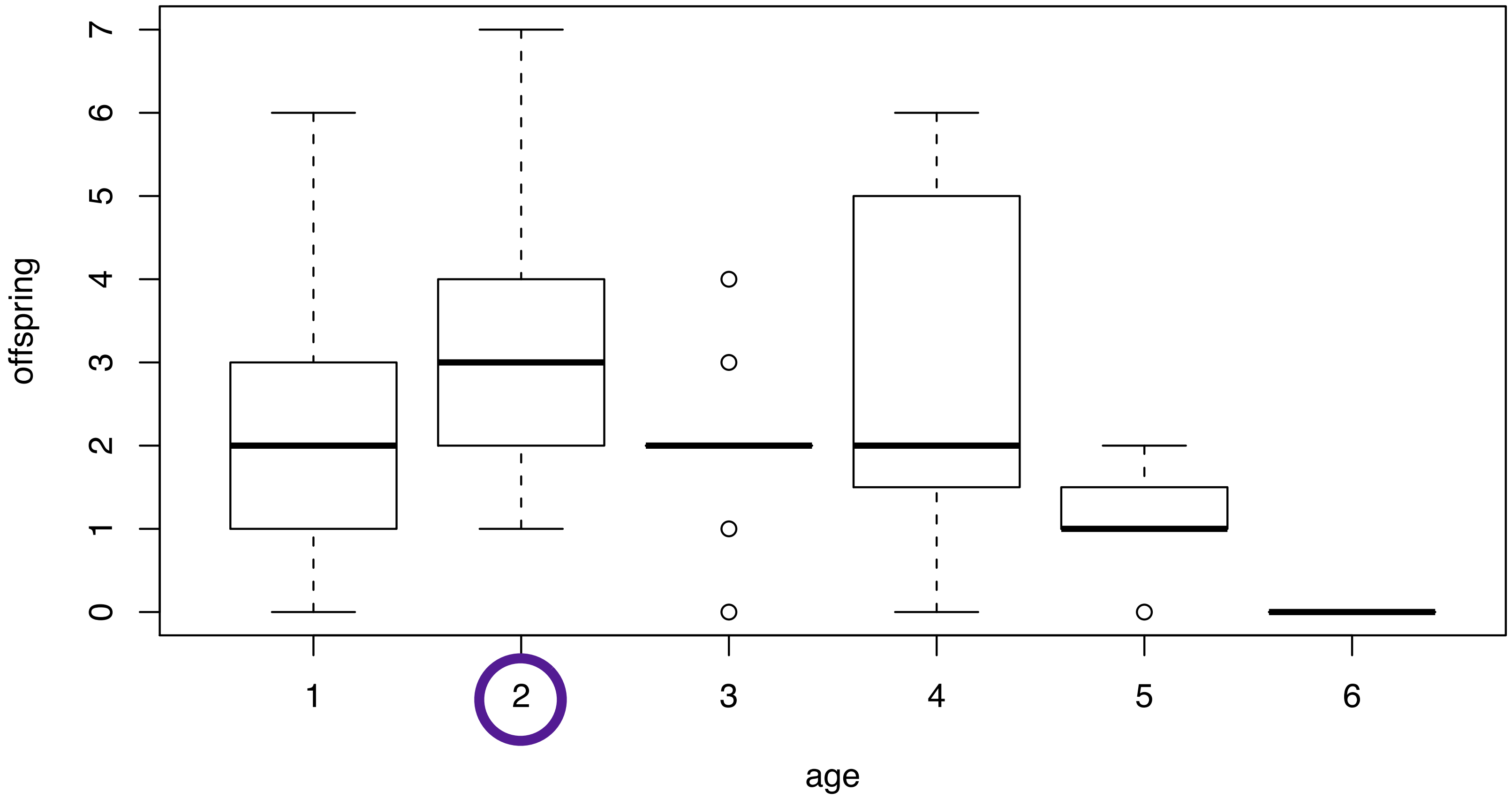
Part 8:
GLMs and Hierarchical LMs and GLMs

Example: Song sparrow reproductive success

Arcese et al., (1992) provide data on a sample from a population of 52 female song sparrows studied over the course of a summer, during which their reproductive activities were recorded



2-year-old birds had the highest median reproductive success, declining thereafter

Example: Poisson model

Since the number of offspring for each bird is a non-negative integer $\{0, 1, 2, \ldots\}$, a simple probability model for

$$\begin{aligned} Y &= \text{the number of offspring} \\ \text{conditional on} \quad x &= \text{age} \end{aligned}$$

would be a Poisson model

$$\{Y|x\} \sim \mathrm{Pois}(\theta_x)$$

One possibility would be to estimate θ_x separately for each age group

Example: Adding stability

However, the number of birds of each age is small and so the estimates of θ_x would be imprecise

To add stability to the estimation we will assume that the mean number of offspring is a smooth function of age

We will want to allow this function to be quadratic so that we can represent

- the increase in mean offspring while birds mature
- and the decline they experience thereafter

Example: A linear model?

One possibility would be to express θ_x as

$$\theta_x = \beta_1 + \beta_2 x + \beta_3 x^2$$

However, this might allow some values of θ_x to be negative, which is not physically possible

As an alternative, we will model the log-mean of Y in terms of this regression so that

$$\log \mathbb{E}\{Y|x\} = \log \theta_x = \beta_1 + \beta_2 x + \beta_3 x^2$$

which means that, for all x and β

$$\mathbb{E}\{Y|x\} = \exp\{\beta_1 + \beta_2 x + \beta_3 x^2\} > 0$$

Poisson regression

The resulting model

$$\{Y|x\} \sim \mathrm{Pois}(\exp\{x^\top \beta\})$$

is called a Poisson regression model, or log-linear model

The term $x^\top \beta$ is called the linear predictor

In the regression model the linear predictor is linked to $\mathbb{E}\{Y|x\}$ via the $\log$ function, and so we say that this model has a $\log$ link

Generalized linear model

The Poisson regression/log-linear model is a type of generalized linear model (GLM), a model which

- allows more general response distributions for Y than the normal distribution

- relates a function of the expectation $\mu = \mathbb{E}\{Y\}$ to a linear predictor $\eta = x^\top \beta$ through the link $g(\mu) = \eta$

These two choices define the GLM

Priors

As in the case of ordinary regression, a natural class of prior distributions for β is MVN

However, it is not the case that, when combined with the GLM likelihood (e.g., Poisson sampling model and log link), the resulting posterior distribution would be MVN

- a notable exception is when the Normal sampling model is used with the identity link, recovering the standard Bayesian LM

Conjugate priors for a GLM are not generally available, except in the above special case

Inference by MCMC

Therefore, our only recourse will be to proceed by the MH algorithm in the general case of the GLM

E.g., for our motivating Poisson regression example

So we have that $\log \mathbb{E}\{Y_i|x_i\} = \beta_1 + \beta_2 x_i + \beta_3 x_i^2$ where x_i is the age of sparrow i

We will abuse notation slightly by writing $x_i \equiv (1, x_i, x_i^2)$ so that we may use the simplified expression

$$\log \mathbb{E}\{Y_i|x_i\} = x_i^\top \beta$$

MH acceptance ratio

Suppose we take the prior $\beta \sim \mathcal{N}_3(0, 10 I_3)$

Given the current value of $\beta^{(s)}$ and a value of β^* generated from a symmetric proposal $q(\beta^{(s)}, \beta^*)$, the acceptance probably for the Metropolis algorithm is $\min\{1, A\}$ where

$$A = \frac{p(\beta^*|X, y)}{p(\beta^{(s)}|X, y)}$$

$$= \frac{\prod_{i=1}^n \text{Pois}(y_i; x_i^\top \beta^*)}{\prod_{i=1}^n \text{Pois}(y_i; x_i^\top \beta^{(s)})} \times \frac{\prod_{j=1}^3 \mathcal{N}(\beta_j^*; 0, 10)}{\prod_{j=1}^3 \mathcal{N}(\beta_j^{(s)}; 0, 10)}$$

Choosing a proposal

All that remains is to specify the proposal distribution $q(\beta^{(s)}, \beta^{*})$

A convenient choice is a MVN with mean $\beta^{(s)}$

In many problems, the posterior variance can be an efficient choice of a proposal variance

Although we do not know the posterior variance before running the MH algorithm, it is often sufficient just to use a rough approximation

Proposal variance

In a Bayesian LM, the posterior variance of β will be close to $\sigma^2(X^\top X)^{-1}$, where σ^2 is the variance of Y

