GPH-GU2372/3372
Applied Bayesian Analysis in Public Health

Lesson 10: Hierarchical regression models

Hai Shu, PhD

11/28/2022

Topics

- Hierarchical linear regression models
- Hierarchical generalized linear regression models

Introduction

- In your homework 4, problem 9.1, you were asked to fit a separate linear regression model for each swimmer, that is: given the 6 measurements y_{ij}, $i = 1, \ldots, 6$ for swimmer $j = 1, \ldots, 4$ you were asked to fit the model

$$y_{ij} = \beta_{0,j} + \beta_{1,j} * week_i + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma_j^2)$$

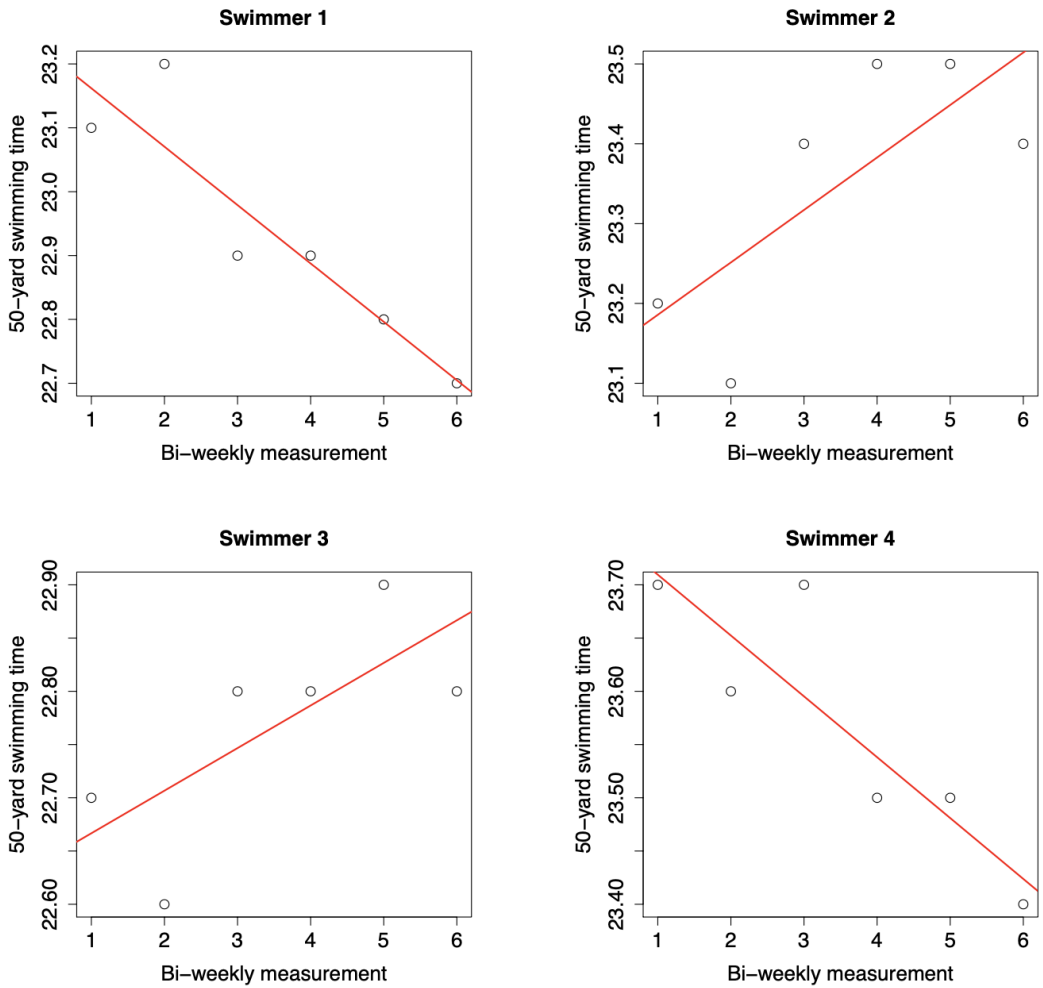
- Thus, at the end you were supposed to have 4 sets of two regression coefficient estimates: $\hat{\beta}_{0,1}, \hat{\beta}_{1,1}$, $\hat{\beta}_{0,2}, \hat{\beta}_{1,2}$, $\hat{\beta}_{0,3}, \hat{\beta}_{1,3}$, $\hat{\beta}_{0,4}, \hat{\beta}_{1,4}$.
- What do each of these sets of estimates represent?
- If we fit one linear regression model to all the data (i.e. to all 24 measurements), instead we would have:

$$y_{ij} = \beta_0 + \beta_1 * week_i + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$

  and we would have only 1 set of regression coefficient estimates: $\hat{\beta}_0, \hat{\beta}_1$.
- What is the difference between these two models? What do β_0, β_1 represent?

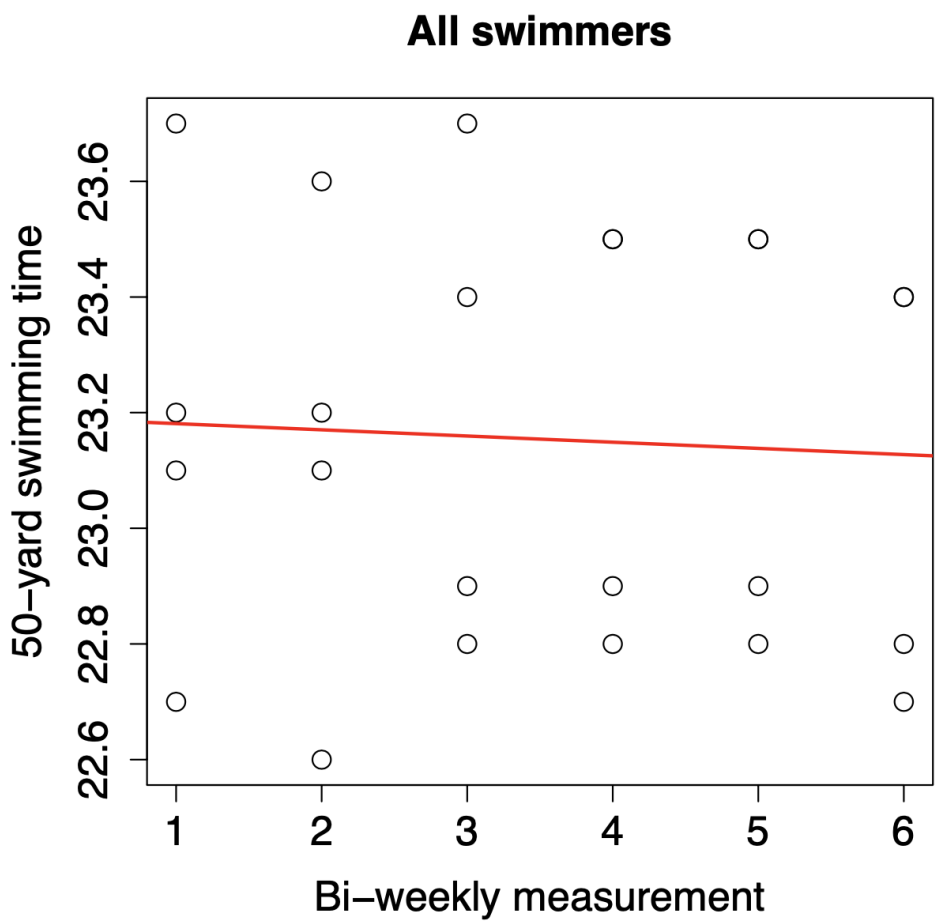
Introduction

- See the difference in plots: these are the estimated regression lines (via OLS) when we fit linear regression models separately for each swimmer



Introduction

- This is the estimated regression line when we fit a single linear regression model to all swimmers' data.



Introduction

- We can see that there is quite a difference between the estimated regression lines in the first case and the second case.
- They also represent different things/relationships.
- Today we will look more into fitting these type of models, without having to fit separate linear regressions for each swimmer.
- We will introduce these type of models in the context of hierarchical models, looking at the same example we considered previosuly.

Hierarchical linear regression models

- We have seen previously the concept of hierarchical modeling, a modeling approach that is appropriate when we have multiple measurements within each of several groups.
- Hierarchical linear regression models are useful when we want to describe how relationships between variables might differ between groups.
- In particular, hierarchical linear regression models are useful as soon as there is covariate information at different levels of variation.
- For example, in studying scholastic achievement we may have information about individual students (for example, family background), class-level information (characteristics of the teacher), and also information about the school (educational policy, type of neighborhood).
- Bayesian methods allow to handle this type of situations rather easily.

