Bayesian Hierarchical Modelling using WinBUGS

Nicky Best, Alexina Mason and Philip Li

Short Course, Feb 17–18, 2011

http://www.bias-project.org.uk

Lecture 1. Introduction to Bayesian hierarchical models

Outline

- What are hierarchical models?
- Different modelling assumptions for hierarchical data
- Examples of simple normal and Poisson hierarchical models
- Implementation in WinBUGS

What are hierarchical models?

'Hierarchical model' is a very broad term that refers to wide range of model set-ups

- Multilevel models
- Random effects models
- Random coefficient models
- Variance-component models
- Mixed effect models
- ....

Key feature: Hierarchical models are statistical models that provide a formal framework for analysis with a complexity of structure that matches the system being studied.

(from `www.cmm.bristol.ac.uk/MLwiN/tech-support/workshops/materials/multilevel-m.shtml`)

Four Key Notions

- Modelling data with a complex structure
  - large range of structures that can be handled routinely using hierarchical models, e.g. pupils nested in schools, houses nested in neighbourhoods
- Modelling heterogeneity
  - standard regression 'averages' (i.e. the general relationship)
  - hierarchical models additionally model variances, e.g. house prices vary from neighbourhood to neighbourhood
- Modelling dependent data
  - potentially complex dependencies in the outcome over time, over space, over context, e.g. house prices within a neighbourhood tend to be similar
- Modelling contextuality: micro and macro relations
  - e.g. individual house prices depend on individual property characteristics and on neighbourhood characteristics

(from `www.cmm.bristol.ac.uk/MLwiN/tech-support/workshops/materials/multilevel-m.shtml`)

Some motivating examples

- Hospital mortality rates (Surgical)
  - Structure: Patients nested in hospitals
  - Questions: Which hospitals have particularly high or low mortality rates?
- Environmental exposures (THM)
  - Structure: Measurements of tap water THM concentrations in different water-supply zones
  - Questions: Obtain estimate of average THM concentration in each water zone
  - Questions: How much variation is there in THM concentrations within and between water zones?
- Longitudinal clinical trial (HAMD)
  - Structure: Repeated measurements within Patients
    $\rightarrow$ data correlated within person
  - Questions: Are there differences between treatments?
  - Questions: Is there heterogeneity between patients in their response to treatment?

Motivating examples (continued)

- Educational outcomes (Schools)
  - Structure: Exam scores at age 16 for pupils in Scotland
    - ⋆ Pupils grouped by primary school and secondary school (not nested)
  - Questions: How much of the variation in children's exam scores at age 16 is due to primary school,and how much to their secondary school?
- N-of-1 trials
  - Structure: Repeated within-person crossover trials
    - ⋆ Repeated measurements of treatment difference for each subject
    - → Subject-specific treatment effect *and* residual variance
  - Questions: Is there a beneficial effect of treatment?

The hierarchical approach

- Attempts to capture ('model') and understand the structure of the data
- Is flexible
  - all sources of correlation and heterogeneity can be incorporated in a modular fashion, in particular by the introduction of *unit-specific* parameters
  - can be combined with other types of models, e.g. for missing values or measurement error (see lecture 6)
- Unit specific parameters will borrow strength from the corresponding parameters associated with the other units
- Graphical representation of hierarchical models can be helpful in understanding structure and building complex models

Modelling data with complex structure

Previous examples all present the same basic modelling issues: we wish to make inferences on models with many parameters $\theta_1, ..., \theta_N$ measured on N 'units' (individuals, subsets, areas, time-points, trials, etc) *which are related or connected by the structure of the problem*
We can identify three different modelling assumptions:

1. **Identical parameters:** All the θ's are identical, in which case all the data can be pooled and the individual units ignored.
2. **Independent parameters:** All the θ's are entirely unrelated, in which case the results from each unit can be analysed independently (for example using a fully specified prior distribution within each unit)
   $\rightarrow$ individual estimates of θ_i are likely to be highly variable
3. **Exchangeable parameters:** The θ's are assumed to be 'similar' in the sense that the 'labels' convey no information

Under broad conditions an assumption of exchangeable units is mathematically equivalent to assuming that $\theta_1, ..., \theta_N$ are drawn from a *common prior distribution with unknown parameters*

Example: Hierarchical model for THM concentrations

- Regional water companies in the UK are required to take routine measurements of trihalomethane (THM) concentrations in tap water samples for regulatory purposes
- Samples tested throughout year in each water supply zone
- We want to estimate the average THM concentration for each of 70 water zones, as an exposure estimate for an epidemiological study
- We assume a normal likelihood for the data in each zone

$$x_{iz} \sim \text{Normal}(\theta_z, \sigma^2_{[e]}); \qquad i = 1, ..., n_z; \; z = 1, ..., 70$$

- Notice that we have 70 distinct mean parameters θ_z
- What prior should we specify for each θ_z?
- We assume a vague prior for the inverse of the residual error variance, $1/\sigma^2_{[e]} \sim \text{Gamma}(0.001, 0.001)$ (more on priors in Lecture 2)

Identical parameters

- Assume that the mean THM levels are the same in all zones, $\theta_z = \theta$ for all z
- Assign a prior

$$\theta \sim \text{Normal}(\mu, \sigma^2_{[z]})$$

with **specified values** of μ and $\sigma^2_{[z]}$ (note that '$[z]$' is a label not a subscript), e.g.

$$\theta \sim \text{Normal}(0, 100000)$$

$\rightarrow$ conjugate normal-normal model

- But, assuming $\theta_z = \theta$ is not really sensible since we do not expect zones supplied by different water sources to have identical THM levels

Independent parameters

- Instead, we might assume independent vague priors for each zone mean, e.g.

$$\theta_z \sim \text{Normal}(0, 100000), \qquad z = 1, ..., 70$$

- This will give posterior estimates $E(\theta_z|\boldsymbol{x}_z) \approx \overline{x}_z$ (the empirical zone mean, which is the MLE)
  - $\rightarrow$ each θ_z estimated independently
  - $\rightarrow$ no 'pooling' or 'borrowing' of information across zones
  - $\rightarrow$ no smoothing of estimates
  - $\rightarrow$ imprecise estimates if sparse data per zone
  - $\rightarrow$ how do we choose (and justify) values for the parameters of the Normal prior?

Similar (exchangeable) parameters

- Rather than specifying independent priors for each θ_z, we could specify a **hierarchical** prior:

$$\theta_z \sim \text{Normal}(\mu, \sigma^2_{[z]}), \qquad z = 1, ..., 70$$

where μ and $\sigma^2_{[z]}$ **are unknown parameters** to also be **estimated**
(Note: subscripts in [] are labels, not indices)

- Assign (hyper)prior distributions to μ and $\sigma^2_{[z]}$, e.g.

$$\begin{aligned} \mu &\sim \text{Normal}(0, 100000) \\ 1/\sigma^2_{[z]} &\sim \text{Gamma}(0.001, 0.001) \end{aligned}$$

- *joint prior distribution* for the entire set of parameters

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

Similar (exchangeable) parameters

- Joint posterior distribution of all the unknown quantities:

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

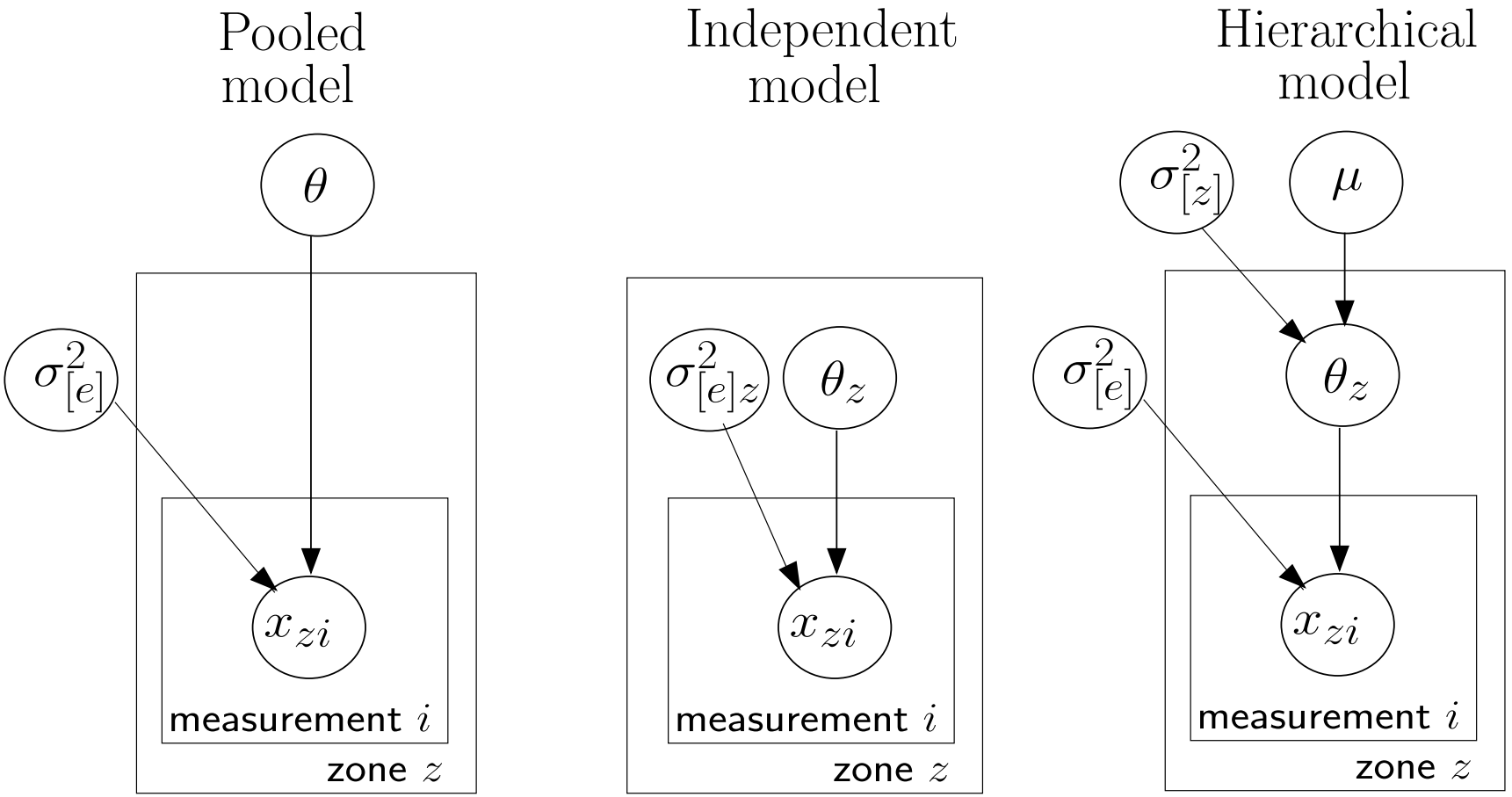
- Marginal posterior for each zone mean parameter θ_z is obtained by integrating the joint posterior $p(\boldsymbol{\theta}, \sigma^2_{[e]}\mu, \sigma^2_{[z]}|\boldsymbol{x})$ over the other parameters ($\sigma^2_{[e]}\mu, \sigma^2_{[z]}$, other θ_j, $j \neq z$) [easy using MCMC]

Posterior for each θ_z

- *'borrows strength'* from the likelihood contributions for *all* of the zones, via their joint influence on the estimate of the unknown hyperparameters μ and $\sigma^2_{[z]}$
- leads to *global smoothing* of the zone mean THM levels
- leads to improved precision of zone mean estimates
- reflects our full uncertainty about the true values of μ and $\sigma^2_{[z]}$