In our Poisson regression, the model is that the log of Y has expectation equal to $x^\top\beta$, so it is sensible to try a proposal variance of $\hat{\sigma}^2(X^\top X)^{-1}$ where $\hat{\sigma}^2$ is the sample variance of

$$\{\log(y_i + 1/2), \ldots, \log(y_n + 1/2)\}$$

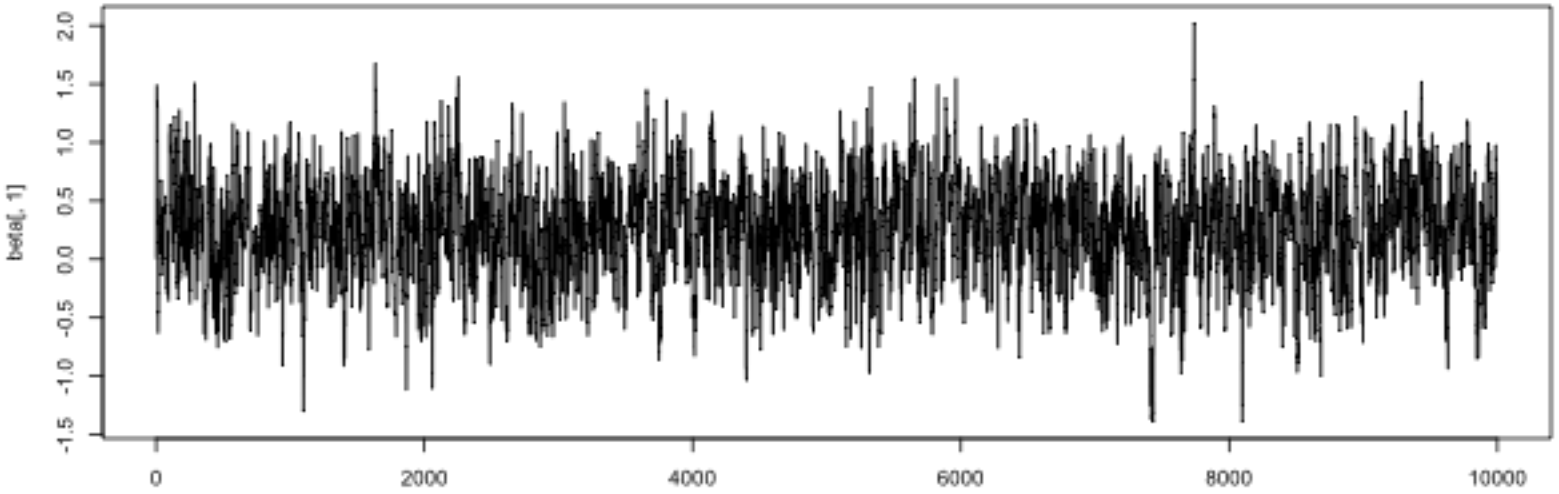
If this results in an acceptance rate that is too high or too low, we can always adjust the proposal variance accordingly