Hierarchical models: an example

- Example: in Lecture 8 we considered the math scores in a standardized test for 10th grade children in $m = 100$ urban schools.
- We modeled that data using the following hierarchical model:

$$
\begin{aligned}
y_{1,j}, \ldots, y_{n_j,j} \mid \theta_j, \sigma^2 &\overset{iid}{\sim} N(\theta_j, \sigma^2) \\
\theta_1, \ldots, \theta_m &\perp \sigma^2 \\
\theta_1, \ldots, \theta_m \mid \mu, \tau^2 &\overset{iid}{\sim} N(\mu, \tau^2) \\
\sigma^2 &\sim \text{InverseGamma}\left(\tfrac{\nu_0}{2}, \tfrac{\nu_0 \sigma_0^2}{2}\right) \\
\mu &\sim N(\mu_0, \tau_0^2) \\
\tau^2 &\sim \text{InverseGamma}\left(\tfrac{\eta_0}{2}, \tfrac{\eta_0 \tau_0^2}{2}\right)
\end{aligned}
$$

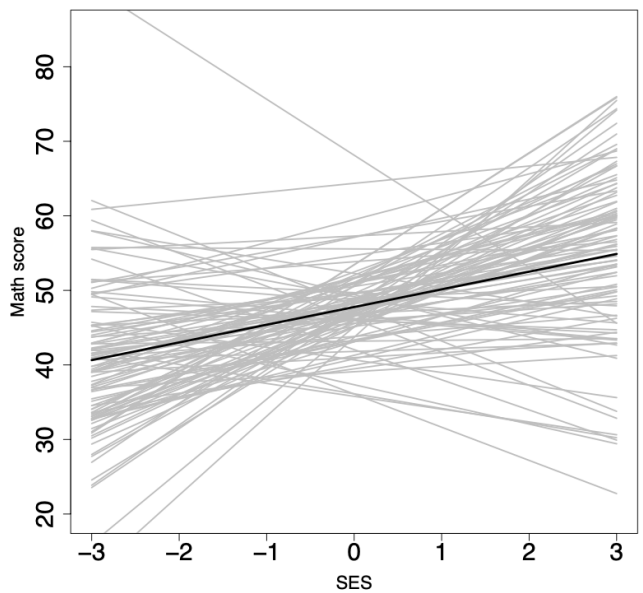
where the first level (the likelihood) models the within-group variability, and the second level (the population distribution) models the between-group variability.

Hierarchical linear regression model

- Now suppose that we are interested in the relationship between math score and socio-economic status.
- In our dataset, the socio-economic status for each student was derived from parental income and from the educational level of the parents for each student.
- It is possible that the relationship between a student's socio-economic status and his math score varies from school to school.
- A first simple way to assess this is to fit a separate linear regression model of math score on student socio-economic status for each school.
- If the relationship between the two variables is the same across schools, what will we expect for the separate linear regressions?

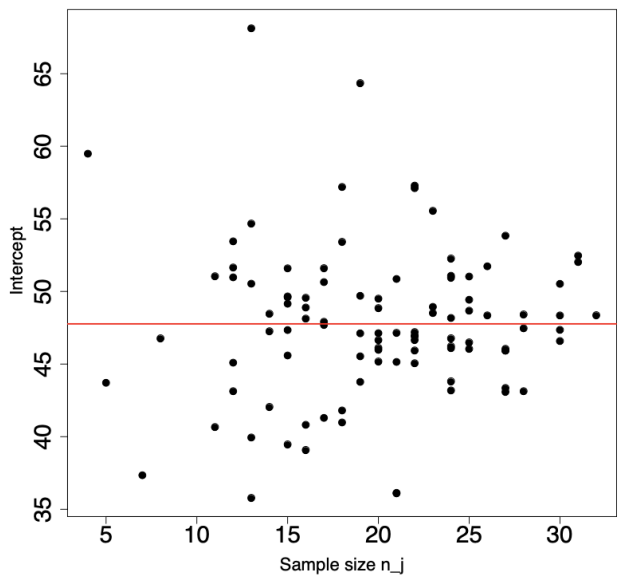
Hierarchical linear regression model

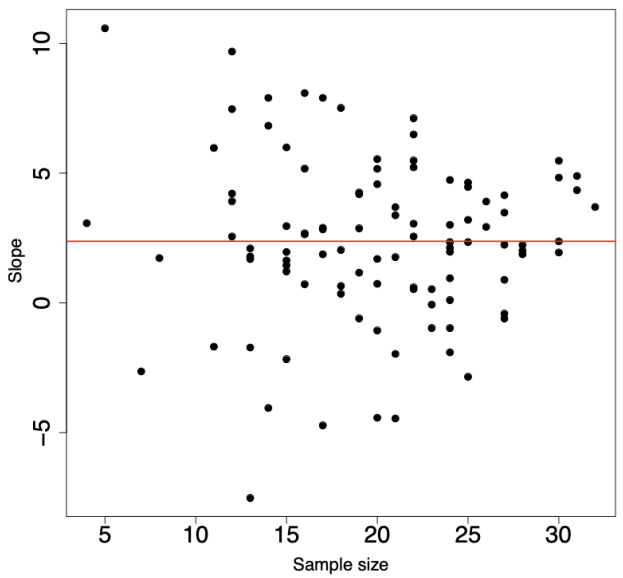
- We took the $\text{SES}_{i,j}$ socio-economic status for each student, standardized it so to have mean 0 and variance 1 and, for each school, we regressed the students' math scores on their SES.
- The estimated least squares regression lines $\hat{\beta}_{1,j} + \hat{\beta}_{2,j}\text{SES}_{i,j}$ for schools $j = 1, \ldots, m$ are displayed below.
- Each grey line refers to a different school, while the black line has been obtained by averaging the 100 regression lines.



Hierarchical linear regression model

- The estimates of the regression coefficients are very sensitive to the number of observations available.
- The plots below show the ordinary least squares estimates $\hat{\beta}_{1,j}$ and $\hat{\beta}_{2,j}$ of $\beta_{1,j}$ and $\beta_{2,j}$, respectively, as a function of the sample size n_j





Hierarchical linear regression model

- The estimates of $\beta_{1,j}$ and $\beta_{2,j}$ for schools with smaller sample sizes tend to be more unstable and extreme.
- To address this problem, we proceed as earlier when we were interested in estimating the school-specific mean θ_j. Hence, we allow sharing of information across schools to obtain more stable estimates of the regression coefficients.
- Following what we did for the hierarchical model in Lecture 8 , we model the within-group and the across-groups variability as follows:

$$\begin{aligned} y_{i,j} &= \beta_{1,j}x_{1,ij} + \beta_{2,j}x_{2,ij} + \ldots + \beta_{p,j}x_{p,ij} + \varepsilon_{i,j} \\ &= \beta_j'\mathbf{x}_{ij} + \varepsilon_{i,j} \qquad\qquad \varepsilon_{ij} \overset{iid}{\sim} N(0,\sigma^2) \end{aligned}$$

  - $\beta_j = (\beta_{1,j}\ \beta_{2,j}\ \ldots\ \beta_{p,j})'$ is the $p \times 1$ vector of regression coefficients for school j
  - $\mathbf{x}_{ij}$ is the $p \times 1$ vector with the p covariate information for student i in school j.
  - ε_{ij} is the error term for student i in school j.

Hierarchical linear regression model

- Collecting all the observations for students in the same school in a $n_j \times 1$ vector $\mathbf{Y}_j$ and collecting all the covariate information for students in school j into a $n_j \times p$ matrix $\mathbf{X}_j$, we have that the first stage of the hierarchical linear regression model is

$$\mathbf{Y}_j = \mathbf{X}_j \beta_j + \varepsilon_j$$

$$\implies \quad \mathbf{Y}_j | \mathbf{X}_j, \beta_j, \sigma^2 \sim N_{n_j}(\mathbf{X}_j \beta_j, \sigma^2 \mathbf{I}_{n_j})$$

  for $j = 1, \ldots, m$ where $\varepsilon_j = (\varepsilon_{1,j} \ldots \varepsilon_{n_j,j})'$.

- At the second stage of the model, we want to allow for sharing of information across groups in the regression coefficients. Thus we model:

$$\beta_1, \beta_2, \ldots, \beta_m | \theta, \mathbf{\Sigma} \overset{iid}{\sim} N_p(\theta, \mathbf{\Sigma})$$

- This model captures the variability between groups in the β_j's.