This is a *hierarchical model*

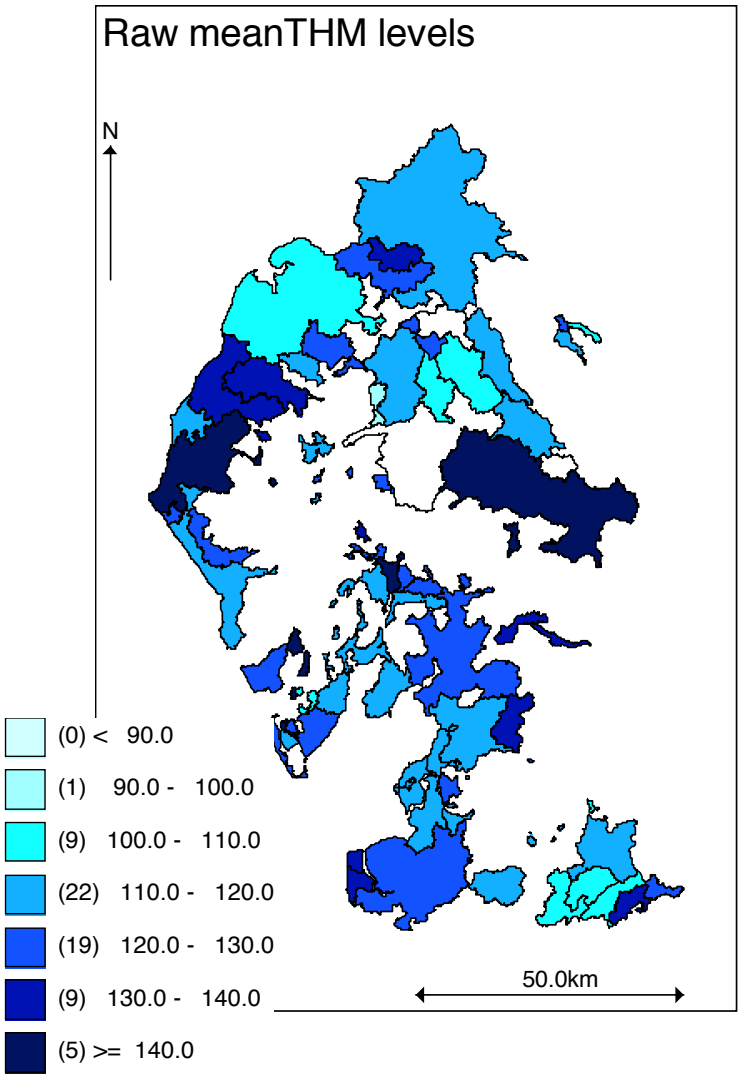
Graphical representation of THM models

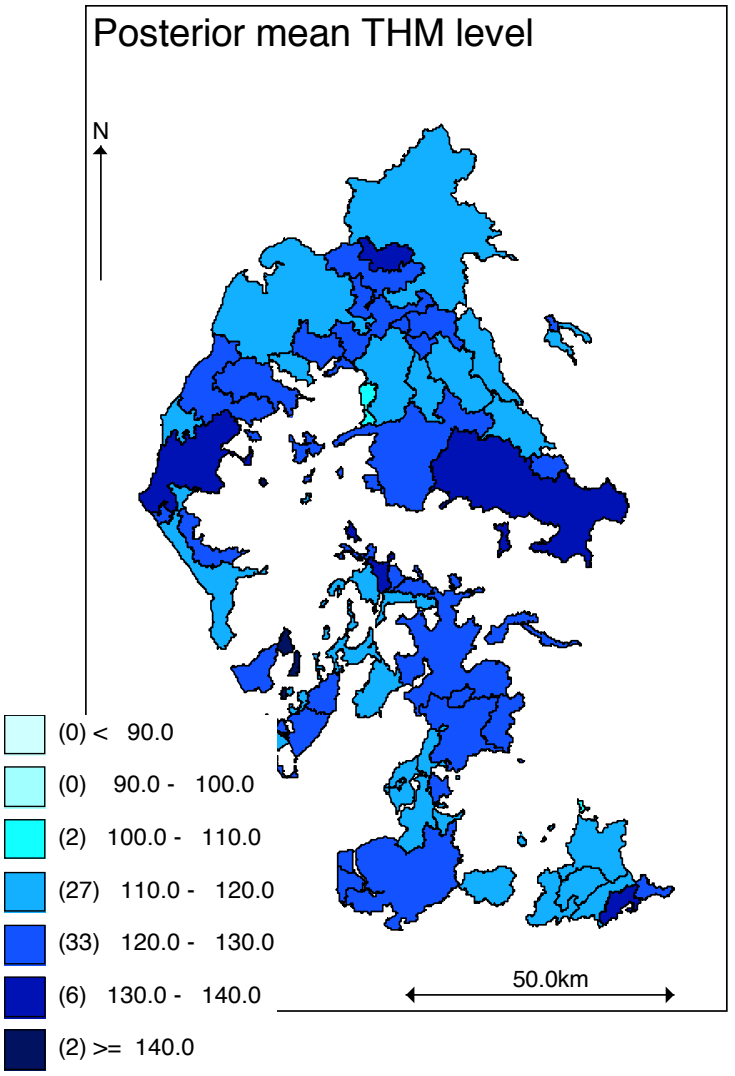


Parameter Interpretation for hierarchical model

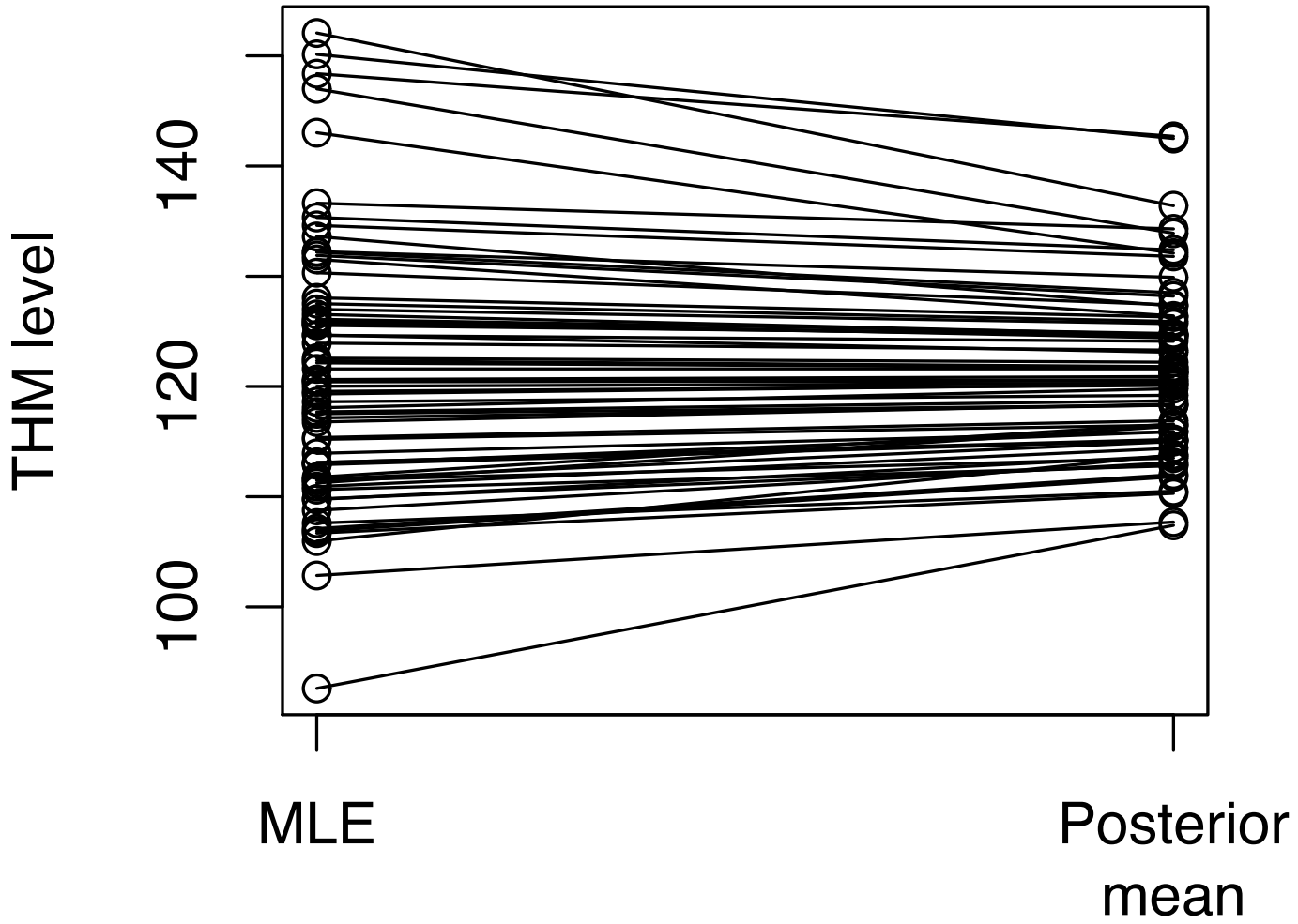
- θ_z is mean THM concentration in zone z for the study period
- μ is the overall mean THM concentration across all zones for the study period
- $\sigma^2_{[z]}$ is the between-zone variance in THM concentrations
- $\sigma^2_{[e]}$ is the residual variance in THM concentrations (reflects measurement error and true within-zone variation in THM levels)

Maps of estimated mean THM level in each zone

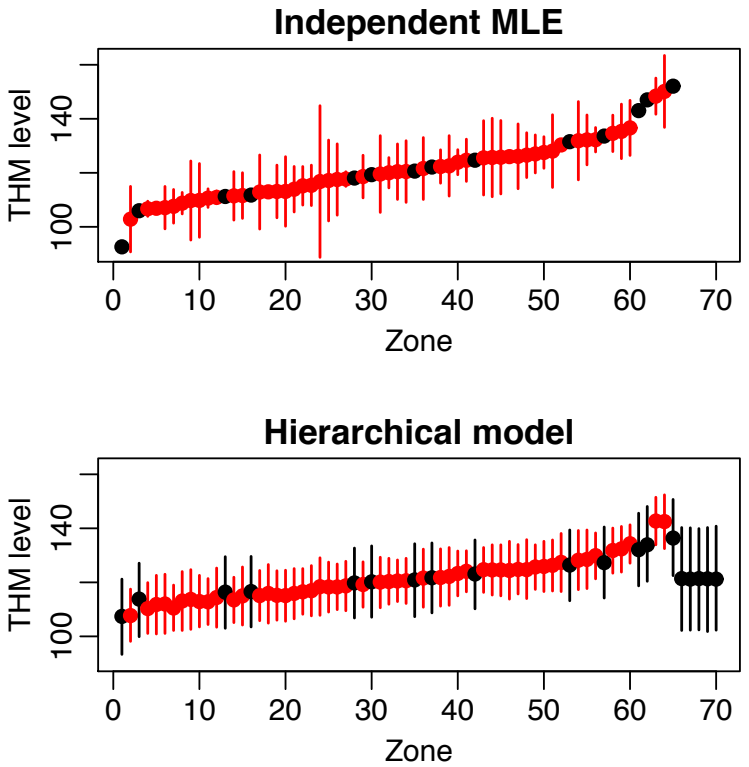




Shrinkage (smoothing) of zone mean THM levels in hierarchical model



Point estimates and 95% intervals for zone mean THM levels



- Note estimates for 5 zones with no data under hierarchical model
- Independent model also assumes independent zone-specific variances — hence no CI for zones with only 1 measurement
- Hierarchical model assumes common error variance — might be preferable to assume hierarchical prior on zone-specific variances (see Lecture 5)

Implementation of THM hierarchical model in WinBUGS

Data contain between 0 and 6 observations per zone $\rightarrow$ 'ragged array'

Zone	THM level
1	111.3, 112.9, 112.9, 105.5
2	122.6, 124.6, 135.4, 135.7, 156.7, 144.8
3	133.1, 116.6, 106.2, 126
4	111.6, 112.5, 98.6, 107.7
5	–
6	124.7
..	....

Three alternative ways to code model and data in BUGS

Method 1 — Offsets

Model code

```
for(z in 1:Nzone){
  for(i in offset[z]:(offset[z+1]-1)) {
     thm[i] ~ dnorm(theta[z], tau.e) # likelihood
  }
  theta[z] ~ dnorm(mu, tau.z) # zone mean (random effects)
}

# hyperpriors on random effects mean and variance
mu ~ dnorm(0, 0.000001)
tau.z ~ dgamma(0.001, 0.001)
sigma2.z <- 1/tau.z   # random effects variance

tau.e ~ dgamma(0.001, 0.001)
sigma2.e <- 1/tau.e  # residual error variance
```

Method 1 — Offsets (continued)

Data

```
list(Nzone=70,
     thm=c(111.3, 112.9, 112.9, 105.5, 122.6, 124.6,
           135.4, 135.7, 156.7, 144.8, 133.1, 116.6,
           106.2, 126, 111.6, 112.5, 98.6, .....),
     offset = c(1, 5, 11, 15, 19, 19, 20.....),
)
```

Method 2 — Nested index

Model code

```
for(i in 1:Nobs) {
 thm[i] ~ dnorm(theta[zone[i]], tau.e) # likelihood
}

for(z in 1:Nzone) {
 theta[z] ~ dnorm(mu, tau.z) # zone means (random effects)
}

# hyperpriors on random effects mean and variance
mu ~ dnorm(0, 0.000001)
tau.z ~ dgamma(0.001, 0.001)
sigma2.z <- 1/tau.z   # random effects variance

tau.e ~ dgamma(0.001, 0.001)
sigma2.e <- 1/tau.e  # residual error variance
```

Method 2 — Nested index (continued)

Data

```
list(Nobs = 173, Nzone = 70,
     thm = c(111.3, 112.9, 112.9, 105.5, 122.6, 124.6,
             135.4, 135.7, 156.7, 144.8, 133.1, 116.6,
             106.2, 126, 111.6, 112.5, 98.6, ...),
     zone = c(1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3,
              4, 4, 4, 4, 6,.....))
```

Alternative data format:

```
list(Nobs=173, Nzone=70)

thm[]     zone[]
111.3     1
112.9     1
112.9     1
105.5     1
122.6     2
..... .....
END
```

Method 3 — Pad out data with NA's

Model code

```
for(z in 1:Nzone) {
  for(i in 1:6) {
     y[z,i] ~ dnorm(theta[z],tau.e)  # likelihood
  }
  theta[z] ~ dnorm(mu, tau.z) # zone means (random effects)
}

# hyperpriors on random effects mean and variance
mu ~ dnorm(0, 0.000001)
tau.z ~ dgamma(0.001, 0.001)
sigma2.z <- 1/tau.z   # random effects variance

tau.e ~ dgamma(0.001, 0.001)
sigma2.e <- 1/tau.e  # residual error variance
```

Method 3 — Pad out data with NA's (cont.)

Data

```
list(Nzone=70,
    thm=structure(.Data=
        c(111.3, 112.9, 112.9, 105.5,    NA,    NA,
          122.6, 124.6, 135.4, 135.7, 156.7, 144.8,
          133.1, 116.6, 106.2, 126.0,    NA,    NA,
          111.6, 112.5,  98.6, 107.7,    NA,    NA,
             NA,    NA,    NA,    NA,    NA,    NA,
          .......), .Dim=c(70, 6)))
```

Alternative data format:

```
list(Nzone=70)

thm[,1]  thm[,2]  thm[,3]  thm[,4]  thm[,5]   thm[,6]
  111.3    112.9    112.9    105.5       NA        NA
  122.6    124.6    135.4    135.7    156.7     144.8
  133.1    116.6    106.2    126.0       NA        NA
  .....
END
```

Variance Partition Coefficient (VPC)

- In hierarchical or multilevel models, the residual variation in the response variable is split into components attributed to different levels
- Often of interest to quantify percentage of total variation attributable to higher level units
- In simple 2-level Normal linear models, can use VPC or intra-cluster correlation (ICC) coefficient

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

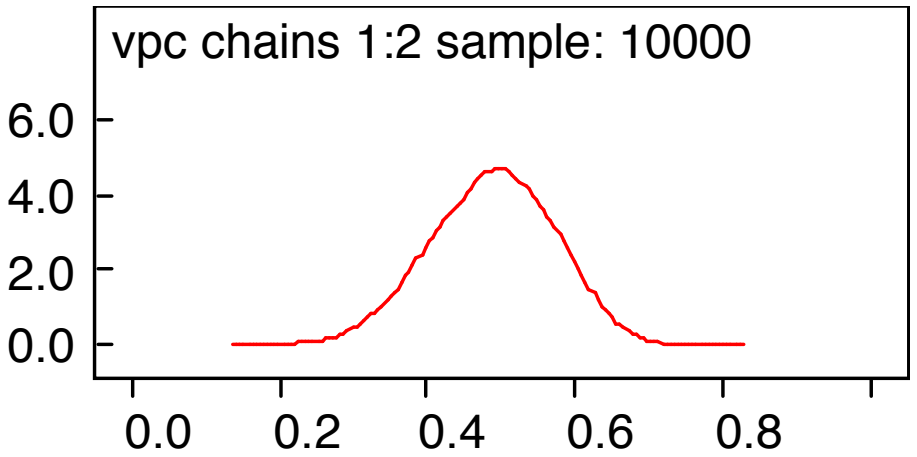
  - $\sigma^2_{[e]}$ is the 'level 1' variance (i.e. variance of Normal likelihood)
  - $\sigma^2_{[z]}$ is the 'level 2' variance (i.e. random effects variance)
- In WinBUGS, add extra line in model code to calculate VPC, e.g.

```
vpc <- sigma2.z / (sigma2.z + sigma2.e)
```

then monitor posterior samples of `vpc` to obtain point estimate and uncertainty interval

VPC for THM example

Posterior distribution of VPC



Posterior mean = 0.49

95% CI (0.32, 0.64)

So approximately half the total variation in THM levels is between water zones, and half is within water zones

Example: Disease mapping

- In many applications, we may be interested in modelling data collected on each of a number of geographical areas within a study region, e.g.
  - Counts of disease cases in different geographical areas
  - Counts of burglaries in different neighbourhoods
  - Number of fish caught in each subregion of a river
- Here we consider data on the observed number of cases of childhood leukaemia, y_i, diagnosed in a 10 year period in each of $i = 1, ..., 879$ areas (electoral wards) in London (data from Thames Cancer Registry)
- Using national age/sex-standardised reference rates for leukaemia and Census population counts, we can also calculate the expected number of cases, E_i, in each area
- Assume a Poisson likelihood for the disease count in each area:

$$y_i \sim \text{Poisson}(\lambda_i E_i), \qquad i = 1, ..., 879$$

- We have 879 *distinct* relative risk parameters λ_i
- What prior should we specify for each λ_i?

Different modelling assumptions

Identical parameters

Assume $\lambda_i = \lambda$ for all i and assign a prior

$$\lambda \sim \text{Gamma}(a, b)$$

with *specified values* of a and b, e.g.

$$\lambda \sim \text{Gamma}(1, 1)$$

$\rightarrow$ conjugate Poisson-gamma model

Independent parameters

Assume independent vague priors for each relative risk, e.g.

$$\lambda_i \sim \text{Gamma}(0.1, 0.1), \qquad i = 1, ..., 879$$

$\rightarrow$ This will give estimates of the posterior mean for $\lambda_i \approx y_i / E_i$, which is the MLE (also termed standardised morbidity ratio, SMR)

Different modelling assumptions (continued)

Similar (exchangeable) parameters
Specify a **hierarchical** random effects prior:

$$\lambda_i \sim \text{Gamma}(a, b), \qquad i = 1, ..., 879$$

where a and b *are unknown parameters* to also be **estimated**

$\rightarrow$ assign hyperprior distributions to a and b

$\rightarrow$ what is a suitable hyperprior for these parameters?

A more flexible hierarchical prior for the relative risks

- A gamma random effects prior for the λ_i is mathematically convenient, but may be restrictive:
  - Covariate adjustment (regression) is difficult
  - No possibility for allowing spatial correlation between risks in nearby areas
- A normal random effects prior for $\log \lambda_i$ is more flexible:

$$\begin{aligned} y_i &\sim \text{Poisson}(E_i \lambda_i) \\ \log \lambda_i &= \alpha + \theta_i \\ \theta_i &\sim \text{Normal}(0, \sigma^2) \end{aligned}$$

- Need to specify hyperprior distributions for

  σ^2 (between-area variance), e.g. $\sigma^{-2} \sim \text{Gamma}(0.001, 0.001)$

  α (mean log relative risk), e.g. $\alpha \sim \text{Uniform}(-\infty, \infty)$

  (More on priors in Lecture 2)

Parameter Interpretation

- θ_i are the **random effects**.
- $\exp(\alpha + \theta_i)$ = relative risk in area *i* compared to expected risk based on age and sex of population
- θ_i can also be thought of as a latent variable which captures the effects of unknown or unmeasured area level covariates.
- If these area level covariates are spatially structured (e.g. environmental effects), our model for θ_i should allow for this (i.e. replace normal random effects distribution by spatial distribution — not covered in this course).
- The variance of the random effects (σ^2) reflects the amount of extra-Poisson variation in the data.
- Unclear how to define/calculate VPC for generalised linear hierarchical models
- Alternative summary of random effects variability is to look at ratio of quantiles of their empirical distribution

Ranking in hierarchical models

- Recent trend in UK towards ranking 'institutional' performance e.g. schools, hospitals or areas
- Rank of a point estimate is a highly unreliable summary statistic
  - would like measure of uncertainty about rank
- Bayesian methods provide posterior interval estimates for ranks
- For the leukemia example, at each MCMC iteration, ranking sampled values of $\lambda_1, ..., \lambda_{879}$ gives sample from posterior distribution of ranks for each area
- See Golstein and Spiegelhalter (1996) for further discussion on ranking

WinBUGS contains 'built-in' options for ranks:

- `Rank` option of `Inference` menu monitors the rank of the elements of a specified vector
- `rank(x[],i)` returns the rank of the ith element of `x`
- `ranked(x[],i)` returns the value if the ith-ranked element of x

Quantile ratios for random effects

- A useful summary of variability between units in a hierarchical model is to rank the random effects and calculate the difference or ratio between two units at opposite extremes
- For the leukemia example, suppose we consider the 5^{th} and 95^{th} percentiles of the area relative risk distribution
  - let $\lambda_{5\%}$ denote the relative risk of leukemia for the area ranked at the 5^{th} percentile
  - let $\lambda_{95\%}$ denote the relative risk of leukemia for the area ranked at the 95^{th} percentile
  - then $\text{QR}_{90} = \frac{\lambda_{95\%}}{\lambda_{5\%}} =$ ratio of relative risks of leukemia between the top and bottom 5% of areas
- Using MCMC, we can calculate the ranks, and hence the QR_{90}, at each iteration, and hence obtain a posterior distribution for QR_{90}

BUGS code

```
model {
 for(i in 1 : N) {
  Y[i]  ~ dpois(mu[i])
  log(mu[i]) <- log(E[i]) + alpha + theta[i]
  theta[i] ~ dnorm(0, tau)  # area random effects
  lambda[i] <- exp(alpha + theta[i]) # area relative risk
 }
 # Priors:
 alpha  ~ dflat()       # uniform prior on overall intercept

 tau ~ dgamma(0.001, 0.001) # precison of area random effects
 sigma <- 1/sqrt(tau)     # between-area sd of random effects

 # 90% quantile ratio for area relative risks
 QR90 <- ranked(lambda[],829)/ranked(lambda[],45)

 #rank
 for(i in 1 : N) {
  rank.lambda[i] <- rank(lambda[], i) # rank of area i
 }
}
```

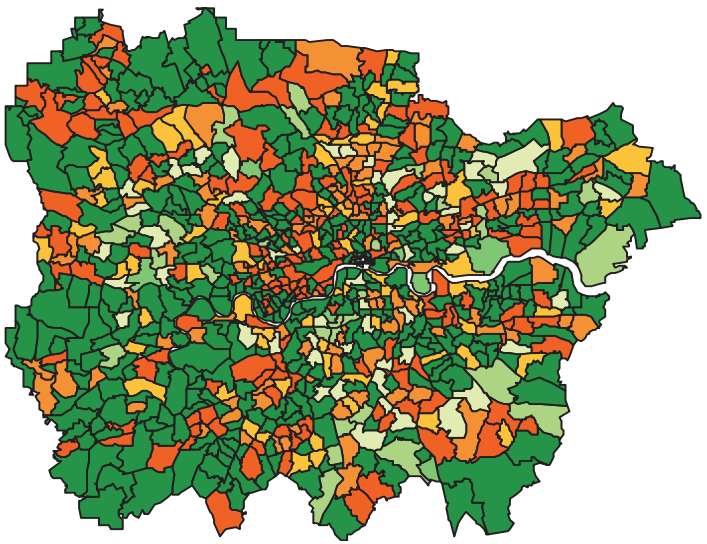
Results for childhood leukaemia example

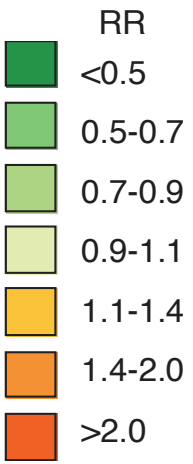
Parameters of interest:

- $e^{\alpha+\theta_i}$ = relative risk of leukaemia in area *i* relative to expected (see map)
- σ = between-area standard deviation of log relative risk of leukaemia
  Posterior mean and 95% interval = 0.46 (0.34, 0.62)
- QR_{90} = 4.7 (95% interval 2.9 to 7.5)
  - So 4.7-fold variation in relative risk of leukemia between top and bottom 5% of areas.

