

# Biostatistics 682: Applied Bayesian Inference

## Lecture 11: Beyond Linear Regression

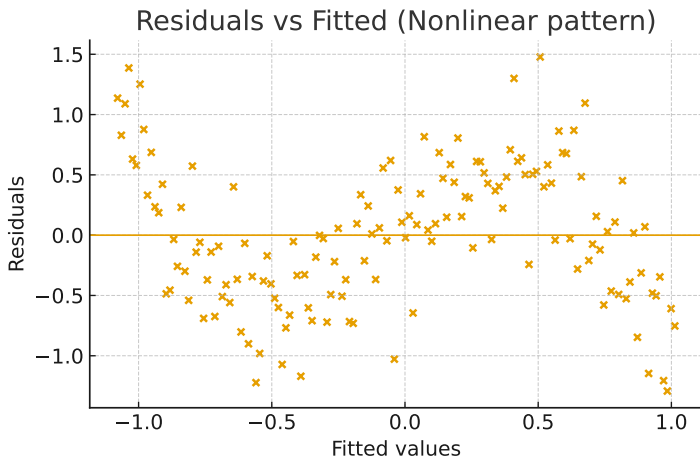
**Jian Kang**

Department of Biostatistics  
University of Michigan, Ann Arbor

# Why Go Beyond Linear Regression?

- Classical linear models assume:
  - Linear relationship between predictors and response
  - Homoscedastic errors and independence
  - Gaussian likelihood
- In practice, we often encounter:
  - Clustered or hierarchical data (e.g., repeated measures)
  - Nonlinear relationships
  - Binary or count outcomes
- Bayesian methods provide a unified way to handle these complexities.

# Limitations of Linear Models



Nonlinear trends or grouped patterns indicate model misspecification.

# 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

# Hierarchical models

- Often Bayesian models can be 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  $\theta$  that define the latent data generating process
- **Prior layer:**  $[\alpha]$  prior for hyperparameters

# One-way random-effects model

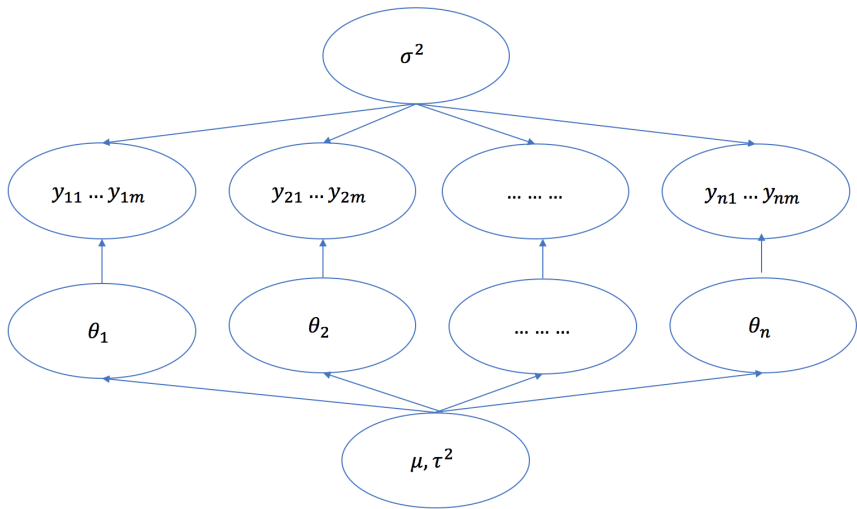
- Consider the classical one-way random effects model: for  $i = 1, \dots, n$  and  $j = 1, \dots, m$ ,

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

where  $y_{ij}$  is the  $j$ th 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)

# One-way random-effects model



# 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
- - 1  $[\theta_i \mid \cdot]$  follows a normal distribution
  - 2  $[\mu \mid \cdot]$  follows a normal distribution
  - 3  $[\sigma^{-2} \mid \cdot]$  follows a Gamma distribution
  - 4  $[\tau^{-2} \mid \cdot]$  follows a Gamma distribution
- Each of these updates is a draw from a standard one-dimensional normal or gamma distribution

# Two-way random effects model

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

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

- $\mu$  is the overall mean.
- $\alpha_i$  is the random effect for location  $i$ .
- $\gamma_j$  is the random effect of day  $j$ .

# Two-way random-effects model

- Model:

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

- Priors for the fixed-effects model:

$$\alpha_j \sim N(0, 10^4), \quad \gamma_j \sim N(0, 10^4).$$

- Priors for the random-effects model:

$$\alpha_j \sim N(0, \sigma_\alpha^2), \quad \gamma_j \sim N(0, \sigma_\gamma^2).$$

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

- What is the difference between these two prior settings?

# Random slopes model

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

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

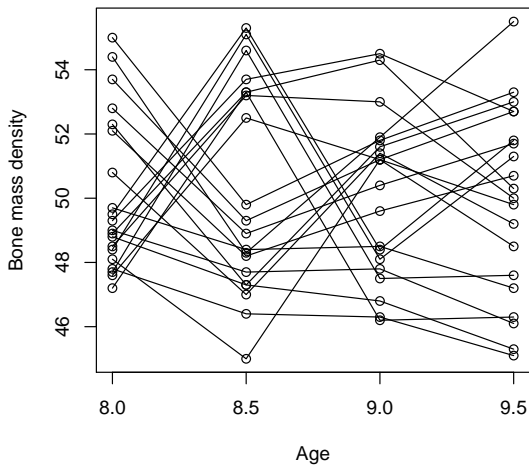
- $\gamma_i = (\gamma_{i0}, \gamma_{i1})^T$  controls the growth curve for child  $i$ .
- These separate regression are tied together in the prior

$$\gamma_i \sim N(\beta, \Sigma),$$

which borrows strength across children.