Hierarchical linear regression model

- Note that if $\mathbf{Z} \sim N_p(\mu, \mathbf{\Lambda})$, we can express $\mathbf{Z}$ as

$$\mathbf{Z} = \mu + \xi \qquad \xi \sim N_p(\mathbf{0}, \mathbf{\Lambda})$$

- Applying this to the β_j for $j = 1, \ldots, m$, we have

$$\beta_j = \theta + \xi_j \qquad \xi_j \overset{iid}{\sim} N_p(\mathbf{0}, \mathbf{\Sigma})$$

- The operation or rewriting $\beta_j | \theta, \mathbf{\Sigma} \sim N_p(\theta, \mathbf{\Sigma})$ as

$$\begin{array}{rcl} \beta_j & = & \theta + \xi_j \\ \xi_j | \mathbf{\Sigma} & \overset{iid}{\sim} & N_p(\mathbf{0}, \mathbf{\Sigma}) \end{array}$$

is called hierarchical centering and is one of the suggested methods to improve convergence in MCMC algorithms.

Hierarchical linear regression model as mixed linear model

- Substituting the expression of β_j as $\theta + \xi_j$ into the linear regression model for $\mathbf{Y}_j$, we obtain

$$\begin{aligned} \mathbf{Y}_j &= \mathbf{X}_j\beta_j + \varepsilon_j \\ &= \mathbf{X}_j(\theta + \xi_j) + \varepsilon_j \\ &= \mathbf{X}_j\theta + \mathbf{X}_j\xi_j + \varepsilon_j \end{aligned} \tag{1}$$

- Since in this parametrization, θ is not a random variable, θ represent the fixed effect of the p covariates on the response. θ refers to the entire population.
- On the other hand, as for each $j = 1, \ldots, m$, ξ_j is a random variable, ξ_j represent the school-specific random effects of the p covariates on the response for school j.
  The ξ_j, $j = 1, \ldots, m$, represent the deviation from the overall population effect θ for each group j.
- In light of this, the model (1) above is called a linear mixed model.

Hierarchical linear regression model as mixed linear model

- In the previous model, the covariates in the fixed and random part are the same. This does not have to be always the case.
  For example, a more general linear mixed model could be

$$Y_{i,j} = \theta' \mathbf{x}_{i,j} + \gamma_j' \mathbf{z}_{i,j} + \varepsilon_{i,j} \qquad \varepsilon_{i,j} \sim N(0, \sigma^2)$$

- Going back to the hierarchical linear regression model for the 100 math scores. At the first two stages we have:

$$\begin{aligned} \mathbf{Y}_j &= \mathbf{X}_j \beta_j + \varepsilon_j \qquad \varepsilon_j \sim N_{n_j}(\mathbf{0}, \sigma^2 \mathbf{I}_{n_j}) \qquad j = 1, \ldots, m \\ \beta_j | \theta, \mathbf{\Sigma} &\overset{iid}{\sim} N_p(\theta, \mathbf{\Sigma}) \end{aligned}$$

- The model is completely specified once we provide priors for all the model parameters, that is, we provide priors for $\theta, \mathbf{\Sigma}, \sigma^2$. For θ and σ^2, we can use the usual prior distributions:

$$\begin{aligned} \theta &\sim N_p(\mu_0, \mathbf{\Lambda}_0) \\ \sigma^2 &\sim \text{InverseGamma}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right) \end{aligned}$$

- What about $\mathbf{\Sigma}$?

Prior for a (co)variance matrix

- $\mathbf{\Sigma}$ is the prior (co)variance matrix of the β_j's. As such, $\mathbf{\Sigma}$ is a square $p \times p$ matrix and
  - its diagonal elements are positive, that is $\mathbf{\Sigma}_{kk} > 0$ for $k = 1, \ldots, p$
  - it is symmetrix, that is $\mathbf{\Sigma}_{lk} = \mathbf{\Sigma}_{kl}$ for $k, l = 1, \ldots, p$
  - it is positive definite

Inverse Wishart distribution

- If a $p \times p$ random matrix $\mathbf{\Sigma}$ has an Inverse Wishart $(\nu_0, \mathbf{\Phi}_0^{-1})$ distribution, then:

$$E[\mathbf{\Sigma}] = \frac{\mathbf{\Phi}_0}{\nu_0 - p - 1} \qquad \text{Mode}[\mathbf{\Sigma}] = \frac{\mathbf{\Phi}_0}{\nu_0 + p + 1}$$

- The parameter ν_0 is called the degrees of freedom and it should be a real number greater than $p-1$, that is $\nu_0 > p-1$.
- Given the form of the expectation of an Inverse Wishart distribution, in constructing a prior distribution for $\mathbf{\Sigma}$ we can proceed as follows:
  - if we have a prior idea on what $\mathbf{\Sigma}$ should be, say $\mathbf{\Sigma}_0$, then we can take ν_0 large and $\mathbf{\Phi}_0 = (\nu_0 - p - 1) \cdot \mathbf{\Sigma}_0$. This will make the distribution of $\mathbf{\Sigma}$ concentrated around $\mathbf{\Sigma}_0$
  - if we choose $\nu_0 = p + 2$ and $\mathbf{\Phi}_0 = \mathbf{\Sigma}_0$, then we will have a distribution that is loosely centered around $\mathbf{\Sigma}_0$.

Wishart distribution

- Finally, if the $p \times p$ random matrix $\mathbf{\Sigma}$ has an Inverse Wishart $(\nu_0, \mathbf{\Phi}_0^{-1})$ distribution, then its inverse, $\mathbf{\Sigma}^{-1}$ has a Wishart $(\nu_0, \mathbf{\Phi}_0^{-1})$ distribution and

$$E\left[\mathbf{\Sigma}^{-1}\right] = \nu_0 \mathbf{\Phi}_0^{-1} \qquad \text{Mode}\left[\mathbf{\Sigma}^{-1}\right] = (\nu_0 - p - 1)\mathbf{\Phi}_0^{-1}$$

- Hence, we place the following prior on $\mathbf{\Sigma}$:

$$\mathbf{\Sigma} \sim \text{InverseWishart}(\eta_0, \mathbf{S}_0^{-1})$$

Joint posterior distribution

- To infer upon this model, we need to derive the joint posterior distribution of the model parameters, that is, $\beta_1, \ldots, \beta_m, \sigma^2, \theta, \mathbf{\Sigma}$
- The joint posterior distribution $p(\beta_1, \ldots, \beta_m, \sigma^2, \theta, \mathbf{\Sigma} | \mathbf{y}_1, \ldots, \mathbf{y}_m, \mathbf{X}_1, \ldots, \mathbf{X}_m)$ is:

$$
\begin{aligned}
p(\beta_1, \ldots, \beta_m, \sigma^2, \theta, \mathbf{\Sigma} | \mathbf{y}_1, \ldots, \mathbf{y}_m, \mathbf{X}_1, \ldots, \mathbf{X}_m) \quad \propto \quad & \left\{ \prod_{j=1}^{m} p(\mathbf{y}_j; \mathbf{X}_j, \beta_j, \sigma^2) \right\} \\
\cdot \quad & \left\{ \prod_{j=1}^{m} p(\beta_j; \theta, \mathbf{\Sigma}) \right\} \\
\cdot \quad & p(\theta; \mu_0, \mathbf{\Lambda}_0) \cdot p(\mathbf{\Sigma}; \eta_0, \mathbf{S}_0^{-1}) \\
\cdot \quad & p(\sigma^2; \nu_0, \sigma_0^2)
\end{aligned}
$$

- We approximate this posterior distribution by drawing samples from it using a Gibbs sampling algorithm.

Full conditional of β_j

- The full conditional distribution of β_j, $j = 1, \ldots, m$, is given by:

$$\begin{aligned} p(\beta_j | \mathbf{Y}_1, \ldots, \mathbf{Y}_m, \beta_{-j}, \sigma^2, \theta, \mathbf{\Sigma}) &\propto p(\mathbf{y}_j; \mathbf{X}_j, \beta_j, \sigma^2) \cdot p(\beta_j; \theta, \mathbf{\Sigma}) \\ &= N_{n_j}(\mathbf{y}_j; \mathbf{X}_j \beta_j, \sigma^2 \mathbf{I}_{n_j}) \cdot N_p(\beta_j; \theta, \mathbf{\Sigma}) \end{aligned}$$

- Hence, the full conditional distribution of β_j is $N_p(\mathbf{m}_j, \mathbf{V}_j)$ where:

$$\begin{aligned} \mathbf{V}_j &= \left(\mathbf{\Sigma}^{-1} + \frac{\mathbf{x}_j' \cdot \mathbf{x}_j}{\sigma^2} \right)^{-1} \\ \mathbf{m}_j &= \left(\mathbf{\Sigma}^{-1} + \frac{\mathbf{x}_j' \cdot \mathbf{x}_j}{\sigma^2} \right)^{-1} \cdot \left(\mathbf{\Sigma}^{-1} \theta + \frac{\mathbf{x}_j' \mathbf{y}_j}{\sigma^2} \right) \\ &= \mathbf{V}_j \cdot \left(\mathbf{\Sigma}^{-1} \theta + \frac{\mathbf{x}_j' \mathbf{y}_j}{\sigma^2} \right) \end{aligned}$$