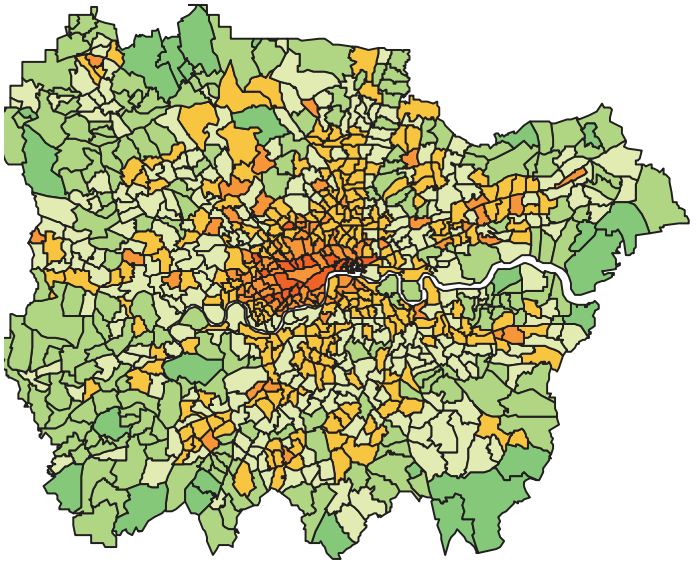
Maps of estimated area-specific RR of leukaemia

SMR

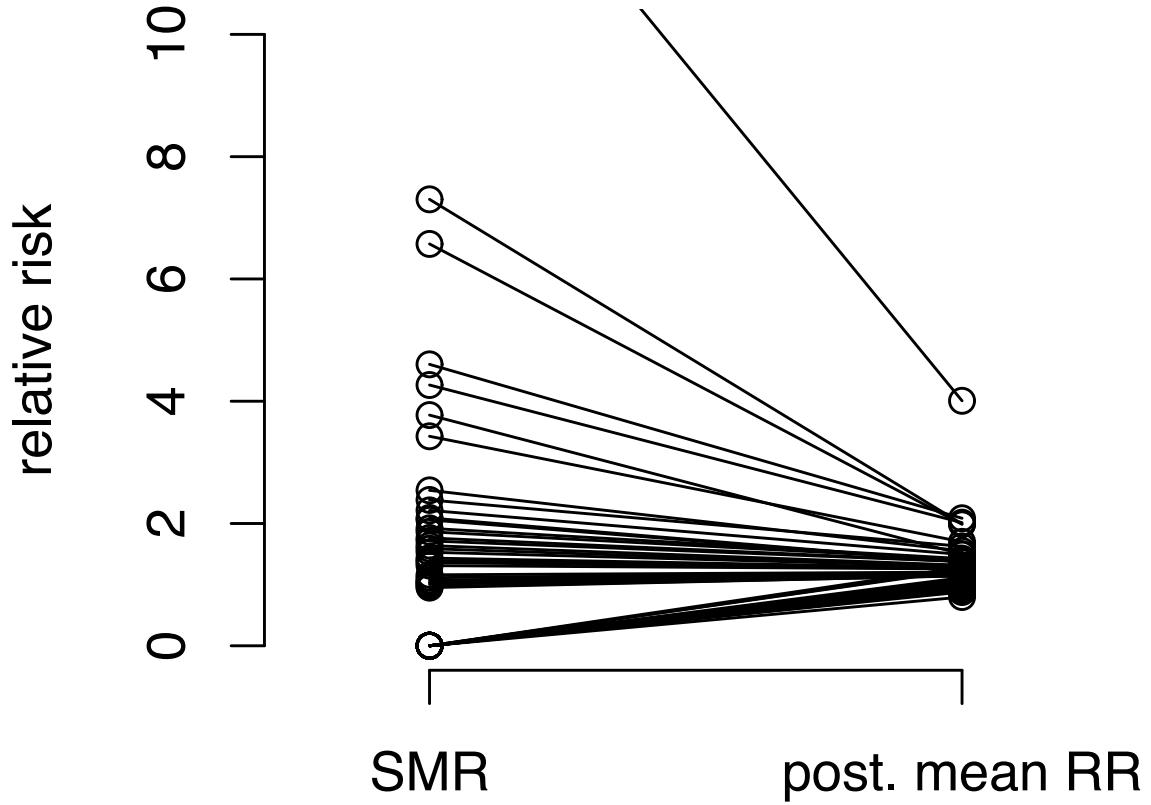




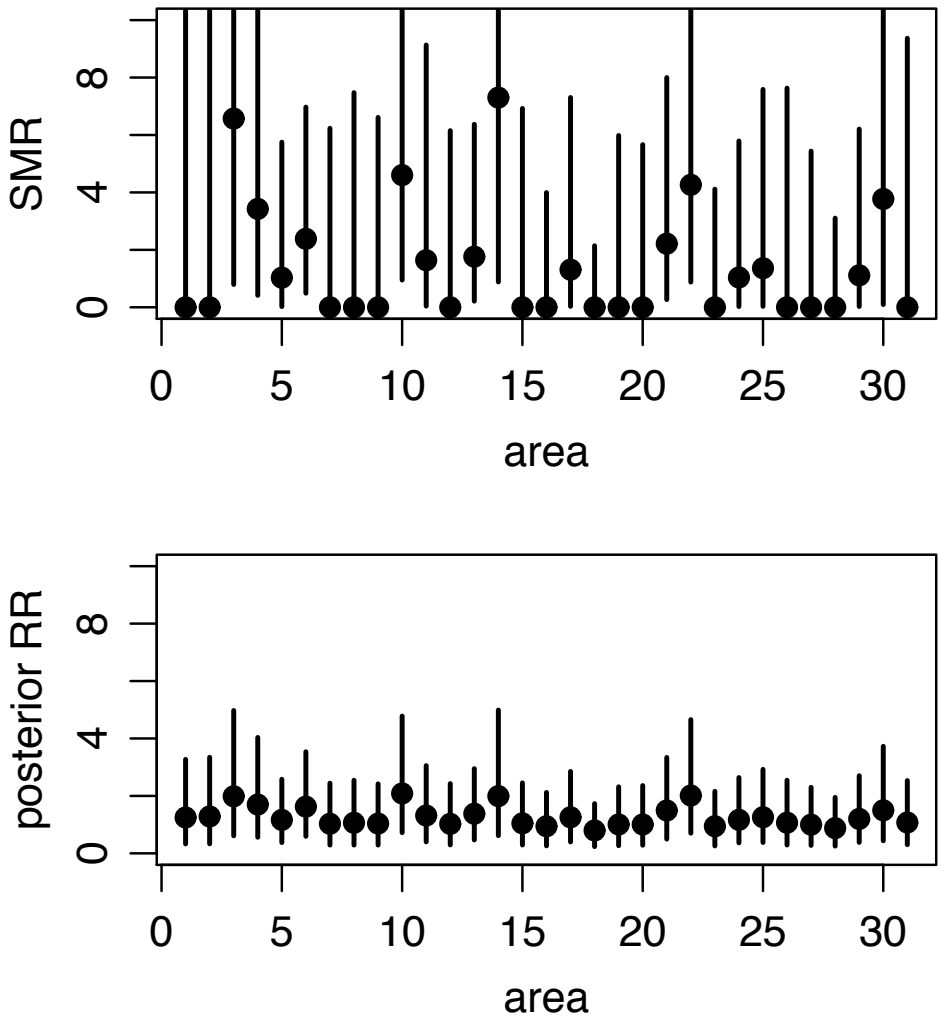
Smoothed RR



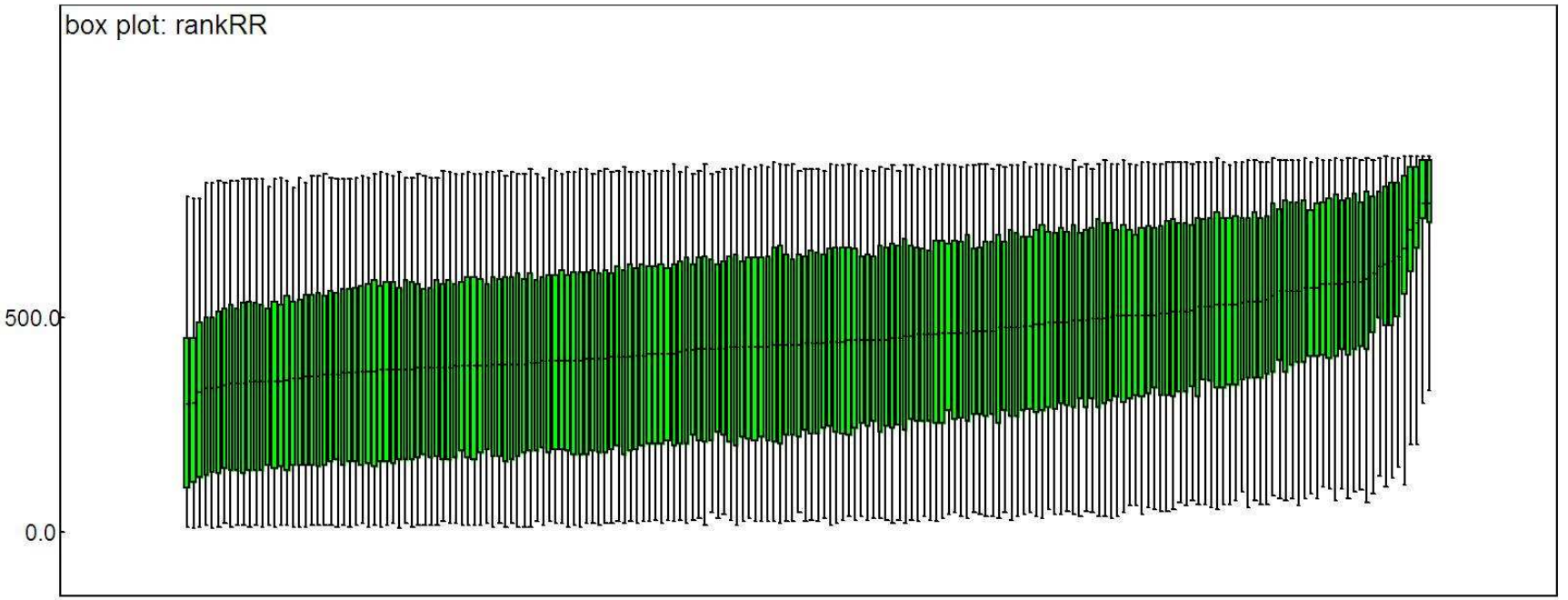
SMR versus posterior mean RR for selected areas



Point estimate and 95% interval for relative risk in selected areas



Posterior distribution of area ranks



General comments on hierarchical models

Hierarchical models allow "borrowing of strength" across units

- posterior distribution of θ_i for each unit borrows strength from the likelihood contributions for *all* the units, via their joint influence on the posterior estimates of the unknown hyper-parameters

$\rightarrow$ improved efficiency

- MCMC allows considerable flexibility over choice of random effects distribution (not restricted to normal random effects)
- Judgements of exchangeability need careful assessment
  - units suspected a priori to be systematically different might be modelled by including relevant covariates so that residual variability more plausibly reflects exchangeability
  - subgroups of prior interest should be considered separately

Lecture 2.
Prior distributions for hierarchical models

Outline

- Priors for location parameters
- Priors for level variances
- Priors for random effects variances
- Sensitivity analysis to priors
- Examples

Some general recommendations for specifying priors

- Distinguish
  - *primary* parameters of interest in which one may want minimal influence of priors
  - *secondary* structure used for smoothing *etc.* in which (moderately) informative priors may be more acceptable
- In hierarchical models, we may be primarily interested in the level 1 parameters (regression coefficients, etc.), or in the variance components (random effects and their variances), or in both
- Prior best placed on interpretable parameters
- Great caution needed in complex models that an apparently innocuous uniform prior is not introducing substantial information
- *'There is no such thing as a 'noninformative' prior. Even improper priors give information: all possible values are equally likely'* (Fisher, 1996)

Priors for location parameters

'Location' parameters are quantities such as means, regression coefficients,...

- Uniform prior on a wide range, or a Normal prior with a large variance can be used, e.g.

$\theta \sim \text{Unif}(-100, 100)$	`theta ~ dunif(-100, 100)`
$\theta \sim \text{Normal}(0, 100000)$	`theta ~ dnorm(0, 0.00001)`

  Prior will be locally uniform over the region supported by the likelihood

  - ! remember that `WinBUGS` parameterises the Normal in terms of mean and *precision* so a vague Normal prior will have a *small* precision
  - ! 'wide' range and 'small' precision depend on the scale of measurement of θ

Priors for Level 1 scale parameters

- Sample variance σ^2: standard 'reference' (Jeffreys') prior is the 'inverse' prior

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

- This is equivalent to a flat (uniform) prior on the log scale:

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

- This prior makes intuitive sense: if totally ignorant about the scale (order of magnitude) of a parameter, then it is equally likely to lie in the interval 1–10 as it is to lie in the interval 10–100, etc.

Priors for Level 1 scale parameters (continued)

- Jeffreys' prior on the inverse variance (precision, $\tau = \sigma^{-2}$) is

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

  which may be approximated by a 'just proper' prior

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

- This is also the conjugate prior and so is widely used as a 'vague' proper prior for the precision of a Normal likelihood
- In BUGS language: `tau ~ dgamma(0.001, 0.001)`

Priors for hyper-parameters in hierarchical models

Consider a hierarchical model with exchangeable random effects

$$\theta_i \sim \mathsf{N}(\mu, \sigma^2) \quad i = 1, ..., I$$

What priors should we assume for the hyper-parameters μ and σ^2?

