第 9 讲：贝叶斯层次模型

张伟平

目录

1.1 贝叶斯层次模型

- 层次模型（Hierachical model）并没有一个公认的定义
- Gelman et al., 2004
  - “Estimating the population distribution of unonobserved parameters”
  - “Multiple parameters related by the structure of the problem”
- 知道一个“试验”的一些信息就知道另外一个试验的一些信息
  - 多个相似的试验
  - 不同位置的相似测量
  - 在同一组图像上执行多个任务

- “Sharing statistical strength.”这里想法是一些我们能从一组数据中容易推断出的东西，可以帮助我们解决在另外一组数据中不能方便推断的东西。比如，我们有 A 试验很多数据，但类似试验 B 的数据很少。那么我们从试验 A 的数据中学习出的东西能够帮助我们分析试验 B 的数据。

- **关键想法** 对一个未知量的推断，影响对另外一个未知量的推断。

  - 多种称法: Multilevel models, 随机效应模型, 混合效应模型等
  - 包括许多传统的层次模型
  - 先验/似然不称为层次模型
  - 包含一些并没有表示成层次形式的模型，例如 HMMs，Kalman filters，高斯混合模型等

四个关键观点

- 对具有复杂结构的数据建模
  - 有许多层次模型可以处理的结构, 例如学生嵌套在学校中; 房屋嵌套在社区中等
- 对异质性建模
  - 标准回归考察” 平均”, 层次模型额外对方差进行建模, 例如房屋价格的波动性从一个社区到另一个社区
- 对相依数据建模
  - 结果随时间, 空间, 背景等变化而存在潜在复杂相依关系
- 对背景因素建模: 微观和宏观关系
  - 例如当个房子价格依赖于其特点和社区的特点

一些例子

- 医院手术死亡率
  - 结构: 病人嵌套在不同医院中
  - 问题: 那个医院死亡率特别高或特别低?
- 环境暴露 (THM)
  - 结构: 测量不同供水区域自来水中的三卤甲烷 (THM) 浓度
  - 问题: 估计每个区域平均的 THM 浓度, THM 浓度在各区域内和之间的波动性如何?
- 纵向临床试验
  - 结构: 每位病人重复观测

  - 问题: 治疗效果上是否有差异? 不同治疗方案下病人的响应是否有异质性?

- 教育成果
  - 结构: 苏格兰学生 16 岁考试成绩, 学生分为小学和中学 (没有嵌套)
  - 问题: 学生考试成绩中有多少波动性来自小学, 多少来自中学?

- 单病例随机对照 (N-of-1) 试验
  - 结构: 对单个病人重复观测多种干预的差异
  - 问题: 疗法是否有效

对具有复杂结构的数据建模

上述例子都说明了一个基本建模问题: 我们希望基于模型推断 n 个" 单元"(个体, 子集, 地区, 时间点, 试验等) 的各自参数 $\theta_1, \ldots, \theta_n$, 它们之间通过问题背景被联系起来. 那么我们考虑下面三种不同的模型假设

- **相同参数** 所有 θ 都相同, 此时数据可以混合起来, 个体单元可以被忽略
- **独立参数** 所有 θ 是完全不相关的, 此时每个单元的数据可以独立分析
- **参数可交换** 所有 θ 被假设是" 相似的", 即它们的标签没有信息. 数学上来说, 就是假设 $\theta_1, \ldots, \theta_n$ 来自同一个具有未知参数的先验分布

Exchangeability

如果除了观测数据外, 没有其他信息来区分 $\theta_j, j = 1, \ldots, n$, 参数也没有顺序或者组别的假设, 那么我们需要假设参数在其联合分布里是对称的.

可交换性: 称 $X_1, \ldots, X_n$ 是可交换的，如果对 $1, \ldots, n$ 的任一置换 $\sigma(1), \ldots, \sigma(n)$，有

$$(X_1, \ldots, X_n) \stackrel{d}{=} (X_{\sigma(1)}, \ldots, X_{\sigma(n)})$$

Definition

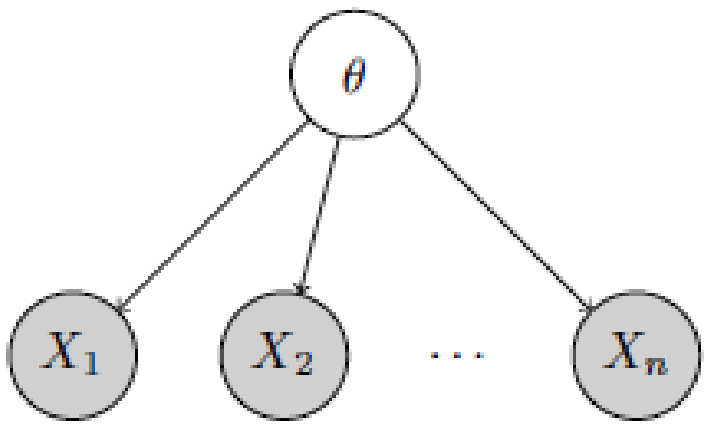
无穷个变量称为是无穷可交换的，如果其任意有限个满足可交换性.

定理 1. *(de Finetti's theorem,roughly)* $X_1, X_2, \ldots$ 为无穷可交换的,

当且仅当对任意有限 n 和某个分布 P, 使得

$$p(X_1, \ldots, X_n) = \int_\theta \left[\prod_{j=1}^{n} p(X_j|\theta) \right] P(d\theta)$$

即, 如果一个序列是无穷可交换的, 则必存在某个参数 θ 与其上的概率分布 P, 使得给定 θ 后, 所有数据是 i.i.d 的. 使用图模型表示如下



考察英国自来水 THM 浓度问题: 70 个区域, 每年在供水区域里随机采样. 考察每个区域的平均浓度

↑Example

↓Example

对每个区域假设正态似然

$$x_{iz} \sim N(\theta_z, \sigma^2_{[e]}), i = 1, \ldots, n_z, z = 1, \ldots, 70$$

- 我们有 70 个 θ_z, 应该用什么先验?
- 对方差, 使用模糊先验, $\sigma^2_{[e]} \sim InvGamma(0.001, 0.001)$

相同参数

所有 $\theta_z = \theta$, $z = 1, \ldots, 70$ 指定正态先验

$$\theta \sim N(\theta_0, \sigma^2_0)$$

比如 $\theta \sim N(0, 100000)$.

假设所有均值参数都相同显然不合理, 我们期望不同供水区域的 THM 是不同的.

独立参数

假设所有的 θ_z 相互独立, 每个赋予模糊先验, 比如

$$\theta_z \sim N(0, 100000), z = 1, \ldots, 70$$

此时每个 θ_z 的估计是独立的, 其他数据对其估计没有影响, 无法使用其他地区的数据.

可交换参数

我们使用如下的分层先验:

$$\theta_z \sim N(\mu, \sigma^2_{[z]}), z = 1, \ldots, 70$$

$$\mu \sim N(0, 100000)$$

$$\sigma^2_{[z]} \sim InvGamma(0.001, 0.001)$$

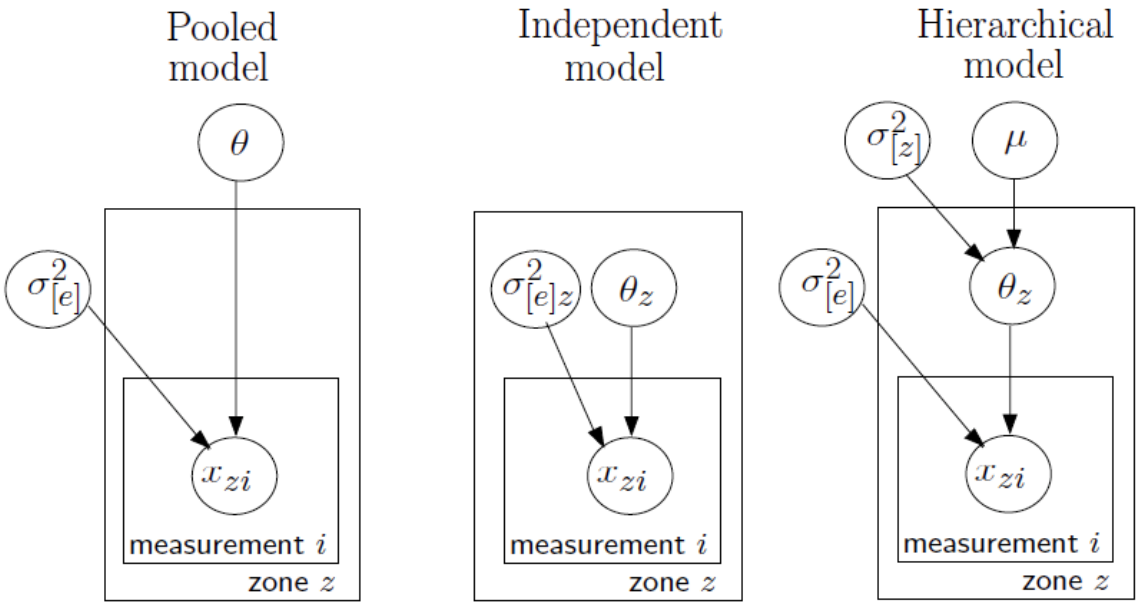
其中 $\sigma^2_{[z]}$ 表示不同供水区域之间的 THM 浓度方差, $\sigma^2_{[e]}$ 表示了 THM 浓度残差的方差; θ 是总的平均值.

- 联合先验分布

$$p\left(\theta_1, \ldots, \theta_{70}, \sigma^2_{[e]}, \mu, \sigma^2_{[z]}\right) \propto \left\{\prod_{z=1}^{70} p\left(\theta_z | \mu, \sigma^2_{[z]}\right)\right\} p\left(\sigma^2_{[e]}\right) p(\mu) p\left(\sigma^2_{[z]}\right)$$

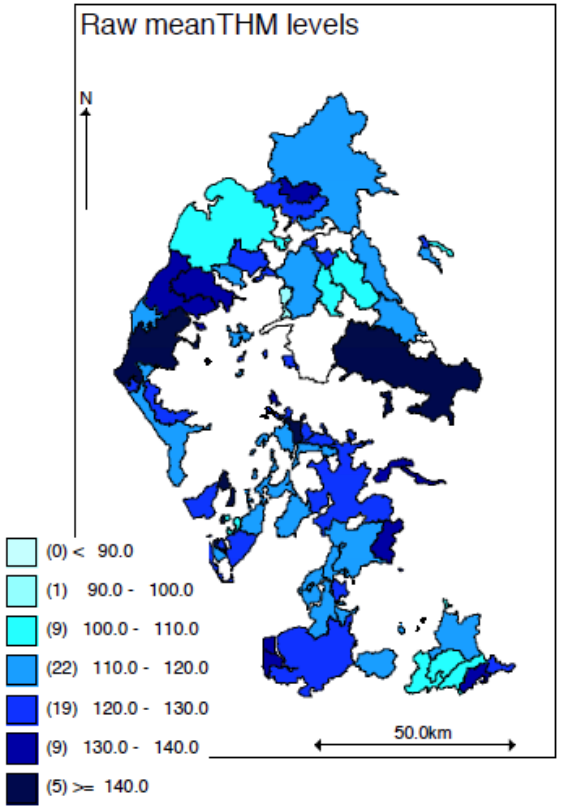
- 每个地区均值 θ_z 的边际后验分布就是从联合后验分布中积分掉其他变量.
- 从而可以" 借力" 其他地区的信息
- 得到每个地区均值的总的光滑, 每个地区均值估计的精度得到提高

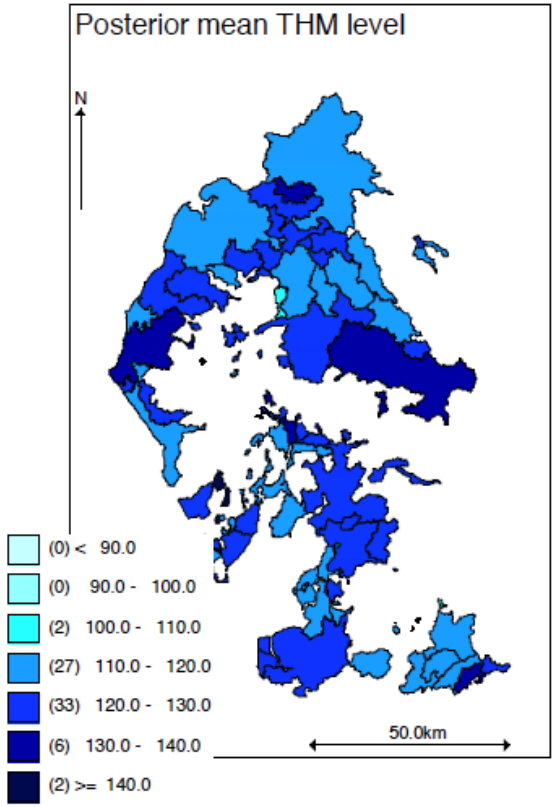
THM 模型的图模型表示:



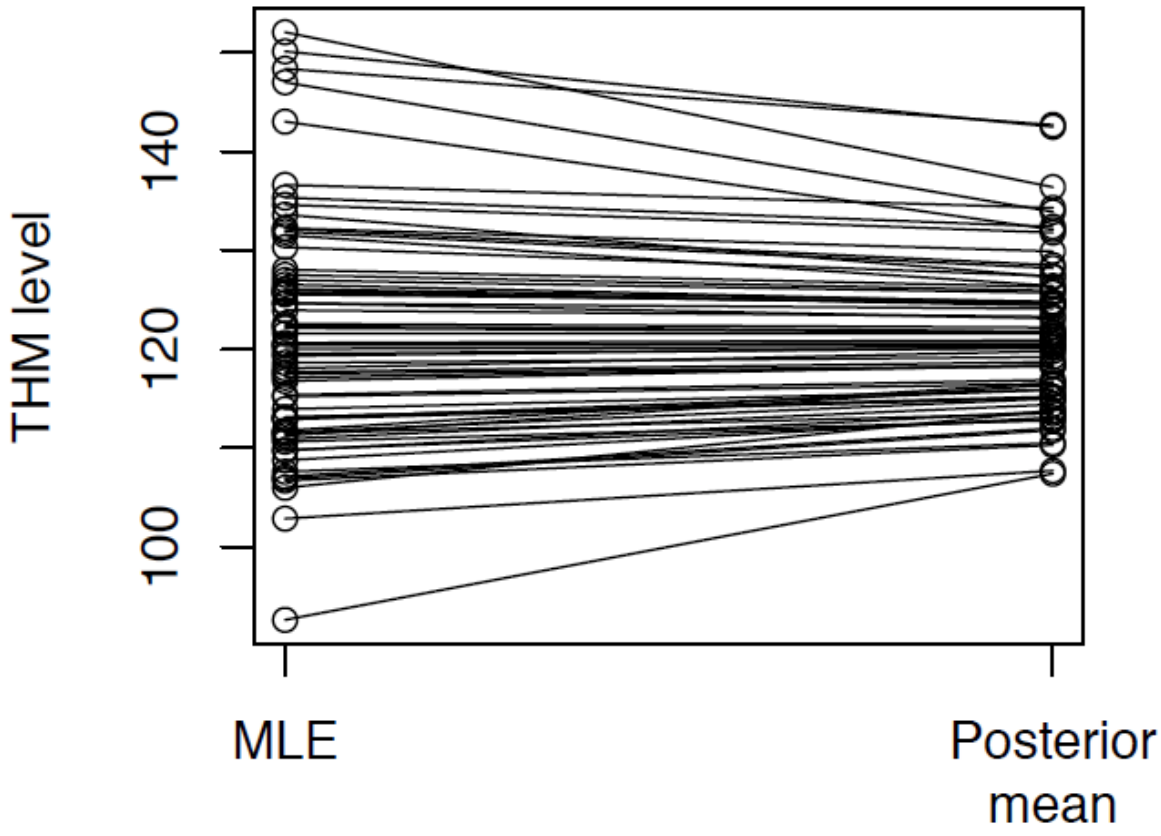
- θ_z 是第 z 个地区的均值
- μ 是所有区域的总平均
- $\sigma^2_{[z]}$ 是不同区域浓度间的波动
- $\sigma^2_{[e]}$ 是 THM 浓度的残差方差