Full conditional of θ

- The full conditional distribution of θ is given by:

$$\begin{aligned} p(\theta|\mathbf{Y}_1,\ldots,\mathbf{Y}_m,\beta_1,\ldots,\beta_m,\sigma^2,\boldsymbol{\Sigma}) &\propto \left\{\prod_{j=1}^m p(\beta_j;\theta,\boldsymbol{\Sigma})\right\}\cdot p(\theta;\mu_0,\boldsymbol{\Lambda}_0) \\ &= \left\{\prod_{j=1}^m N_p(\beta_j;\theta,\boldsymbol{\Sigma})\right\}\cdot N_p(\theta;\mu_0,\boldsymbol{\Lambda}_0) \end{aligned}$$

- Hence, the full conditional distribution of θ is $N_p(\mu_n,\boldsymbol{\Lambda}_n)$ where:

$$\begin{aligned} \boldsymbol{\Lambda}_n &= \left(\boldsymbol{\Lambda}_0^{-1}+m\boldsymbol{\Sigma}^{-1}\right)^{-1} \\ \mu_n &= \left(\boldsymbol{\Lambda}_0^{-1}+m\boldsymbol{\Sigma}^{-1}\right)^{-1}\cdot\left(\boldsymbol{\Lambda}_0^{-1}\mu_0+\boldsymbol{\Sigma}^{-1}\left(\textstyle\sum_{j=1}^m\cdot\beta_j\right)\right) \\ &= \boldsymbol{\Lambda}_n\cdot\left(\boldsymbol{\Lambda}_0^{-1}\mu_0+m\boldsymbol{\Sigma}^{-1}\bar{\beta}\right) \end{aligned}$$

Full conditional of Σ

- The full conditional distribution of $\mathbf{\Sigma}$ is given by:

$$p(\mathbf{\Sigma}|\mathbf{Y}_1,\ldots,\mathbf{Y}_m,\beta_1,\ldots,\beta_m,\sigma^2,\theta) \propto \left\{\prod_{j=1}^m p(\beta_j;\theta,\mathbf{\Sigma})\right\}\cdot p(\mathbf{\Sigma};\eta_0,\mathbf{S}_0^{-1})$$

$$= \left\{\prod_{j=1}^m N_p(\beta_j;\theta,\mathbf{\Sigma})\right\}\cdot \text{InvWishart}(\mathbf{\Sigma};\eta_0,\mathbf{S}_0^{-1})$$

- Hence, the full conditional distribution of $\mathbf{\Sigma}$ is $\text{InverseWishart}(\eta_n,\mathbf{S}_n^{-1})$ where:

$$\begin{aligned}\eta_n &= \eta_0 + m\\ \mathbf{S}_n &= \mathbf{S}_0 + \textstyle\sum_{j=1}^m (\beta_j-\theta)(\beta_j-\theta)'\\ &= \mathbf{S}_0 + \mathbf{S}_\theta\end{aligned}$$

Full conditional of σ^2

- The full conditional distribution of σ^2 is given by:

$$\begin{aligned} p(\sigma^2|\mathbf{Y}_1,\ldots,\mathbf{Y}_m,\beta_1,\ldots,\beta_m,\theta,\mathbf{\Sigma}) &\propto \left\{\prod_{j=1}^m p(\mathbf{y}_j;\mathbf{X}_j,\beta_j,\sigma^2)\right\}\cdot p(\sigma^2;\nu_0,\sigma_0^2) \\ &= \left\{\prod_{j=1}^m N_{n_j}(\mathbf{y}_j;\mathbf{X}_j\beta_j,\sigma^2\mathbf{I}_{n_j})\right\} \\ &\cdot \text{InvGamma}(\sigma^2;\tfrac{\nu_0}{2},\tfrac{\nu_0\sigma_0^2}{2}) \end{aligned}$$

- Hence, the full conditional distribution of σ^2 is $\text{InverseGamma}(\frac{\nu_n}{2},\frac{\nu_n\sigma_n^2}{2})$ where:

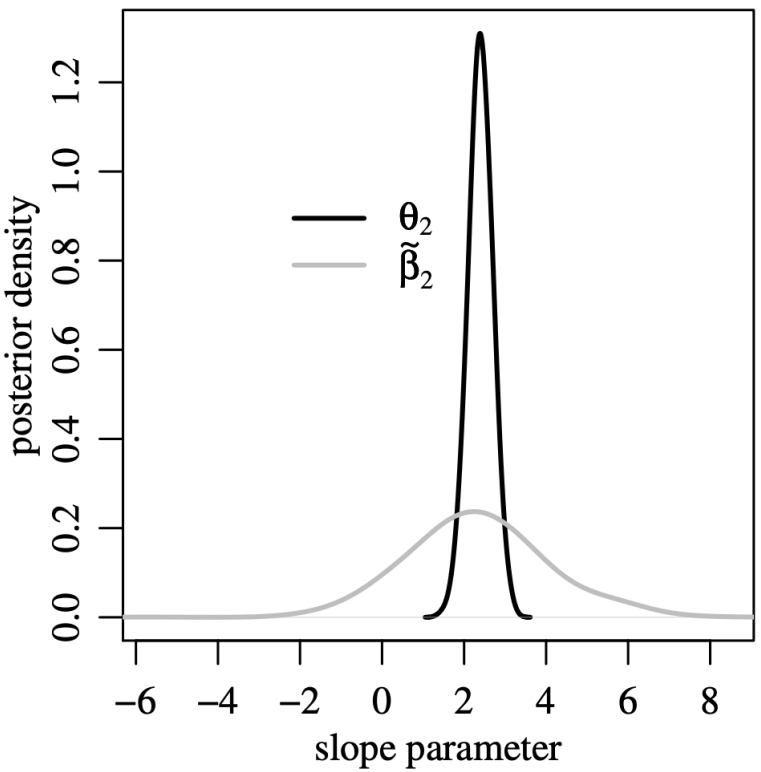
$$\begin{aligned} \nu_n &= \nu_0+\textstyle\sum_{j=1}^m n_j \\ \sigma_n^2 &= \frac{1}{\nu_n}\cdot\left[\nu_0\sigma_0^2+\textstyle\sum_{j=1}^m (\mathbf{y}_j-\mathbf{X}_j\beta_j)'(\mathbf{y}_j-\mathbf{X}_j\beta_j)\right] \\ &= \frac{1}{\nu_n}\cdot\left[\nu_0\sigma_0^2+\textstyle\sum_{j=1}^m\sum_{i=1}^{n_j}(y_{i,j}-\beta_j'\mathbf{x}_{i,j})^2)\right] \end{aligned}$$

Analysis of math score data

- We analyze the math scores data using a hierarchical linear regression model.
- We need to specify the prior parameters
  - μ_0 and $\mathbf{\Lambda}_0$: prior mean and prior (co)variance matrix of θ
  - η_0 and $\mathbf{S}_0$: prior degrees of freedom and matrix parameter in the prior distribution of $\mathbf{\Sigma}$
  - ν_0 and σ_0^2: prior degrees of freedom and prior mean squares in the prior distribution of σ^2
- In determining these prior parameters, we use an approach similar to the unit information prior (that is, we use the data to derive the prior distributions). Thus we take:
  - μ_0 equal to the average of the OLS estimates of the school-specific regression coefficients β_j, $j = 1, \ldots, m = 100$, and $\mathbf{\Lambda}_0$ equal to their sample (co)variance matrix
  - $\mathbf{S}_0$ equal to the sample (co)variance matrix of the OLS estimate $\hat{\beta}_j$, $j = 1, \ldots, m = 100$ and $\eta_0 = p + 2 = 4$
  - σ_0^2 equal to the average of the school-specific sample variances in math-score and $\nu_0 = 1$

