Introduction to Bayesian Data Analysis

Statistical Models, Model Evaluation

Alix I. Gitelman

Statistics Department
gitelman@science.oregonstate.edu

July 27, 2016

Overview

Some Resources

- *Bayesian Data Analysis*, 3rd Edition by Gelman, Carlin, Stern, and Rubin (rather heavy on mathematics, but (and?) a great resource.
- *Introduction to Bayesian Statistics*, 2nd Edition by Bolstad (introductory, does not get into MCMC)
- *Data Analysis Using Regression and Multilevel/Hierarchical Models* by Gelman and Hill (intermediate, some chapters on MCMC and hierarchical models)
- *Matrix Algebra: Theory, Computations, and Applications in Statistics* by Gentle

Statistical Models

A statistical (or stochastic) model is usually characterized by a systematic part and a random parts (or several random parts):

$$y = g(x; \theta) + \epsilon.$$

Here y is a response variable; $g(x; \theta)$ is some function of explanatory information, x, and unknown parameters, θ; and ϵ is a mean zero error term, with some unknown variance, σ^2.

- Essentially, $g(x; \theta)$ helps us learn about the mean of y as some function of explanatory information. and ϵ helps us learn about the variance/covariance of y after we've accounted for that part of it due to $g(x; \theta)$.
- If $\epsilon \sim N(0, \sigma^2)$, then we can write:

$$y \sim N(g(x; \theta), \sigma^2)$$

Statistical Models

What might the function $g(x;\theta)$ look like?

- Multiple linear regression models:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + \epsilon$$

- Non-linear regression models; for example:

$$y = \alpha_1 + \alpha_2 e^{-\alpha_3 x} + \epsilon$$

- There are non-parametric models (e.g., gams, splines, mixtures)

Statistical Models

When the response, y isn't Normally distributed, we turn to generalized linear models; some examples:

- Logistic regression:

$$\begin{aligned} Y &\sim \mathit{binomial}(n, \pi) \\ \log\left(\frac{\pi}{1-\pi}\right) &= \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} \end{aligned}$$

- Log-linear regression:

$$\begin{aligned} Y &\sim \mathit{Poisson}(\lambda) \\ \log(\lambda) &= \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} \end{aligned}$$

Statistical Models

Of course in realistic situations, we have a whole sample of data:

- Responses: $y_1, y_2, \ldots, y_n$
- And, for each observation $i = 1, 2, \ldots, n$ we mights have $p - 1$ explanatory variables, $x_{i1}, x_{i2}, \ldots, x_{i,p-1}$.
- In the case of the Normal regression model, we also have $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$

Matrix Notation

A multiple linear regression model in matrix form:

$$\begin{array}{cccccc} \mathbf{Y} & = & \mathbf{X} & \boldsymbol{\beta} & + & \boldsymbol{\epsilon} \\ \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix} & = & \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{bmatrix} & \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} & + & \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix} \end{array}$$

Matrix Notation

To summarize, you'll often see statistical models written in matrix notation. Some examples:

- $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \sigma^2\mathbf{I})$
- $\mathbf{Y} \sim Bin(\mathbf{n}, \boldsymbol{\pi})$, with $\log\left(\frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}}\right) = \mathbf{X}\boldsymbol{\beta}$
- $\mathbf{Y} = \alpha_1\mathbf{1} + \alpha_2 e^{-\alpha_3\mathbf{X}} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \sigma^2\mathbf{I})$

More to come when we talk about hierarchical models...

Model Evaluation

This quote is usually attributed to George Box:

> *"All models are wrong, but some are useful."*

For our models to be useful, we should evaluate them for how well the data fit them (and/or how well they predict data) and whether the assumptions we make are reasonable for the data in hand.

- Is the exponential decay model a good one for ice core CO_2 data?
- Specifically, is the assumption of independent error terms reasonable?

Examining Model Residuals

A common way to evaluate model fit is to look at plots of the fitted values and the residuals from that model.

- Residuals are: $e_i = y_i - \tilde{y}_i$, for $i = 1, 2, \ldots, n$, where $\tilde{y}_i$ is the fitted values from the model corresponding to observation i.
- For our non-linear regression model, we could take

$$\tilde{y}_i = \tilde{\alpha}_1 + \tilde{\alpha_2} e^{-\tilde{\alpha}_3 x_i}$$

  where $\tilde{\alpha}_1, \tilde{\alpha}_2$ and $\tilde{\alpha}_3$ are posterior estimates of the parameters (e.g., posterior means)
- Or, we could use **posterior predictive samples** and then average those

Posterior Predictive Checks

The idea is that we use several draws from the posterior distribution to generate predictive samples—these are just replicated data, but generated based on draws of the parameters from the posterior distribution.