每个区域估计均值地图

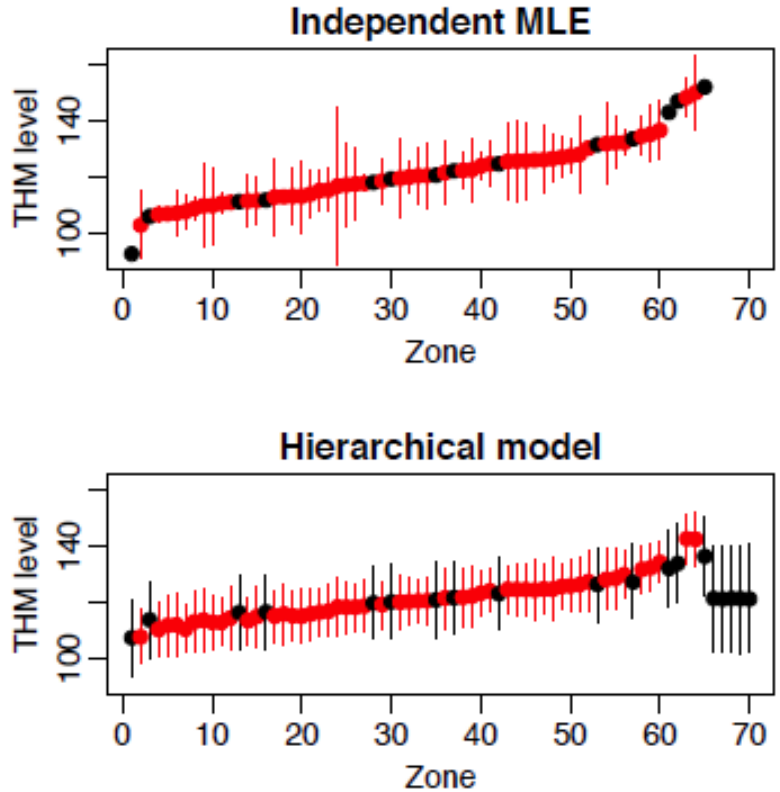

Raw meanTHM levels
N
(0) < 90.0
(1) 90.0 - 100.0
(9) 100.0 - 110.0
(22) 110.0 - 120.0
(19) 120.0 - 130.0
(9) 130.0 - 140.0
(5) >= 140.0
50.0km


Posterior mean THM level
N
(0) < 90.0
(0) 90.0 - 100.0
(2) 100.0 - 110.0
(27) 110.0 - 120.0
(33) 120.0 - 130.0
(6) 130.0 - 140.0
(2) >= 140.0
50.0km

BHM 对均值进行了压缩 (光滑)

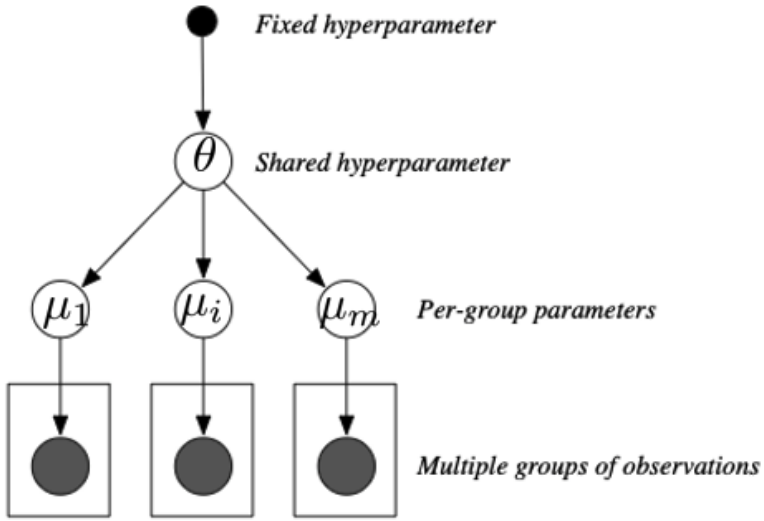


不同供水区域 THM 浓度的点估计和 95% 置信区间



- 注意在 BHM 下给出了 5 个没有数据区域的估计
- 独立 MLE 模型每个区域的方差也是独立的，因此如果只有一个观测，就没有方差的估计
- BHM 假设了相同的误差方差 (因此也可以假设层次先验

- 经典层次模型看起来如下



$$X_{ij} \sim F_2(\mu_j)$$

$$\mu_j \sim F_1(\theta)$$

$$\theta \sim F_0(\alpha)$$

- 我们观测到多组数据，每个组有自己的参数
- 参数的分布共享一个（些）超参数，超参数有着自己的固定超参数。
- 对一个组的参数进行推断会影响推断其他组的参数
- 考虑下面的一个生成过程, 固定超参数 α:

- 抽取 $\theta \sim F_0(\alpha)$
- 对每个组 $i \in \{1, \ldots, m\}$:
  - 抽取 $\mu_i|\theta \sim F_1(\theta)$
  - 对每个数据点 $j \in \{1, \ldots, n_j\}$:
    - 抽取 $x_{ij} \sim F_2(\mu_j)$

- 考虑第 i 组参数的后验分布

$$p(\mu_i|\mathscr{D}) \propto \int p(\theta|\alpha) p(\mu_i|\theta) p(\mathbf{x}_i|\mu_i) \left(\prod_{j \neq i} \int p(\mu_j|\theta) p(\mathbf{x}_j|\mu_j) d\mu_j \right) d\theta$$

- 记 $\mathbf{x}_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$,$\mu_{-i} = (\mu_1, \ldots, \mu_{i-1}, \mu_{i+1}, \ldots, \mu_n)$, 上述积分中第一项和第四项一起为 $p(\mathbf{x}_{-i}, \theta|\alpha)$, 注意到

$$p(\mathbf{x}_{-i}, \theta|\alpha) = p(\mathbf{x}_{-i}|\alpha) p(\theta|\mathbf{x}_{-i}, \alpha)$$

因此

$$p\left(\mu_i|\mathscr{D}\right) \propto \int p\left(\theta|\mathbf{x}_{-i},\alpha\right) p\left(\mu_i|\theta\right) p\left(\mathbf{x}_i|\mu_i\right) d\theta$$

即其他组数据通过超参数的分布影响了第 i 组的参数推断

- 如果超参数 θ 固定，则显然此时其他组数据对第 i 组参数推断没有影响
- 下面我们看看其他组数据是怎么影响 θ 的. 从图模型中我们可以看出它们通过 μ_j 来影响 θ. 但是也可以通过打开后验分布 $p(\theta|\mathbf{x}_{-i})$ 来看
- θ 给定 $\mathbf{x}_{-i}$ 的后验分布为

$$p\left(\theta|\mathbf{x}_{-i},\alpha\right) \propto p(\theta|\alpha) \prod_{j\neq i} \int p\left(\mu_j|\theta\right) p\left(\mathbf{x}_j|\mu_j\right) d\mu_j$$

- 视每个积分为 $p(\mu_j|\theta)$ 的加权平均, 对每个 μ_j, 权重为 $p(\mathbf{x}_j|\mu_j)$. 假设每个积分被在一个点处的值所控制, 记其为 $\hat{\mu}_j$, 则可以视后验为

$$p\left(\theta|\mathbf{x}_{-i},\alpha\right) \propto p(\theta|\alpha)\prod_{j\neq i} p\left(\hat{\mu}_j|\theta\right)$$

  即, 其他组数据对超参数的影响通过它们各自组的代表来进行的.

- 这看起来有些像先验/似然设置. 即如果 $p(\theta|\alpha)$ 和 $p(\hat{\mu}_j|\theta)$ 形成一个共轭对, 则后验分布也是共轭的.

- 这不是实际的推断, 但是表明了信息是如何传递的:

  – 其它组告诉我们它们的参数

  – 这些参数告诉我们关于超参数的一些东西

  – 超参数告诉我们关于当前研究数据组的一些东西

- 考虑其他计算问题, 在给定数据条件下
  - 一个已知组的预测分布 $p\left(x_i^{\text{new}} \mid \mathscr{D}\right)$ 将依赖其他组 (对超参数) 和该组内的其他数据 (对组内参数)
  - 一个完全未知组参数 μ^{new} 的分布将依赖于 θ 给定所有数据的后验分布

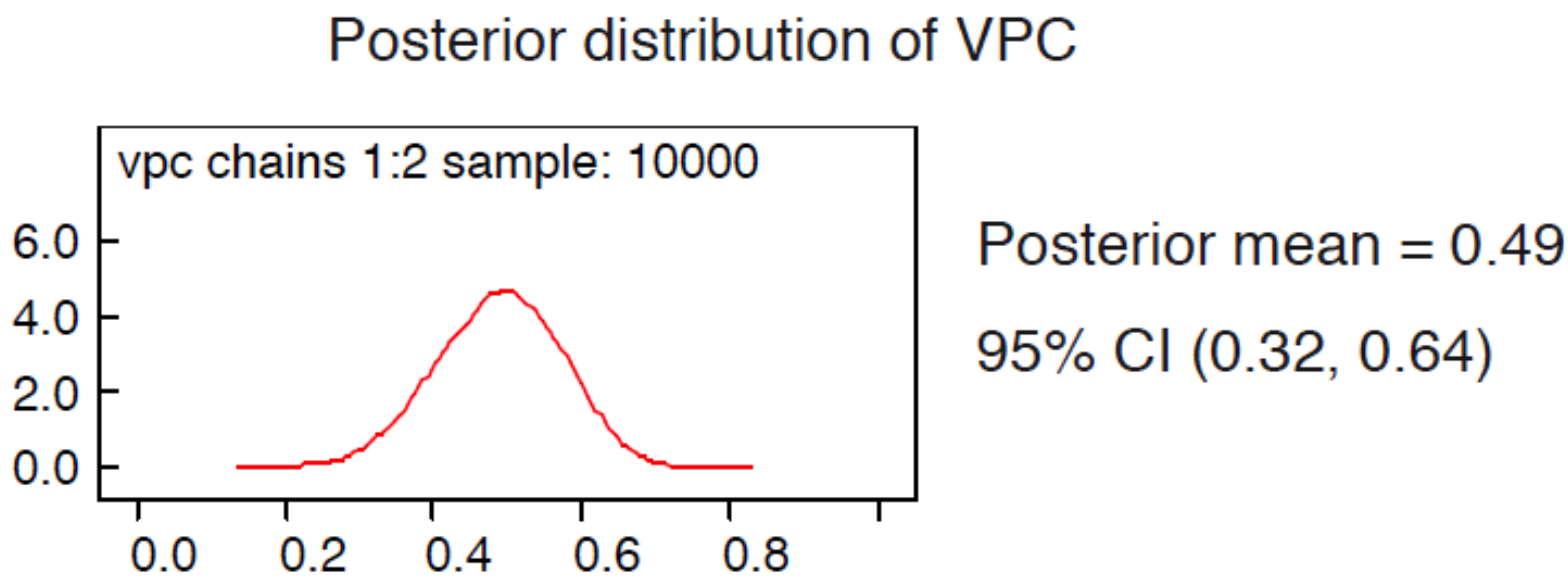
方差分解系数 (VPC)

- 在层次模型中, 响应变量的残差波动性被分成对应不同水平的不同部分
- 经常感兴趣的是高一层水平单元波动占总波动的百分比
- 在简单 2-水平正态线性模型中, 我们可以用 VPC 或类内相关系数

$$VPC = \frac{\sigma^2_{[z]}}{\sigma^2_{[z]} + \sigma^2_{[e]}}$$

  - $\sigma^2_{[e]}$ 是"level 1" 方差 (正态似然的方差)
  - $\sigma^2_{[z]}$ 是"level 2" 方差 (随机效应方差)
- 在 WINBUGS 中, 添加一行计算 VPC 的代码, 同时监视 VPC 的后验样本来得到点估计等.

THM 例子的 VPC



因此, 大约一半的总波动性是来自不同区域之间的波动性, 另一半是供水区域内的波动性

在很多问题中, 我们感兴趣的是对来自不同地区的数据进行建模, 我们这里考虑伦敦 10 年期间不同地区观测到的儿童白血病人数, y_i, $i = 1, \ldots, 879$ 个地区. 使用全国年龄/性别标准化参考发病率, 以及人口量, 我们可以计算出每个地区期望的病例个数 E_i. 使用贝叶斯层次模型进行分析.

↑Example

↓Example

对每个地区病例数假设 Poisson 似然

$$y_i \sim Pois(\lambda_i E_i), i = 1, \ldots, 879$$

- **相同参数** 假设所有 $\lambda_i = \lambda$, 然后赋先验

$$\lambda \sim Gamma(a, b)$$

其中 a, b 为指定的值, 比如 $a = b = 1$, 此时模型为 Poisson-Gamma 共轭系统.

- **独立参数** 对每个 λ_i 假设独立先验, 例如

$$\lambda_i \sim Gamma(0.1, 0.1), i = 1, \ldots, 879$$

此时后验均值估计 $\hat{\lambda}_i \approx y_i/E_i$, 而后者为 MLE(也称为标准化的发病率, SMR)

- **可交换参数** 假设一个层次随机效应先验

$$\lambda_i \sim Gamma(a, b), a \sim G_1, b \sim G_2$$

什么样的超先验 G_1, G_2 是合适的? 这种做法不允许相邻地区的空间相依关系

- $\log \lambda_i$ 正态随机效应模型更加灵活

$$y_i \sim Pois\left(E_i \lambda_i\right)$$

$$\log \lambda_i = \alpha + \theta_i$$

$$\theta_i \sim N\left(0, \sigma^2\right)$$

超参数 σ^2 和 α 可以取无信息先验, 例如

$$\sigma^2 \sim InvGamma(0.001, 0.001), \alpha \sim N(0, 10000)$$

- θ_i 为随机效应
- $\lambda_i = exp(\alpha + \theta_i) =$ 第 i 个地区相对基于人群中年龄和性别的期望风险的相对风险
- θ_i 也可以被视为一个隐变量, 其代表了未知或没有观测地区的水平协变量效应
- 如果这些地区水平协变量具有空间相依关系 (比如环境效应), 则我们对 θ_i 的模型应该允许这一点 (即使用空间分布代替正态随机效应模型)
- 随机效应的方差 σ^2 反映了数据中 extra-Poisson 波动性
- 对广义线性层次模型尚不清楚如何定义/计算 VPC