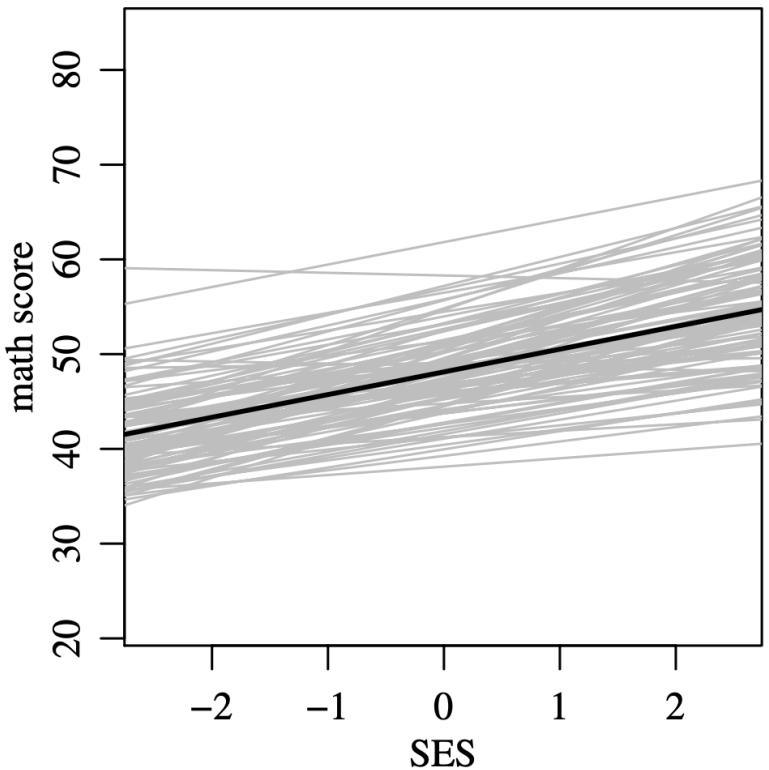
Analysis of math score data

- We run a Gibbs sampling algorithm for 10,000 iterations. We base our posterior inference on 1,000 MCMC samples obtained by using a thinning of 10
- Marginal posterior density for θ_2 with superimposed the marginal posterior predictive density for $\tilde{\beta}_2$ for a school yet to be sampled.



Analysis of math score data

- Plot of the posterior estimates of individual regression lines of math scores on socio-economic status (SES) for each school. In the plot the black line represents the average regression line.



Hierarchical generalized linear regression models

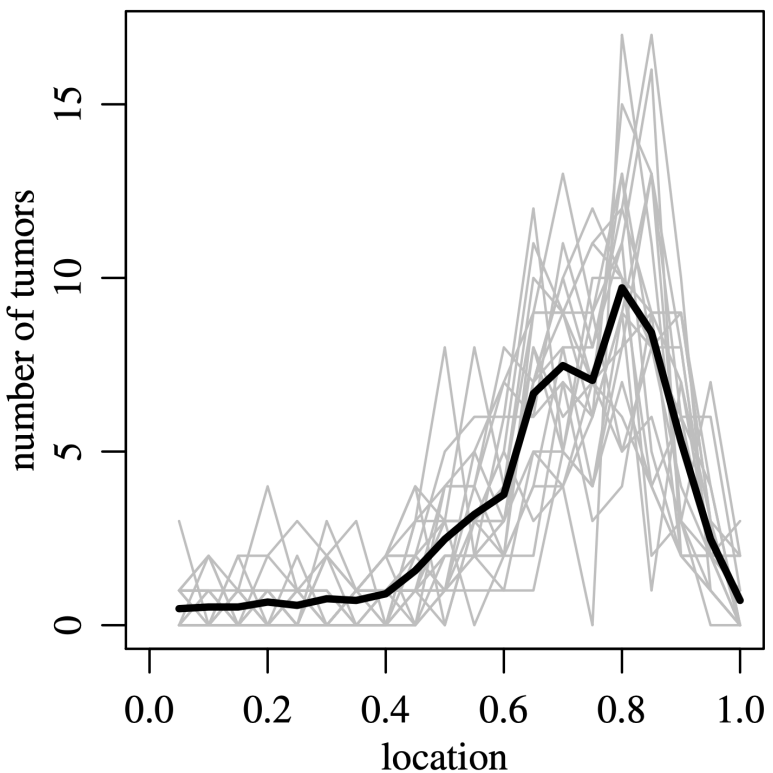
- In hierarchical linear regression models, the response variable is modeled using a normal distribution.
- In hierarchical generalized linear regression model, the response variable cannot be appropriately described using a normal distribution, even though the hierarchical structure of the model is still useful to capture the between-groups variation.
- In hierarchical generalized linear regression model, the model for the response variable at the first stage belongs to the class of generalized linear models.
- Example: a certain strain of laboratory mice experiences a high rate of intestinal tumor growth. One question that researchers are interested to answer is if the rate of tumor growth varies along the locationof the intestine. For this purpose, researchers took 21 mice, whose intestine was partitioned into 20 sections. The number of tumors in each of these 20 sections for each mice was recorded.

Hierarchical generalized linear regression models

- The data in this case is: y_{ij}, number of tumors at location $x_i = \frac{i}{20}$ with $i = 1, \ldots, n_j = 20$ for mouse j where $j = 1, \ldots, m = 21$.
- Clearly, modeling the number of tumors, y_{ij} as the observation of a normal random variable is not appropriate. A better model would be to assume that $Y_{ij} \sim \text{Poisson}(\theta_{ij})$, for $i = 1, \ldots, n_j$ and $j = 1, \ldots, m = 21$.
- As the interest is in seeing whether the variability in the number of tumors depends on the tumor location, we might want to relate θ_{ij} to x_i, the location of tumor.

Tumor number versus location

- Plot of the number of tumors y_{ij} at location $x_i = \frac{i}{20}$, $i = 1, \ldots, n_j = 20$, for mouse j, $j = 1, \ldots, m = 21$, versus tumor location: each grey line represents one of the 21 mice.
- The black line represents the average line: for each location $x_i = \frac{i}{20}$ along the intestine, the average number $\bar{y}_i = \frac{1}{m} \sum_{j=1}^{m} y_{ij}$ of tumors at that location is plotted against the tumor location x_i for $i = 1, \ldots, 20$.



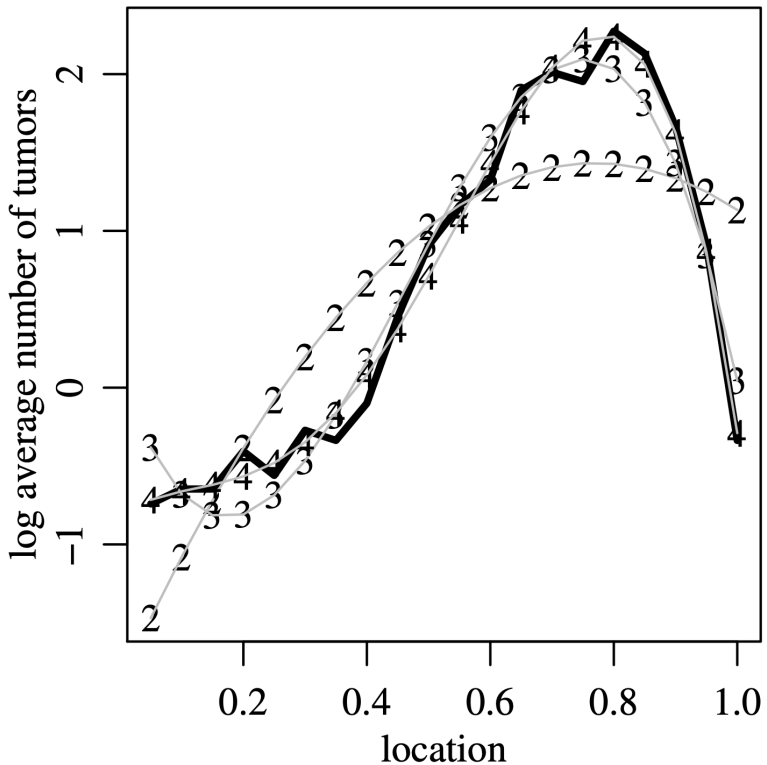
Tumor number versus location

- It seems that there is an association between the number of tumors and the location of the tumor for all the mice, though the particular shape of the functional form varies from mouse to mouse.
- The association between number of tumors and location is definitely not linear: it can probably be described using a polynomial of at least degree 2.
- We do some exploratory analysis of the relationship using the fact that in a Poisson regression model the expected value of the Poisson random variable, here θ_{ij}, is linked to explanatory variables via a log link, thus:

$$\log(E[Y_{ij}|\theta_{ij}]) = \beta_1 + \beta_2 x_{ij} + \beta_3 x_{ij}^2 + \ldots + \beta_p x_{ij}^{p-1}$$

Tumor number versus location

- Plot of the log of the average number of tumors, $\log(\bar{y}_i)$, at each location versus the location x_i with fitted a second, third and fourth degree polynomial on tumor location.



- Based on the plot we chose a fourth degree polynomial.

Tumor number versus location

- Based on the exploratory analysis, we model the number of tumors Y_{ij} of mouse j at location i as a Poisson random variable with mean θ_{ij} where

$$\begin{aligned} \log[E(Y_{ij}|x_i)] &= \log(\theta_{ij}) = \beta_{1,j} + \beta_{2,j}x_i + \beta_{3,j}x_i^2 + \ldots + \beta_{p,j}x_i^{p-1} \\ &= \beta_{1,j} + \beta_{2,j}x_i + \beta_{3,j}x_i^2 + \beta_{4,j}x_i^3 + \beta_{5,j}x_i^4 \end{aligned}$$

  with mouse-specific coefficients $\beta_{1,j}, \ldots, \beta_{5,j}$.