Often want to be reasonably non-informative about the mean and (possibly) the variance of the random effects

- For location parameters (i.e. random effects mean, μ), a uniform prior on a wide range, or a Normal prior with a large variance can be used (see previous slides)
- 'Non-informative' priors for random effects variances are more tricky

Priors for random effects variances

- As noted above, standard 'non-informative' (Jeffreys) prior for a Normal variance, σ^2, is the inverse prior

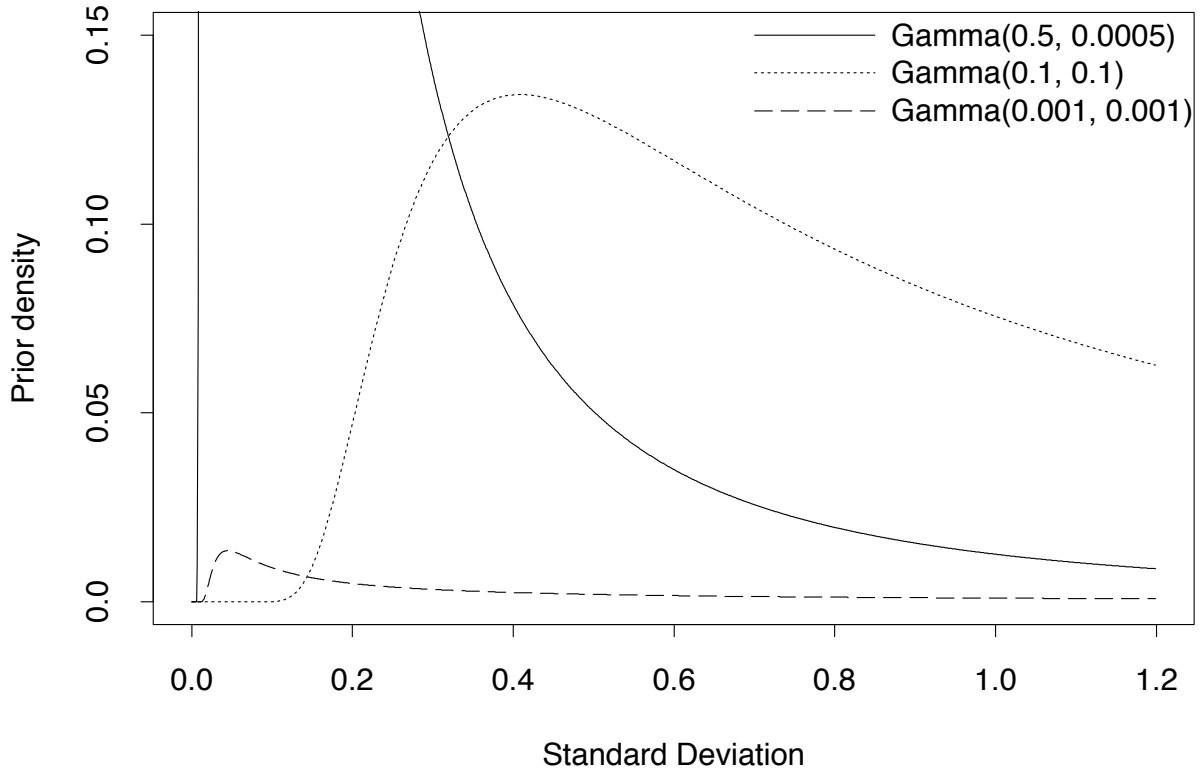
$$p(\sigma^2) \propto \frac{1}{\sigma^2} \propto \text{Gamma}(0,0)$$

- This prior is improper (doesn't integrate to 1)
  - OK as prior on sampling variance (level 1) as still gives proper posterior
  - If used as prior for random effects variance, can lead to *improper* posterior
  - prior has infinite mass at zero, but $\sigma^2 = 0$ (no level 2 variation) may be supported by non-negligible likelihood

Priors for random effects variances

- Gamma(ϵ, ϵ)— with ϵ small and positive — is 'just proper' form of Jeffreys prior
  - A Gamma(0.001, 0.001) prior for the random effects *precision* is often used, as it also has nice conjugacy properties with the Normal distribution for the random effects
  - But inference may still be sensitive to choice of ϵ
  - Sensitivity particularly a problem if data (likelihood) supports small values of σ^2 (i.e. little evidence of heterogeneity between units)
  - See Gelman (2006) for further discussion

Priors for random effects variances



Some different gamma(a, b) priors for the precision, shown on the transformed scale of the standard deviation

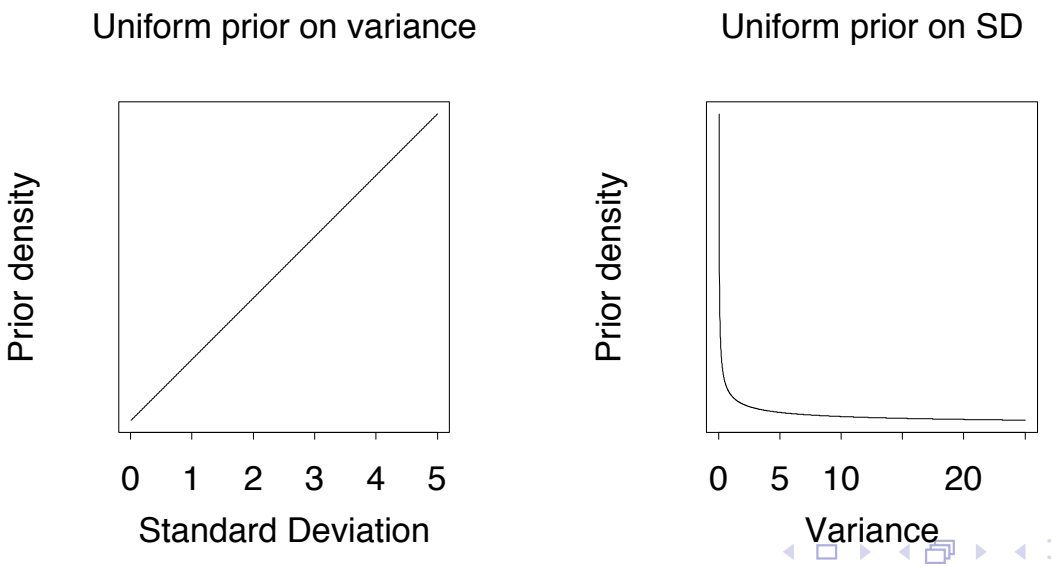
Priors for random effects variances

Other options for 'vague' or 'weakly informative' priors on variance components:

- Uniform priors over a finite range on standard deviation, e.g.

$$\sigma \sim \text{Uniform}(0, 1000)$$

  Appropriate upper bound will depend on scale of measurement of random effects
- Uniform prior on variance not generally recommended, as implies unrealistic prior information about SD



Priors for random effects variances

- Another option is a half-normal or half-t prior on the standard deviation, e.g.

$$\sigma \sim \text{Normal}(0, 100)I(0,)$$

  In WinBUGS : `sigma ~ dnorm(0, 0.01)I(0, )`
- Again, value chosen for variance of half-normal will depend on scale of measurements for continuous data

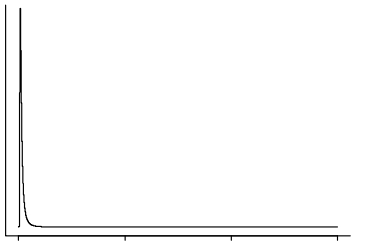
Example: Comparing performance on Aptitude Test in schools

(Gelman 2006)

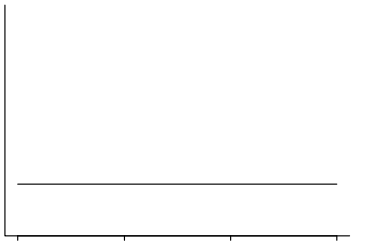
- Two-level normal hierarchical model with school-specific random effect
- Observed effects for 8 schools range from -2.75 (SE 16.3) to 28.4 (SE 14.9)
- Likelihood supports wide range of values for between-school SD
  - $\rightarrow$ plausible that all schools have same underlying effect (i.e. between-school SD = 0)
  - $\rightarrow$ data also support large between-school variation, although implausible that SD $>$ 50
- Consider subset of first 3 schools ($\rightarrow$ very sparse data) and compare Gamma(0.001, 0.001) prior on random effects precision with Uniform(0, 1000) prior and half-t(0, 25, 2) prior on random effects sd

Sensitivity to priors on random effects variance in schools example

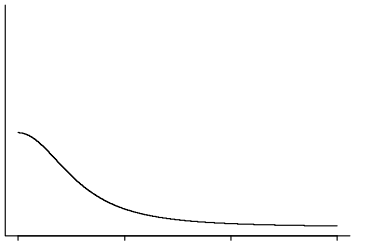
Gamma(0.001, 0.001)
prior on precision



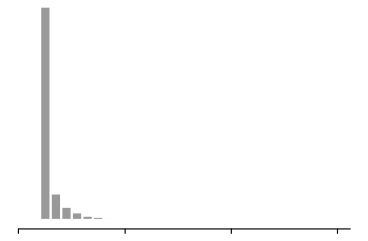
Unif(0, 1000)
prior on SD



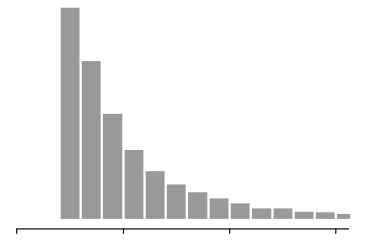
Half t(0, 25, 2)
prior on SD



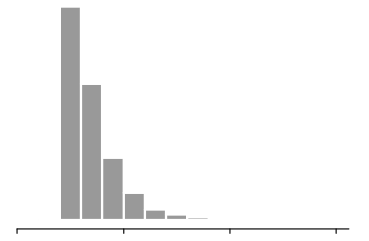
Posterior



Posterior



Posterior



Interpreting the variance of log normal random effects

In hierarchical GLMs, we typically assume a normal random effects distribution on the log odds ratio or log rate or log relative risk scale

- Assume that log ORs or log rates or log RRs, $\theta_i \sim \text{N}(\mu, \sigma^2)$

$\Rightarrow$ *a priori*, 90% of values of θ believed to lie in interval $\mu \pm 1.645\sigma$

$\Rightarrow$ $\theta_{95\%} - \theta_{5\%} = 2 \times 1.645 \times \sigma = 3.29\sigma$

$\Rightarrow$ $\frac{\exp(\theta_{95\%})}{\exp(\theta_{5\%})} = \exp(3.29\sigma)$

$= 90\%$ prior range of variation in ORs or RRs or rates across units

(cf Quantile ratio (QR_{90}) for summarising empirical variability of posterior distribution of random effects)

Prior range of variation of log normal random effects

σ	95% range ($e^{3.92\sigma}$)	90% range ($e^{3.29\sigma}$)
0.0	1.00	1.00
0.1	1.48	1.39
0.2	2.19	1.93
0.3	3.24	2.68
0.4	4.80	3.73
0.5	7.10	5.18
0.75	18.92	11.79
1.0	50.40	26.84
2.0	2540.20	720.54

- σ around 0.1 to 0.5 may appear reasonable in many contexts
- σ around 0.5 to 1.0 might be considered as fairly high
- σ around 1.0 would represent fairly extreme heterogeneity

So might specify moderately informative half-normal prior, say $\sigma \sim$ N(0, 0.4^2) I(0,), which puts very little prior mass on $\sigma > 1$.

Similar approach can be used to 'calibrate' σ in terms of, say, intra-class correlation $\sigma^2_{\text{btw}}/(\sigma^2_{\text{btw}} + \sigma^2_{\text{wth}})$ in normal hierarchical models

Example: Surgical — sensitivity to priors

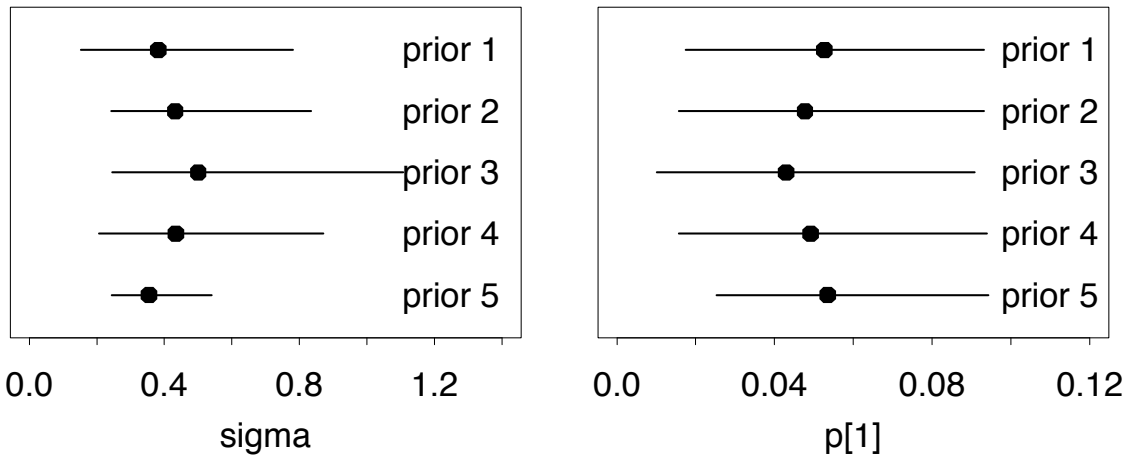
- Recall the surgical example from Practical 1.
- Hierarchical model for surgical data:

$$\begin{aligned} r_i &\sim \text{Binomial}(n_i, \pi_i) \\ \text{logit}\, \pi_i &\sim \text{Normal}(\mu, \sigma^2) \\ \mu &\sim \text{Uniform}(-1000, 1000) \end{aligned}$$

- Consider 5 alternative priors for random effects variance:
  1. $\sigma^{-2} \sim \text{Gamma}(0.001, 0.001)$
  2. $\sigma^{-2} \sim \text{Gamma}(0.1, 0.1)$
  3. $\sigma^2 \sim \text{Uniform}(0, 100)$ (not generally recommended)
  4. $\sigma \sim \text{Uniform}(0, 100)$
  5. $\sigma \sim \text{N}(0, 0.4^2)I(0,)$

Results of Sensitivity analysis

Posterior median and 95% intervals for the random effects standard deviation (`sigma`) and the mortality rate in hospital 1 (`p[1]`), under each prior



- Posterior variability of σ is somewhat sensitive to the prior
  - the rather inappropriate uniform prior on the variance (prior 3) results in the highest posterior estimates of σ (recall this prior puts increasing weight on larger values of σ
  - surprisingly, the moderately informative prior (prior 5) gives the narrowest posterior intervals
- The posterior estimates of the hospital mortality rates, p_i, are quite robust to the prior on the random effects variability

Lecture 3: Model Criticism and Comparison for Bayesian Hierarchical Models

Outline

- Bayesian approaches to model choice
- Deviance Information Criterion (DIC)
- Model criticism for non-hierarchical models (residuals, predictive checks)
- Model criticism for hierarchical models (predictive checks)
- Examples and WinBUGS implementation

General framework

Need to distinguish three stages

1. *Criticism:* exploratory checking of a single model, which may suggest -
2. *Extensions:* embed initial model in list of alternatives, which leads to -
3. *Comparison:* assess candidates in terms of their evidential support and influence on conclusions of interest.

There should be iteration between these stages (see O'Hagan, 2003)

Classical model choice

- Most widely used approach is the hypothesis test for comparing nested models, e.g. deviance (likelihood ratio) test in GLMs
- Alternatives include Akaike Information Criterion

$$\text{AIC} = -2 \log p(y|\hat{\theta}) + 2k$$

where

  - $p(y|\hat{\theta})$ is the likelihood evaluated at the MLEs of the model parameters, $\hat{\theta}$
  - k is the number of parameters in the model (dimension of Θ)

- Note that, asymptotically, AIC is equivalent to leave-one-out cross validation

Bayesian approaches to model choice

- Traditional Bayesian comparison of models M_0 and M_1 is based on hypothesis tests using the **Bayes factor**
  - If both models (hypotheses) are equally likely a priori, the Bayes factor is the posterior odds in favour of model M_0 being true
- But, Bayes factors have a number of practical and theoretical limitations:
  - Difficult to compute in practice except for simple models
  - Need informative priors
  - Objective is identification of 'true' model
- Alternatively, we can adopt a **predictive approach** to model comparison
  - Would like to chose model that produces good predictions of the quantities we are interested in (e.g. future observations on the same sampling units or other similar units)
  - Objective is to select the model that best approximates the data generating mechanism rather than to identify the 'true' model

Predictive approach to model comparison

- Could use cross-validation and choose model that best predicts (e.g. minimises MSE) independent validation data
  - appropriate validation data rarely available
  - leave-one-out approaches possible, but can be computationally intensive
- Alternatively, assess (predictive) fit to observed data
  - need some form of penalty for having used the same data to estimate model parameters
  - natural to penalize models by 'lack of fit' + 'complexity'
- Complexity reflects the number of parameters in a model
  - well-defined in classical non-hierarchical models (= degrees of freedom or number of free parameters)
  - in (Bayesian) hierarchical models, prior effectively acts to 'restrict' the freedom of the model parameters, so appropriate degrees of freedom ('effective number of parameters') is less clear

What is the 'deviance'?

- For a likelihood $p(y|\theta)$, we define the deviance as

$$D(\theta) = -2 \log p(y|\theta)$$

- WinBUGS automatically calculates `deviance` (for most models) based on specified sampling distribution of the data
  - e.g. for Binomial data, `y[i] ~ dbin(theta[i],n[i])`, the `deviance` is

$$-2 \left[\sum_i y_i \log \theta_i + (n_i - y_i) \log(1 - \theta_i) + \log \binom{n_i}{r_i} \right]$$

- Full normalising constants for $p(y|\theta)$ are included in `deviance`
- `deviance` is evaluated at each MCMC iteration given the current sampled values of the parameters θ
- Type `deviance` in the WinBUGS samples monitor tool, set a monitor, update model as normal and inspect history plot, density plots, posterior summary statistics of `deviance` in the same way as for other monitored parameters

Deviance Information Criterion, DIC

Spiegelhalter et al (2002) proposed a Bayesian predictive model comparison criterion based on trade off between model fit and complexity:

1. Summarise fit by posterior mean deviance, $E_{\theta|y}[D]$
2. Summarise complexity by

$$\begin{aligned} p_D &= E_{\theta|y}[D] - D(E_{\theta|y}[\theta]) \\ &= \overline{D} - D(\overline{\theta}); \end{aligned}$$

   the posterior mean deviance minus the deviance evaluated at the posterior means of the parameters
3. Combine into a *Deviance Information Criterion*

$$\begin{aligned} DIC &= \overline{D} + p_D \\ &= D(\overline{\theta}) + 2p_D; \end{aligned}$$

This can be seen as a generalisation of Akaike's criterion: for non-hierarchical models, $\overline{\theta} \approx \hat{\theta}$, $p_D \approx p$ and hence $DIC \approx AIC$.