- Typically, we might generate 100 or 1000 replicated (predictive) samples.
- Then evaluate how similar the predictive samples are to the observed sample.
- A good correspondence between posterior predictive samples and the observed data suggests a good fit.

Posterior Predictive Checks

In posterior predictive checking, obtaining additional samples is easy—usually an additional few lines of code.

- But how do we make a comparison between the original data and the predictive samples?
  - Typically we try to devise a statistic that might highlight differences between observations and predictions?
  - Any thoughts for such a statistic in our non-linear regression example? [think about the assumption that might be concerning]

Posterior Predictive Checks

Over to R and NIMBLE:

1. Devise a statistic for comparing observations to predictions
2. Obtain posterior predictive draws and calculate the statistic
3. Plot and/or summarize results.

The statistic(s) you use for comparison and how you evaluate them are problem-dependent.

Other Methods

There are other methods for model evaluation and/or model comparison:

- Training set/validation set
- Cross-validation
- Information criteria
- Posterior model probabilities

Fixed and Random Effects

Broadly speaking, here are a few ways to think about fixed versus random effects:

> *Fixed effects are what we want to make inference about; random effects are what we should account for in making inference about fixed effects.*
>
> *Fixed effects influence the mean, random effects influence the variance.*
>
> *If observations are collected repeatedly for some experimental or observational unit, or if observations are grouped or clustered together somehow: think random effects to account for the grouping/clustering.*

Examples

Fixed effects:

- fertilizer treatments
- gender
- light versus shade
- logged versus unlogged

Random Effects:

- individual (thinking about repeated measurements)
- forest stand
- time
- space

Our Non-linear Regression Model

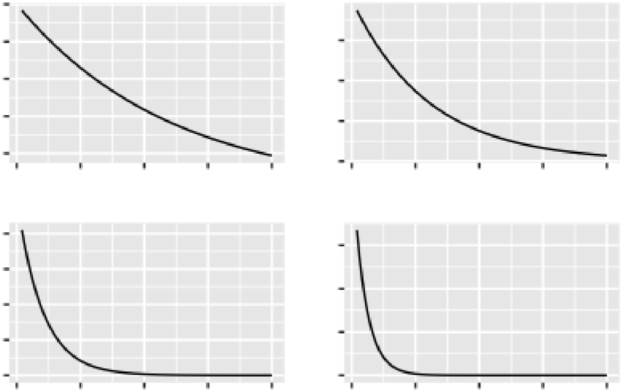
Let's consider a mixed effects model with a random effect for time:

$$\mathbf{y} = \alpha_1 \mathbf{1} + \alpha_2 e^{-\alpha_3 \mathbf{x}} + \boldsymbol{\eta} + \boldsymbol{\epsilon}$$

($\mathbf{1}$ is an $n \times 1$ vector of 1's).

- Here, $\boldsymbol{\eta}$ is an $n \times 1$ vector of random effects. We assume that it follows a multivariate Normal distribution, with mean zero and variance-covariance matrix Σ, where each entry, $\Sigma[i,j] = \tau^2 e^{-|x_i - x_j|/\theta}$ (exponential covariance).
- The $\boldsymbol{\epsilon}$ vector contains, as before, independent $N(0, \sigma^2)$, and we assume that $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$ are independent.

Exponential Covariance



Mixed Effects Non-linear Regression Model

I've added two parameters, τ^2 and θ, to our model in an effort to better represent the serial correlation in the residuals:

$$\mathbf{y} = \alpha_1 \mathbf{1} + \alpha_2 e^{-\alpha_3 \mathbf{x}} + \boldsymbol{\eta} + \boldsymbol{\epsilon}$$

with $\boldsymbol{\eta} \sim MVN(\mathbf{0}, \Sigma), \boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I})$ ($\mathbf{I}$ is the $n \times n$ identity matrix) and $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$ independent.

- The likelihood is still Normal, since both of the random effect and the independent errors are Normal. The means of the y_i's are still $\alpha_1 + \alpha_2 e^{-\alpha_3 x_i}$, but now the variance-covariance of the y_i's is more complicated.

Next Steps

How do you have to modify your NIMBLE code to deal with this random effect?

- The likelihood has changed because the variance-covariance has changed
- Then, in the way that I specified a function for the variance-covariance, you need to specify priors for τ and θ