- After collecting the $\beta_{1,j}, \ldots, \beta_{5,j}$ into a 5×1 vector β_j, for $j = 1, \ldots, m$ we model the variability between mice in the regression coefficients vector β_j by assuming that the $\beta_1, \ldots, \beta_m$ are conditionally independent and identically distributed.
  Hence:

$$\beta_1, \ldots, \beta_m | \mu, \mathbf{\Sigma} \overset{iid}{\sim} N_5(\mu, \mathbf{\Sigma})$$

Generalized linear mixed model

- As for the hierarchical linear regression model, we can write each β_j as

$$\beta_j = \mu + \varepsilon_j \qquad \varepsilon_j \sim N_5(\mathbf{0}, \mathbf{\Sigma})$$

- Therefore, we have:

$$\begin{aligned} \log\left[E\left(Y_{ij}|x_i\right)\right] &= \log\left(\theta_{ij}\right) = \beta_{1,j} + \beta_{2,j}x_i + \beta_{3,j}x_i^2 + \beta_{4,j}x_i^3 + \beta_{5,j}x_i^4 \\ &= \mu_1 + \mu_2 x_i + \mu_3 x_i^2 + \mu_4 x_i^3 + \mu_5 x_i^4 \\ &+ \varepsilon_{1,j} + \varepsilon_{2,j}x_i + \varepsilon_{3,j}x_i^2 + \varepsilon_{4,j}x_i^3 + \varepsilon_{5,j}x_i^4 \end{aligned}$$

showing that a generalized hierarchical linear model is equivalent to a generalized linear mixed model.

Fixed and random effects

$$\begin{aligned}\log\left[E\left(Y_{ij}|x_i\right)\right] &= \log\left(\theta_{ij}\right) = \mu_1 + \mu_2 x_i + \mu_3 x_i^2 + \mu_4 x_i^3 + \mu_5 x_i^4 \\ &+ \varepsilon_{1,j} + \varepsilon_{2,j} x_i + \varepsilon_{3,j} x_i^2 + \varepsilon_{4,j} x_i^3 + \varepsilon_{5,j} x_i^4\end{aligned}$$

- $\mu = \left(\mu_1\ \mu_2\ \mu_3\ \mu_4\ \mu_5\right)'$ represents the overall or fixed effect of the tumor location on the log of the expected number of tumors.
- $\varepsilon_j = \left(\varepsilon_{1,j}\ \varepsilon_{2,j}\ \varepsilon_{3,j}\ \varepsilon_{4,j}\ \varepsilon_{5,j}\right)'$ is the random effect of the tumor location on the log of the expected number of tumors for mouse j: this explains the variation of the curve for the log expected number of tumors for mouse j from the population average curve.

Complete model specification

- We have modeled the regression coefficients $\beta_1, \ldots, \beta_m$ as:

$$\beta_1, \ldots, \beta_m | \mu, \mathbf{\Sigma} \overset{iid}{\sim} N_5(\mu, \mathbf{\Sigma})$$

- We complete the specification of the model by specifying priors on μ and $\mathbf{\Sigma}$:

$$\mu \sim N_5(\mu_0, \mathbf{\Lambda}_0) \qquad \mathbf{\Sigma} \sim \text{InverseWishart}(\eta_0, \mathbf{S}_0^{-1})$$

- The complete model is then:

$$\begin{aligned}
Y_{ij} \mid \theta_{ij} &\overset{iid}{\sim} \text{Poisson}(\theta_{ij}) && j = 1, \ldots, m \\
\log(\theta_{ij}) &= \beta_j' \mathbf{x}_i && \mathbf{x}_i = \left(1 \; x_i \; x_i^2 \; x_i^3 \; x_i^4\right)' \\
\beta_1, \ldots, \beta_m \mid \mu, \mathbf{\Sigma} &\overset{iid}{\sim} N_5(\mu, \mathbf{\Sigma}) \\
\mu &\sim N_5(\mu_0, \mathbf{\Lambda}_0) \\
\mathbf{\Sigma} &\sim \text{InverseWishart}(\eta_0, \mathbf{S}_0^{-1})
\end{aligned}$$

Posterior inference

- We can rewrite the model also as:

$$
\begin{aligned}
Y_{ij} \mid \beta_j, \mathbf{x}_i &\overset{iid}{\sim} \text{Poisson}\left(\exp\left(\beta_j' \mathbf{x}_i\right)\right) \qquad j = 1, \ldots, m \\
\beta_1, \ldots, \beta_m \mid \mu, \mathbf{\Sigma} &\overset{iid}{\sim} N_5(\mu, \mathbf{\Sigma}) \\
\mu &\sim N_5(\mu_0, \mathbf{\Lambda}_0) \\
\mathbf{\Sigma} &\sim \text{InverseWishart}(\eta_0, \mathbf{S}_0^{-1})
\end{aligned}
$$

- We want to infer upon the fixed effect of tumor location μ, the random effect $\beta_1, \ldots, \beta_m$ of tumor location for each mouse, and the covariance $\mathbf{\Sigma}$ among the β's **given** the data.
- The joint posterior distribution is:

$$
\begin{aligned}
p(\beta_1, \ldots, \beta_m, \mu, \mathbf{\Sigma} | y_{1,1}, \ldots, y_{20,21}) &\propto \left\{ \prod_{j=1}^{m} \left[\prod_{i=1}^{n_j} p(y_{ij}; \beta_j, \mathbf{x}_i) \right] \right\} \\
&\cdot \left\{ \prod_{j=1}^{m} p(\beta_j; \mu, \mathbf{\Sigma}) \right\} \\
&\cdot p(\mu; \mu_0, \mathbf{\Lambda}_0) \cdot p(\mathbf{\Sigma}; \eta_0, \mathbf{S}_0^{-1})
\end{aligned}
$$

Posterior inference

- We approximate the posterior distribution by devising an MCMC algorithm that generates a Markov chain with stationary distribution the posterior distribution $p(\beta_1, \ldots, \beta_m, \mu, \mathbf{\Sigma} | \text{data})$
- We use a Gibbs sampling algorithm as MCMC algorithm: we need to derive all the full conditionals.
- The full conditional of $\beta_1, \ldots, \beta_m$ is given by: for $j = 1, \ldots, m$

$$
\begin{aligned}
p(\beta_j | \beta_{-j}, \mu, \mathbf{\Sigma}, y_{1,1}, \ldots, y_{20,21}, \mathbf{x}_1, \ldots, \mathbf{x}_{20}) &\propto \left[\prod_{i=1}^{n_j} p(y_{ij}; \beta_j, \mathbf{x}_i) \right] \\
&\cdot \; p(\beta_j; \mu, \mathbf{\Sigma}) \\
&= \left[\prod_{i=1}^{n_j} \text{Poisson}\left(y_{ij}; \exp\left(\beta_j' \mathbf{x}_i\right)\right) \right] \\
&\cdot \; N_5(\beta_j; \mu, \mathbf{\Sigma})
\end{aligned}
$$

Full conditional distribution of β_j

- The full conditional distribution of β_j doesn't have a closed form for $j = 1, \ldots, m$: we will use the Metropolis algorithm to sample from the full conditional of β_j.
- At the k-th iteration, if $\beta_j^{(k-1)}$ is the current value of β_j, we will use as jumping distribution

$$J_k(\beta_j^\star | \beta_j^{(k-1)}) = N_5(\beta_j^{(k-1)}, \mathbf{V}^{(k)})$$

- This is an example of a random-walk Metropolis algorithm. Note that this is not the only choice. We could have chosen to update each component of β_j individually. In that case, the jumping/proposal distribution would be a univariate distribution (for example, a normal distribution).
- In order for the Metropolis algorithm to be efficient, we need to choose $\mathbf{V}^{(k)}$ so that we have an acceptance rate between $20\% - 50\%$. A good choice of $\mathbf{V}^{(k)}$ would be if it is proportional to the posterior variance of β_j.