Comments on DIC and p_D

- These quantities are easy to compute in an MCMC run
- Aiming for Akaike-like, cross-validatory, behaviour based on ability to make short-term predictions of a repeat set of similar data.
- The minimum DIC estimates the model that will make the best short-term predictions
  - Very roughly, differences in DIC $>$ 10 are substantial; differences in DIC $<$ 5 may be negligible
- p_D measures the dimensionality of the model, i.e. effective number of free parameters (degrees of freedom) that can be identified by the data
- p_D is not invariant to reparameterisation (see later)
- p_D can be contrived to be negative (see later)
- Note: negative DIC is not a problem

Example: Scottish lip cancer data

- Counts of cases of lip cancer, Y_i, in each of 56 districts in Scotland.

$$Y_i \sim \text{Poisson}(\lambda_i E_i)$$

- λ_i represents the underlying district-specific relative risk of lip cancer and E_i are the 'expected' number of cases based on the age-sex structure of the population in district i.
- x_i is a covariate representing the proportion of the population in 'outdoor' occupations (proxy for sunlight exposure)

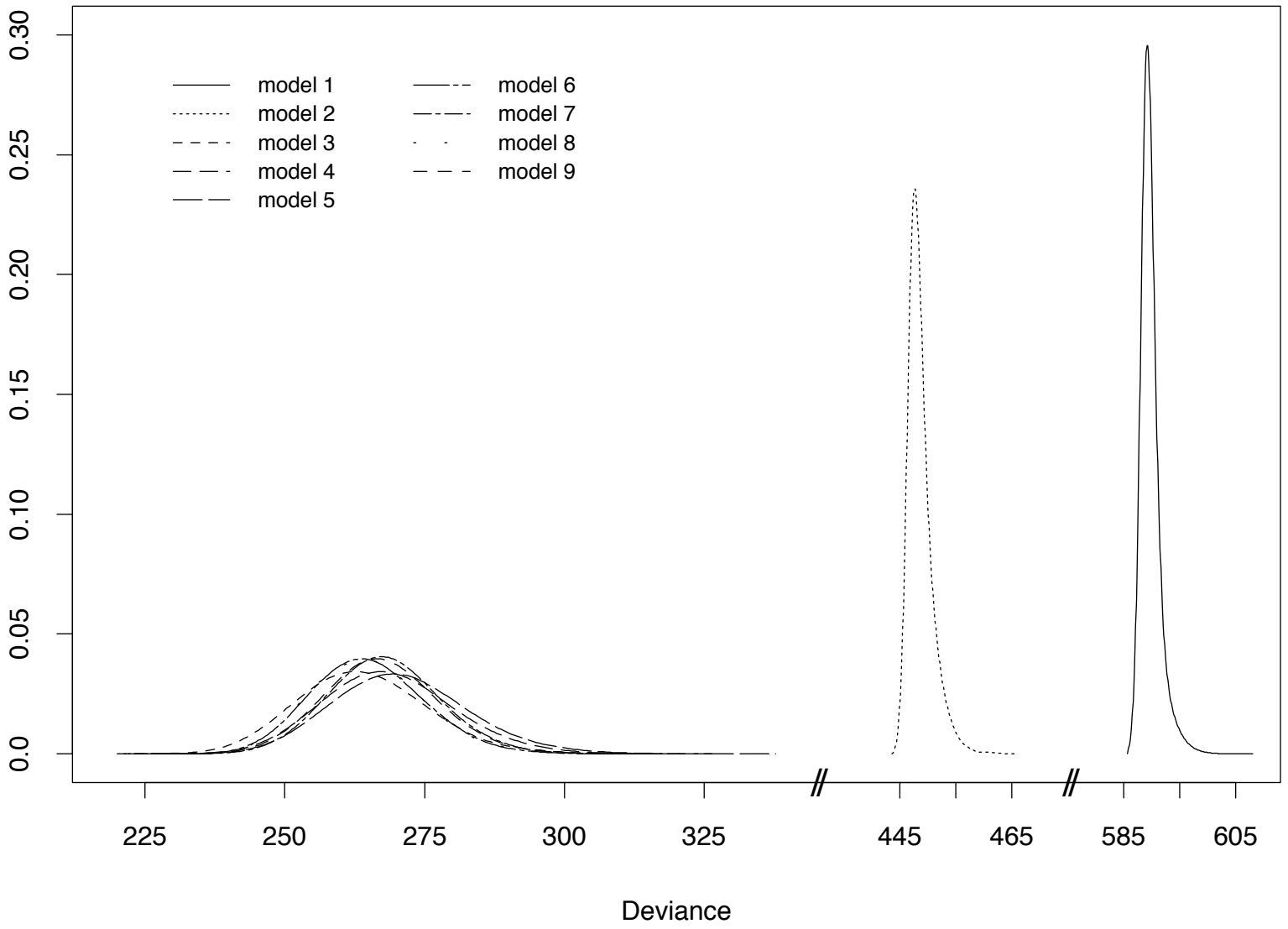
Lip cancer example (continued)

Compare the following models for log relative risk of lip cancer, $\log \lambda_i$

Pooled models:		
const	[1]	α
cov	[2]	$\alpha + \beta x_i$
Random effects :		
exch	[3]	$\alpha + \psi_i$
exch + cov	[4]	$\alpha + \beta x_i + \psi_i$
CAR	[5]	ϕ_i
CAR + cov	[6]	$\beta x_i + \phi_i$
exch + CAR	[7]	$\psi_i + \phi_i$
exch + CAR + cov	[8]	$\beta x_i + \psi_i + \phi_i$
Saturated model:	[9]	α_i

where ϕ_i and ψ_i are random effects representing spatially structured and unstructured heterogeneity, respectively

Posterior distribution of deviance for lip cancer models



DIC results for lip cancer example

Model	$\overline{D}$	$D(\overline{\theta})$	p_D	DIC	$\overline{D}$	$D(\overline{\theta})$	p_D	DIC	
1	589.7	588.7	1.0	590.7	6	267.6	238.2	29.4	297.0
2	448.7	446.6	2.1	450.8	7	264.5	232.0	32.5	297.0
3	268.6	224.8	43.8	312.4	8	267.0	236.7	30.3	297.3
4	270.2	230.5	39.7	309.9	9	264.0	211.1	52.9	316.9
5	264.9	233.3	31.6	296.5					

- Correct degrees of freedom for Model 1 = 1; $p_D = 1.0$
- Correct degrees of freedom for Model 2 = 2; $p_D = 2.1$
- Correct degrees of freedom for Model 9 = 56; $p_D = 52.9$
- Models with random effects have lower DIC than those without (3-8 versus 1, 2, 9)
- Models with spatial random effects (5-8) are virtually indistinguishable, and have lower DIC than models with only unstructured random effects (3, 4)
- Adding heterogeneity random effects to the spatial models has virtually no impact on DIC or on $p_D \Rightarrow$ spatial random effects are accounting for virtually all the residual variation between areas
- Looking at model 5, the 56 spatial random effects are 'worth' approximately 32 degrees of freedom

Which plug-in estimate to use in p_D?

- Calculation of p_D involves calculating a point estimate of the parameters in the sampling distribution (denoted generically by $\tilde{\theta}$) in order to calculate the plug-in deviance $D(\tilde{\theta})$
- So far, we have specified $\tilde{\theta} = \overline{\theta}$, the posterior mean of θ
- Posterior median or mode of θ might also represent reasonable point estimates
- Could also use functions of point estimates, depending on how the sampling distribution is parameterised, e.g.
  - WINBUGS currently uses posterior mean of stochastic 'parents' of θ, i.e. if there are stochastic nodes ψ such that $\theta = f(\psi)$, then $D(\tilde{\theta}) = D(f(\overline{\psi}))$ — see next slide
- **p_D is not invariant to these different reparameterisations**
- p_D can be be negative if posterior of ψ is very non-normal and so $f(\overline{\psi})$ does not provide a very good estimate of θ.
- Also can get negative p_D if non-log-concave sampling distribution and strong prior-data conflict

Example of non-invariant p_D

```
y1 ~ dbin(theta[1],n)  # sampling distribution
y2 ~ dbin(theta[2],n)  # sampling distribution

# theta[1] is stochastic
theta[1] ~ dunif(0, 1)

# psi is stochastic parent of theta[2]
theta[2] <- exp(psi)/(1+exp(psi)) # i.e. logit(theta[2])=psi
psi ~ dlogis(0, 1) # equivalent to uniform prior on theta[2]
```

To calculate the plug-in deviance $D(\tilde{\theta})$, WinBUGS will use

- posterior mean of `theta[1]` to evaluate binomial deviance for `y1`
- the inverse logit transformed value of the posterior mean of `psi` to evaluate the binomial deviance for `y2`

Example of non-invariant p_D (continued)

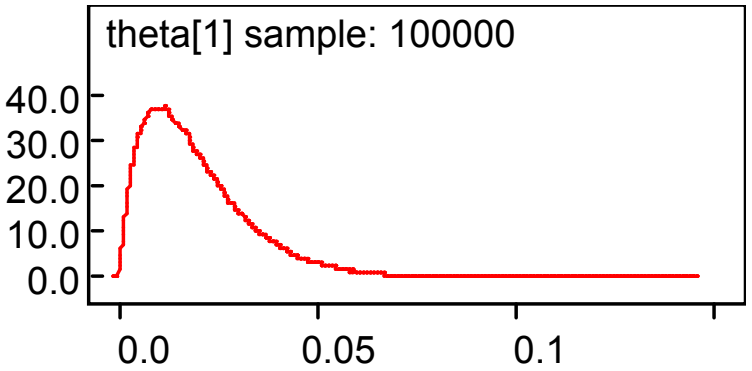
- Suppose we observe $y = 1$ successes out of $n = 100$ Bernoulli trials, so that $y \sim \text{Binomial}(\theta, n)$
- Setting `y1=1`, `y2=1` and `n=100` in above WinBUGS model produces the following output

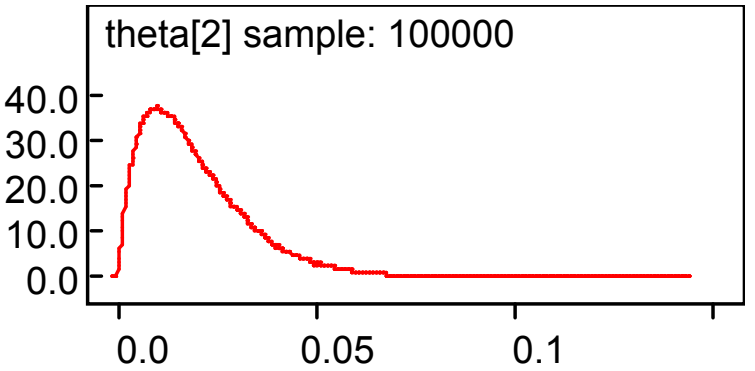
	Dbar	Dhat	pD	DIC
y1	3.12	2.57	0.55	3.67
y2	3.12	2.18	0.94	4.07

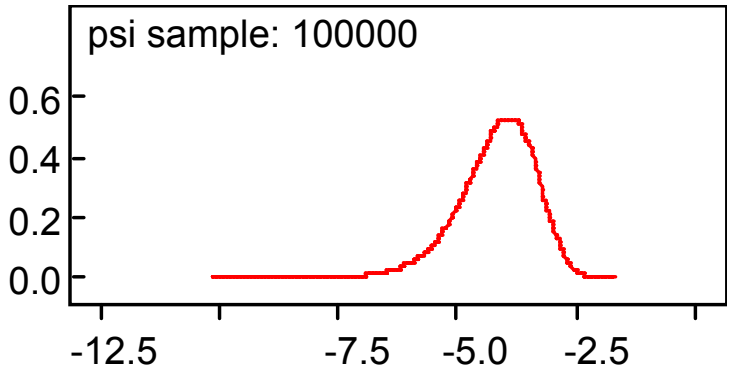
node	mean	sd	2.5%	median	97.5%
psi	-4.18	0.81	-6.032	-4.10	-2.86
theta[1]	0.02	0.01	0.0024	0.016	0.054
theta[2]	0.02	0.01	0.0024	0.016	0.054

- Mean deviances (`Dbar`) and posteriors for `theta[1]` and `theta[2]` are the same, but mean of `theta[1]` is poor plug-in estimate since posterior is skewed

Posterior distributions of `theta` and `psi`







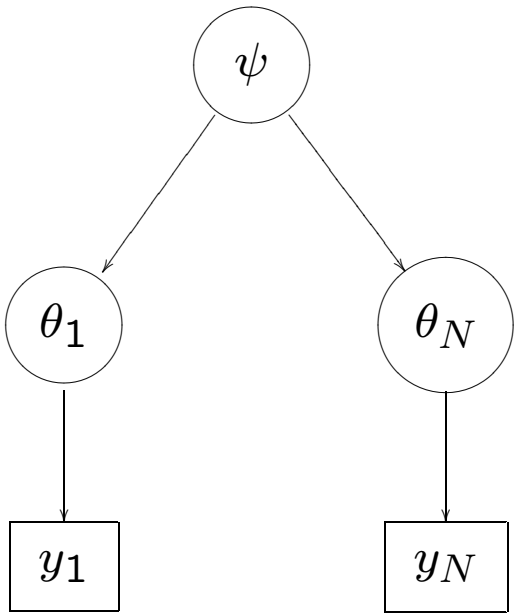
Summary of DIC

- DIC works well for most non-hierarchical and hierarchical linear and generalized linear modelling problems
- Need to be careful with highly non-linear models, where posterior means may not lead to good plug-in estimates
  - WINBUGS users are free to calculate plug-in deviance themselves: could dump out posterior means or medians of 'direct' parameters in likelihood, then calculate deviance outside WINBUGS or by reading posterior means/medians in as data and checking `deviance` in `node info`
- Same problem arises when using mixture models as sampling distribution (ok as prior) and other models involving discrete parameters in the sampling distribution
  - not clear what plug-in estimate to use for discrete parameter
  - WINBUGS does not calculate DIC for such models
- Care needed when using DIC for models with missing data as likelihood not uniquely defined

See also

`www.mrc-bsu.cam.ac.uk/bugs/winbugs/dicpage.shtml`

Changing the 'focus' of a hierarchical model



Hierarchical model:

$$p(y, \theta, \psi) = p(y|\theta)p(\theta|\psi)p(\psi)$$

Marginal distribution of data:

$$p(y) = \int_{\Theta} p(y|\theta)p(\theta)d\theta = \int_{\Psi} p(y|\psi)p(\psi)d\psi$$

depending on whether 'focus' is Θ or Ψ

- No unique definition of likelihood, prior or complexity in hierarchical models
- Prediction is not well-defined in a hierarchical model without stating the 'focus', i.e. what remains fixed
  - predictions on same units (subject-specific) $\rightarrow p(y|\theta)$ is focus
  - predictions on new/'typical' units (pop-average) $\rightarrow p(y|\psi)$ is focus

Example: Changing the model focus

Suppose we fit a 2-level hierarchical model concerning classes (level 1) within schools (level 2)

- If we were interested in predicting results of future classes in those actual schools, then Θ is the focus and DIC is appropriate for model comparison
- If we were interested in predicting results of future schools then Ψ is the focus and marginal-likelihood (with θ's integrated out) methods such as AIC are appropriate for model comparison

Note: 'focus' is also an issue for non-Bayesian hierarchical models. Vaida and Blanchard (2005) developed a 'conditional' AIC for when focus is random effects — this counts parameters using the 'hat' matrix dimensionality, $p = tr(H)$, and so is restricted to normal linear models.

Model criticism for non-hierarchical models

'Standard' checks based on fitted model, such as

- *residuals:* plot versus covariates, checks for auto-correlations and so on
- *prediction:* check accuracy on external validation set, or cross validation
- etc...

All this applies in Bayesian modelling, but in addition:

- parameters have distributions and so residuals are variables
- should check for conflict between prior and data
- should check for unintended sensitivity to the prior
- using MCMC, have ability to generate replicate parameters and data.

Residuals

- Standardised Pearson residuals

$$r = (y - \mu)/\sigma$$

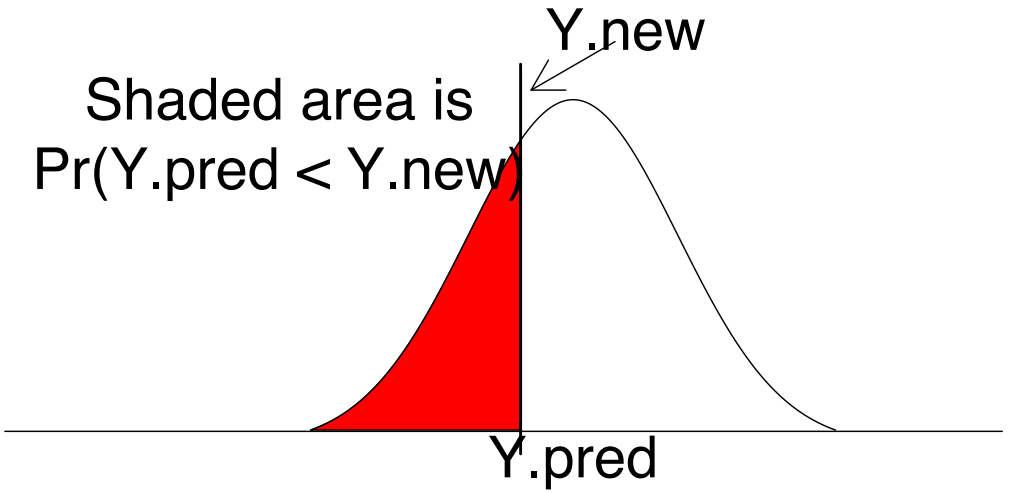
  where μ = E[y], σ^2 = V[y]
- In Bayesian analysis these are random quantities, with distributions
- If assuming Normality, then

$$P(Y) = \Phi[(y - \mu)/\sigma]$$

  has a Uniform[0,1] distribution under true μ and σ
- If we want single value for testing distributional shape, or plotting against covariates, fitted values etc., could use:
  - posterior mean of the standardised residuals, $\mathbb{E}(r)$
  - plug-in posterior means, $r = [y - \mathbb{E}(\mu)]/\mathbb{E}(\sigma)$ (calculate external to WinBUGS)