Example: MCMC for Sparrows log-linear model

Trace Plots

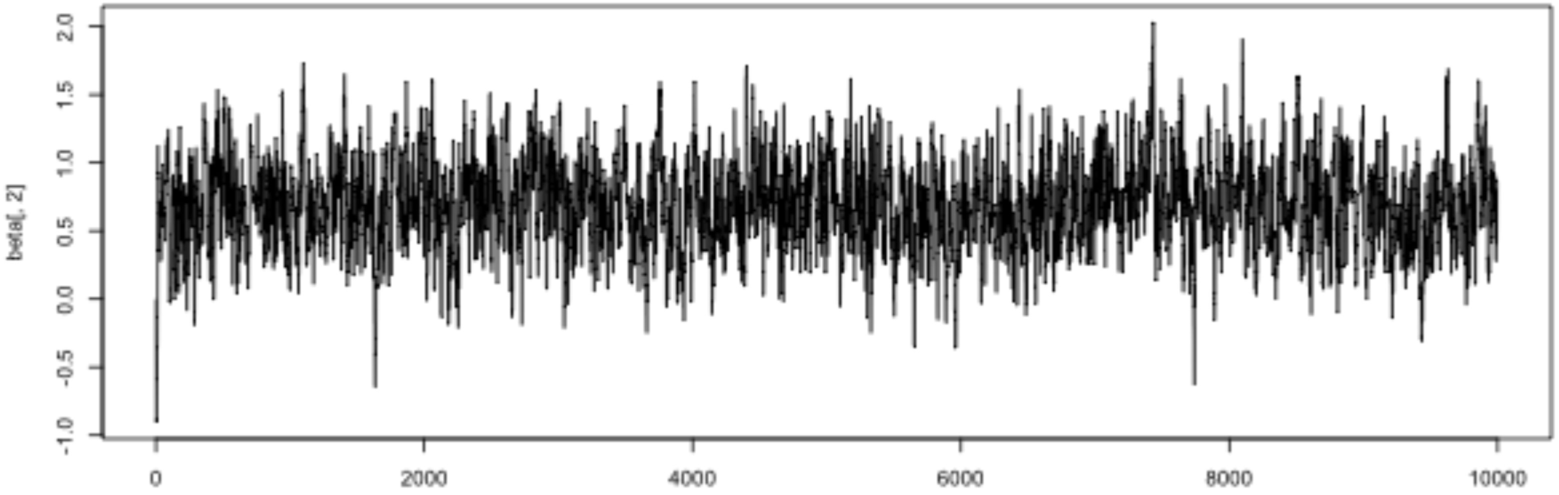
ESS

β_1



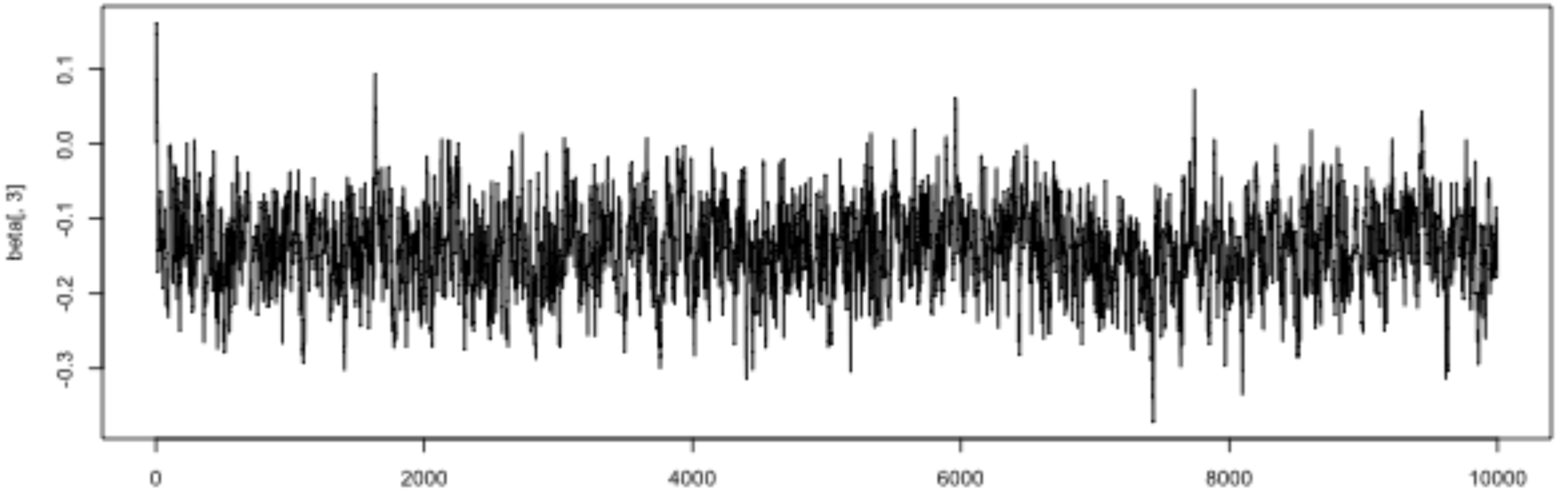
808

β_2



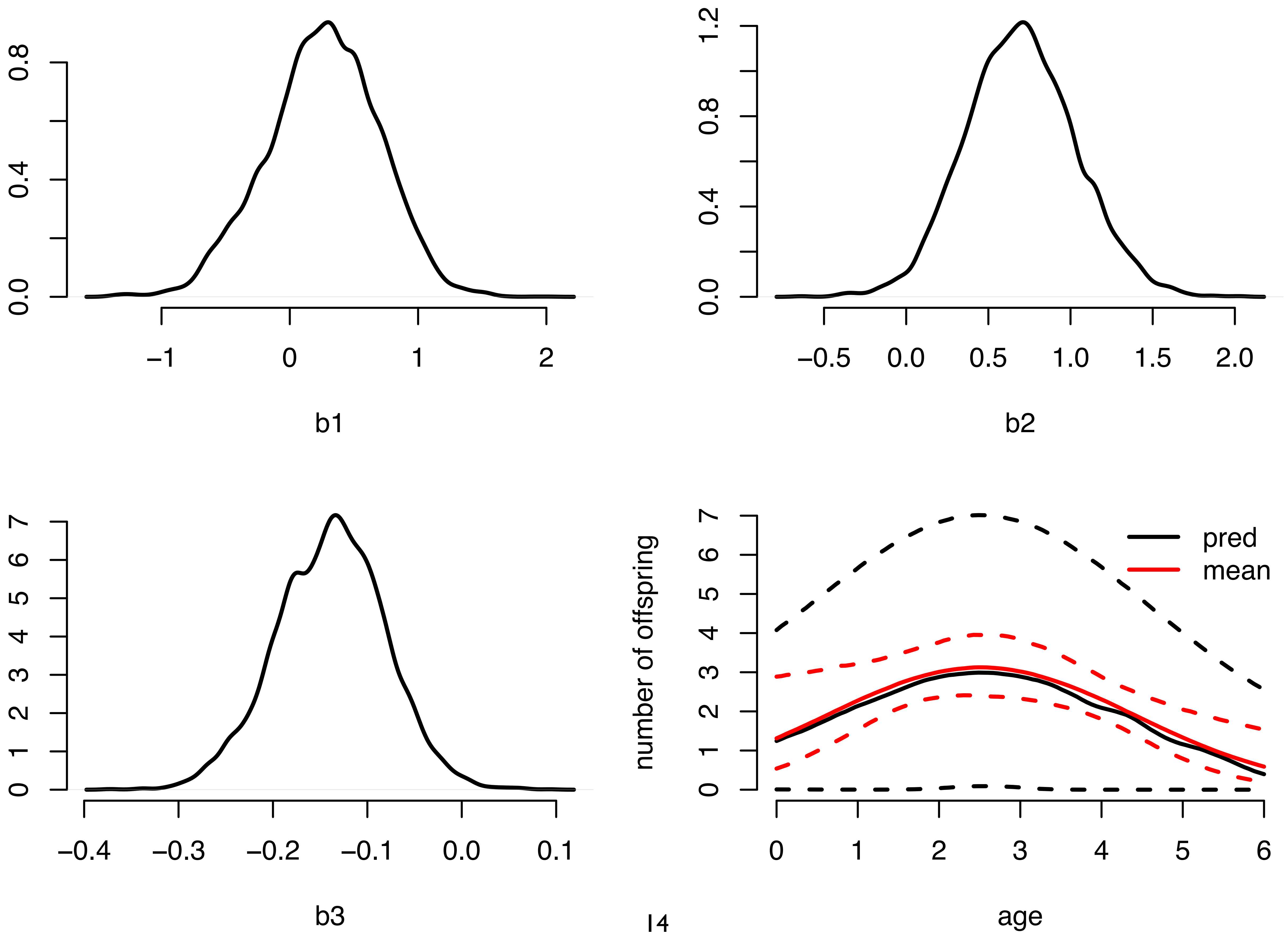
725

β_3



559

Example: Posterior marginals and predictive



Hierarchical regression

Here we shall extend the concept of hierarchical modeling to regression problems

- the regression model shall be used to describe within-group variation
- and a MVN model will be used to describe heterogeneity between groups
- first with LMs, then with GLMs

Example: Math score data

Let us return to the math score data which included scores of 10th grade children from 100 different large urban public high schools

- we estimated school-specific expected math scores, as well as how these expected values varied from school to school

Now suppose that we are interested in examining the relationship between math score and another variable, socioeconomic status (SES), which was calculated from parental income and education levels for each student in the dataset

Example: Math score data

With our hierarchical normal model we quantified the between-school heterogeneity in expected math score

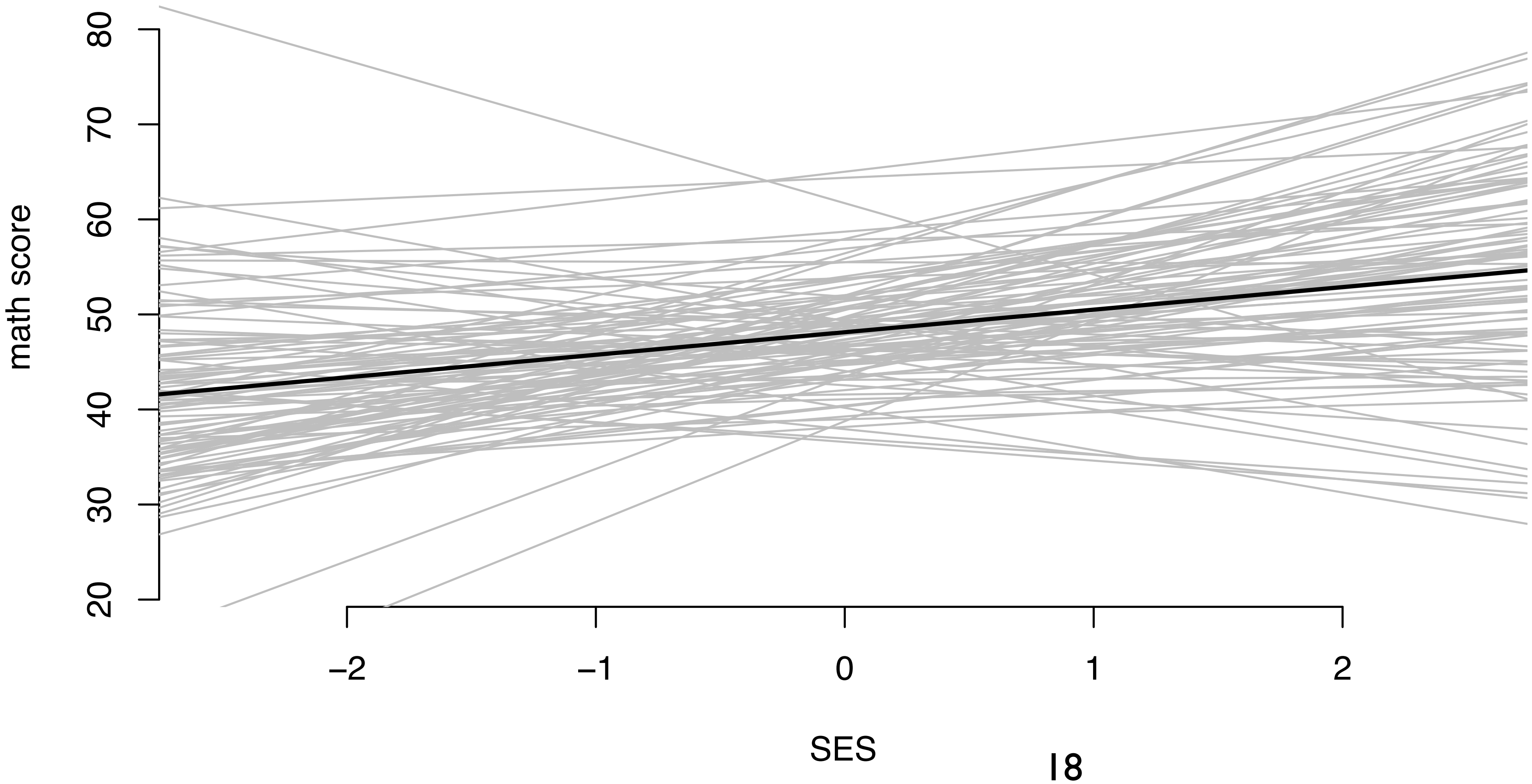
Given the amount of variation we observed it seems possible that the relationship between math score and SES might vary from school to school as well

A quick and easy way to assess this possibility is to fit a linear regression model of math score as a function of SES for each of the 100 schools in the dataset

Example: Math score data

To make the parameters more interpretable we center the SES scores within each school separately, so the sample average SES score within each school is zero