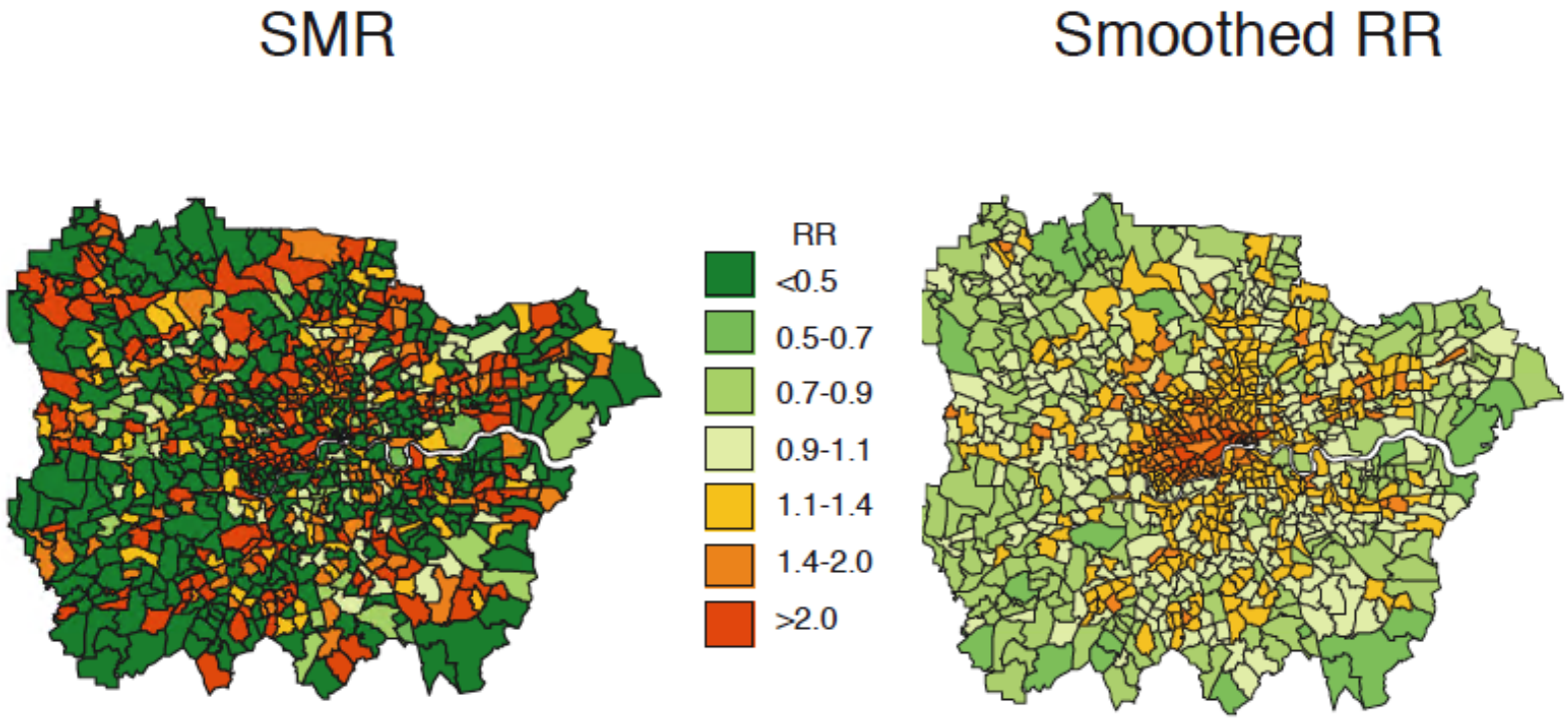
- 对随机效应波动性进行汇总时候, 则替代研究它们经验分布的分位数之比.
- 层次模型中一种有用的汇总个体波动性的方法, 是对随机效应进行排序, 然后计算两个相反极值的差或比. 比如, 我们考虑地区相对风险分布的 5% 和 95% 分位数, 则 $\lambda_{5\%}$ 表示相对风险处于第 5 百分位数的地区的相对风险, $\lambda_{95\%}$ 表示相对风险处于第 95 百分位数的地区的相对风险, 则 $QR_{90} = \lambda_{95\%}/\lambda_{5\%} =$ 位于所有地区上下 5% 处的两个地区相对风险之比
- WINBUGS 中使用 *Inference* 中的 *Rank* 选项来监视一个向量的元素排序: rank(x[],i) 返回 x 第 i 个元素的秩; ranked(x[],i) 返回 x 中秩为 i 的元素

结果分析

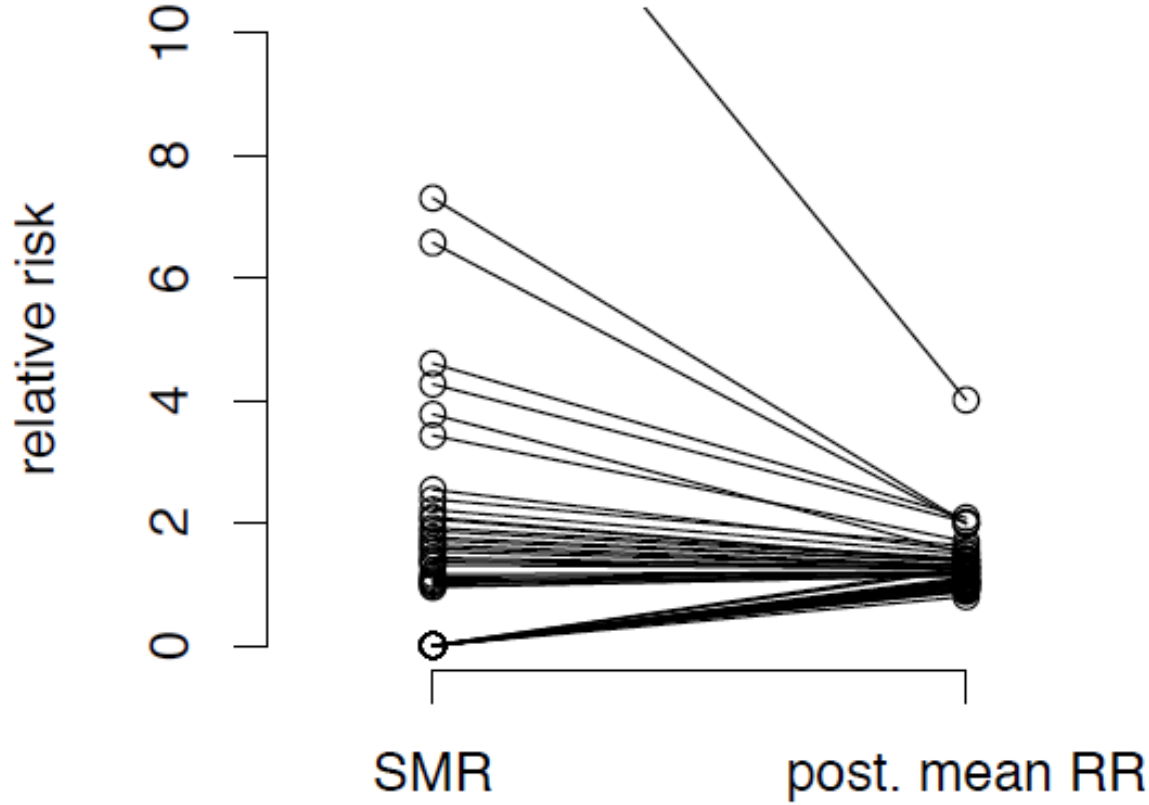
感兴趣的参数包括

- $\lambda_i = exp(\alpha + \theta_i)$: 第 i 个地区相对于期望风险的相对风险 (见下图)
- σ: 对数相对风险在不同地区之间的标准偏差, 后验均值和 95% 区间为 0.46(0.34,0.62)
- $QR_{90} = 4.7$, 95% 区间为 (2.9,7.5), 因此在所有地区上下 5% 的两个地区相对风险为 4.7 倍

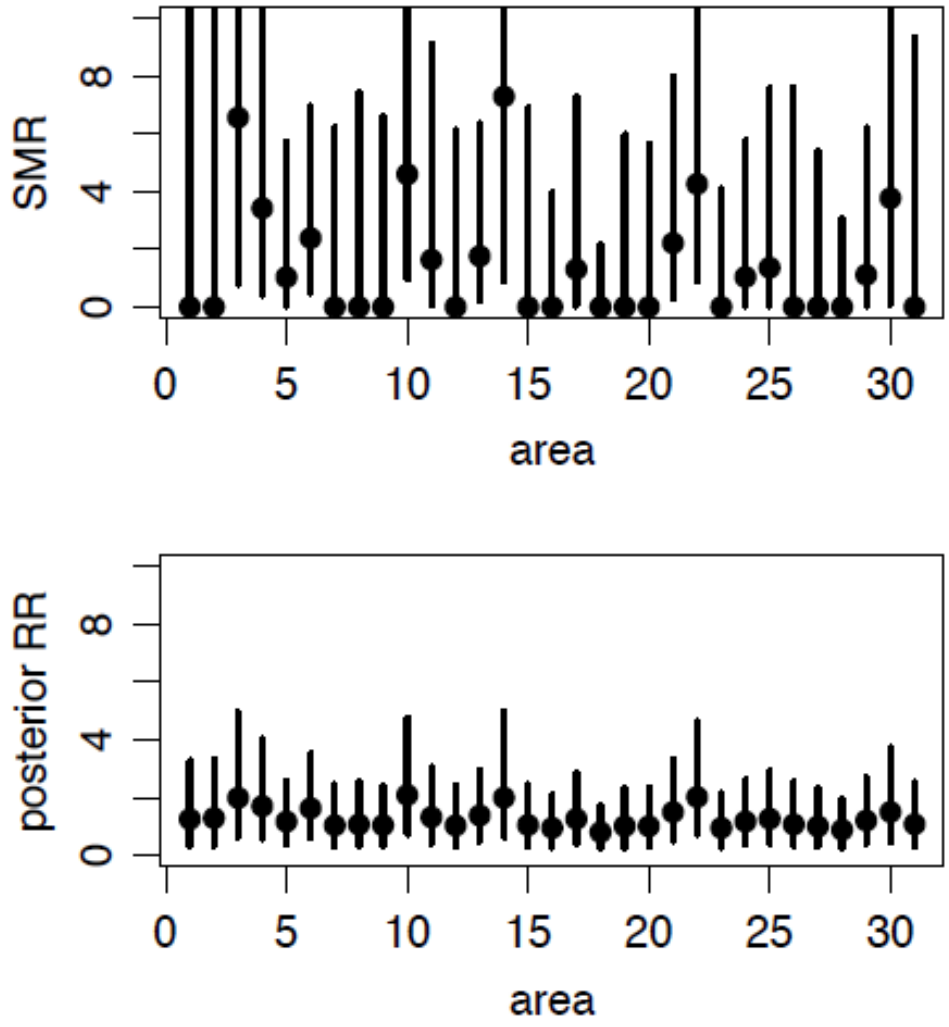
估计的地区水平白血病相对风险地图:



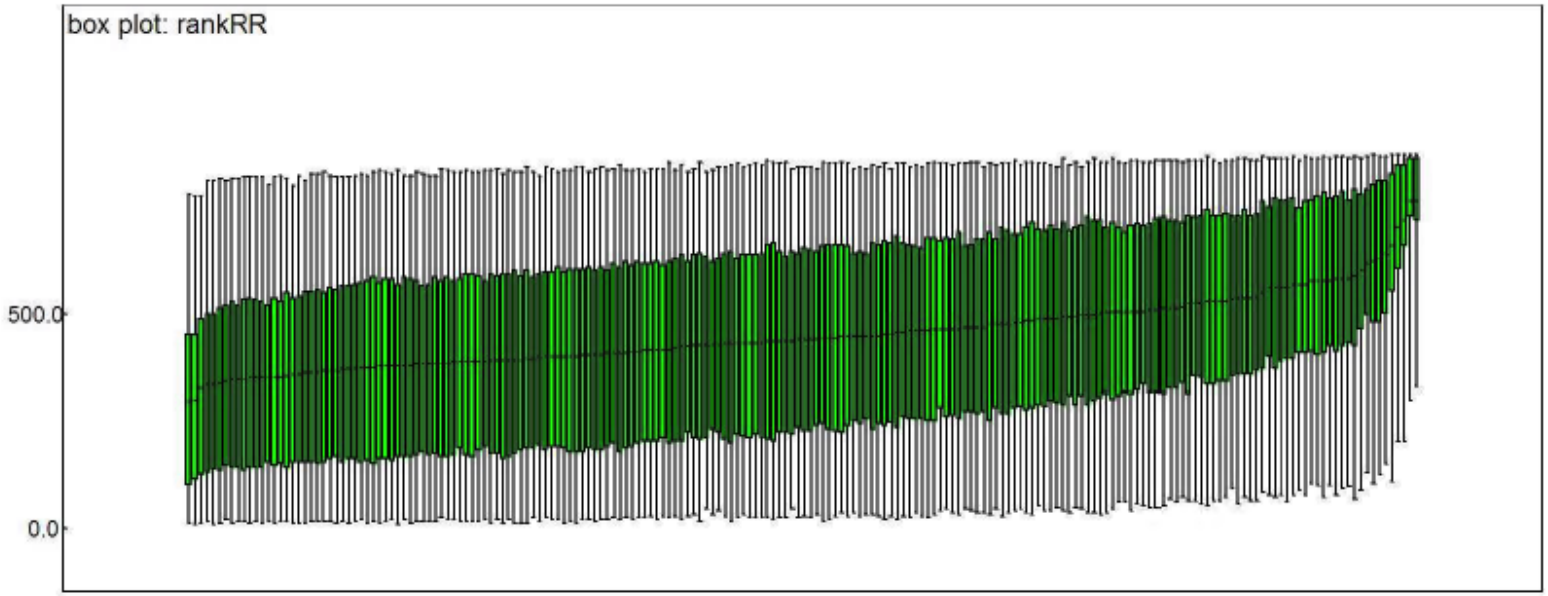
部分地区的 SMR 与相对风险的后验均值



对部分地区的点估计和 95% 区间估计



地区相对风险的后验分布排序



层次模型的一般性备注

层次模型允许从不同单元" 借力"

- θ_i 的后验分布从所有单元的似然贡献中" 借力", 通过它们对未知超参数后验估计的联合影响, 从而提升效率
- MCMC 允许随机效应分布灵活选择, 不必要限制于正态随机效应
- 可交换性的判断需要仔细评估
  - 对预先怀疑具有系统性差异的单元, 可以通过引入相关协变量使得残差波动性能更合理反映可交换性来建模
  - 具有先验性兴趣的子群体需要单独考虑

1.2 层次模型的先验

先验指定的一般性建议

- 区分
  - 对主要感兴趣的参数, 我们或许希望减少先验的影响
  - 对用于光滑的次要结构, 一个 (适度) 有信息先验更合适
- 在层次模型中, 我们可能主要感兴趣第一层参数 (回归系数等), 或者方差分量 (随机效应和它们的方差), 或者两者都是
- 先验最好施加在可解释参数上
- 要**特别小心** ” 在复杂模型中一个明显没有什么影响的均匀先验不会带来大量信息” 这一观念
- ” 根本就没有什么东西可以作为一个” 无信息” 先验, 即便不正常先验也具有信息: 所有可能值是等可能的”(Fisher, 1996)

位置参数的先验

位置参数指均值, 回归系数等

- 在较大范围取值的均匀分布, 或者方差很大的正态分布, 例如

$$\theta \sim U(-100, 100), \theta \sim dunif(-100, 100)$$

$$\theta \sim N(0, 1e5), \theta \sim dnorm(0, 0.00001)$$

先验在似然支撑区域是局部均匀的

- 注意 WINBUGS 对正态分布使用均值和精度 (方差倒数)
- 注意" 宽" 范围和" 小" 精度依赖于 θ 的尺度

第一层参数的先验

- 样本的方差 σ^2: 标准'reference(Jeffreys)' 先验

$$p\left(\sigma^2\right) \propto \frac{1}{\sigma^2} \propto \text{Gamma}(0,0)$$

- 这等价于对数尺度下的扁平 (均匀) 先验

$$p\left(\log\left(\sigma^2\right)\right) \propto \text{ Uniform } (-\infty, \infty)$$

- 此先验直观上合理: 它完全忽略了参数的尺度 (量级), 则它位于区间 1-10, 和位于 10-100 是等可能的
- 对精度 $\tau = \sigma^{-2}$, Jeffreys 先验为

$$p(\tau) \propto \frac{1}{\tau} \propto \text{Gamma}(0,0)$$

这等价于

$$\tau \sim \text{Gamma}(\epsilon, \epsilon) \quad (\text{with } \epsilon \text{ small })$$

- 这也是一个共轭先验, 在正态似然场合常常使用.
- 在 WINBUGS 中, $tau \sim dgamma(0.001, 0.001)$

超参数的先验

假设一个层次模型具有可交换的随机效应

$$\theta_i \sim \mathrm{N}\left(\mu, \sigma^2\right) \quad i = 1, \ldots, l$$

则对 μ, σ^2 应该选择什么先验?