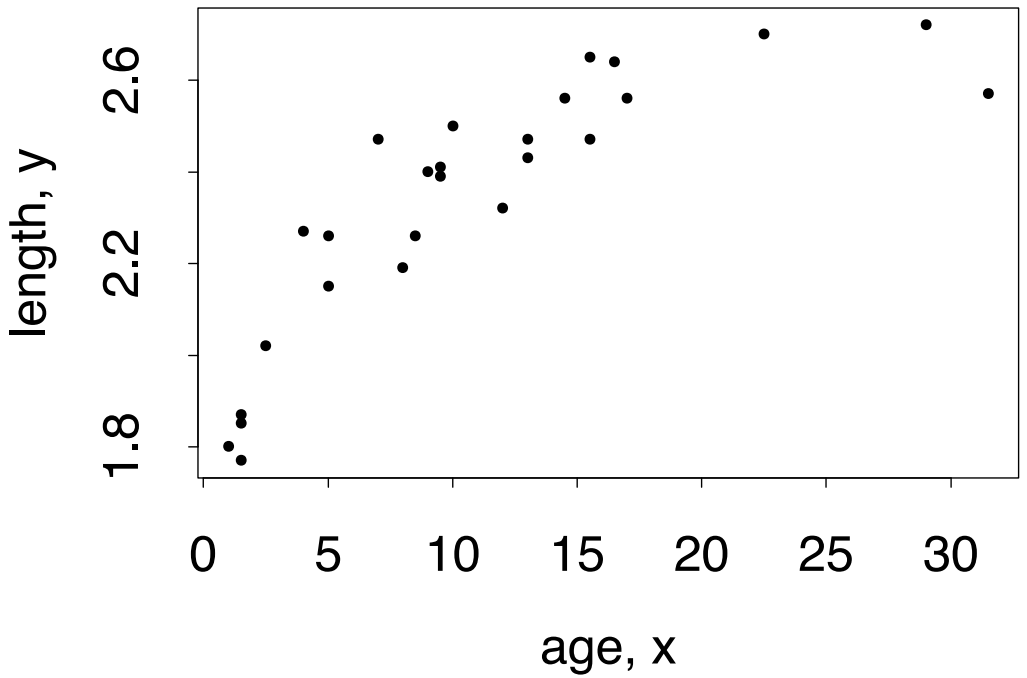
Cross validation and predictive checks

- Ideally would like to test fit on new data y^{new} (cross-validation)
  - Could then calculate expectation of *P*-value $\Phi[(y^{\text{new}} - \mu)/\sigma]$ under current posterior distribution, and see whether extreme
  - Equivalently, could **predict** new observation y^{pred}, and estimate *P*-value by posterior probability that $y^{\text{pred}} < y^{\text{new}}$
- Second approach more general (doesn't require Normality)
  - Each can be approximated by just using current data instead of new data (conservative)



Example: Dugongs

Data on length (y_i) and age (x_i) measurements for 27 dugongs (sea cows) captured off the coast of Queensland



Dugongs model

A frequently used nonlinear growth curve with no inflection point and an asymptote as x_i tends to infinity is

$$\begin{aligned} y_i &\sim \text{Normal}(\mu_i, \sigma^2) \\ \mu_i &= \alpha - \beta\gamma^{x_i} \end{aligned}$$

where $\alpha, \beta > 0$ and $\gamma \in (0, 1)$

Vague prior distributions with suitable constraints may be specified:

$$\begin{aligned} \alpha &\sim \text{Uniform}(0, 100) \\ \beta &\sim \text{Uniform}(0, 100) \\ \gamma &\sim \text{Uniform}(0, 1) \\ 1/\sigma^2 &\sim \text{Gamma}(0.001, 0.001) \end{aligned}$$

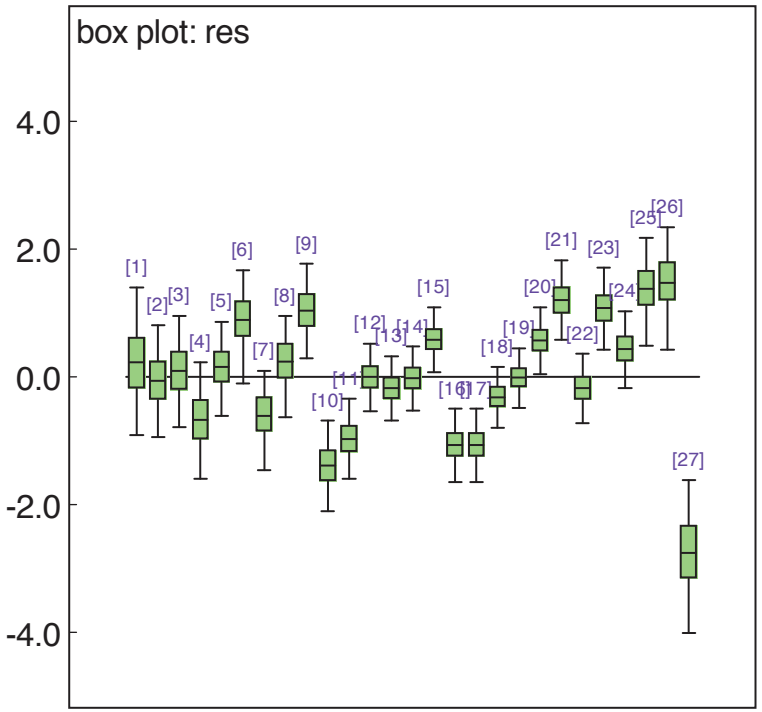
BUGS code for calculating residuals

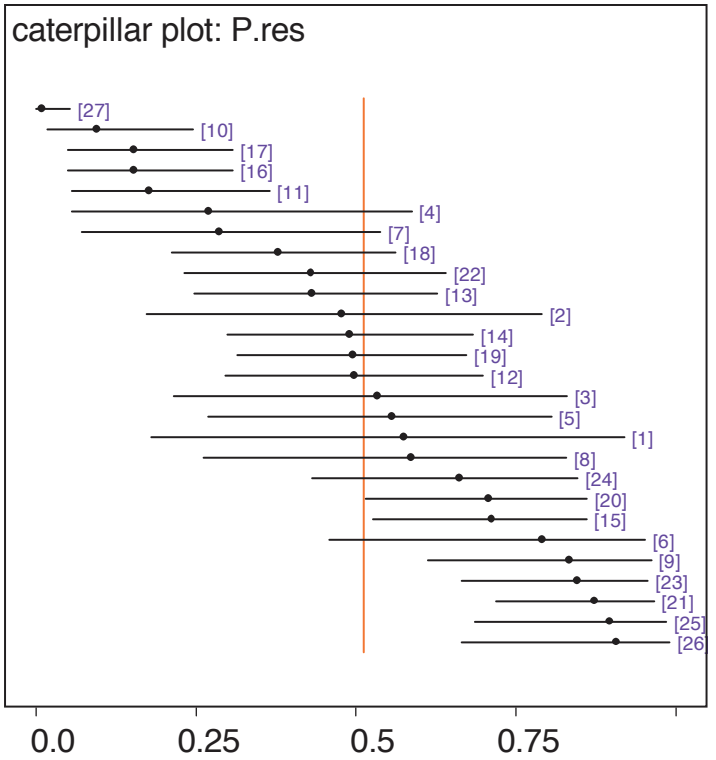
```
for (i in 1:N){
 y[i]     ~   dnorm(mu[i], inv.sigma2)
 mu[i]  <-    alpha - beta * pow(gamma, x[i])
 res[i] <- (y[i] - mu[i])/sigma  # standardised residual
 p.res[i] <- phi(res[i])         # p-value

 # predictive p-value
 y.pred[i]  ~ dnorm(mu[i], inv.sigma2) # re-predict each obs.
 p.pred[i] <- step(y[i] - y.pred[i])   # post. mean = p-value
}
alpha        ~ dunif(0,100)
beta         ~ dunif(0,100)
gamma        ~ dunif(0, 1)
inv.sigma2   ~ dgamma(0.001, 0.001)
```

- Monitor `res`, `p.res`, `y.pred` and `p.pred`

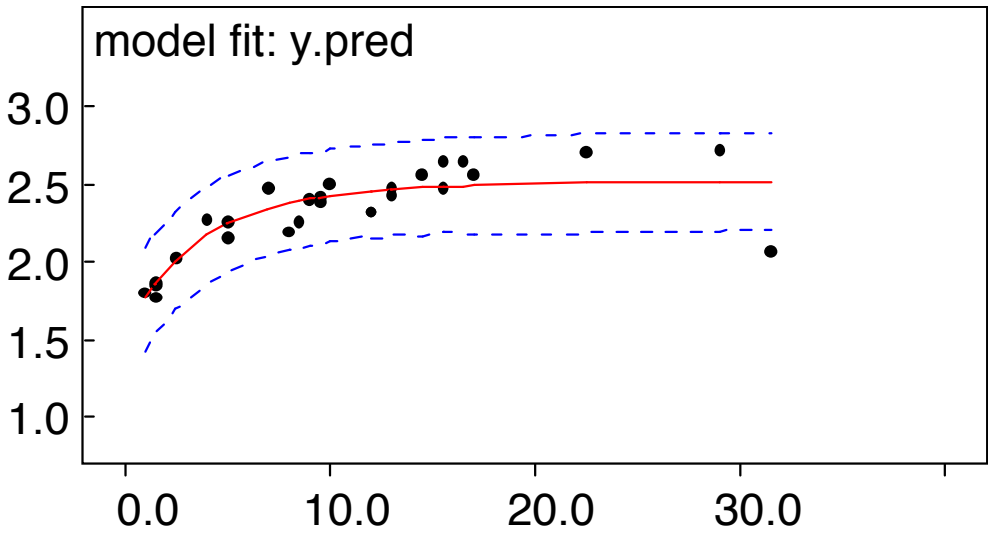
Dugongs: residuals

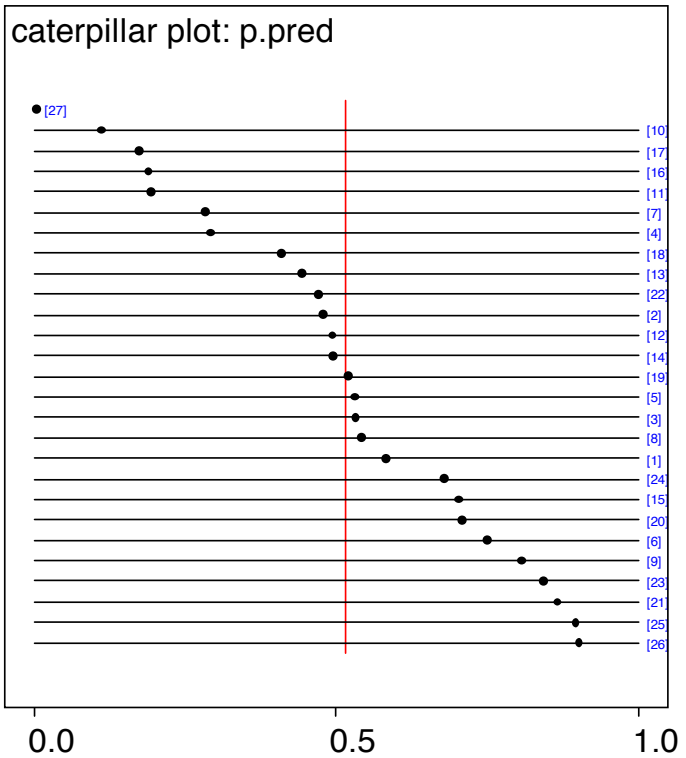
Note — last data point has been deliberately altered to be outlier





Dugongs: prediction as model checking





Model criticism for hierarchical models

Usually interested in 2 things:

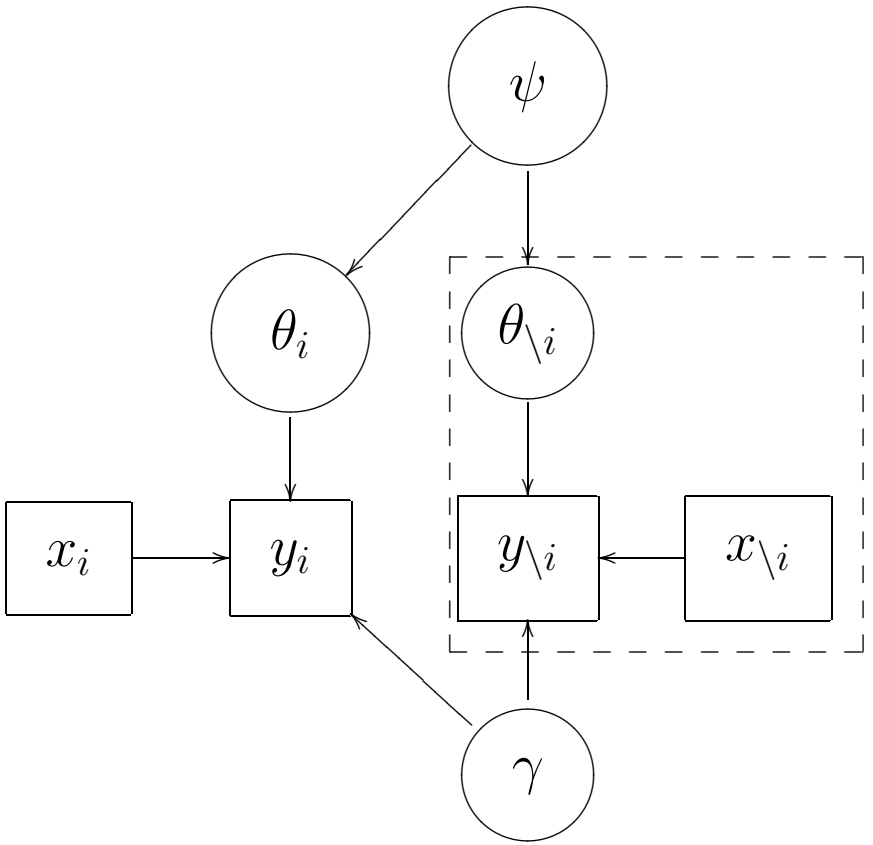
1. How well does the model fit the data?
   - can use standard residual checks as for non-hierarchical models
2. How well do the random effects fit their assumed distribution?
   - can produce q-q plots of random effects (single realisations from posterior; posterior mean)

- But, confounding across levels means it can be difficult to isolate the cause of any apparent problems.
  - a few unusual observations may cause a unit to appear strange
  - conversely, sparse data within units may be labelled as individually outlying when in reality the entire unit is at fault.
- Can also use cross-validation/predictive checks to detect unusual level 2 units

‘Simple’ hierarchies

Suppose data on each unit are grouped (one obs. per unit)

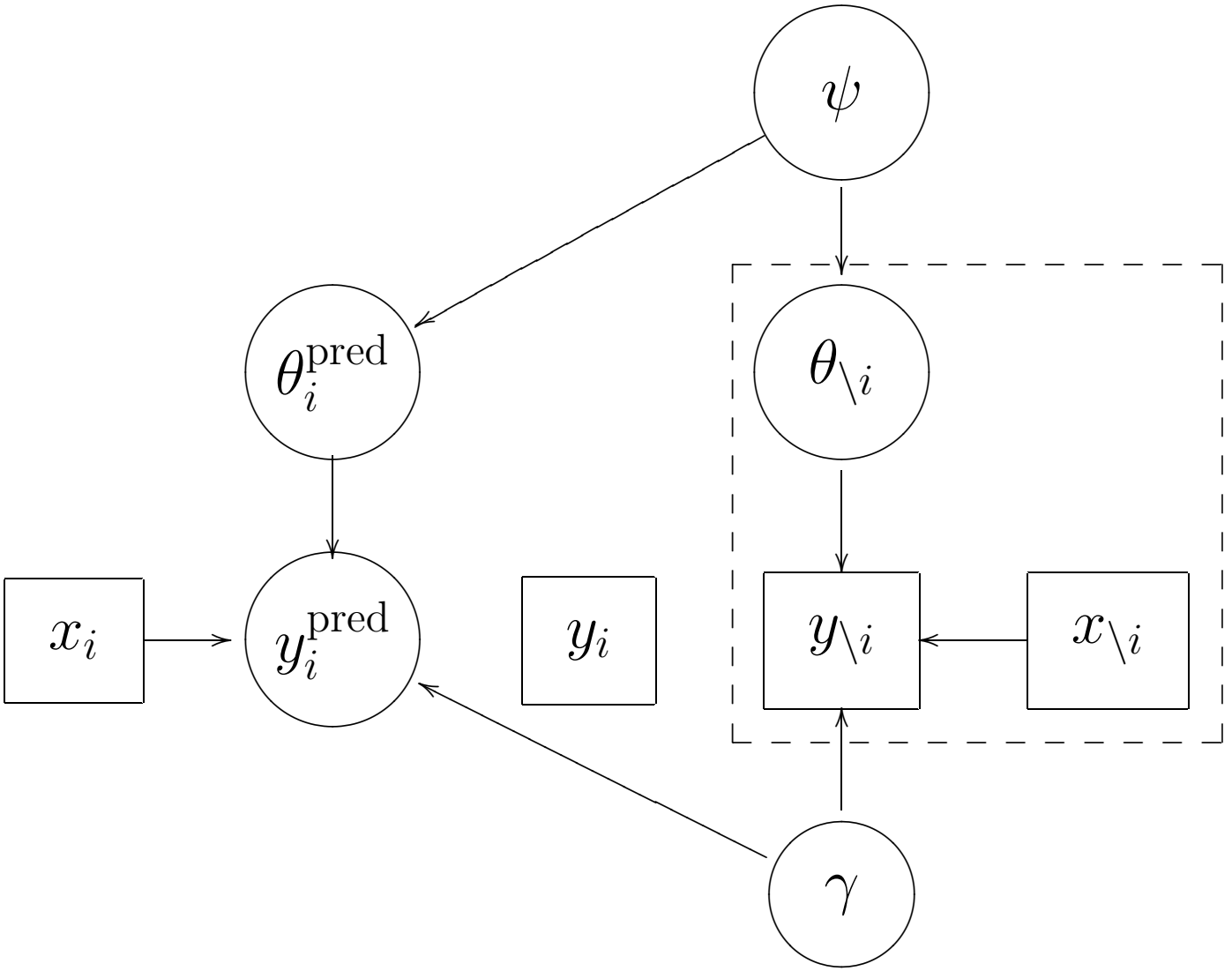
- Simultaneous criticism of Levels 1 and 2
- Confounding means we do not know whether problems are due to mis-specification of likelihood (level 1) or prior (level 2).