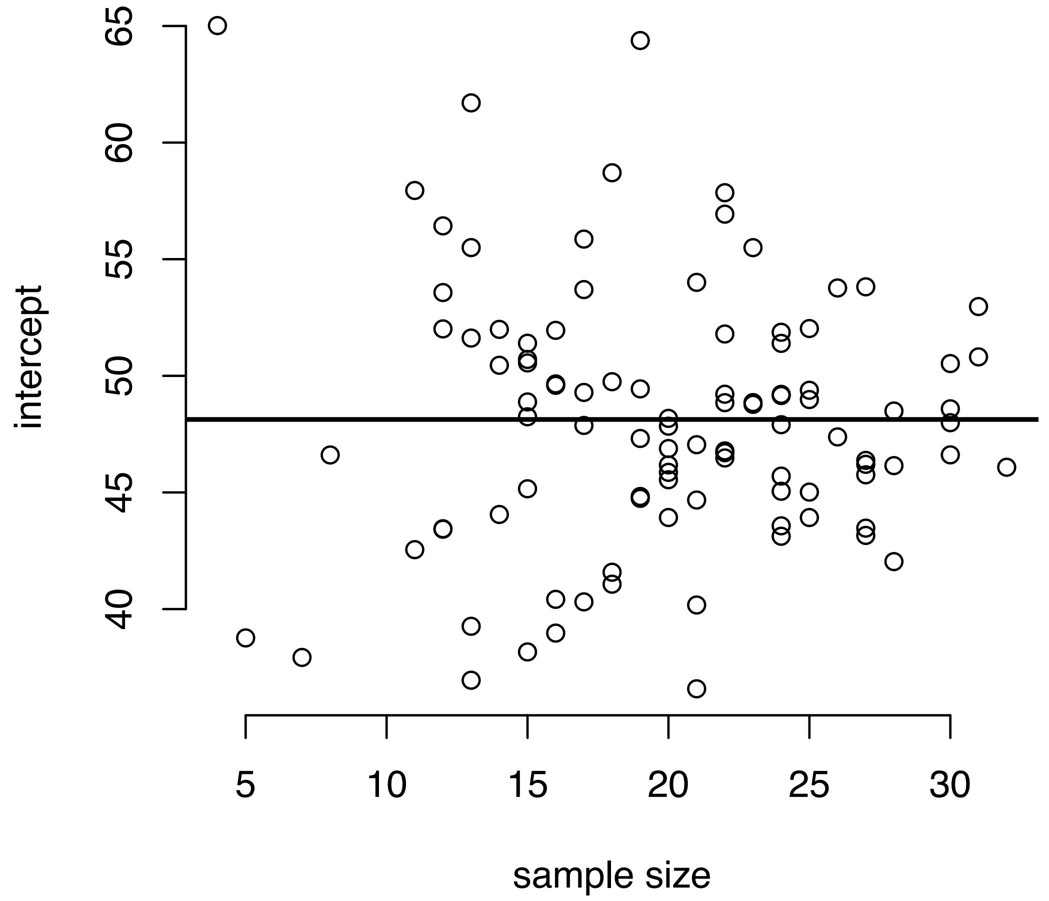
- as a result, the intercept of the regression line can be interpreted as the school-level average

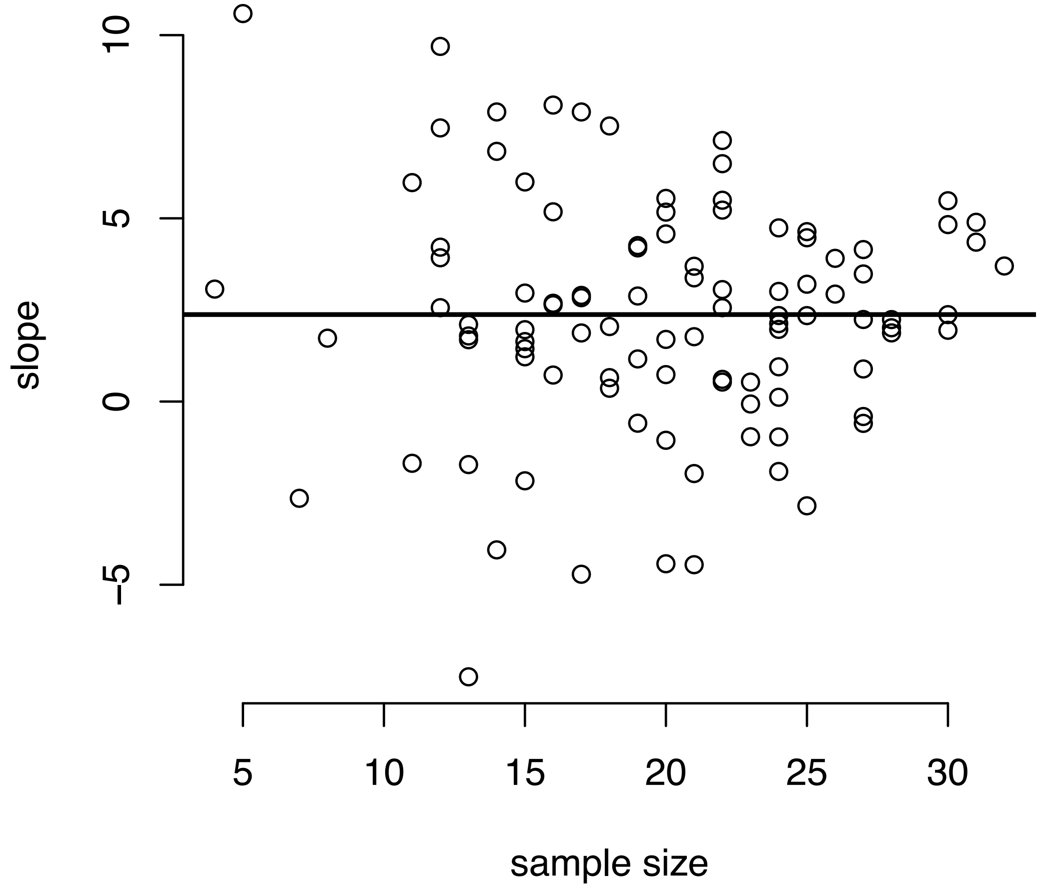


OLS regression lines for each school, and the average

Example: Math score data

It is also informative to plot the OLS slope and intercept as a function of the sample size





Schools with the highest sample sizes have regression coefficients that are generally close to the average

- extreme coefficients correspond to low sample sizes

Example: Pooling data

The smaller the sample size for the group, the more probable that unrepresentative data are sampled and an extreme OLS estimate is produced

Our remedy to this problem will be to stabilize the estimates for small sample size schools by sharing information across groups

- using a hierarchical model

Hierarchical LM

The hierarchical model in the linear regression setting is a conceptually straightforward generalization of the normal hierarchical model

- we use an ordinary regression model to describe the within-group heterogeneity

- then we describe the between-group heterogeneity using a sampling model for the group-specific regression parameters

Within-group model

Symbolically, our within-group sampling model is

$$Y_{i,j} = x_{i,j}^\top \beta_j + \varepsilon_{i,j}, \;\; \{\varepsilon_{i,j}\} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

where $x_{i,j}$ is a $p \times 1$ vector of covariates for observation i in group j

Expressing $Y_{1,j}, \ldots, Y_{n_j,j}$ as a vector Y_j and combining $x_{1,j}, \ldots, x_{n_j,j}$ into an $n \times p$ matrix, the within-group sampling model can be expressed equivalently as $Y_j \sim \mathcal{N}_{n_j}(X_j \beta_j, \sigma^2 I_{n_j})$

The group-specific data vectors $Y_1, \ldots, Y_m$ are conditionally independent given $\beta_1, \ldots, \beta_m$ and σ^2