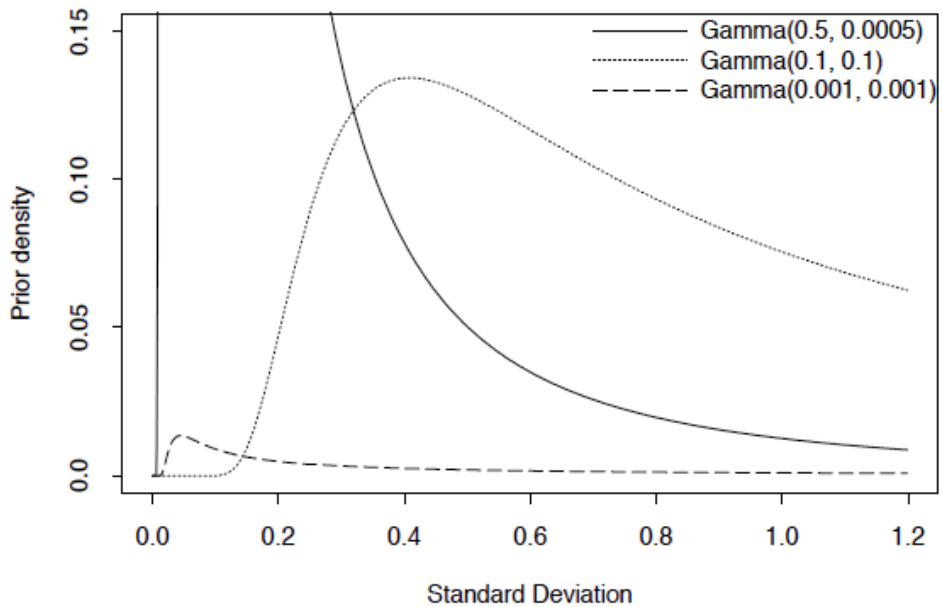
常常希望对均值和随机效应方差指定合理的无信息先验

- 对位置参数, 取值范围较大的均匀先验, 或者具有大方差的正态先验, 都可以使用
- 对随机效应方差的无信息先验则比较有技巧性
  - Jeffreys 无信息先验

$$p\left(\sigma^2\right) \propto \frac{1}{\sigma^2} \propto \mathrm{Gamma}(0, 0)$$

是不正常先验, 但是作为第一层方差先验是可以使得后验正常

- 作为随机效应方差的先验, 有可能导致后验不正常
- 先验在 0 点有无穷质量, 但是 $\sigma^2 = 0$(没有第 2 层波动性), 支持这一现象的似然不能忽略
- $Gamma(\epsilon, \epsilon)$, 其中 $\epsilon > 0$ 非常小, 是一个正常先验. Gamma(0.001,0.001) 常作为随机效应精度的先验, 因为对正态分布具有很好的共轭性质, 但是推断结果仍然可能对 ϵ 的值敏感



对精度使用不同的 gamma(a, b) 先验

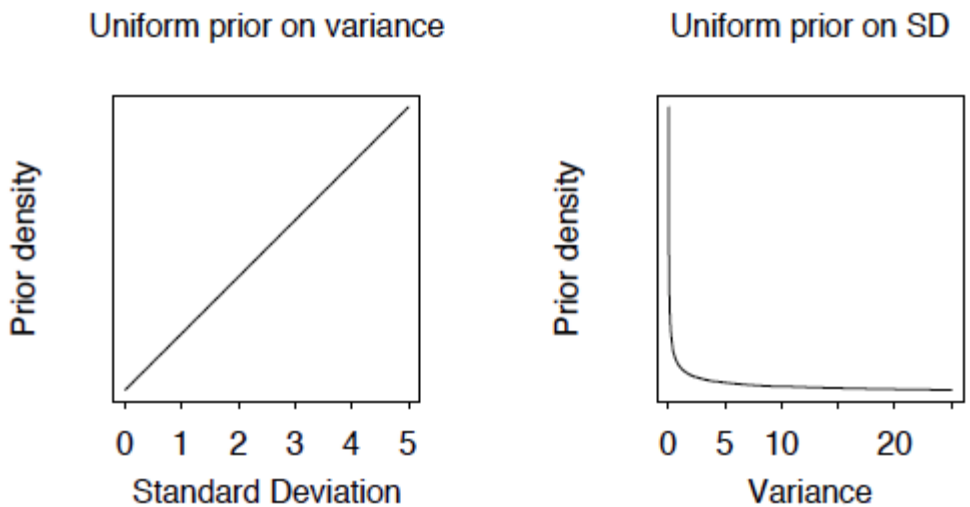
对方差的其他” 模糊” 或弱信息先验有

- 对标准偏差指定一个有限区间上的均匀分布, e.g.

$$\sigma \sim U(0, 1000)$$

显然, 上限的选择依赖于参数的尺度

- 对方差指定均匀先验一般不推荐, 因为其暗含关于标准差的不实际先验信息



- 其他选择还有对标准差赋予半正态或半 t 分布

$$\sigma \sim N(0, 100)I(0,)$$

在 WINBUGS 中, $sigma \sim dnorm(0, 0.01)I(0,)$

- 当然, 半正态方差的值依赖于连续数据的测量尺度

Gelman(2006) 比较学生天赋测试

↑Example
↓Example

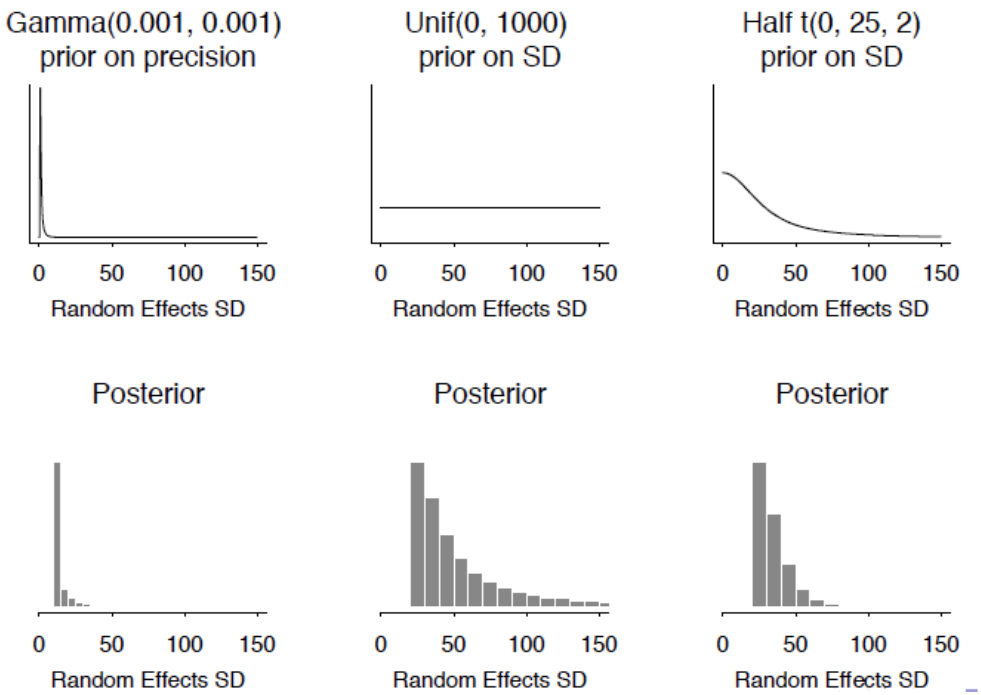
两层正态层次模型

$$X_{ij} \sim N(\theta_i, \sigma^2), \theta_i \sim N(\theta, \sigma_2^2)$$

θ_i 为关于学校的随机效应

- 8 所学校观测的效应值从-2.75(SE 16.3) 到 28.4(SE 14.9)

- 似然支持学校间的标准差取值范围较大
  - 所有学校具有相同的效应 (校间 SD=0) 是合理的
  - 数据支持学校间的波动很大, 尽管 $SD > 50$ 不合理
- 考虑前三个学校 (数据非常稀疏), 比较随机效应精度取先验 $Gamma(0.001, 0.00$ 与随机效应标准差取先验 $U(0, 1000)$ 以及半 t 分布 $t(0, 25, 2)I(0,)$:



1.3 纵向数据中的层次模型

- 层次模型的一个主要应用领域是回归
- 层次 (或多水平,multilevel) 建模允许我们在复杂数据集上使用回归
  - 分组回归问题 (i.e., 嵌套结构)
  - 重叠分组问题 (i.e., 非嵌套结构)
  - 每组有不同系数的问题
  - 随机效应模型
- 本节我们使用两个数据集讨论对纵向数据进行建模的方法:
  - 比较一个抗抑郁临床试验中不同处理
  - 评估认知随年龄减退

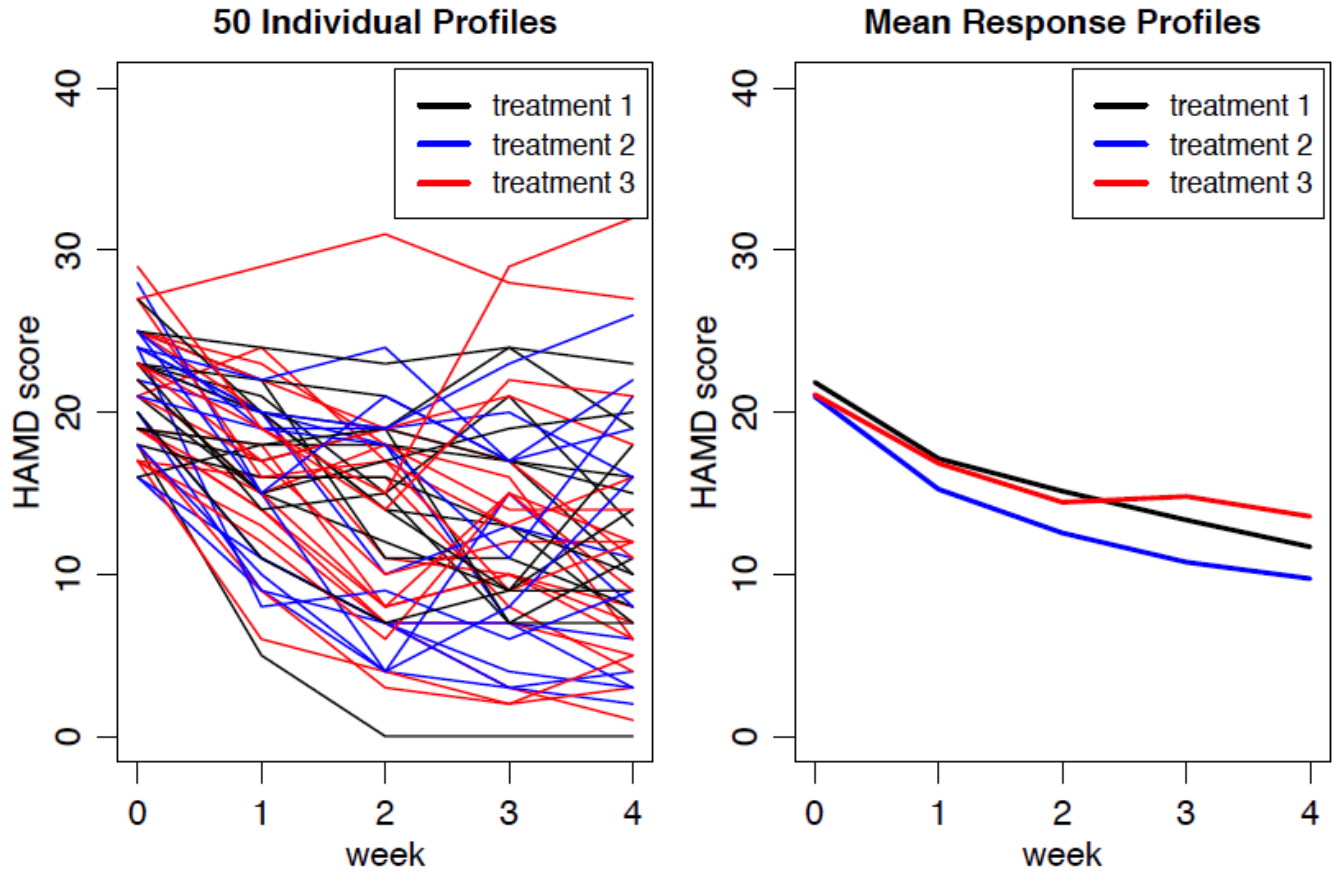
纵向数据

- 个体被依时间顺序重复观测所得到的数据, 同一个体的观测之间是相关的
- 纵向数据有不同的形式
  - 连续或离散的响应变量
  - 等间距或不等间距观测
  - 每个个体观测时间点相同或不同
  - 有或没有缺失数据
  - 许多或很少时间点, T
  - 许多或很少个体, n
- 我们仅关注随机效应线性模型和自回归模型

HAMD: 抗抑郁症试验

- 6 个临床试验中心, 比较 3 种抑郁症疗法
- 367 个个体随机分成 3 组
- 使用 Hamilton depression score(HAMD) 评分对每个个体治疗效果进行评分, 每个个体每周来访一次, 共 5 次
  - 第 0 周是参加临床试验前
  - 第 1-4 周是治疗期间
  - HAMD 得分为 0-50 分, 分值越高抑郁症越严重
- 有些个体第 2 周起退出试验 (dropout), 我们忽略掉退出的个体, 分析其他 246 位完成试验的个体数据 (Diggle and Kenward, 1994)

HAMD 例子: 数据


50 Individual Profiles
Mean Response Profiles
treatment 1
treatment 2
treatment 3
HAMD score
week
0
10
20
30
40
0
1
2
3
4

HAMD 试验目标

- 三种治疗方式中, HAMD 得分随时间变化有差异吗?
- 变量:
  - y: HAMD 得分
  - t: 处理 (treatment)
  - w: 星期
- 为简单起见, 我们忽略试验中心的差异, 假设线性关系变化. 考虑三种模型
  - 标准线性模型 (非层次模型)
  - 随机效应模型 (层次模型)
  - 自回归模型 (AR1)

贝叶斯线性模型 (非层次模型)(LM)

假设

$$y_{iw} \sim N\left(\mu_{iw}, \sigma^2\right)$$

以及线性关系

$$\mu_{iw} = \alpha + \beta_{treat(i)} W$$

- $treat(i)$: 第 i 个个体接收的处理 (疗法) 标签变量, 因此取值 1,2,3
- w: 来访的星期, 取值 0,1-4.

此模型下重复观测结构被忽略了. 先验分布取为

$$\begin{aligned} \alpha, \beta_1, \beta_2, \beta_3 &\sim N(0, 10000) \\ \frac{1}{\sigma^2} &\sim Gamma(0.001, 0.001) \end{aligned}$$

层次模型 假设

$$y_{iw} \sim N\left(\mu_{iw}, \sigma^2\right)$$

以及线性关系

$$\mu_{iw} = \alpha_i + \beta_{treat(i)} W$$

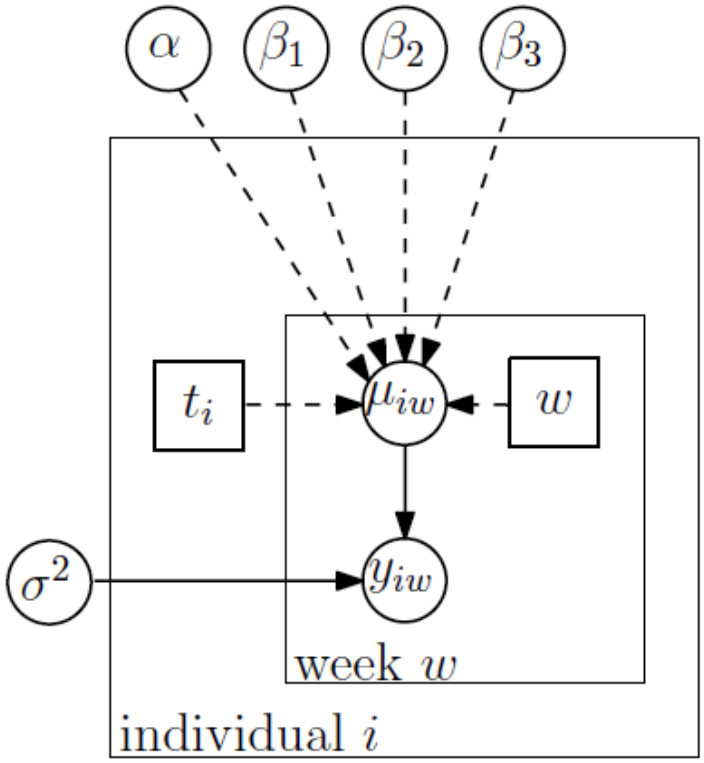
- 假设给定 α_i 时, $\{y_{iw}, w = 0, 1, \ldots, 4\}$ 相互独立
- 假设 $\alpha_i \sim N\left(\mu_\alpha, \sigma_\alpha^2\right)$ $i = 1, \ldots, 246$; 假设可交换性成立
- 假设先验分布

$$\mu_\alpha \sim N(0, 10000) \qquad \sigma_\alpha \sim Uniform(0, 100)$$

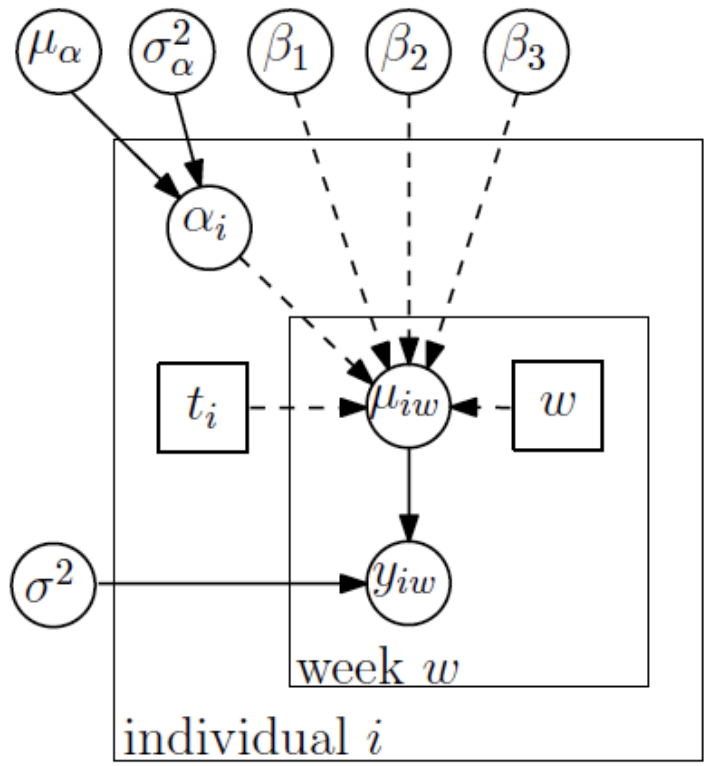
这是层次线性模型或线性混合效应模型 (LMM) 或随机系数模型的例子

HAMD 例子: LM 和 LMM 的图模型表示

non-hierarchical model (LM)