- This is a linear mixed-effects model:  $\gamma_i$  are random-effects specific to one child and  $\beta$  are fixed-effects common to all children

# Bone mass density



# Prior for a covariance matrix

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

# Non- and Semi-parametric modeling

- Nonparametric (NP) methods attempt to analyze the data by making the fewest number of assumptions as possible
- NP methods are generally are robust and flexible, but less powerful than correctly specified parametric models
- Most frequentist NP methods completely avoid specifying a model
- For example, a rank or sign test to compare two means

# Non- and Semi- parametric modeling

- Bayesian methods need a likelihood in order to obtain a posterior, so you cannot completely avoid specifying a model
- Bayesian NP (BNP) then attempts to specify a model that is so flexible that it almost certainly captures the true model
- One definition of the BNP model is one that has infinitely-many parameters
- In some cases, NP models are difficult conceptually and computationally, and so semi-parametric models with a large but finite number of parameters are useful approximations.

# Parametric simple linear regression

- Consider the classical parametric model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2).$$

Assumptions:

- $\epsilon_i$  are independent
- The mean of  $y_i$  is linear in  $x_i$
- The residual distribution does not depend on  $x_i$

Alternatives:

- Parametric alternatives such as a time series model
- Let  $\epsilon_i \sim F$ , and place a prior on the distribution  $F$ .
- Let  $E(y_i | x_i) = g(x_i)$  put a prior on the function  $g$ .
- Heteroskedastic regression  $\text{Var}(\epsilon_i) = \exp\{\alpha_0 + \alpha_1 x_i\}$ .

# Bayesian Nonparametric regression

- The mean of  $y_i$  is  $g(x_i)$ , where  $g$  is a function
- Parametric models include
  - Linear:  $g(x) = \beta_0 + \beta_1 x$ ,
  - Quadratic:  $g(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ ,
  - Logistic:  $g(x) = \beta_0 + \beta_1 \frac{\beta_2 + \beta_3 x}{1 + \exp(\beta_2 + \beta_3 x)}$ .
- NP regression puts a prior on the curve  $g(x)$ , rather than parameters  $\beta_1, \dots, \beta_p$  that determine the parametric model.
  - For example, Gaussian process priors:

$$g \sim \text{GP}(\mu, \kappa),$$

where  $\mathbb{E}\{g(x)\} = \mu(x)$  and  $\text{Cov}\{g(x), g(x')\} = \kappa(x, x')$ .

- Gaussian processes: a stochastic process for which any finite linear combination of samples has a joint Gaussian

$$[g(x_1), \dots, g(x_n)] \sim \text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where  $\boldsymbol{\mu} = \{\mu(x_1), \dots, \mu(x_n)\}$  and  $\boldsymbol{\Sigma} = \{\kappa(x_i, x_j)\}_{1 \leq i, j \leq n}$ .

# Bayesian Semiparametric regression

- Semiparametric regression approximates the function  $g$  using a finite basis expansion

$$g(x) = \sum_{j=1}^J B_j(x) \beta_j,$$

where  $B_j(x)$  are known basis functions and  $\beta_j$  are unknown coefficients that determine the shape of  $g$

- Example: the cubic spline basis functions are

$$B_j(x) = (x - \nu_j)_+^3,$$

where  $\nu_j$  are fixed knots that span the range of  $x$ .

- Many other expansions exist: wavelets; Fourier, etc
- Fact: A basis expansion of  $J$  terms can match the true curve  $g$  at any  $J$  points  $x_1, \dots, x_J$ .
- So increasing  $J$  gives an arbitrarily flexible model

- The model is  $y_i \sim \text{N}(\mathbf{B}_i^T \boldsymbol{\beta}, \sigma^2)$ , where  $\beta_j \sim \text{N}(0, \tau^2)$  and  $B_i$  is comprised of the known basis functions  $B_j(x_i)$ , where  $\mathbf{B}_i = \{B_1(x_i), \dots, B_J(x_i)\}^T$ .
- Therefore, the model is usual linear regression model and is straightforward to fit using MCMC.
- How to pick  $J$ ?
- Can we have more basis functions than observations?
- What would you do if your prior was that  $g$  was probably quadratic, but you are not 100% sure about this. That is, your prior is that  $g(x) \approx \beta_0 + \beta_1 x + \beta_2 x^2$ .

# Bayesian 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

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

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

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

# Bayesian logistic regression

- Bayesian logistic regression requires a prior for  $\beta$
- 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.

# When to Use Which Model?

- Hierarchical data  $\rightarrow$  Random effects / slopes
- Smooth nonlinear effects  $\rightarrow$  Semiparametric or GP
- Binary / count outcomes  $\rightarrow$  Logistic or Poisson
- Combine multiple extensions in one hierarchical framework