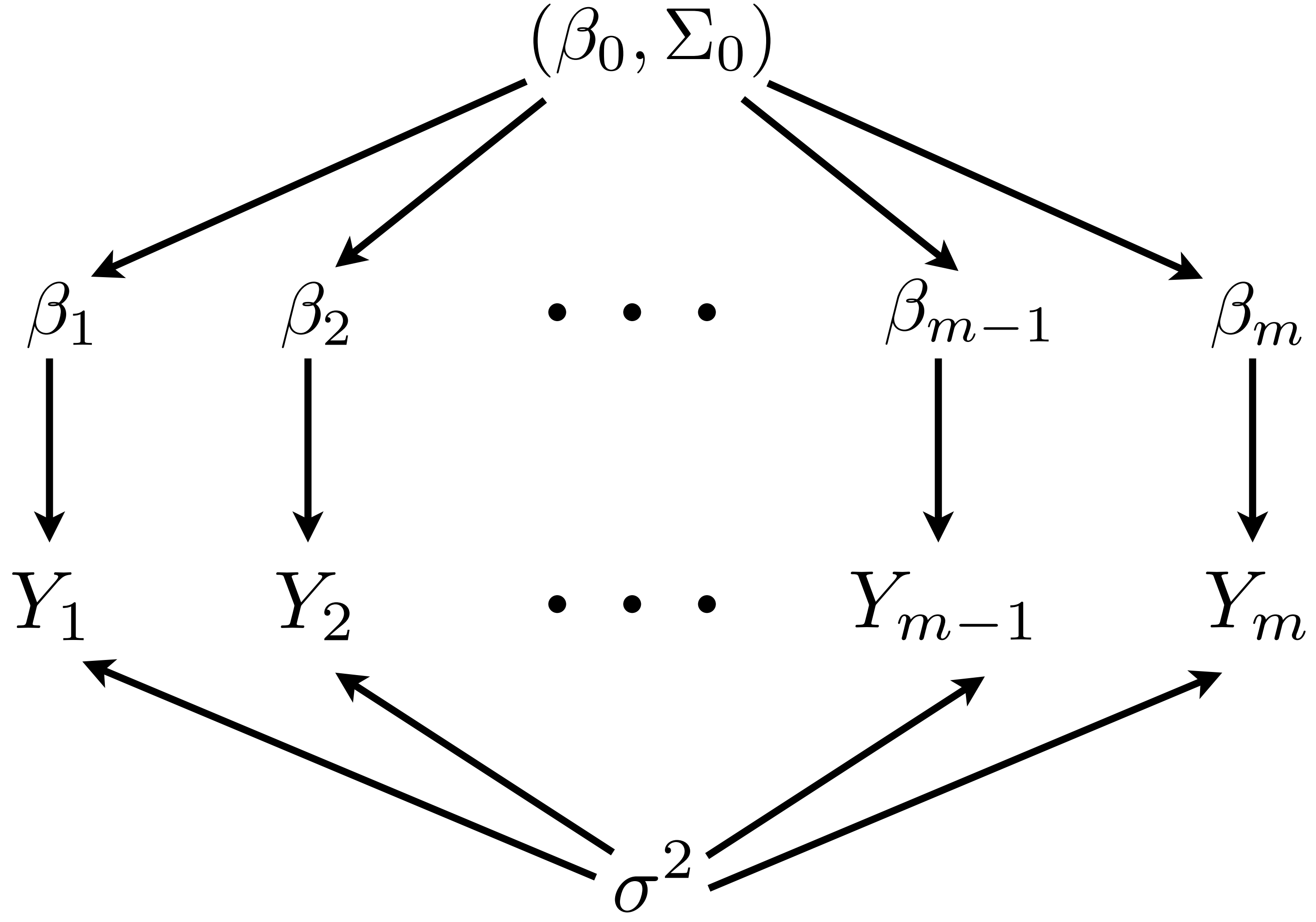
Between-group model

The heterogeneity among the regression coefficients $\beta_1, \ldots, \beta_m$ will be described with a between-group sampling model

The normal hierarchical regression model describes the across-group heterogeneity with a multivariate normal distribution, so that

$$\beta_1, \ldots, \beta_m \overset{\text{iid}}{\sim} \mathcal{N}_p(\beta_0, \Sigma_0)$$

Hierarchical diagram



The values of β_0 and Σ_0 are fixed but unknown parameters to be estimated

This hierarchical regression model is sometimes called a linear mixed effects model

Full conditionals

While computing the posterior distribution for so many parameters may seem daunting, the calculations involved in computing the full conditional distributions have the same mathematical structure as many examples we have come across before

Once we have the full conditional distributions we can iteratively sample from them to approximate the joint posterior distribution by GS

Full conditional of $\beta_1, \ldots, \beta_m$

The hierarchical LM shares information across groups via the parameters β_0, Σ_0 and σ^2

As a result, conditional on $\beta_0, \Sigma_0, \sigma^2$, the regression coefficients $\beta_1, \ldots, \beta_m$ are independent

Therefore, $\{\beta_j | y_j, X_j, \sigma^2, \beta_0, \Sigma_0\}$ is MVN with

$$\mathrm{Var}[\beta_j | y_j, X_j, \sigma^2, \beta_0, \Sigma_0] = (\Sigma_0^{-1} + X_j^\top X_j / \sigma^2)^{-1}$$
$$\mathbb{E}\{\beta_j | y_j, X_j, \sigma^2, \beta_0, \Sigma_0\} = (\Sigma_0^{-1} + X_j^\top X_j / \sigma^2)^{-1} (\Sigma_0^{-1} \beta_0 + X_j^\top y_j / \sigma^2)$$

Full conditionals of (β_0, Σ_0)

Our sampling model for the β_j's is that they are IID samples from a MVN with mean β_0 and covariance Σ_0

Therefore, if $\beta_0 \sim \mathcal{N}_p(\mu_0, \Lambda_0)$ then our previous result for MVN posterior conditionals gives that

$$\{\beta_0 | \beta_1, \ldots, \beta_m, \Sigma_0\} \sim \mathcal{N}_p(\mu_m, \Lambda_m)$$

where

$$\Lambda_m = (\Lambda_0 + m\Sigma_0^{-1})^{-1}$$

$$\mu_m = \Lambda_m(\Lambda_0\mu_0 + m\Sigma_0^{-1}\bar{\beta})$$

and $\bar{\beta} = \frac{1}{m}\sum \beta_j$

... continued ...

Likewise, if $\Sigma_0 \sim \mathrm{IW}(\nu_0, S_0^{-1})$ then

$$\{\Sigma_0 | \beta_0, \beta_1, \ldots, \beta_m\} \sim \mathrm{IW}(\nu_0 + m, [S_0 + S_\beta]^{-1})$$

where $$S_\beta = \sum_{j=1}^{m} (\beta_j - \beta_0)(\beta_j - \beta_0)^\top$$

Full conditional of σ^2

The parameter σ^2 represents the error variance, assumed to be common across all groups

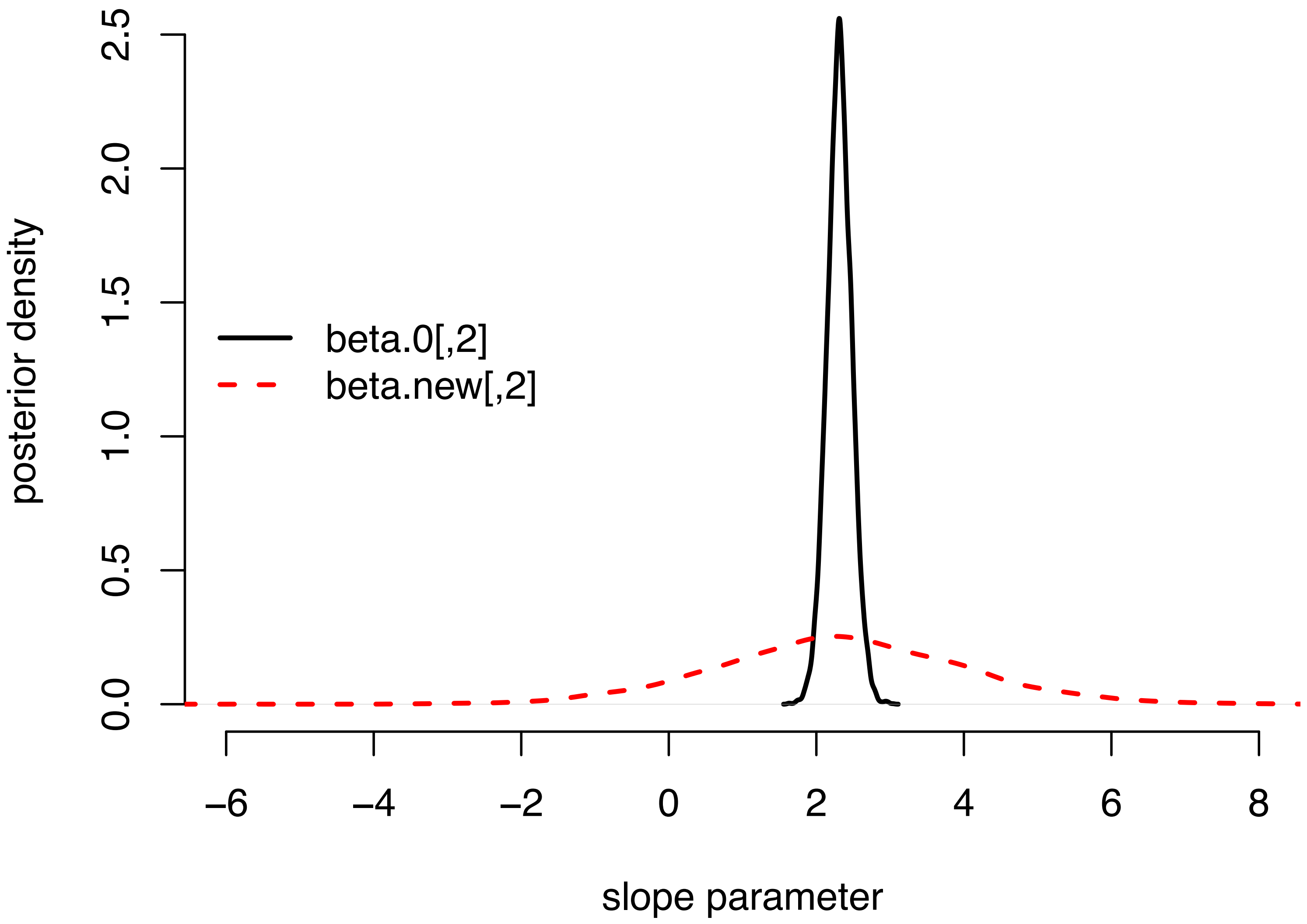
As such, conditional on $\beta_1, \ldots, \beta_m$, the data provide information about σ^2 via the sum of squared residuals from each group