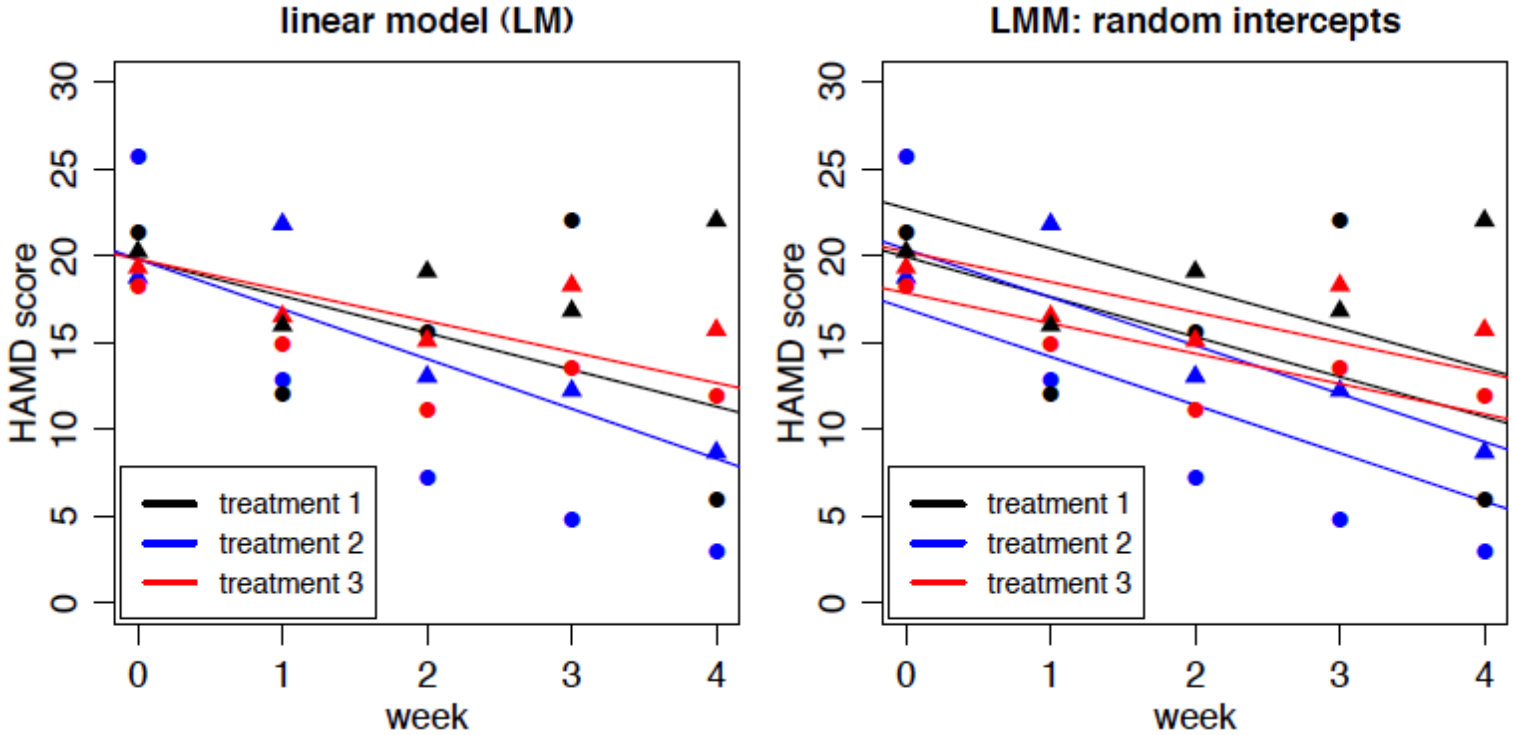


hierarchical model (LMM)



t_i represents the treatment indicator of individual i

HAMD 例子: LM 和 LMM 拟合直线



圆点和三角形点表示 6 个个体的得分 (每个处理 2 个)

- LM: 拟合 3 条回归线, 每个处理 1 条; 截距相同, 但斜率不同
- LMM: 每个个体有不同的回归直线; 每种处理中, 个体直线斜率相同

HAMD 例子: LM 和 LMM 拟合结果

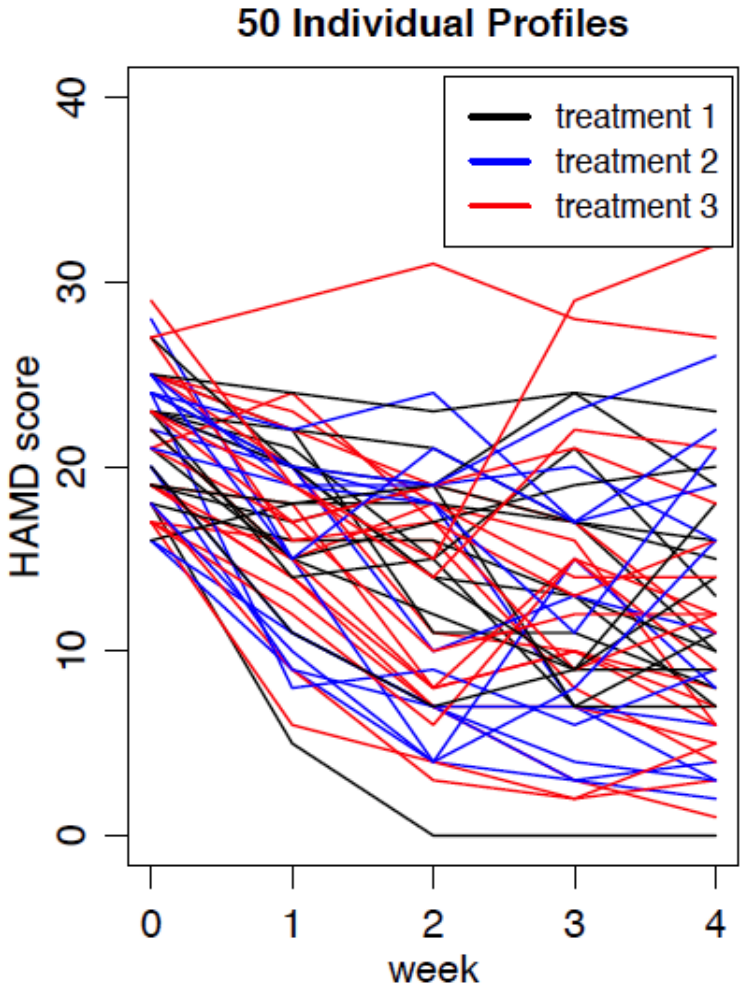
Table: posterior mean (95% credible interval) for the non-hierarchical and hierarchical models fitted to the HAMD data

non-hierarchical model			hierarchical model		
α	19.79	(19.20,20.35)	μ_α	19.81	(19.14,20.47)
			σ^2_α	17.62	(14.03,21.92)
β_1	-2.12	(-2.42,-1.79)	β_1	-2.30	(-2.58,-2.02)
β_2	-2.87	(-3.18,-2.56)	β_2	-2.77	(-3.03,-2.50)
β_3	-1.78	(-2.07,-1.50)	β_3	-1.74	(-1.99,-1.48)
σ^2	35.41	(32.64,38.48)	σ^2	18.17	(16.62,19.85)

注意

- 层次模型斜率的变化性
- 残差方差 (σ^2) 在引入随机效应后减低了

HAMD 例子: 数据再观察



从原始数据的图形上可以看出

- 不同截距是合适的
- 同时建议不同斜率

因此我们在层次模型中加入随机斜率

层次模型: 增加随机斜率 假设

$$y_{iw} \sim N\left(\mu_{iw}, \sigma^2\right)$$

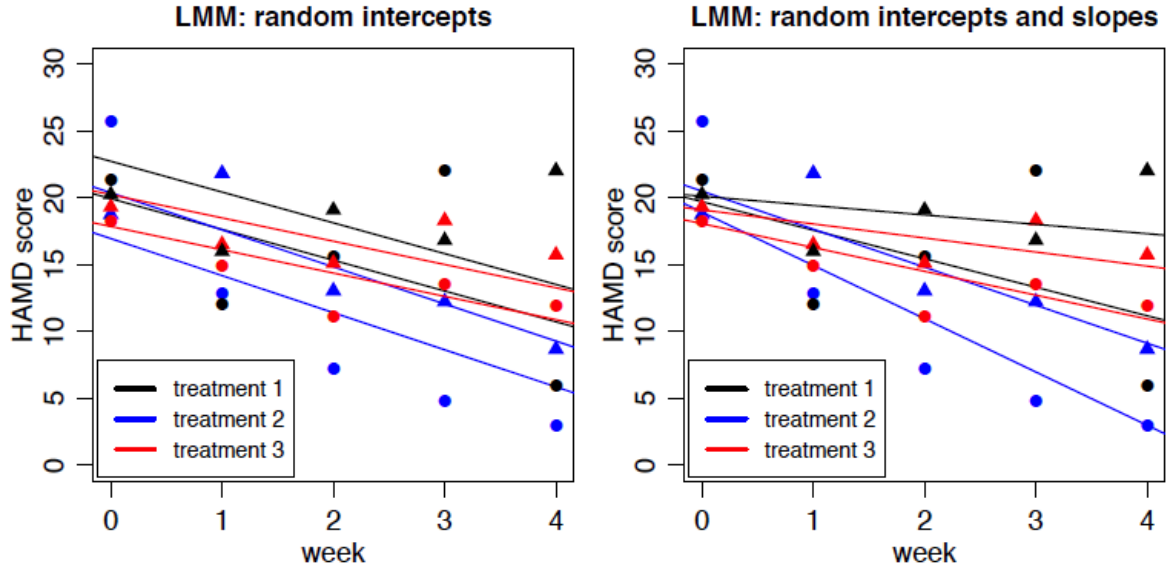
以及线性关系

$$\mu_{iw} = \alpha_i + \beta_{treat(i),i} W$$

- 假设给定 $\alpha_i, \beta_{treat(i),i}$ 时, $\{y_{iw}, w = 0, 1, \ldots, 4\}$ 相互独立
- 假设 $\alpha_i \sim N\left(\mu_\alpha, \sigma_\alpha^2\right) \quad i = 1, \ldots, 246$; $\beta_{1,i}, \beta_{2,i}, \beta_{3,i}$ 服从相同分布; 假设可交换性成立
- 假设先验分布

$$\mu_\alpha \sim N(0, 10000) \qquad \beta_{l,i} \sim N(\beta_\mu, \beta_\tau) \quad \sigma_\alpha \sim Uniform(0, 100)$$

HAMD 例子: LM 和 LMM 拟合直线



圆点和三角形点表示 6 个个体的得分 (每个处理 2 个)

- LMM: 只有随机截距时每个个体回归直线不同; 但同一处理中, 只有截距不同
- LMM: 随机截距和随机斜率: 每个个体的截距和斜率都不同, 更好的拟合每个个体

HAMD 例子: 估计结果比较

Table: posterior mean (95% credible interval) for the non-hierarchical and hierarchical models fitted to the HAMD data

linear model			hierarchical model 1⋆			hierarchical model 2†		
α	19.79	(19.20,20.35)	μ_α	19.81	(19.14,20.47)	μ_α	19.81	(19.23,20.39)
			σ^2_α	17.62	(14.03,21.92)	σ^2_α	11.09	(8.38,14.35)
β_1	-2.12	(-2.42,-1.79)	β_1	-2.30	(-2.58,-2.02)	μ_{β_1}	-2.29	(-2.70,-1.90)
						$\sigma^2_{\beta_1}$	1.96	(1.15,3.02)
β_2	-2.87	(-3.18,-2.56)	β_2	-2.77	(-3.03,-2.50)	μ_{β_2}	-2.79	(-3.15,-2.45)
						$\sigma^2_{\beta_2}$	1.18	(0.53,2.02)
β_3	-1.78	(-2.07,-1.50)	β_3	-1.74	(-1.99,-1.48)	μ_{β_3}	-1.73	(-2.11,-1.38)
						$\sigma^2_{\beta_3}$	1.91	(1.11,2.93)
σ^2	35.41	(32.64,38.48)	σ^2	18.17	(16.62,19.85)	σ^2	14.39	(13.05,15.92)

⋆ random intercepts only
† random intercepts and random slopes

自回归模型-没有协变量

- 记 $\mathbf{y} = (y_1, \ldots, y_T)$ 为依时间观测
- 自回归高斯模型

$$y_t | y_{t-1}, y_{t-2}, \ldots y_{t-p} \sim \mathrm{N}\left(\sum_{j=1}^{p} \gamma_j y_{t-j}, \sigma_\epsilon^2\right), \quad t = p{+}1, \ldots, T$$

- 简单情形: AR(1) 模型

$$y_t = \gamma_1 y_{t-1} + \epsilon_t; \quad \epsilon_t \sim N\left(0, \sigma_\epsilon^2\right)$$

或者等价地,

$$= y_t \sim N\left(\gamma_1 y_{t-1}, \sigma_\epsilon^2\right)$$

自回归模型-有协变量

- 协变量可以通过假设残差是顺次相依的来引入
- 例如, 对 AR(1) 模型, 一个解释变量, x, $t = 1, \ldots, n$
  - 令 $\mu_t = \beta x_t$
  - 然后 $y_t|y_{t-1} = \mu_t + \gamma\left(y_{t-1} - \mu_{t-1}\right) + \epsilon_t; \quad \epsilon_t \sim \mathrm{N}\left(0, \sigma^2\right)$ 可以表示为

$$
\begin{aligned}
y_t|y_{t-1} &= \mu_t + \mathcal{R}_t \\
\mathcal{R}_1 &= \epsilon_1 \\
\mathcal{R}_t &= \gamma\mathcal{R}_{t-1} + \epsilon_t \quad t > 1 \\
\epsilon_t &\sim \mathrm{Normal}\left(0, \sigma^2\right)
\end{aligned}
$$

HAMD 例子: 自回归模型

- 指定

$$y_{iw} = \mu_{iw} + \mathcal{R}_{iw}$$
$$\mu_{iw} = \alpha_i + \beta_{treat(i)} w$$