Full conditional distribution of μ

- We take $\mathbf{V}^{(k)}$ to be a scaled version of $\mathbf{\Sigma}^{(k)}$, i.e. $\mathbf{V}^{(k)} = \delta\mathbf{\Sigma}^{(k)}$ where $\mathbf{\Sigma}^{(k)}$ is the value of $\mathbf{\Sigma}$ at the k-th iteration of the MCMC algorithm.
- The full conditional distribution of μ is given by:

$$\begin{aligned} p(\mu|\beta_1,\ldots,\beta_m,\mathbf{\Sigma},y_{1,1},\ldots,y_{20,21},\mathbf{x}_1,\ldots,\mathbf{x}_{20}) &\propto \left\{\prod_{j=1}^m p(\beta_j;\mu,\mathbf{\Sigma})\right\} \\ &\cdot\ p(\mu;\mu_0,\mathbf{\Lambda}_0) \\ &= \left\{\prod_{j=1}^m N_5(\beta_j;\mu,\mathbf{\Sigma})\right\} \\ &\cdot\ N_5(\mu;\mu_0,\mathbf{\Lambda}_0) \end{aligned}$$

- Thus, the full conditional of μ is $N_5(\mu_n,\mathbf{\Lambda}_n)$ where

$$\begin{aligned} \mathbf{\Lambda}_n &= \left(\mathbf{\Lambda}_0^{-1} + m\mathbf{\Sigma}^{-1}\right)^{-1} \\ \mu_n &= \mathbf{\Lambda}_n \cdot \left(\mathbf{\Lambda}_0^{-1}\cdot\mu_0 + \mathbf{\Sigma}^{-1}\bar{\beta}\right) \end{aligned}$$

- We can sample μ directly from its full conditional distribution.

Full conditional distribution of Σ

- The full conditional distribution of $\boldsymbol{\Sigma}$ is given by:

$$\begin{aligned} p(\boldsymbol{\Sigma}|\beta_1,\ldots,\beta_m,\mu,y_{1,1},\ldots,y_{20,21},\mathbf{x}_1,\ldots,\mathbf{x}_{20}) &\propto \left\{\prod_{j=1}^m p(\beta_j;\mu,\boldsymbol{\Sigma})\right\} \\ &\cdot\ p(\boldsymbol{\Sigma};\eta_0,\mathbf{S}_0^{-1}) \\ &= \left\{\prod_{j=1}^m N_5(\beta_j;\mu,\boldsymbol{\Sigma})\right\} \\ &\cdot\ \text{InverseWishart}(\boldsymbol{\Sigma};\eta_0,\mathbf{S}_0^{-1}) \end{aligned}$$

- Thus, the full conditional of $\boldsymbol{\Sigma}$ is $\text{InverseWishart}(\eta_n,\mathbf{S}_n^{-1})$ where

$$\begin{aligned} \eta_n &= \eta_0 + m \\ \mathbf{S}_n &= \mathbf{S}_0 + \sum_{j=1}^m (\beta_j-\mu)(\beta_j-\mu)' \\ &= \mathbf{S}_0 + \mathbf{S}_\mu \end{aligned}$$

- We can sample $\boldsymbol{\Sigma}$ directly from its full conditional distribution.

The MCMC algorithm

- Thus, the MCMC algorithm that we are going to use is a Gibbs sampling algorithm with a Metropolis step. The algorithm will proceed as follows:
  - Choose a number S of iterations.
  - For iteration $k = 1, \ldots, S$, repeat the following steps:
    1. Sample $\mu^{(k)}$ from its full conditional distribution $p(\mu | \beta_1^{(k-1)}, \ldots, \beta_m^{(k-1)}, \boldsymbol{\Sigma}^{(k-1)}, \textbf{data})$
    2. Sample $\boldsymbol{\Sigma}^{(k)}$ from its full conditional distribution $p(\boldsymbol{\Sigma} | \beta_1^{(k-1)}, \ldots, \beta_m^{(k-1)}, \mu^{(k)}, \textbf{data})$
    3. Update each β_j for $j = 1, \ldots, m$ using a Metropolis algorithm: for $j = 1, \ldots, m$
       - propose a value $\beta_j^\star$ from $J_k(\beta_j^\star | \beta_j^{(k-1)}) = N_5(\beta_j^\star; \beta_j^{(k-1)}, \mathbf{V}^{(k)})$
       - compute the ratio

$$r = \frac{p(\beta_j^\star | \textbf{data, other pars})}{p(\beta_j^{(k-1)} | \textbf{data, other pars})} = \frac{p(\mathbf{y}_j; \beta_j^\star, \mathbf{X}) \cdot p(\beta_j^\star; \mu^{(k)}, \boldsymbol{\Sigma}^{(k)})}{p(\mathbf{y}_j; \beta_j^{(k-1)}, \mathbf{X}) \cdot p(\beta_j^{(k-1)}; \mu^{(k)}, \boldsymbol{\Sigma}^{(k)})}$$

       - sample $u \sim \text{Uniform}(0, 1)$. If $r > 1$ or $u < r$, set $\beta_j^{(k)} = \beta_j^\star$ otherwise set $\beta_j^{(k)} = \beta_j^{(k-1)}$

Application to tumor data

- We fit the hierarchical Poisson regression model on the tumor data.
- Here we have: $m = 21$ mice and for each mice we have $n_j = 20$ measurements.
- The covariates $\mathbf{x}_i$ for number of tumors y_{ij} at location $x_i = \frac{i}{20}$, $i = 1, \ldots, 20$ of mouse $j = 1, \ldots, 21$ are:

$$\mathbf{x}_i = \begin{pmatrix} 1 & \frac{i}{20} & \left(\frac{i}{20}\right)^2 & \left(\frac{i}{20}\right)^3 & \left(\frac{i}{20}\right)^4 \end{pmatrix}$$

- To get values for the parameters in the priors, that is:

$$\mu \sim N_5(\mu_0, \mathbf{\Lambda}_0) \qquad \mathbf{\Sigma} \sim \text{InverseWishart}(\eta_0, \mathbf{S}_0^{-1})$$

  we use a unit information prior type of approach.
- We fit a separate linear regression of $\log(y_{1,j} + \frac{1}{20}), \ldots, \log(y_{20,j} + \frac{1}{20})$ on $\mathbf{x}_1, \ldots, \mathbf{x}_{20}$ and we obtain the OLS estimates of $\hat{\beta}_1, \ldots, \hat{\beta}_{21}$. We take the average of those estimates and set $\mu_0 = \frac{1}{21} \cdot \sum_{j=1}^{21} \hat{\beta}_j$.

Application to tumor data

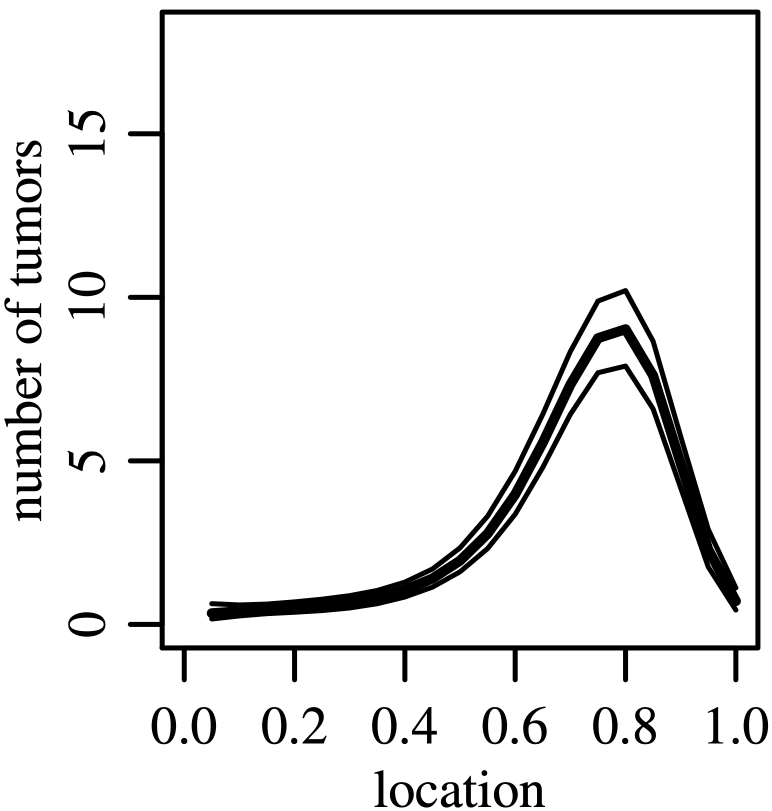
- We compute the sample covariance $\hat{\mathbf{S}}_\beta$ matrix for the $\hat{\beta}_1, \ldots, \hat{\beta}_{21}$ and we set $\mathbf{\Lambda}_0 = \hat{\mathbf{S}}_\beta$.
- Similarly, we set $\mathbf{S}_0 = \hat{\mathbf{S}}_\beta$ and we take $\eta_0 = p + 2 = 5 + 2 = 7$.
- We chose the scaling factor in the jumping distribution of the β's to $\delta = 0.25$: this leads to an acceptance rate of approximately 31%.
- We ran the algorithm for 50,000 iterations; we used the first 10,000 as burnin and based our posterior inference on 4,000 samples obtained by using a thinning of 10 on the MCMC samples.