With prior $\sigma^2 \sim \mathrm{IG}(\nu_0/2, \nu_0\sigma_0^2/2)$ we have

$$\{\sigma^2 | \beta_1, \ldots, \beta_m, \ldots\} \sim \mathrm{IG}\left(\frac{\nu_0 + \sum n_j}{2}, \frac{\nu_0\sigma_0^2 + \mathrm{SSR}}{2}\right)$$

$$\mathrm{SSR} = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (y_{i,j} - x_{i,j}^\top \beta_j)^2$$

Example: Analysis of math score data

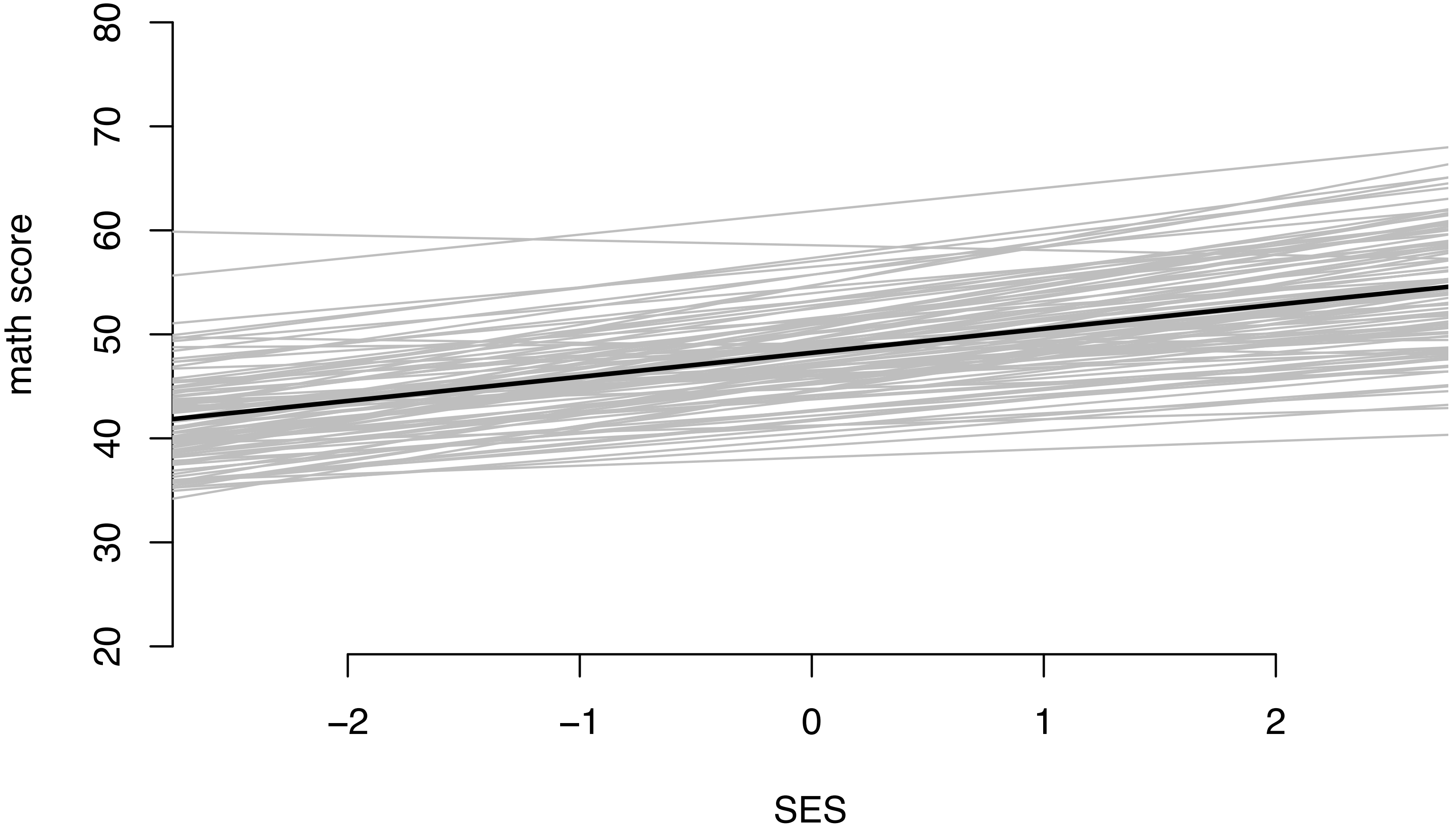


$\beta_0|y$ is extremely unlikely to be negative, but a new $\beta_{\text{new}}|\beta_0, y$ may indeed be negative

$$p(\beta_{\text{new}} < 0|y) \approx 0.0861$$

- therefore the population average slope is positive: higher SES yields higher score
- but it is not unlikely for a particular school to observe a reverse trend

Example: Analysis of math score data



hardly any slopes are negative

- compared to the plot of OLS lines that we saw before, these are more homogeneous
- this indicates how the hierarchical model is able to share information across groups, shrinking extreme regression lines towards the average

Hierarchical GLMs

The same Bayesian hierarchical modeling techniques extend to GLMs

- sometimes called a generalized linear mixed effects model

Such models are useful when we have a hierarchical data structure but the normal model for the within-group variation is not appropriate

For example, if the variable Y were binary or a count, then more appropriate models for within-group variation would be logistic or log-linear models, respectively

Basic model

A basic hierarchical GLM is

$$\beta_1, \ldots, \beta_m \overset{\text{iid}}{\sim} \mathcal{N}_p(\beta_0, \Sigma_0)$$

$$p(y_j | X_j, \beta_j, \gamma) = \prod_{i=1}^{n_j} p(y_{i,j} | x_{i,j}^\top \beta_j, \gamma)$$

with observations from different groups also being conditionally independent

In this formulation $p(y|x^\top \beta, \gamma)$ is a density whose mean depends on $x^\top \beta$, and γ is an additional parameter often representing variance or scale

For example

In the normal model $p(y|x^\top\beta,\gamma) = \mathcal{N}(y; x^\top\beta,\gamma)$ where γ represents the variance

In the Poisson model

$$p(y|x^\top\beta,\gamma) = \mathrm{Pois}(y; \exp\{x^\top\beta\})$$

and there is no γ parameter

Likewise, in the Binomial model

$$p(y|x^\top\beta,\gamma) = \mathrm{Bin}\left(y; n, \frac{\exp\{x^\top\beta\}}{1+\exp\{x^\top\beta\}}\right)$$

and there is no γ parameter

Inference

Estimation for the Hierarchical LM was straightforward because the full conditional distribution of each parameter was standard, allowing for GS

In contrast, for non-normal GLMs, typically only β_0 and Σ_0 have standard full conditional distributions

This suggests using the Metropolis-within-Gibbs algorithm to approximate the posterior distribution of the parameters