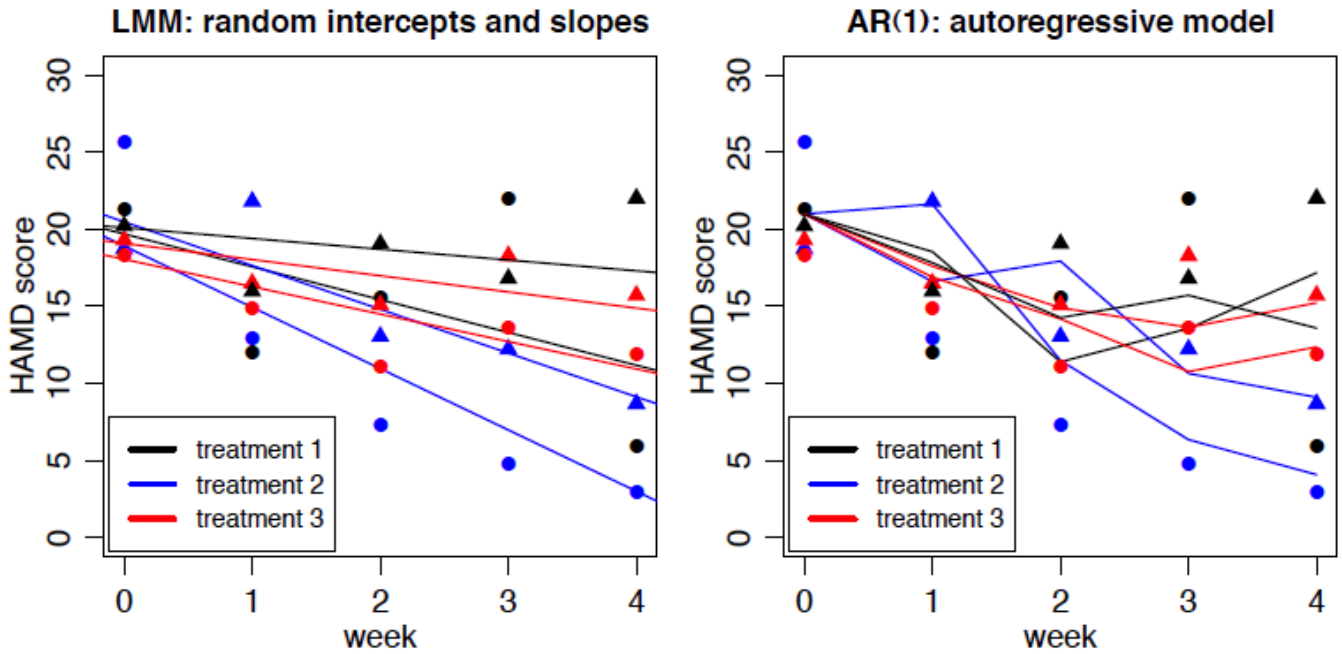
其中

$$\mathcal{R}_{i0} = \epsilon_{i0}$$
$$\mathcal{R}_{iw} = \gamma \mathcal{R}_{i(w-1)} + \epsilon_{iw} \quad w \geq 1$$
$$\epsilon_{iw} \sim \text{Normal}\left(0, \sigma^2\right) \quad w = 0, \ldots, 4$$

- 在 WINBUGS 里等价表示为

$$y_{iw} \sim \text{Normal}\left(\theta_{iw}, \sigma^2\right)$$
$$\theta_{i0} = \mu_{i0}$$
$$\theta_{iw} = \mu_{iw} + \gamma\left(y_{i(w-1)} - \mu_{i(w-1)}\right) \quad w \geq 1$$
$$\mu_{iw} = \alpha_i + \beta_{treat(i)} w$$

HAMD 例子: LMM 和 AR(1) 拟合直线



圆点和三角形点表示 6 个个体的得分 (每个处理 2 个)

- LMM: 随机截距和随机斜率: 每个个体的截距和斜率都不同, 更好的拟合每个个体
- AR(1): 所有线从一个共同截距出发; 路径根据前面时间点变化

HAMD 例子: 结果比较

Table: posterior mean (95% credible interval) for the parameter estimates from the linear, random effects and autoregressive models fitted to the HAMD data

linear model			random effects model			autoregressive model		
α	19.79	(19.20,20.35)	μ_α	19.81	(19.23,20.39)	α	20.99	(20.39,21.54)
			σ^2_α	11.09	(8.38,14.35)			
β_1	-2.12	(-2.42,-1.79)	μ_{β_1}	-2.29	(-2.70,-1.90)	β_1	-2.46	(-2.83,-2.07)
			$\sigma^2_{\beta_1}$	1.96	(1.15,3.02)			
β_2	-2.87	(-3.18,-2.56)	μ_{β_2}	-2.79	(-3.15,-2.45)	β_2	-2.95	(-3.32,-2.55)
			$\sigma^2_{\beta_2}$	1.18	(0.53,2.02)			
β_3	-1.78	(-2.07,-1.50)	μ_{β_3}	-1.73	(-2.11,-1.38)	β_3	-1.96	(-2.30,-1.62)
			$\sigma^2_{\beta_3}$	1.91	(1.11,2.93)			
						γ	0.72	(0.66,0.77)
σ^2	35.41	(32.64,38.48)	σ^2	14.39	(13.05,15.92)	σ^2	22.53	(20.82,24.39)

HAMD 例子: 结果解释

- 不同处理下 HAMD 得分随时间变化差异体现在 $\beta_1-\beta_2, \beta_1-\beta_3$ 和 $\beta_2-\beta_3$, 或者在随机斜率模型中的 $\mu_{\beta_1}-\mu_{\beta_2}, \mu_{\beta_1}-\mu_{\beta_3}$ 和 $\mu_{\beta_2}-\mu_{\beta_3}$
- 在 WINBUS 中需要监视这些对照

```
    # Calculate contrasts
contrasts[1]<-beta[1]-beta[2]
contrasts[2]<-beta[1]-beta[3]
contrasts[3]<-beta[2]-beta[3]
```

HAMD 例子: 对照

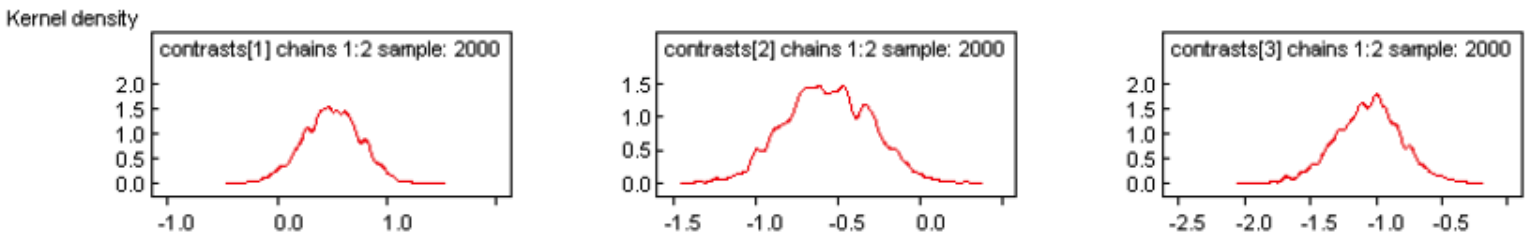
Table: posterior mean (95% credible interval) for the contrasts (treatment comparisons) from models fitted to the HAMD data

treatments	linear model		hierarchical 1*		hierarchical 2†		AR(1)	
1 v 2	0.8	(0.4,1.1)	0.5	(0.1,0.8)	0.5	(0.0,1.0)	0.5	(0.0,1.0)
1 v 3	-0.3	(-0.7,0.0)	-0.6	(-0.9,-0.2)	-0.6	(-1.1,0.0)	-0.5	(-1.0,0.0)
2 v 3	-1.1	(-1.4,-0.8)	-1.0	(-1.4,-0.7)	-1.1	(-1.6,-0.6)	-1.0	(-1.5,-0.5)

* random intercepts only

† random intercepts and random slopes

Density plots for hierarchical 2



HAMD 例子: 模型比较

Table: DIC for the linear, random effects and autoregressive models fitted to the HAMD data

	Dbar	Dhat	pD	DIC
linear	7877	7872	5	7882
random intercepts	7056	6849	207	7263
random intercepts & slopes	6768	6454	314	7082
autoregressive	7320	7314	6	7326

- 考虑结构的模型都比线性模型拟合的好
- 随机效应模型比 AR(1) 模型好

HAMD 例子: 模型扩展

- 允许非线性性: 引入二次项

$$\mu_{iw} = \alpha_i + \beta_{\text{treat}(i)} W + \delta_{\text{treat}(i)} w^2$$

- 包含中心的效应, 允许每个中心截距不同
  - 作为固定效应

$$\mu_{iw} = \alpha_{\text{centre}(i)} + \beta_{\text{treat}(i)} W$$

  - 或者随机效应

$$\mu_{iw} = \alpha_i + \beta_{\text{treat}(i)} W$$
$$\alpha_i \sim \text{Normal}\left(\mu_{\alpha_{\text{cente}(i)}}, \sigma^2_{\alpha_{\text{cente}(i)}}\right)$$

- 允许更复杂的协方差结构, 比如 AR(2) 模型

$$\mathcal{R}_{i0} = \epsilon_{i0}$$

$$\mathcal{R}_{i1} = \gamma_1 \mathcal{R}_{i0} + \epsilon_{i1}$$

$$\mathcal{R}_{iw} = \gamma_1 \mathcal{R}_{i(w-1)} + \gamma_2 \mathcal{R}_{i(w-2)} + \epsilon_{iw} \quad w \geq 2$$

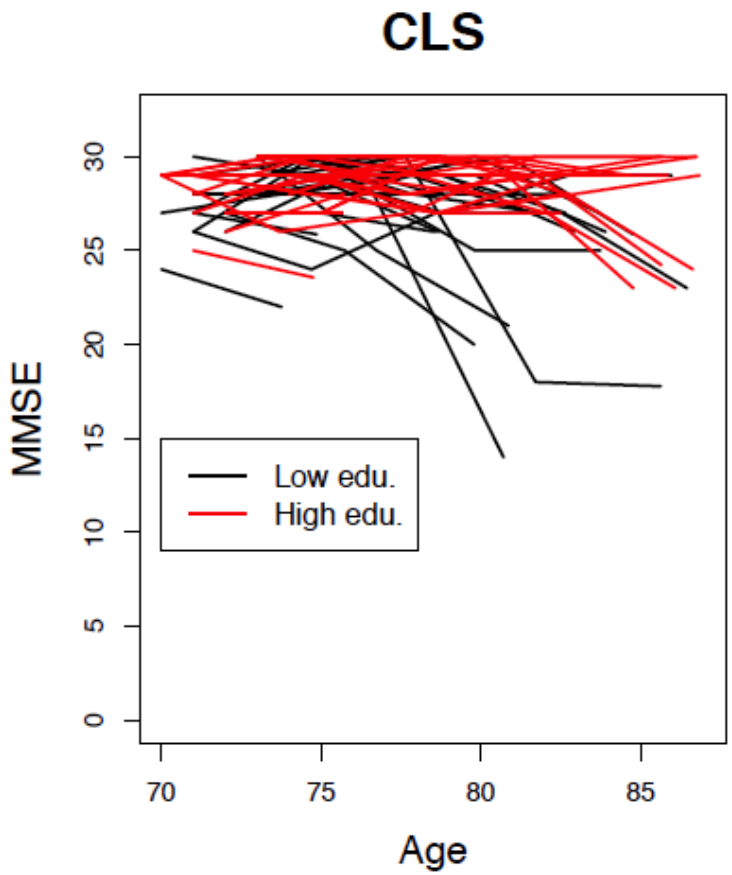
$$\epsilon_{iw} \sim N\left(0, \sigma^2\right)$$

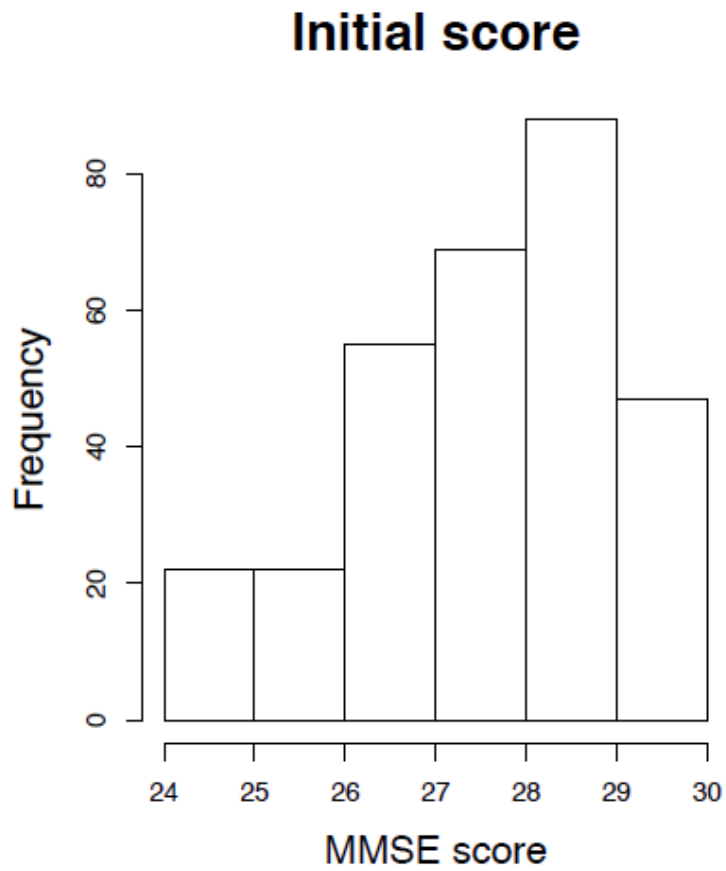
衰老与认知能力研究

- 两个研究目的: 评估年龄增长在认知减退上的效应; 评估教育水平差异如何影响认知随时间变化率的 (称为认知保留)
- 认知保留是指思考对大脑神经病理损害的恢复力。在衰老研究中, 心理学家假设受过高等教育的人往往具有更好的认知能力, 并且减慢了大脑随着年龄的增长而损伤的速度
- 我们使用 Canberra Longitudinal Study (CLS, Australia) 数据, 包含了 586 个个体在度量认知功能的 Mini-Mental State Examination (MMSE) 上的得分值
  - 试验自 1991 年开始, 参加试验人群年龄在 70 到 93 之间 (进入试验时)
  - 跟踪观测了 14 年, 进行了 4 次测试 MMSE 分数, 取值 0-30 分, 值越小认知能力越差

– 其他变量: 测试时年龄, 性别 (这里只考虑男性), 教育 (0= 低; 1= 高)

– 只考虑第一次 MMSE 得分 ≥ 24 分以上的 (避免痴呆)





衰老与认知能力研究例子: 线性隐增长曲线模型

- 为了评估衰老对认知能力的影响, 我们假设 MMSE 得分 y_{ij} 满足

$$y_{ij} \sim N\left(\mu_{ij}, \sigma^2\right)$$
$$\mu_{ij} = \alpha_i + \beta_i \cdot \text{age}^*_{ij}$$

  其中 i 表示个体, j 表示第 j 次 MMSE 测试. age^*_{ij} 表示 $age_{ij} - 75$, β_i 表示个体 i 的认知功能年变化率; $sigma^2$ 为残差方差

- α_i 和 β_i 为随机效应, 假设独立先验

$$\alpha_i \sim N\left(\mu_\alpha, \sigma^2_\alpha\right)$$
$$\beta_i \sim N\left(\mu_\beta, \sigma^2_\beta\right)$$
$$\sigma_\alpha \sim Uniform(0, 30)$$
$$\sigma_\beta \sim Uniform(0, 10)$$

- 或者使用联合先验允许可能的相关性

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \Sigma\right)$$

- Σ 的先验取为

$$\Sigma^{-1} \sim \text{ Wishart } (\mathbf{R}, k)$$

其中 R 为 $p \times p$ 正定矩阵, $k > p-1$ 为自由度. 我们取 $R = diag(20, 2)$.

- $\mu_\alpha, \mu_\beta \sim N(0, 10000)$.