Population average number of tumors vs location

- Our question of interest was whether the tumor location along the intestine was associated to the number of tumors observed in the mice.
- We partitioned the effect of tumor location on the number of tumors in the sum of an overall fixed effect and a mouse-specific random effect.
- We can then look at what is the posterior estimate of the population average curve for the number of tumors as a function of the tumor location.

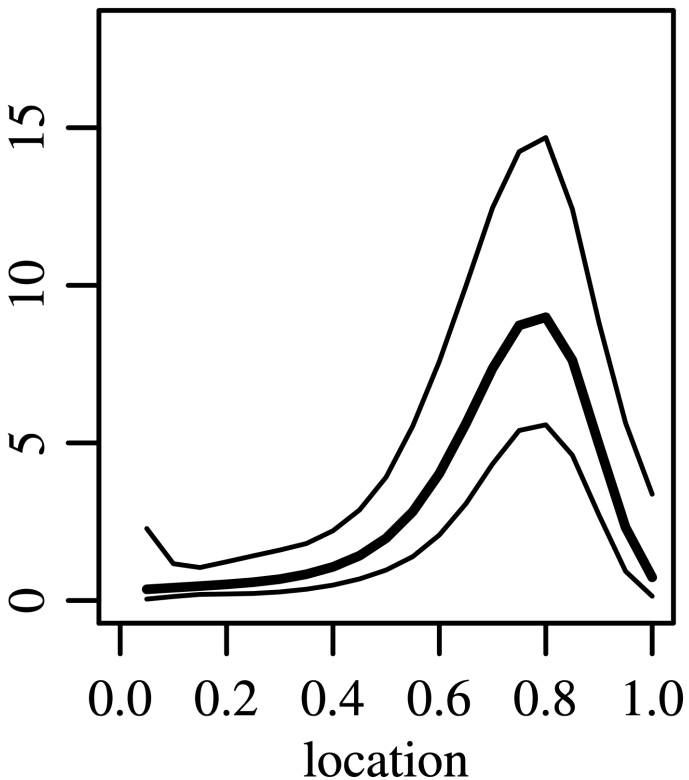
Population average number of tumors vs location

- The plot shows the posterior median curve (thick line) and the 95% credible interval for the curve: this is obtained by just looking at $\exp(\mu' \mathbf{X})$ where $\mathbf{X}$ is the 20×5 matrix with $i-th$ row equal to $\mathbf{x}_i'$.



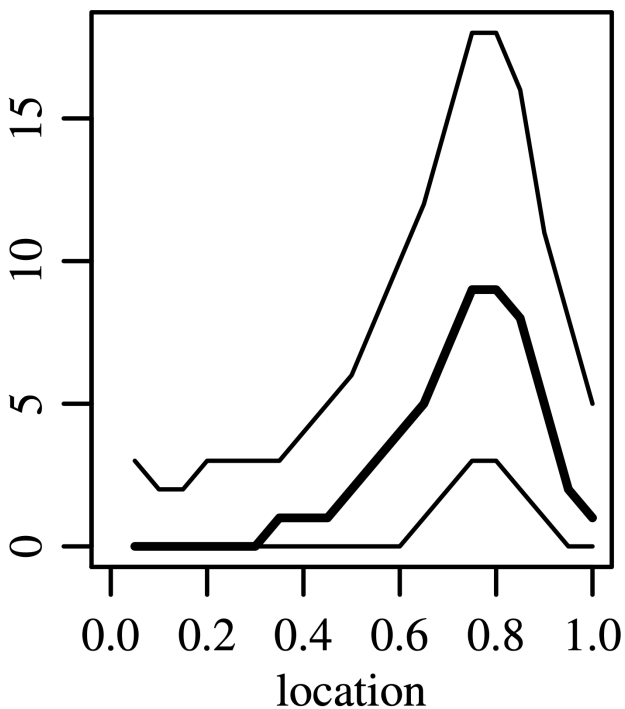
Predicted average number of tumors vs location

- We could also look at what would be the predicted curve for the **expected** number of tumors for a mouse yet to be sampled.
- We can obtain this curve, by sampling $\tilde{\beta}$ from its posterior predictive distribution $p(\tilde{\beta}|\mathbf{data}, \mathbf{X})$ and computing $\exp(\tilde{\beta}'\mathbf{X})$
- The plot shows the posterior predictive median curve (thick line) and the 95% credible interval for the expected number of tumors.



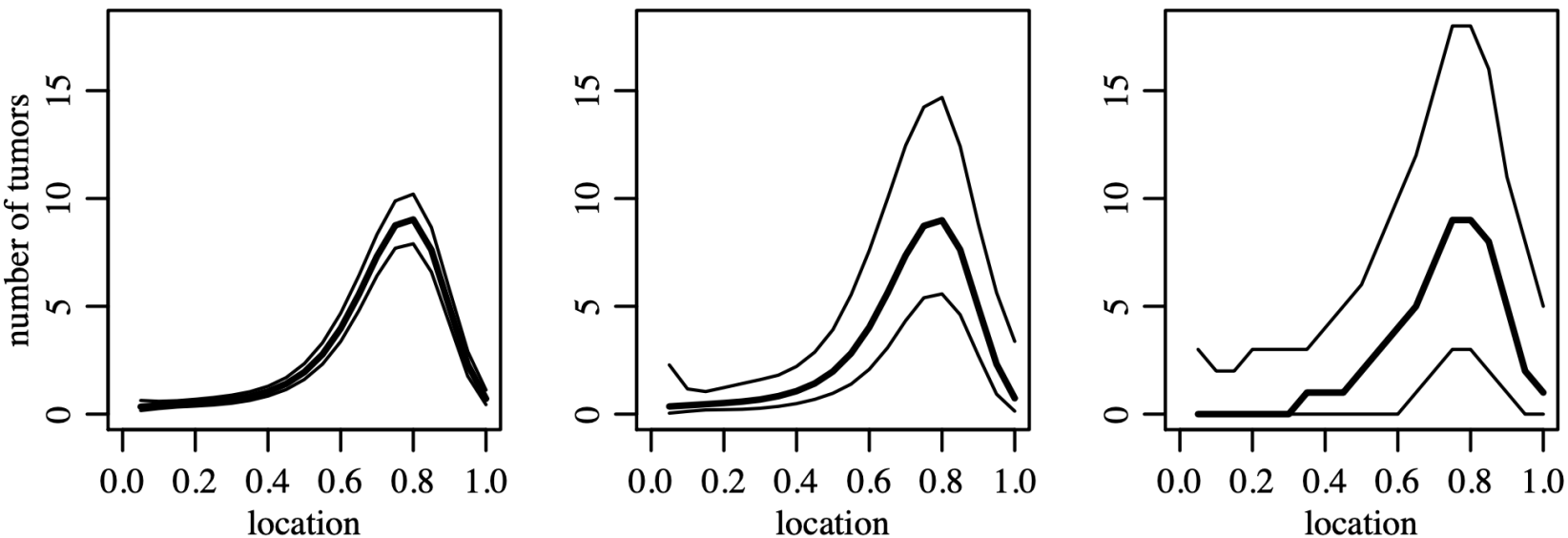
Predicted number of tumors vs location

- We could also look at what would be the predicted curve for the number of tumors for a mouse yet to be sampled.
- We can obtain this curve, by sampling $\tilde{\mathbf{Y}}$ from its posterior predictive distribution $p(\tilde{\mathbf{Y}}|\mathbf{data}, \mathbf{X})$.
- The plot shows the posterior predictive median curve (thick line) and the 95% credible interval for the number of tumors for a mouse yet to be sampled.



Different levels of uncertainty

- Let's look at the three curves altogether:



Different levels of uncertainty

- The three curves display a different degree of uncertainty:
  - the left-most one, the curve for the population average has less uncertainty: the uncertainty here is given by the uncertainty in the fixed effect μ
  - the middle one, the curve for the expected number of tumors for a not-yet-observed mouse has more uncertainty due to the heterogeneity across mice
  - the right-most one, the curve for the number of tumors in a not-yet-observed mouse has the most uncertainty: this is due to the heterogeneity across mice and the additional variability of a Poisson random variable around its expected value.

What to report

- If we want to report the uncertainty in the number of tumors a mouse might have at a certain intestine location, we should use the credible intervals in the right-most plot.
- If we want to report the uncertainty in the expected number of tumors a mouse might have at a certain intestine location, we should use the credible intervals in the middle plot.
- If we want to report the estimated fixed effect of the intestine location on the number of tumors and communicate our uncertainty in that estimate, we should use the credible intervals in the left-most plot.