- y_i is response for unit i
- x_i are covariate values
- γ represents regression coefficients
- θ_i are unit-specific random effects
- ψ represent the hyper-parameters
- ‘$\setminus i$’ denotes all units except i^{th}

Cross-validatory model criticism for simple hierarchies

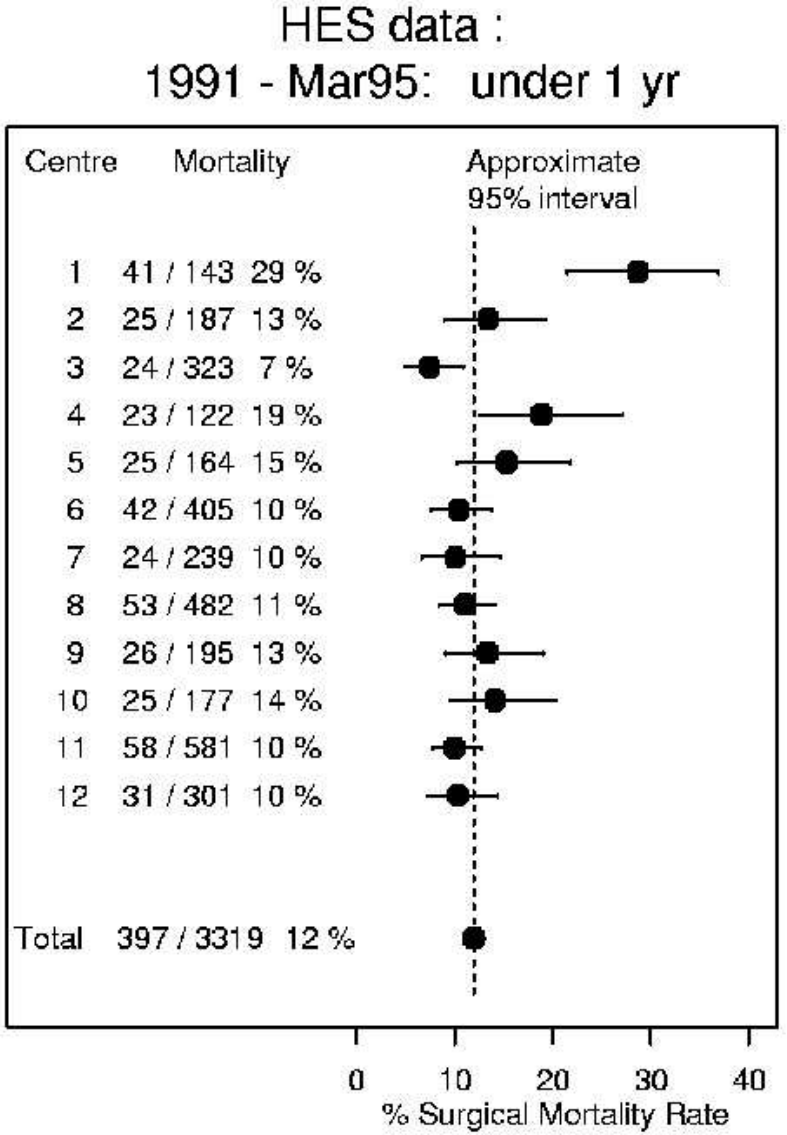
- Delete one **unit** at a time and re-predict



Cross-validatory model criticism for simple hierarchies

- Simulate $\psi, \gamma, \theta_{\setminus i}$ from posterior $p(\psi, \gamma, \theta_{\setminus i} | y_{\setminus i})$
- Simulate θ_i^{pred} from its posterior predictive distribution $p(\theta_i^{\text{pred}} | \psi, \theta_{\setminus i})$
- Simulate replicate observation Y_i^{pred} from its conditional distribution $p(Y_i^{\text{pred}} | \gamma, \theta_i^{\text{pred}}, x_i)$
- Summarise conflict between y_i and y_i^{pred} by Bayesian p-value $\Pr(Y_i^{\text{pred}} > y_i | y_{\setminus i})$
- If data are discrete, use mid p-value $\Pr(Y_i^{\text{pred}} > y_i | y_{\setminus i}) + \frac{1}{2}\Pr(Y_i^{\text{pred}} = y_i | y_{\setminus i})$

Example: Bristol Royal Infirmary data



Model for Bristol Royal Infirmary data

$$\begin{aligned} Y_i &\sim \text{Binomial}(\pi_i, n_i) \\ \text{logit}(\pi_i) &= \phi_i \\ \phi_i &\sim \text{N}(\mu, \sigma^2). \end{aligned}$$

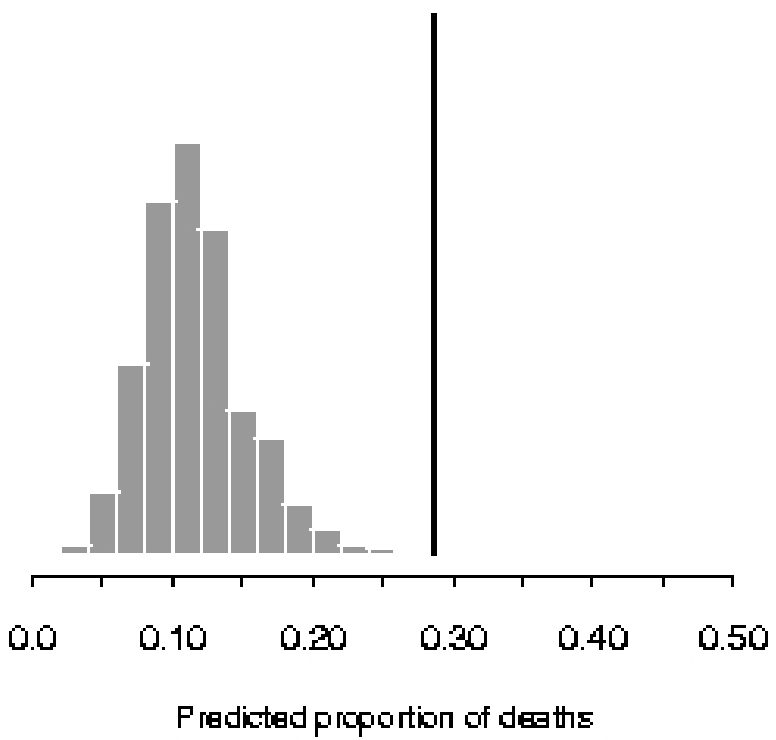
Independent uniform priors are assumed for μ and σ.

- Each hospital is removed in turn from the analysis, the parameters re-estimated, and the observed deaths y_i^{obs} compared to the predictive distribution of $Y_i^{\text{pred}}|y_{\setminus i}$ to give the mid p-value

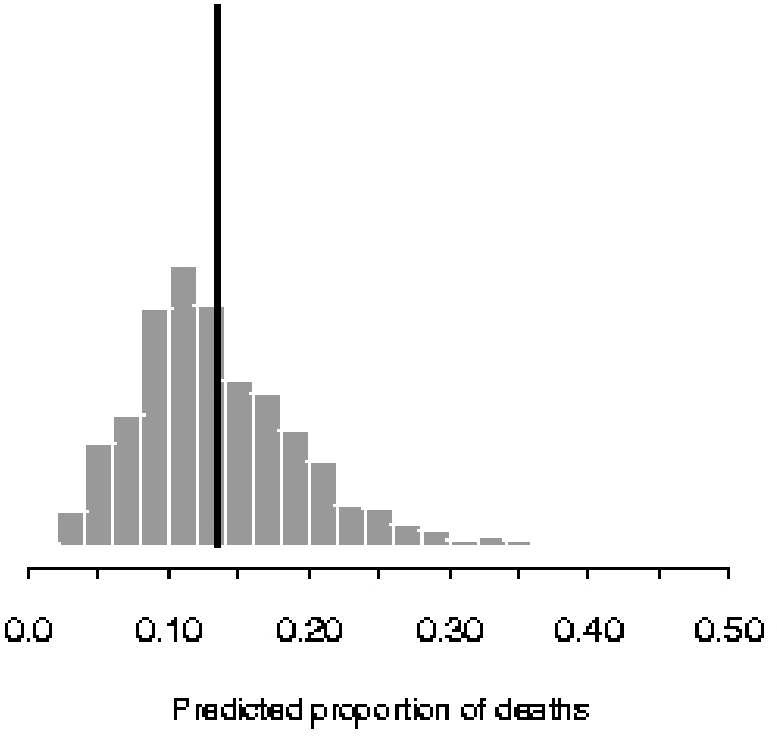
$$p_{\text{mid}} = \Pr(Y_i^{\text{pred}} > y_i^{\text{obs}}|y_{\setminus i}) + \frac{1}{2}\Pr(Y_i^{\text{pred}} = y_i^{\text{obs}}|y_{\setminus i}).$$

Cross-validation predictive distributions

Hospital 1: $p_{\text{mid}} = 0.002$

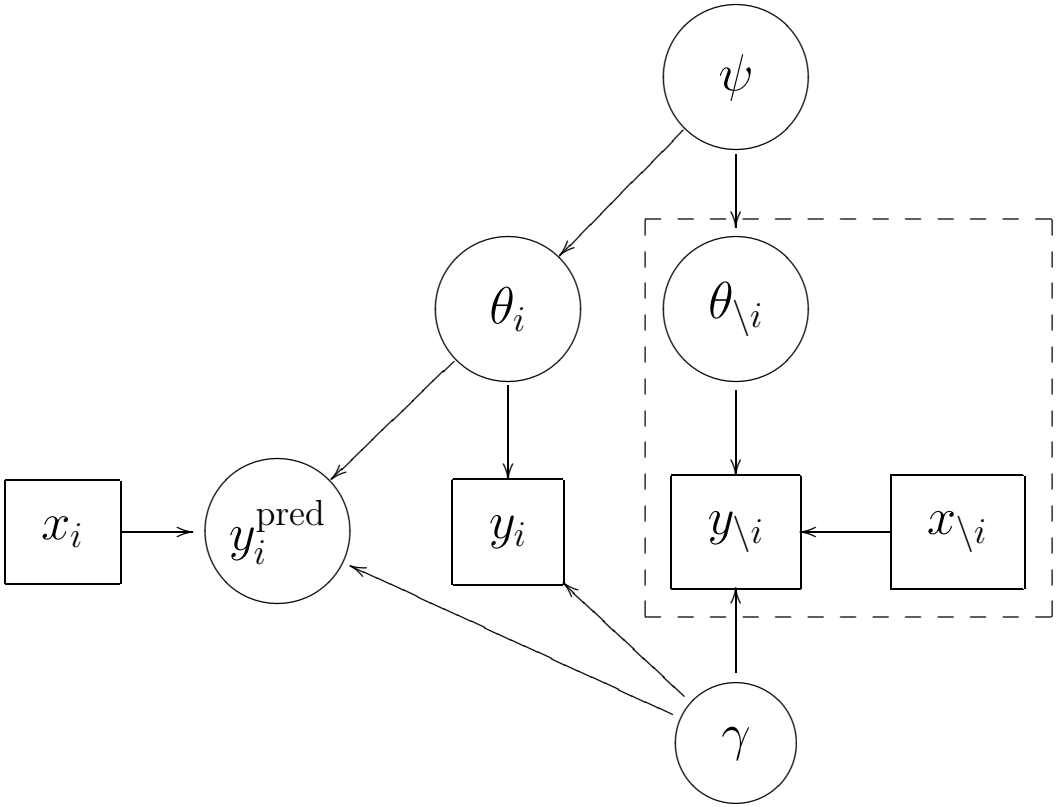


Hospital 2: $p_{\text{mid}} = 0.442$



Approximations to cross-validation I

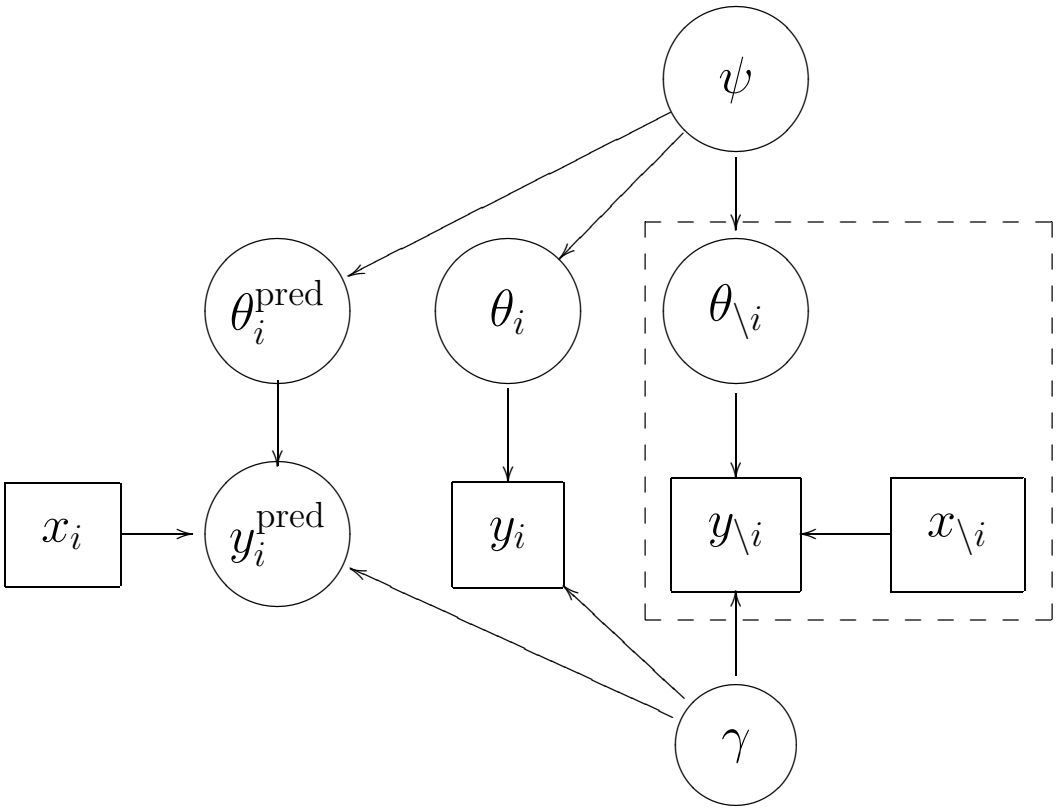
Posterior prediction



- ψ, γ and θ are generated conditional on the full data y
- Y_i^{pred} is generated at each iteration from its conditional distribution $p(Y_i^{\text{pred}}|\theta_i, \gamma)$
- Conservative

Approximations to cross-validation II

'Mixed' (Parameter + Data) prediction (Marshall and Spiegelhalter, 2003)



- ψ, γ and θ are generated conditional on the full data y
- θ_i is ignored and a new replicate θ_i^{pred} is generated from $p(\theta_i^{\text{pred}}|\theta_{\setminus i}, \psi, y)$
- Y_i^{pred} is then generated conditional on θ_i^{pred}

BUGS code for mixed and posterior predictive checks using Bristol data

```
for( i in 1 : N ) {
    r[i] ~ dbin(p[i], n[i])
    logit(p[i]) <- q[i];   q[i] ~ dnorm(mu, tau)
 ### Posterior predictions
    r.pred[i] ~ dbin(p[i], n[i])
    # use mid pvalues as discrete outcomes:
    ppost[i] <- step(r.pred[i] - r[i] - 0.001)
                + 0.5*equals(r.pred[i], r[i])
 ### Mixed predictions
    r.mixed[i] ~ dbin(p.pred[i], n[i])
    logit(p.pred[i]) <- q.pred[i];  q.pred[i] ~ dnorm(mu, tau)
    # use mid pvalues
    pmixed[i] <- step(r.mixed[i] - r[i] - 0.001)
                 + 0.5*equals(r.mixed[i], r[i])
    hospital[i] <- i    # index used for plotting
}
# Hyperpriors
tau ~ dgamma(0.001,0.001);  mu ~ dnorm(0.0,0.000001)
```

Results

node	mean	sd	MC error	2.5%	median	97.5%
ppost[1]	0.1986	0.3872	0.003348	0.0	0.0	1.0
ppost[2]	0.4587	0.4815	0.003535	0.0	0.0	1.0
ppost[3]	0.6777	0.4517	0.003982	0.0	1.0	1.0
ppost[4]	0.3174	0.4479	0.003746	0.0	0.0	1.0
.....						
ppost[11]	0.5583	0.4866	0.003721	0.0	1.0	1.0
ppost[12]	0.5604	0.4825	0.004207	0.0	1.0	1.0
pmixed[1]	0.0164	0.1253	0.001104	0.0	0.0	0.0
pmixed[2]	0.4394	0.4864	0.004186	0.0	0.0	1.0
pmixed[3]	0.8967	0.2983	0.002345	0.0	1.0	1.0
pmixed[4]	0.1514	0.3497	0.003222	0.0	0.0	1.0
....						
pmixed[11]	0.7398	0.4353	0.003597	0.0	1.0	1.0
pmixed[12]	0.7082	0.4472	0.003614	0.0	1.0	1.0

