STA 365: Applied Bayesian Statistics

Boris Babic
Assistant Professor, University of Toronto

Week 10A: Predictions

Predictions

- Say we have a new covariate vector $\mathbf{X}_{\text{new}}$ and we would like to predict the corresponding response Y_{new}.
- A plug-in approach would fix $\boldsymbol{\beta}$ and σ at their posterior means $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ to make predictions
$$Y_{\text{new}} \mid \boldsymbol{\beta}, \hat{\sigma}^2 \sim \mathrm{N}(\mathbf{X}_{\text{new}}\hat{\boldsymbol{\beta}}, \hat{\sigma}^2).$$
- However, this plug-in approach suppresses uncertainty about $\boldsymbol{\beta}$ and σ^2.
- Therefore these prediction intervals will be slightly too narrow leading to under coverage.

Predictions

We should really account for all uncertainty when making predictions, including our uncertainty about $\boldsymbol{\beta}$ and σ^2.

We really want to PPD

$$\begin{aligned} \pi(Y_{\text{new}} \mid \mathbf{Y}) &= \int \pi(Y_{\text{new}}, \boldsymbol{\beta}, \sigma^2 \mid \mathbf{Y}) d\boldsymbol{\beta} d\sigma^2 \\ &= \int \pi(Y_{\text{new}} \mid \boldsymbol{\beta}, \sigma^2) \pi(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{Y}) d\boldsymbol{\beta} d\sigma^2 \end{aligned}$$

Marginalizing over the model parameters accounts for their uncertainty

Predictions

MCMC naturally gives draws from Y_{new}'s PPD

- For MCMC iteration t we have $\boldsymbol{\beta}^{(t)}$ and $\sigma^{2(t)}$.
- For MCMC iteration t we sample

$$Y_{\text{new}}^{(t)} \sim \mathrm{N}\left(\mathbf{X}\boldsymbol{\beta}^{(t)}, \sigma^{2(t)}\right).$$

- $Y_{\text{new}}^{(1)}, \ldots, Y_{\text{new}}^{(S)}$ are samples from the PPD.

Thus, "Bayesian methods" naturally quantify uncertainty.

JAGS can handle it.

Logistic Regression

- Other forms of regression follow naturally from linear regression
- For example, for binary responses $y_i \in \{0, 1\}$, we may use the logistic regression

$$\operatorname{logit}\{\Pr(y_i = 1)\} = \eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

- The logit link is the log-odd $\operatorname{logit}\{x\} = \log[x/(1-x)]$.
- Then β_j represents the increase in the log odds of an event corresponding to a one-unit increase in covariate j
- The expit transformation $\operatorname{expit}(x) = \exp(x)/\{1 + \exp(x)\}$ is the inverse of logit. and

$$\Pr(y_i = 1) = \operatorname{expit}(\eta_i) \in [0, 1].$$

Logistic Regression

- Bayesian logistic regression requires a prior for β
- All of the priors we have discussed for linear regression (Zellner, BLASSO, etc) can apply for logistic regression
- Computationally the full conditional distributions are no longer conjugate and so we must use Metropolis sampling
- It is fast in JAGS.

Hierarchical Modeling

- 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

Hierarchical Modeling

- Often Bayesian models can we written in the following layers of the hierarchy
- **Data layer**: $[y \mid \theta, \alpha]$ is the likelihood for the observed data y
- **Process layer**: $[\theta \mid \alpha]$ is the model for the parameters θ that define the latent data generating process
- **Prior layer**: $[\alpha]$ prior for hyperparameters

Hierarchical Modeling

- Consider the classical one-way random effects model: for $i = 1, \ldots, n$ and $j = 1, \ldots, m$,
$$y_{ij} \sim \mathrm{N}(\theta_i, \sigma^2) \text{ and } \theta_i \sim \mathrm{N}(\mu, \tau^2)$$
where y_{ij} is the jth replicate for unit i and $\alpha = (\mu, \sigma^2, \tau^2)$ has an uninformative prior
- This hierarchy can be written using a directed acyclic graph (DAG; also called Bayesian network or belief network)

Hierarchical Modeling

- 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
- 1. $[\theta_i \mid \cdot]$
  2. $[\mu \mid \cdot]$
  3. $[\sigma^2 \mid \cdot]$
  4. $[\tau^2 \mid \cdot]$
- Each of these updates is a draw from a standard one-dimensional normal or inverse gamma

Hierarchical Modeling

- Data example: national wide daily ozone levels for one month
- Denote by $y_{i,j}$ the ozone measurement at spatial location $i(i = 1, \ldots, 100)$ and day $j(j = 1, \ldots, 31)$
- We consider the model

$$y_{ij} \sim \mathrm{N}(\mu + \alpha_i + \gamma_j, \sigma^2).$$

- μ is the overall mean.
- α_i is the random effect for location i.
- γ_j is the random effect of day j.

Hierarchical Modeling

- Model:

$$y_{i,j} \sim \mathrm{N}(\mu + \alpha_i + \gamma_j, \sigma^2),$$

- Priors for the fixed-effects model:

$$\alpha_j \sim \mathrm{N}(0, 10^4), \qquad \gamma_j \sim \mathrm{N}(0, 10^4).$$

- Priors for the random-effects model:

$$\alpha_j \sim \mathrm{N}(0, \sigma_\alpha^2), \qquad \gamma_j \sim \mathrm{N}(0, \sigma_\gamma^2).$$

$$\sigma_\alpha^2 \sim \mathrm{G}^{-1}(0.001, 0.001), \qquad \sigma_\gamma^2 \sim \mathrm{G}^{-1}(0.001, 0.001).$$

- What is the difference between these two prior settings?

Hierarchical Modeling

- Data example: bone density measurements for children at different ages.
- Let y_{ij} be the jth measurement for child i at the age x_i.

$$y_{ij} \sim \mathrm{N}(\gamma_{i0} + x_i\gamma_{i1}, \sigma^2).$$

- $\boldsymbol{\gamma}_i = (\gamma_{i0}, \gamma_{i1})^{\mathrm{T}}$ controls the growth curve for child i.
- These separate regression are tied together in the prior

$$\boldsymbol{\gamma}_i \sim \mathrm{N}(\boldsymbol{\beta}, \boldsymbol{\Sigma}),$$

  which borrows strength across children.
- This is a linear mixed-effects model: $\boldsymbol{\gamma}_i$ are random-effects specific to one child and $\boldsymbol{\beta}$ are fixed-effects common to all children

Hierarchical Modeling

- The random-effects covariance matrix is $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$
- σ_1^2 is the variance of the intercepts across children
- σ_2^2 is the variance of the slopes across children
- σ_{12} is the covariance between the intercepts and slopes
- Prior 1: $\sigma_1^2, \sigma_2^2 \sim \mathrm{G}^{-1}(0.001, 0.001)$ and $\rho \sim \sigma_{12}/(\sigma_1\sigma_2) \sim \mathrm{U}(-1, 1)$.
- Prior 2: Inverse Wishart works better in higher dimensions