n_i, p_i)$$

Beetles Example

- We consider GLMs with three different link functions
  - Logistic model:
  $$logit(p_i) = \log[p_i/(1-p_i)] = \alpha + X_i\beta \ .$$
  - Probit model:
  $$probit(p_i) = \Phi^{-1}(p_i) = \alpha + X_i\beta \ ,$$
  - Complementary log-log (cloglog) model:
  $$cloglog(p_i) = \log[-\log(1-p_i)] = \alpha + X_i\beta \ .$$
- Prior: flat priors for α and β

Beetles Example

- The regression coefficients will have different interpretations for each link function and are NOT comparable.
- Instead, we compare the fitted values $E(Y_i|X_i)$
- BUGS code ...
- Conclusion: The underlying regression parameters were quite different, but their fitted values are similar.

The Problem of Separation

- Separation happens when a single predictor perfectly predicts a binary outcome.
- In frequentist setting (using iterative weighted least squares), separation will result in infinite MLE estimates, which makes no sense in application.
- Bayesian GLM can easily solve the problem with a weakly informative prior.

Summaries

- The hierarchical models can easily incorporate complex dependence structures by including random effects with hierarchical priors.
- In Bayesian GLM, vague priors will result in similar inference to the frequentist estimation.
- Careful prior specifications can result in sensible inference when frequentist inference is challenging.
- Bayesian GLM can be extended to include various features (e.g. overdispersion, complex correlations) with simple additions to the hierarchy.