- 一种检验教育水平效应的自然方法就是

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{\alpha,i} \\ \mu_{\beta,i} \end{pmatrix}, \Sigma\right)$$

$$\mu_{\alpha,i} = \eta_0 + \eta_1 \cdot edu_i$$

$$\mu_{\beta,i} = \gamma_0 + \gamma_1 \cdot edu_i$$

η_0 为 75 岁低教育水平个体的平均 MMSE 得分; γ_0 为低教育个体认知能力平均变化率; η_1 为 75 岁高教育水平个体的平均 MMSE 得分; γ_1 为高教育个体认知能力平均变化率;

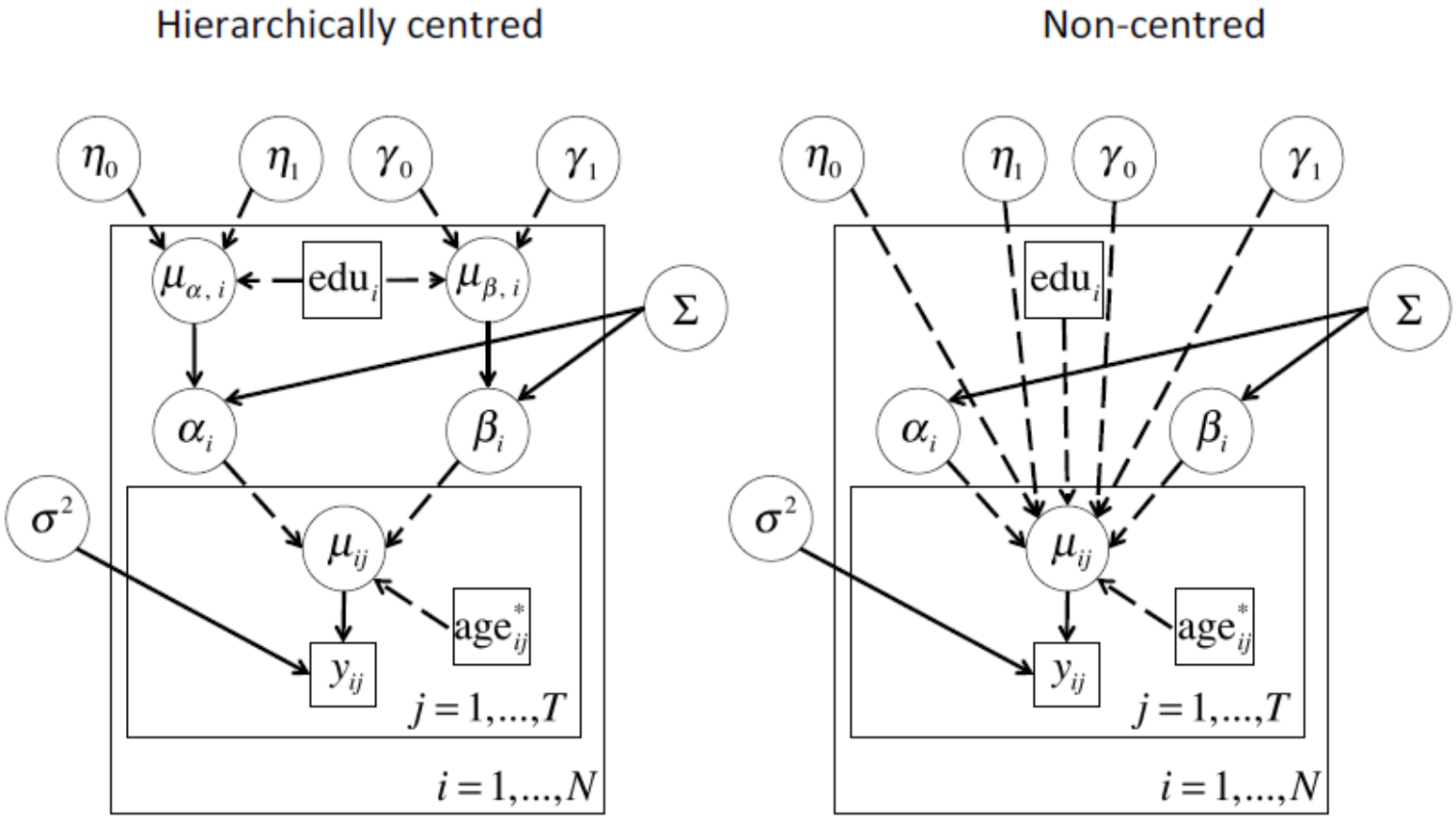
- 另一种建模方法是将 edu 作为一个解释变量直接放入第一层回归模型中

$$\begin{aligned}\mu_{ij} = &(\eta_0 + \alpha_i) + (\gamma_0 + \beta_i) \cdot age^*_{ij} \\ &+ \eta_1 \cdot edu_i + \gamma_1 \cdot age^*_{ij} \cdot edu_i\end{aligned}$$

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma\right).$$

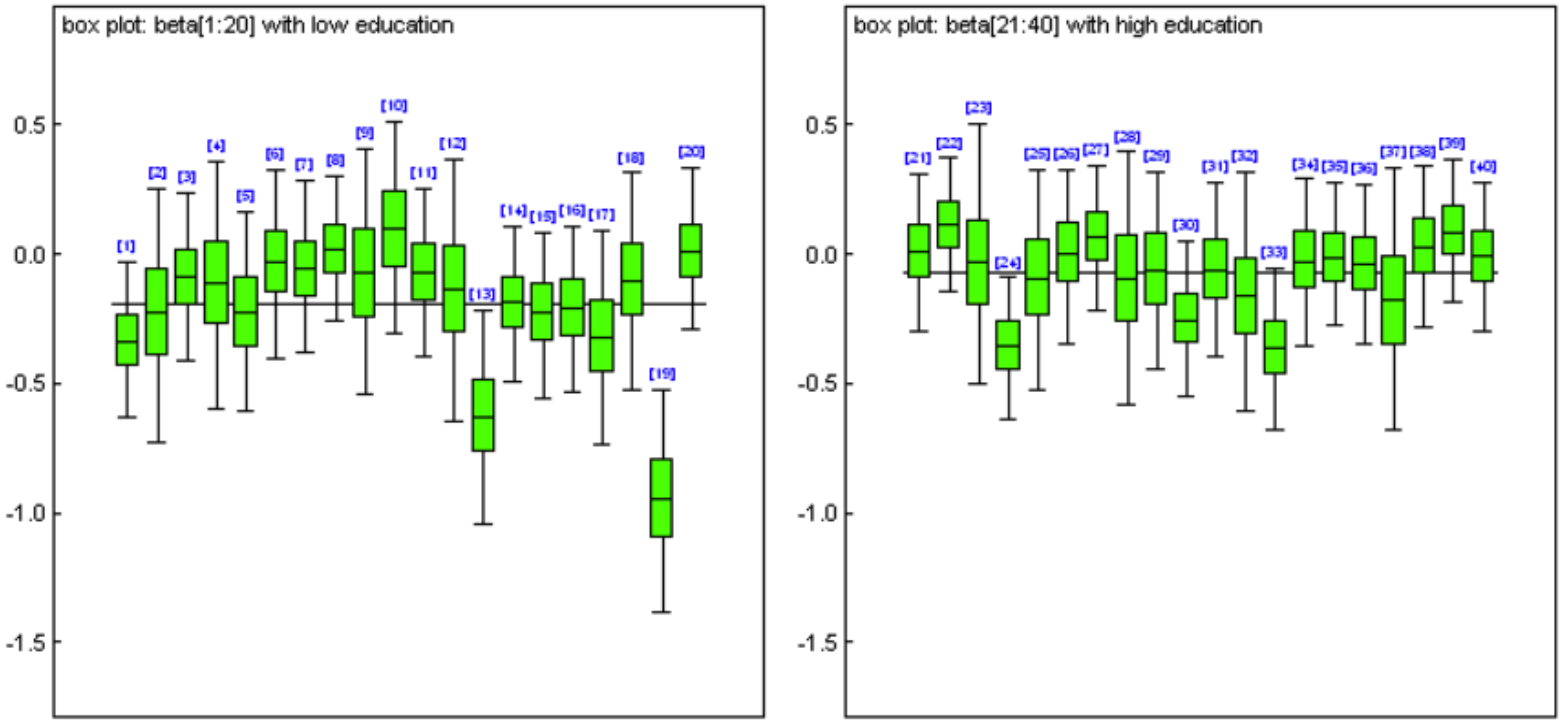
教育的效应通过 $age \times edu$ 的交互项系数 γ_1 来表示; 这种方法是前面模型的" 非中心" 表示形式

DAGs:


Hierarchically centred
Non-centred
η_0
η_1
γ_0
γ_1
$\mu_{\alpha,i}$
edu$_i$
$\mu_{\beta,i}$
Σ
α_i
β_i
σ^2
μ_{ij}
age$^*_{ij}$
y_{ij}
$j = 1,...,T$
$i = 1,...,N$

衰老与认知能力研究例子: 结果

Figure: A box plot of the subject-specific rates of change (β_i) from model without education



衰老与认知能力研究例子: 结果

Table: posterior mean (95% credible intervals) for the parameter estimates from models with/without education

without edu.			with edu.
μ_α	28.24 (28.03, 28.43)	η_0	28.21 (27.96, 28.49)
		$\eta_0 + \eta_1$	28.49 (28.11, 28.87)
		η_1	0.33 (-0.19, 0.81)
μ_β	-0.17 (-0.21, -0.13)	γ_0	-0.19 (-0.24, -0.14)
		$\gamma_0 + \gamma_1$	-0.10 (-0.18, -0.01)
		γ_1	0.09 (-0.01, 0.18)

注意

- MMSE 能力显著的随时间减低
- 教育低的男性显示认知能力减低的比教育高的男性更快, 但是, 差异不是统计显著的

衰老与认知能力研究例子: 结果

Table: posterior mean (95% credible intervals) for the parameter estimates from models with/without education and model comparisons

	without edu.	with edu.
σ^2_α	1.12 (0.73, 1.63)	1.12 (0.74, 1.61)
σ^2_β	0.08 (0.06, 0.10)	0.08 (0.06, 0.10)
cor(α_i, β_i)	-0.12 (-0.35, 0.11)	-0.15 (-0.37, 0.08)
σ^2	3.60 (3.15, 4.03)	3.59 (3.15, 4.03)
$\bar{D}$	3656	3652
pD	283	283
DIC	3939	3935

衰老与认知能力研究例子: 右删失效应

- 因为 MMSE 是为筛查痴呆而设计的, 并不是专门用于认知能力测试. 因此健康人应该容易取得满分, 得分 30 分至少表明在特定时间点该人至少达到了满分—右删失了
- 我们进行如下建模

$$y_{ij}^{*} \sim N\left(\mu_{ij}, \sigma^{2}\right) I\left(lower_{ij}, +\infty\right)$$

其中

$$\text{lower}_{ij} = \begin{cases} 30 & \text{if } y_{ij} = 30 \rightarrow y_{ij}^{*} = NA \\ 0 & \text{if } y_{ij} < 30 \rightarrow y_{ij}^{*} = y_{ij} \end{cases}$$

衰老与认知能力研究例子: 结果

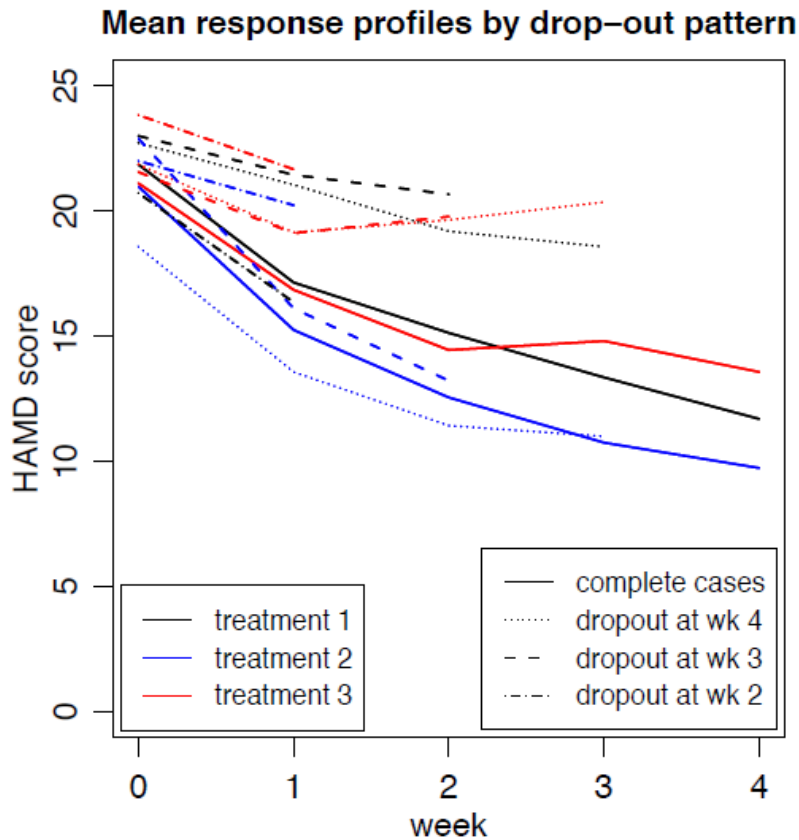
Table: posterior mean (95% credible intervals) for the parameter estimates from models with/without ceiling effects

	Without ceiling effect	With ceiling effect
η_0	28.21 (27.96, 28.49)	28.47 (28.18, 28.78)
$\eta_0 + \eta_1$	28.49 (28.11, 28.87)	28.89 (28.36, 29.42)
η_1	0.33 (-0.19, 0.81)	0.41 (-0.21, 1.01)
γ_0	-0.19 (-0.24, -0.14)	-0.21 (-0.26, -0.15)
$\gamma_0 + \gamma_1$	-0.10 (-0.18, -0.01)	-0.10 (-0.20, -0.01)
γ_1	0.09 (-0.01, 0.18)	0.11 (-0.01, 0.22)
σ^2_α	1.12 (0.74, 1.61)	1.62 (1.02, 2.42)
σ^2_β	0.08 (0.06, 0.10)	0.08 (0.06, 0.11)
$\text{cor}(\alpha_i, \beta_i)$	-0.12 (-0.35, 0.11)	-0.18 (-0.42, 0.08)
σ^2	3.59 (3.15, 4.03)	4.45 (3.84, 5.09)

- 结论不变
- 不确定性轻微增加, 因为允许删失造成信息损失

1.4 缺失数据

- 贝叶斯层次模型很容易处理包含部分观测数据，集成关于缺失原因的假设模型



- 个体在第 1 和 3 组一般有比较高的退出率
- 但是退出和完全数据在第 2 组类似

衰老与认知能力研究例子: 对缺失性建模

- 完整数据包括
  - y: HAMD 得分 (观测的和缺失值)
  - t: 处理
  - 缺失指标

$$m_{iw} = \begin{cases} 0: & y_{iw} \text{ observed} \\ 1: & y_{iw} \text{ missing} \end{cases}$$

- 因为我们
  - y(随机效应模型)
  - y 缺失的概率

$$m_{iw} \sim Bernoulli\left(p_{iw}\right)$$

缺失数据类型

- 完全随机缺失 (MCAR): 缺失不依赖观测或未观测数据, e.g.

$$\log it\left(p_{iw}\right)=\theta_0$$

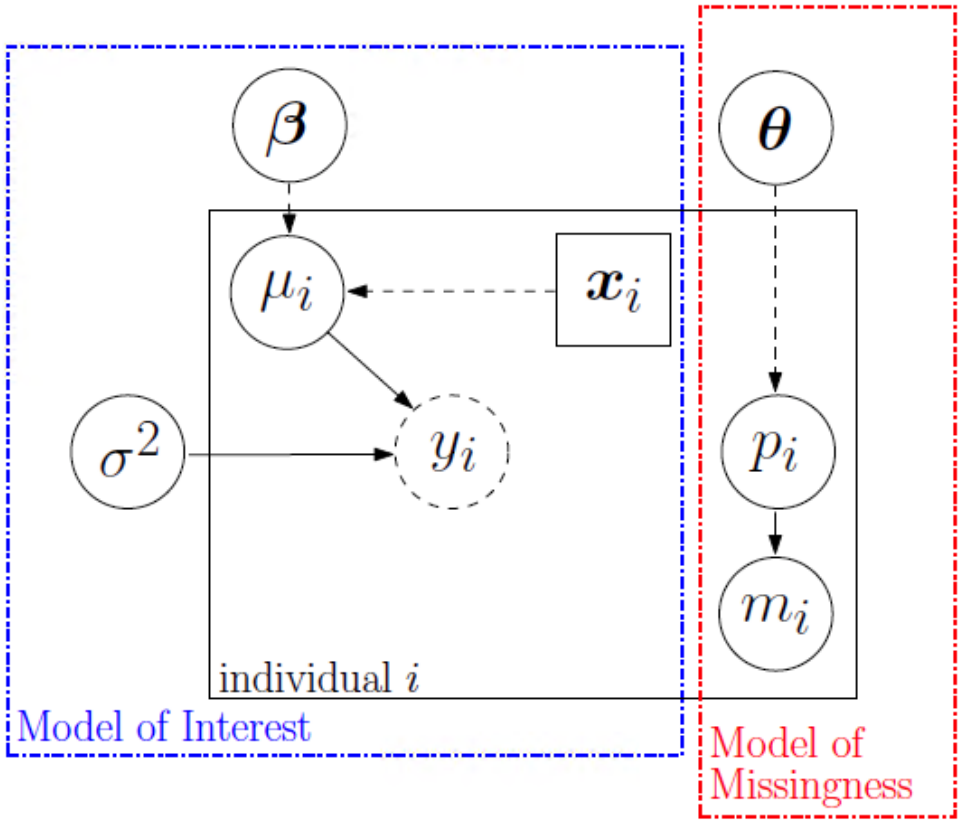
- 随机缺失 (MAR): 缺失依赖于观测数据, e.g.

$$logit\left(p_{iw}\right)=\theta_0+\theta_1 t_i \qquad \text{或者} logit\left(p_{iw}\right)=\theta_0+\theta_2 y_{i0}$$

- 非随机缺失 (MNAR): MCAR 或 MAR 都不成立. e.g.