- using GS for updating (β_0, Σ_0), and
- MH for each β_j

In what follows we assume there is no γ parameter

GS for (β_0, Σ_0)

Just as in the hierarchical GLM, the full conditional distributions of β_0 and Σ_0 depend only on $\beta_1, \ldots, \beta_m$

This means that the form of $p(y|x^\top \beta)$ has no effect on the posterior conditional distributions of β_0 and Σ_0

Therefore, the full conditionals are MVN and IW, respectively

MH for $\beta_1, \ldots, \beta_m$

Updating β_j in a Markov chain can proceed by proposing a new value of β_j^* based on the current parameter values and then accepting or rejecting with the appropriate probability

A standard proposal distribution in this situation would be a MVN with mean equal to the current value of $\beta_j^{(s)}$ and some proposal variance $V_j^{(s)}$

In many cases, setting $V_j^{(s)}$ equal to a scaled version of $\Sigma_0^{(s)}$ produces a well-mixing Markov chain, although the task of finding the right scale might have to proceed by trial and error

The MCMC method

Putting these steps together results in the following MH algorithm for approximating

$$p(\beta_1, \ldots, \beta_m, \beta_0, \Sigma_0 | X_1, \ldots, X_m, y_1, \ldots, y_m)$$

Given the current values at scan s of the Markov chain, we obtain new values as follows

1. Sample $\beta_0^{(s+1)}$ from its full conditional distribution
2. Sample $\Sigma_0^{(s+1)}$ from its full conditional distribution
3. For each $j \in \{1, \ldots, m\}$

    a) propose a new value of β_j^*

    b) set $\beta_j^{(s+1)}$ equal to β_j^* or $\beta_j^{(s)}$ with the appropriate probability

Example: Analysis of tumor location data

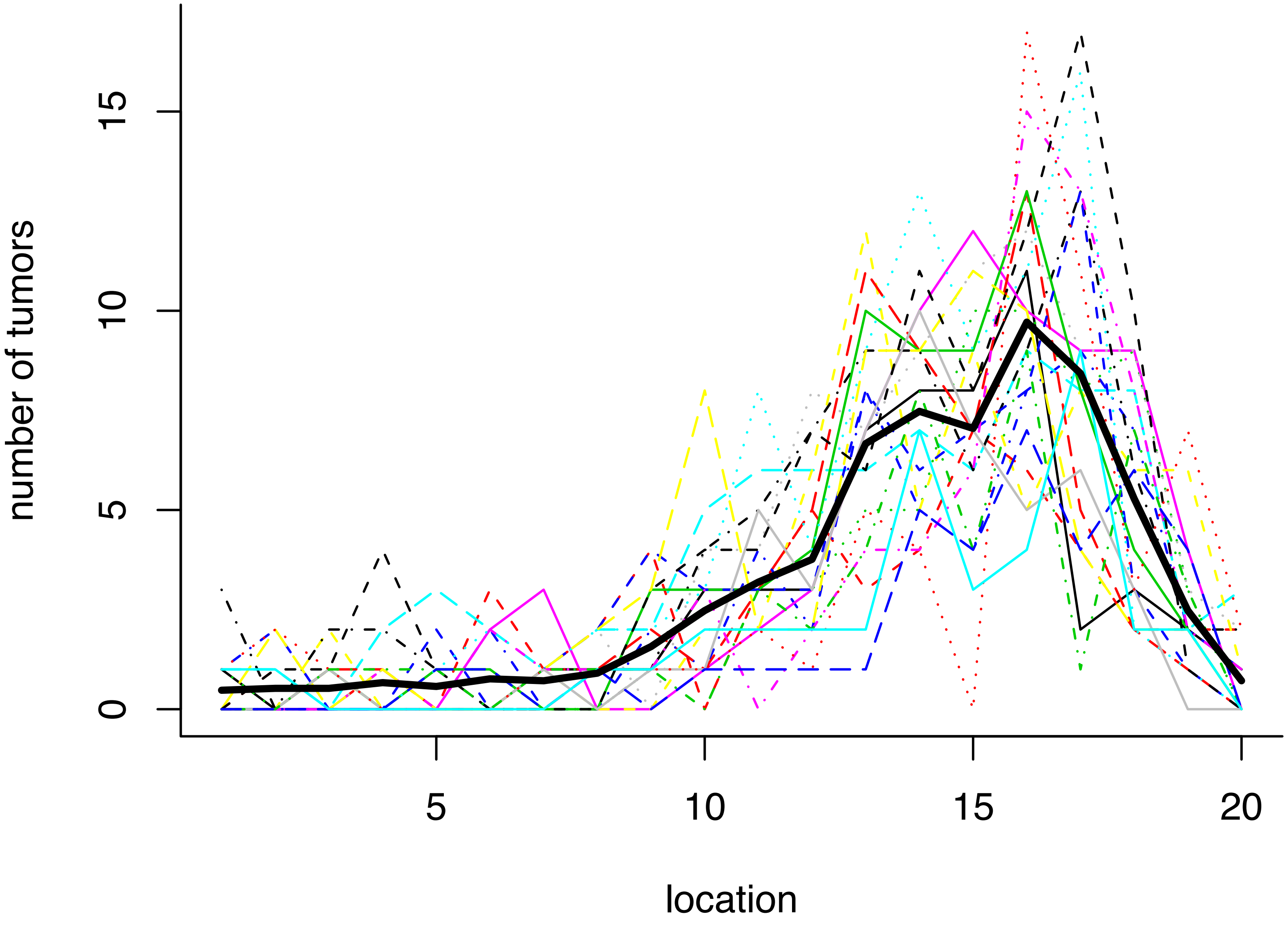
Haigis et al. (2004) report on a certain population of laboratory mice that experiences a high rate of intestinal tumor growth

One item of interest to researchers is how the rate of tumor growth varies along the length of the intestine

To study this, the intestine of each of 21 sample mice was divided into 20 sections and the number of tumors occurring in each section was recorded

Example: Visualizing the data

Each line represents the observed tumor counts of a mouse versus the segment of the intestine



(Although it is hard to tell ...) the lines from some mice are consistently below/above the average

Example: Hierarchical modeling

Therefore, tumor counts may be more similar within a mouse than between mice, and a hierarchical model with mouse-specific effects may be appropriate

A natural model for count data such as these is a Poisson distribution with a log-link

Letting $Y_{x,j}$ be mouse j's tumor count at location x of their intestine, we shall use the model

$$Y_{x,j} \sim \text{Pois}(\exp\{f_j(x)\})$$

where f_j is a smoothly varying function of $x \in [0, 1]$

Example: Polynomial covariates

A simple way to parameterize f_j is as a polynomial, so that

$$f_j(x) = \beta_{1,j} + \beta_{2,j} x + \beta_{3,j} x^2 + \cdots + \beta_{p,j} x^{p-1}$$

for some maximum degree $p - 1$

Such a parameterization allows us to represent each f_j as a regression on $(1, x, x^2, \ldots, x^{p-1})$

For simplicity, we will model each f_j as a fourth-degree polynomial, i.e., $p = 5$

Example: Between-group sampling model

Our between-group sampling model for the β_j's will be as in the previous section, so that

$$\beta_1, \ldots, \beta_m \overset{\text{iid}}{\sim} \mathcal{N}_p(\beta_0, \Sigma_0)$$

Unconditional on β_j, the observations coming from a given mouse are statistically dependent as determined by Σ_0

Estimating Σ_0 in this hierarchical model allows us to account for and describe potential within-mouse dependencies in the data

Example: Between-group sampling model

The unknown parameters in this model are β_0 and Σ_0 for which we need to specify prior distributions

Using conjugate normal and IW priors will allow us to proceed as usual for these parameters

A unit information prior can be constructed using OLS estimators with small sample sizes

e.g., by regressing

$$\{\log(y_{1,j} + 1/20), \ldots, \log(y_{n,j} + 1/20)\}$$

on $\{x_1, \ldots, x_{20}\}$ where $x_i^\top = (1, x_i, x_i^2, x_i^3, x_i^4)$ for $x_i \in (0.05, 0.10, \ldots, 0.95, 2)$

Example: MH-within Gibbs MCMC

The proposal we use for β_j' is

$$\beta_j' \sim \mathcal{N}_p(\beta_j^{(s)}, \Sigma^{(s)}/2)$$

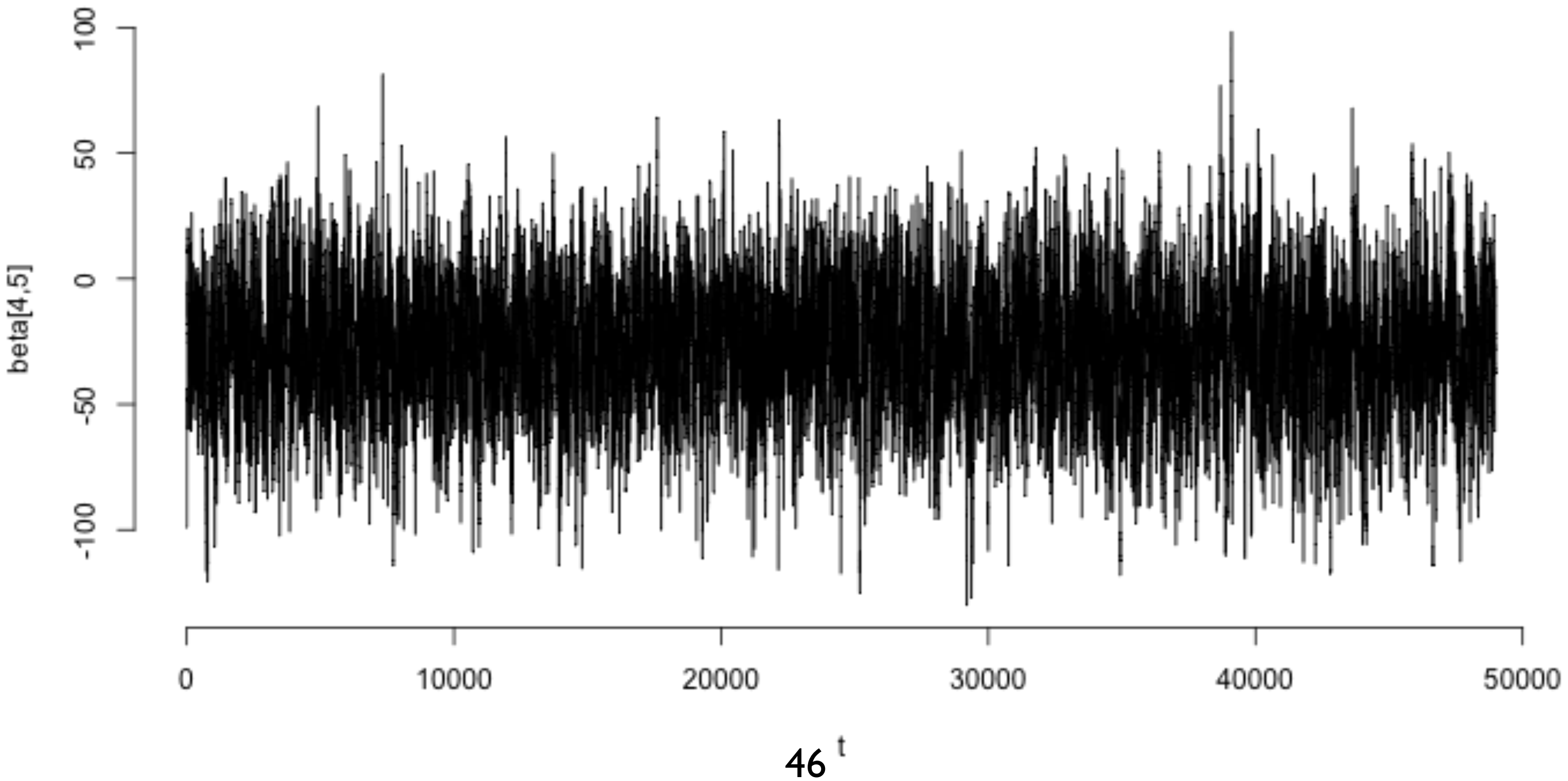
Since the mixing is likely to be worse than the "fully Gibbs" sampler from the hierarchical LM, we will need to obtain many more samples, and check the acceptance rates, autocorrelations, and effective sample sizes carefully

Example: Checking for good mixing

The ESSs obtained for the components of β_0 were

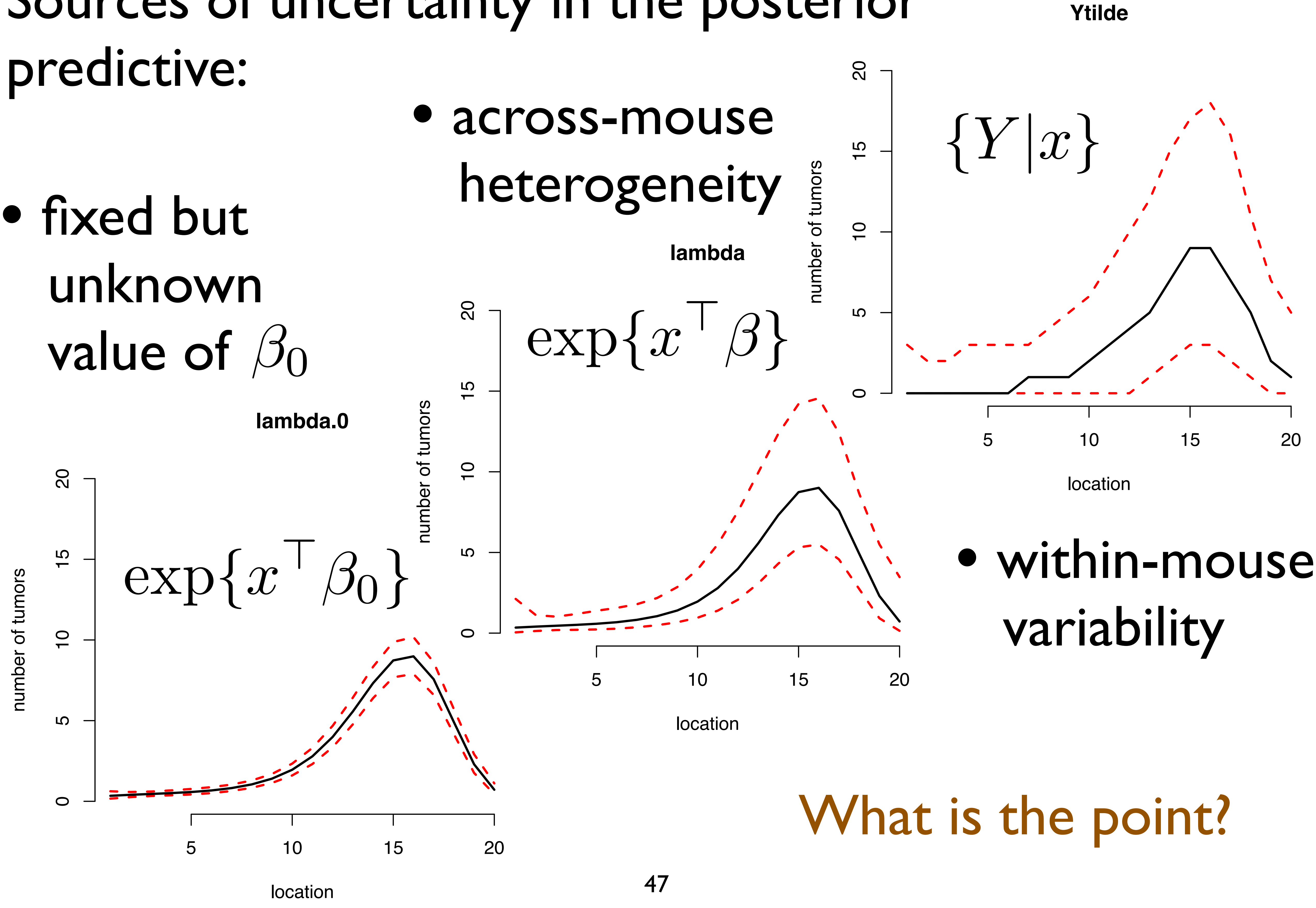
intercept	linear	quadratic	cubic	quartic
880	1310	1499	1616	1696

A trace obtained for $\beta_{4,5}$ was



Example: Predictive distribution(s)

Sources of uncertainty in the posterior predictive:



- fixed but unknown value of β_0
- across-mouse heterogeneity
- within-mouse variability

What is the point?

Example: Understanding uncertainty

Understanding these different sources of uncertainty can be very relevant to inference and decision making

For example, if we want to predict the observed tumor count distribution of a new mouse, we should use the confidence bands for $\{Y|x\}$

Whereas the bands for β_0 would be appropriate if we just wanted to describe the uncertainty in the underlying tumor rate for the population