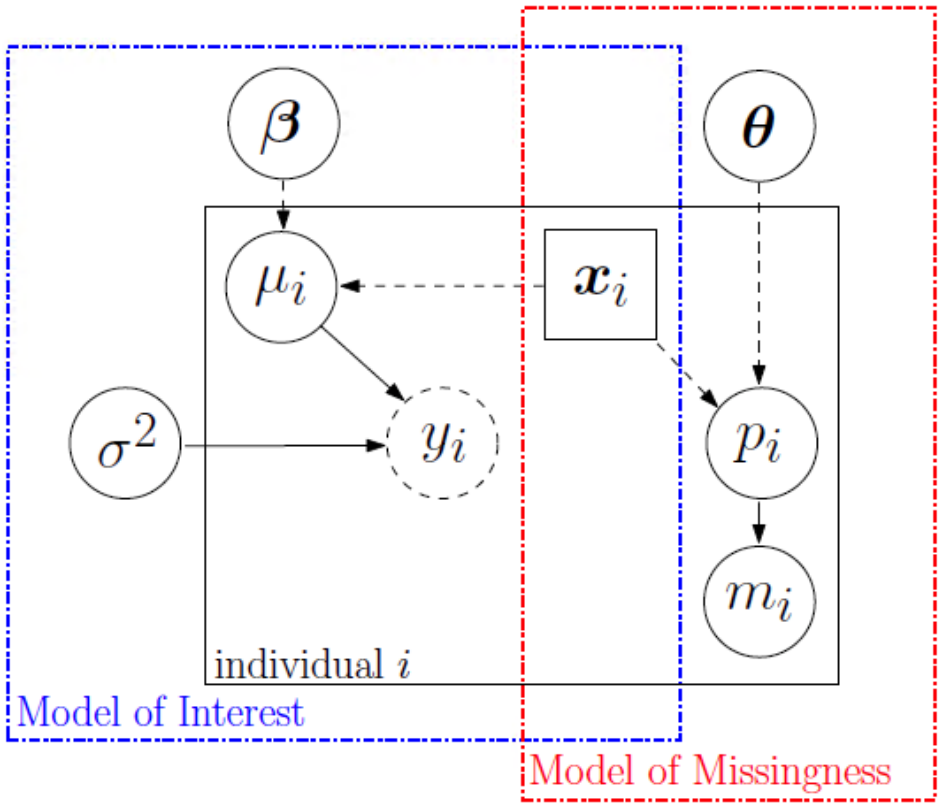
$$logit\left(p_{iw}\right)=\theta_0+\theta_3 y_{iw}$$

DAG: MCAR 缺失



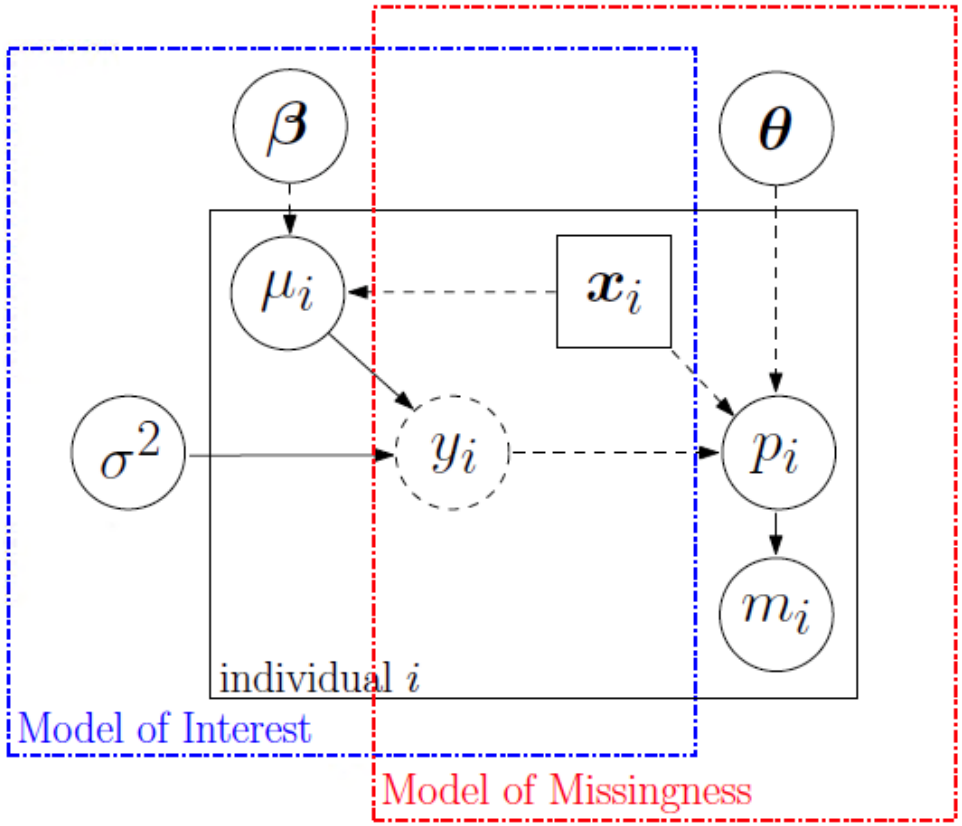
注意:**x** 完全观测, 但是 y 有缺失值

DAG: MAR 缺失



注意:**x** 完全观测, 但是 y 有缺失值

DAG: MNAR 缺失



注意:**x** 完全观测, 但是 y 有缺失值

联合建模

- 令 $\mathbf{z} = (z_{ij})$ 表示数据阵, $i = 1, \ldots, n$ 个体; $j = 1, \ldots, k$ 变量
- 将 $\mathbf{z}$ 划分为观测和缺失部分, $\mathbf{z} = \left(\mathbf{z}^{obs}, \mathbf{z}^{mis}\right)$
- 令 $\boldsymbol{m} = (m_{ij})$ 为缺失变量

$$m_{ij} = \begin{cases} 0: & z_{ij} \text{ observed} \\ 1: & z_{ij} \text{ missing} \end{cases}$$

- 令 β 和 θ 表示未知参数
- 完整数据的联合模型为

$$f(\mathbf{z}, \boldsymbol{m}|\beta, \theta) = f\left(\mathbf{z}^{obs}, \mathbf{z}^{mis}, \boldsymbol{m}|\beta, \theta\right)$$

- 观测数据似然为

$$f\left(\mathbf{z}^{obs}, \boldsymbol{m}|\beta, \theta\right) = \int f\left(\mathbf{z}^{obs}, \mathbf{z}^{mis}, \boldsymbol{m}|\beta, \theta\right) d\mathbf{z}^{mis}$$

- 联合似然可以分解为

$$f\left(\mathbf{z}^{obs}, \mathbf{z}^{mis}, \boldsymbol{m}|\boldsymbol{\beta}, \boldsymbol{\theta}\right) = f\left(\boldsymbol{m}|\mathbf{z}^{obs}, \mathbf{z}^{mis}, \boldsymbol{\beta}, \boldsymbol{\theta}\right) f\left(\mathbf{z}^{obs}, \mathbf{z}^{mis}|\boldsymbol{\beta}, \boldsymbol{\theta}\right)$$

如果假设条件独立性, 则可以进一步简化

$$f\left(\mathbf{z}^{obs}, \mathbf{z}^{mis}, \boldsymbol{m}|\boldsymbol{\beta}, \boldsymbol{\theta}\right) = f\left(\boldsymbol{m}|\mathbf{z}^{obs}, \mathbf{z}^{mis}, \boldsymbol{\theta}\right) f\left(\mathbf{z}^{obs}, \mathbf{z}^{mis}|\boldsymbol{\beta}\right)$$

这种分解方法称为选择模型

- 注意观测似然

$$\begin{aligned} f\left(\mathbf{z}^{\text{obs}}, \boldsymbol{m}|\boldsymbol{\beta}, \boldsymbol{\theta}\right) &= \int f\left(\mathbf{z}^{\text{obs}}, \mathbf{z}^{\text{mis}}, \boldsymbol{m}|\boldsymbol{\beta}, \boldsymbol{\theta}\right) d\mathbf{z}^{\text{mis}} \\ &= \int f\left(\boldsymbol{m}|\mathbf{z}^{\text{obs}}, \mathbf{z}^{\text{mis}}, \boldsymbol{\theta}\right) f\left(\mathbf{z}^{\text{obs}}, \mathbf{z}^{\text{mis}}|\boldsymbol{\beta}\right) d\mathbf{z}^{\text{mis}} \end{aligned}$$

- MAR 机制下

$$f\left(\boldsymbol{m}|\mathbf{z}^{\text{obs}},\mathbf{z}^{\text{mis}},\boldsymbol{\theta}\right)=f\left(\boldsymbol{m}|\mathbf{z}^{\text{obs}},\boldsymbol{\theta}\right)$$

因此

$$\begin{aligned}f\left(z^{obs},m|\beta,\theta\right)&=f\left(m|z^{obs},\theta\right)\int f\left(z^{obs},z^{mis}|\beta\right)dz^{mis}\\&=f\left(m|z^{obs},\theta\right)f\left(z^{obs}|\beta\right)\end{aligned}$$

- MCAR 机制下

$$f\left(\boldsymbol{m}|\mathbf{z}^{obs},\mathbf{z}^{mis},\boldsymbol{\theta}\right)=f(\boldsymbol{m}|\boldsymbol{\theta})$$

因此

$$f\left(\mathbf{z}^{obs},\boldsymbol{m}|\boldsymbol{\beta},\boldsymbol{\theta}\right)=f(\boldsymbol{m}|\boldsymbol{\theta})f\left(\mathbf{z}^{obs}|\boldsymbol{\beta}\right)$$

- 缺失数据机制称为**可忽略的**, 如果
  - 缺失数据是 MCAR 或 MAR
  - 参数 β 和 θ 是不同的

  在贝叶斯模型中, 往往还假设参数 β 和 θ 是独立的
- ” 可忽略” 表示我们可以忽略缺失机制模型, 但不表示我们可以忽略缺失数据. 如果数据机制是**不可忽略的**, 那么就不能忽略缺失机制模型
- 相比于抽样过程 (常常已知), 缺失机制往往未知
- 数据自身往往不能肯定告诉我们抽样过程, 但是可以基于完全观测数据, 使用残差或者诊断方法等, 检查对抽样过程所作假设的合理性

- 类似的, 缺失模式及其与观测之间关系, 并不能肯定识别缺失机制. 很不幸的是, 我们对缺失机制所作假设并不能完全肯定地通过手边的数据来检查

贝叶斯方法处理缺失数据

- 在贝叶斯方法中, 将缺失数据视为额外的未知量, 其可以通过后验分布进行估计. 未知数据与未知参数之间没有本质上区别
- 对观测和缺失数据, 模型参数等**唯一**需要指定联合模型, 估计方法可以使用一般的 MCMC 方法
- 注意区分响应变量缺失和协变量缺失, 可忽略和不可忽略缺失机制.

可忽略缺失机制

此时 $z^{mis} = y^{mis}$, $\mathbf{z}^{obs} = \left(y^{obs}, \mathbf{x}\right)$

- 常常视完全观测的协变量为固定常数而不是随机变量
- 联合模型此时可以通过指定 $f\left(\mathbf{y}^{obs}, \mathbf{y}^{mis} | \mathbf{x}, \beta\right)$, 即完整数据似然函数
- 估计缺失响应 y^{mis} 等价于使用观测数据拟合模型后进行后验预测—在可忽略缺失下, 对缺失响应变量进行插值不影响模型参数的估计
- WINBUGS 中, 使用 NA 表示缺失值, 未知值被视为变量, WINBUGS 自动根据指定的似然分布, 在给定当前所有未知参数值下, 随机生成缺失变量值

HAMD 例子: MCAR 机制对处理比较的影响

Table: posterior mean (95% credible interval) for the contrasts (treatment comparisons) from random effects models fitted to the HAMD data

treatments	complete cases*		all cases†	
1 v 2	0.50	(-0.03,1.00)	0.74	(0.25,1.23)
1 v 3	-0.56	(-1.06,-0.04)	-0.51	(-1.01,-0.01)
2 v 3	-1.06	(-1.56,-0.55)	-1.25	(-1.73,-0.77)

* individuals with missing scores ignored

† individuals with missing scores included under the assumption that the missingness mechanism is ignorable

使用所有数据提供了更强的证据: 处理 2 比处理 1 更有效; 处理 2 比处理 3 更有效.

HAMD 例子: NMAR 机制对处理比较的影响

- 假设退出机制的模型为

$$m_{iw} \sim \text{ Bernoulli } (p_{iw})$$
$$\text{logit}\,(p_{iw}) = \theta_0 + \theta_1\,(y_{iw} - \bar{y})$$
$$\theta_0, \theta_1 \sim \text{mildly informative priors}$$

  $\bar{y}$ 为平均值.

- 一般数据对 θ_1 提供的信息非常少. 信息依赖于参数模型假设和误差分布

HAMD 例子: MAR 与 NMAR

Table: posterior mean (95% credible interval) for the contrasts (treatment comparisons) from random effects models fitted to the HAMD data

treatments	complete cases[1]		all cases (mar)[2]		all cases (mnar)[3]	
1 v 2	0.50	(-0.03,1.00)	0.74	(0.25,1.23)	0.75	(0.26,1.24)
1 v 3	-0.56	(-1.06,-0.04)	-0.51	(-1.01,-0.01)	-0.47	(-0.98,0.05)
2 v 3	-1.06	(-1.56,-0.55)	-1.25	(-1.73,-0.77)	-1.22	(-1.70,-0.75)

[1] individuals with missing scores ignored
[2] individuals with missing scores included under the assumption that the missingness mechanism is ignorable
[3] individuals with missing scores included under the assumption that the missingness mechanism is non-ignorable

允许缺失依赖于当前得分值对处理比较结果有轻微影响, 导致处理 1 和 3 的 95% 区间包含 0

HAMD 例子: 灵敏度分析

- 因为真实的缺失机制是未知的, 因此有必要进行灵敏度分析
- 前面我们通过假设缺失是可忽略的, 以及有信息的缺失机制来评估了结果
- 但是, 我们还需要检查其他的有信息缺失机制, 例如
  - 允许退出概率依赖于得分的变化

$$\text{logit}\left(p_i\right) = \theta_0 + \theta_2\left(y_{i(w-1)} - \bar{y}\right) + \theta_3\left(y_{iw} - y_{i(w-1)}\right)$$

  - 允许 θ 对不同处理不同
  - 使用不同的先验分布

HAMD 例子: 灵敏度分析结果

Table: posterior mean (95% credible interval) for the contrasts (treatment comparisons) from random effects models fitted to all the HAMD data

treatments	mar		mnar1*		mnar2†	
1 v 2	0.74	(0.25,1.23)	0.75	(0.26,1.24)	0.72	(0.23,1.22)
1 v 3	-0.51	(-1.01,-0.01)	-0.47	(-0.98,0.05)	-0.60	(-1.09,-0.11)
2 v 3	-1.25	(-1.73,-0.77)	-1.22	(-1.70,-0.75)	-1.32	(-1.80,-0.84)

* probability of missingness dependent on current score
† probability of missingness dependent on change in score

这是灵敏度分析, 我们并没有选择最好的模型. 对缺失数据下的模型进行比较是非常难办的, 我们不能使用 WINBUGS 自动生成的 DIC(Mason et al. 2010).