- Conservatism of posterior p-values is clear
- Bristol is hospital 1

More 'complex' hierarchies

- Suppose we have repeated observations per unit:

$$y_{ij} \sim p(y_{ij}|\gamma, X_{ij}, \theta_i) \quad \text{level 1}$$

$$\theta_i \sim p(\theta_i|\psi) \quad \text{level 2}$$

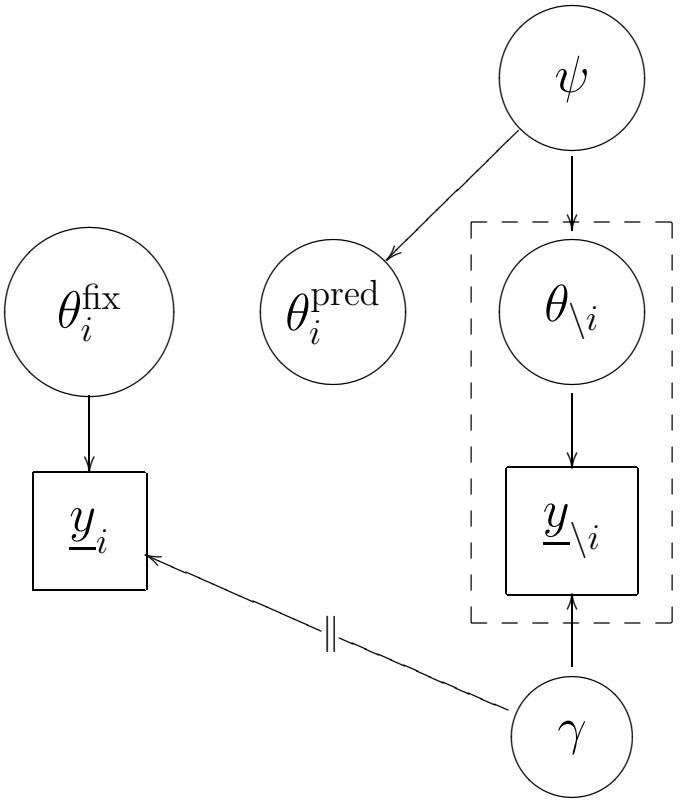
- Enables separate criticism of level 1 (detecting unusual observations using standard residual diagnostics) and criticism of level 2 (detecting unusual units)

Cross validation for more 'complex' hierarchies

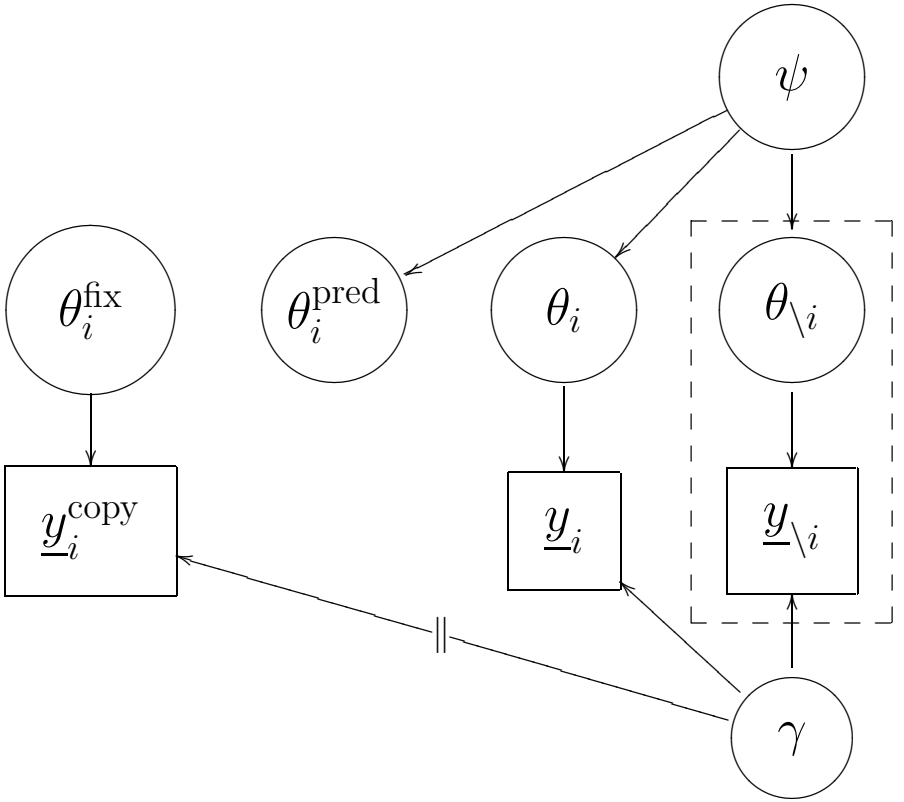
- delete all observations within a given unit, *i*
- fit hierarchical model to data from remaining units and generate $\theta_i^{pred}|\psi, \underline{y}_{\setminus i}$
- generate independent 'fixed effect' estimate θ_i^{fix} using data from unit *i* and non-informative prior on θ_i
- compare posterior distributions of θ_i^{pred} and θ_i^{fix}

Cross validation/Mixed predication for more 'complex' hierarchies

Cross validation



Mixed prediction



Lecture 4. Hierarchical models for Longitudinal data

Introduction and Lecture Outline

- There is huge scope for elaborating the basic hierarchical models discussed in the first lecture to reflect additional complexity in the data, e.g.
  - adding further levels to the hierarchy (pupils within schools within local authorities)
  - adding non-nested (cross-classified) levels (patients within GPs crossed with hospitals)
  - repeated observations on some/all units modelled through linear or generalised linear random effects regressions
  - modelling temporal or spatial structure in data
- In this lecture, we discuss different ways of modelling longitudinal data using two datasets
  - comparison of treatments in an antidepressant clinical trial
  - assessment of cognitive decline with ageing

Longitudinal data

- Arise in studies where individual (or units) are measured repeatedly over time
- For a given individual, observations over time will be typically dependent
- Longitudinal data can arise in various forms:
  - continuous or discrete response
    - discrete response can be binary/binomial, categorical or counts
  - equally spaced or irregularly spaced
  - same or different time points for each individual
  - with or without missing data
  - many or few time points, T
  - many or few individuals or units, n

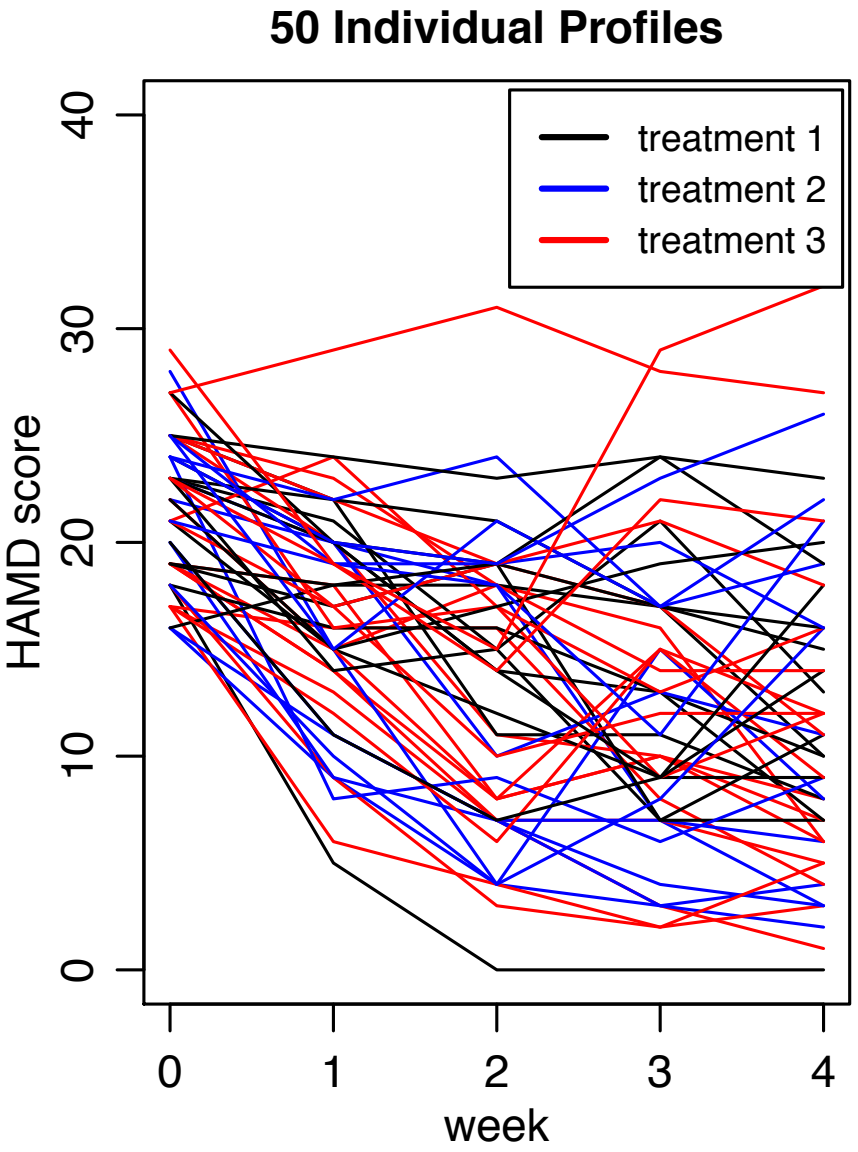
Analysing longitudinal data

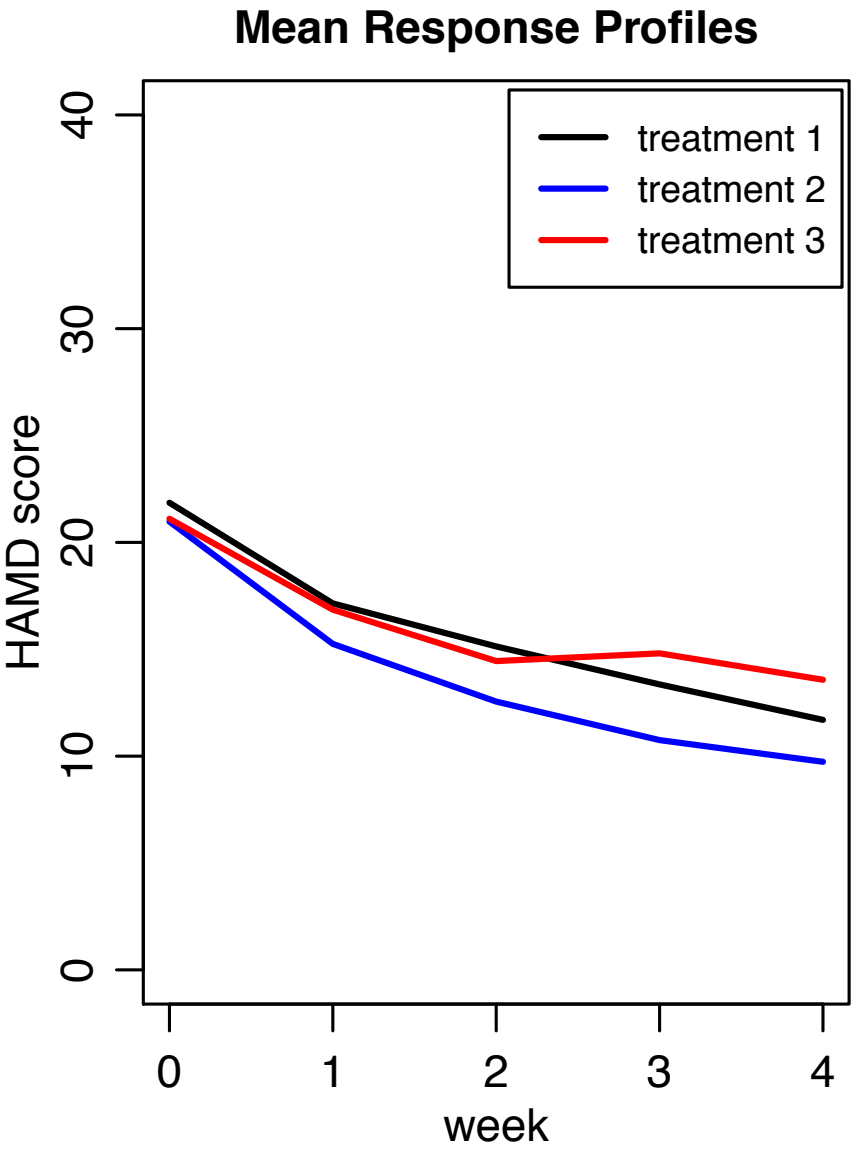
- There are many different ways to analyse longitudinal data
- This is a very big field, so we have to be selective
- The key feature of longitudinal data is the need to account for the dependence structure of the data
- In our first example, we will introduce two common methods:
  - random effects (hierarchical) linear models
  - autoregressive models
- We now introduce the data for the antidepressant clinical trial example (HAMD example)

HAMD Example: antidepressant clinical trial

- 6 centre clinical trial, comparing 3 treatments of depression
- 367 subjects randomised to one of 3 treatments
- Subjects rated on Hamilton depression score (HAMD) on 5 weekly visits
  - week 0 before treatment
  - weeks 1-4 during treatment
- HAMD score takes values 0-50
  - the higher the score, the more severe the depression
- Subjects drop out from week 2 onwards, but for now we
  - ignore the subjects who dropped out
  - analyse the 246 complete cases
- Data was previously analysed by Diggle and Kenward (1994)

HAMD Example: data





HAMD Example: objective

- Study objective: are there any differences in the effects of the 3 treatments on the change in HAMD score over time?
- The variables we will use are:
  - y: Hamilton depression (HAMD) score
  - t: treatment
  - w: week
- For simplicity we will
  - ignore any centre effects
  - assume linear relationships
- The models we will consider are:
  - a non-hierarchical model (standard linear regression)
  - a random effects (hierarchical) model
  - an autoregressive model (AR1)

HAMD Example: a Bayesian (non-hierarchical) linear model (LM)

- Specification:
  - probability distribution for responses:

$$y_{iw} \sim \text{Normal}(\mu_{iw}, \sigma^2)$$

y_{iw} = the HAMD score for individual i in week w (weeks 0,...,4)

  - linear predictor:

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

$treat(i)$ = the treatment indicator of individual i, so it can take values 1, 2 or 3

w = the week of the visit, takes value 0 for visit before treatment and values 1-4 for follow-up visits

- In this model no account is taken of the repeated structure (observations are nested within individuals)

HAMD Example: prior distributions for the unknowns

- Assuming a 'non-informative' (flat) prior distribution for the unknown parameters in a Bayesian LM (or GLM) gives posterior estimates similar to classical maximum likelihood point and interval estimates
- Standard 'non-informative' prior distribution for the unknown parameters in the HAMD model is Uniform on $(\alpha, \beta_1, \beta_2, \beta_3, \log \sigma)$
- This is an *improper* prior density, i.e. it does not integrate to 1
- A vague but *proper* prior for the HAMD model:

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

- Joint posterior is of closed form (Multivariate normal)
  - marginal posteriors may be calculated algebraically

HAMD Example: a Bayesian hierarchical linear model

- Modify LM to allow a separate intercept for each individual:

$$\begin{aligned} y_{iw} &\sim \text{Normal}(\mu_{iw}, \sigma^2) \\ \mu_{iw} &= \alpha_i + \beta_{treat(i)} w \end{aligned}$$

  We are assuming that *conditionally* on α_i, $\{y_{iw}, w = 0, \ldots, 4\}$ are independent

- Assume that all the $\{\alpha_i\}$ follow a *common* prior distribution, e.g.

$$\alpha_i \sim \text{Normal}(\mu_\alpha, \sigma_\alpha^2) \quad i = 1, \ldots, 246$$

  Here we are assuming exchangeability between all the individuals

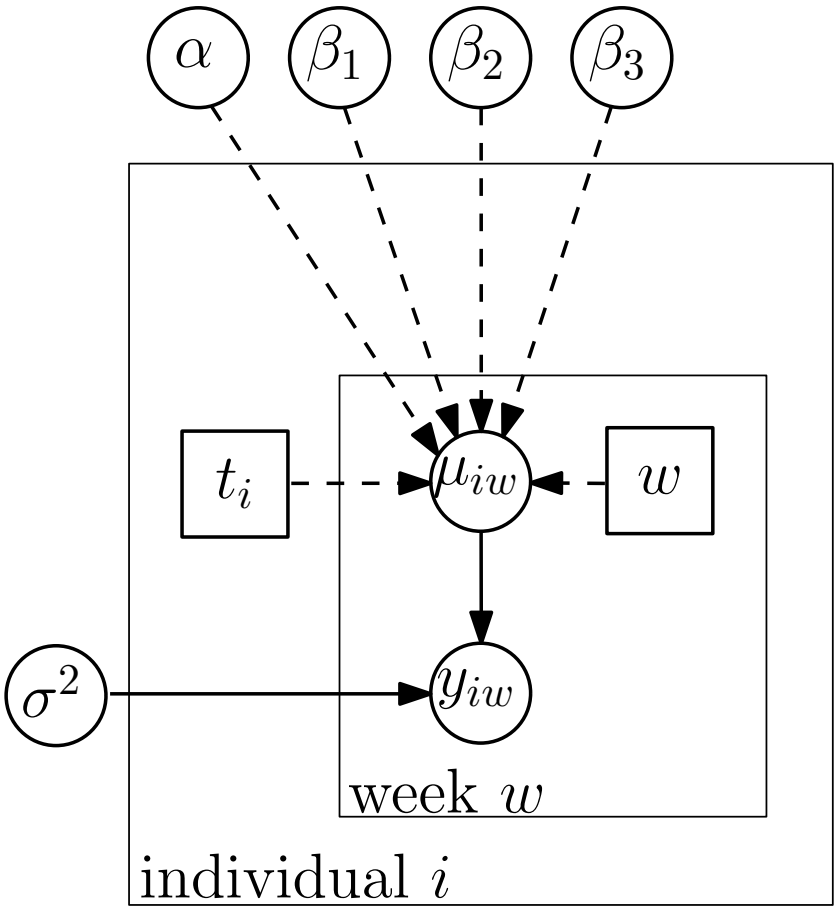
- We may then assume vague priors for the *hyperparameters* of the population distribution:

$$\begin{aligned} \mu_\alpha &\sim \text{Normal}(0, 10000) \\ \sigma_\alpha &\sim \text{Uniform}(0, 100) \end{aligned}$$

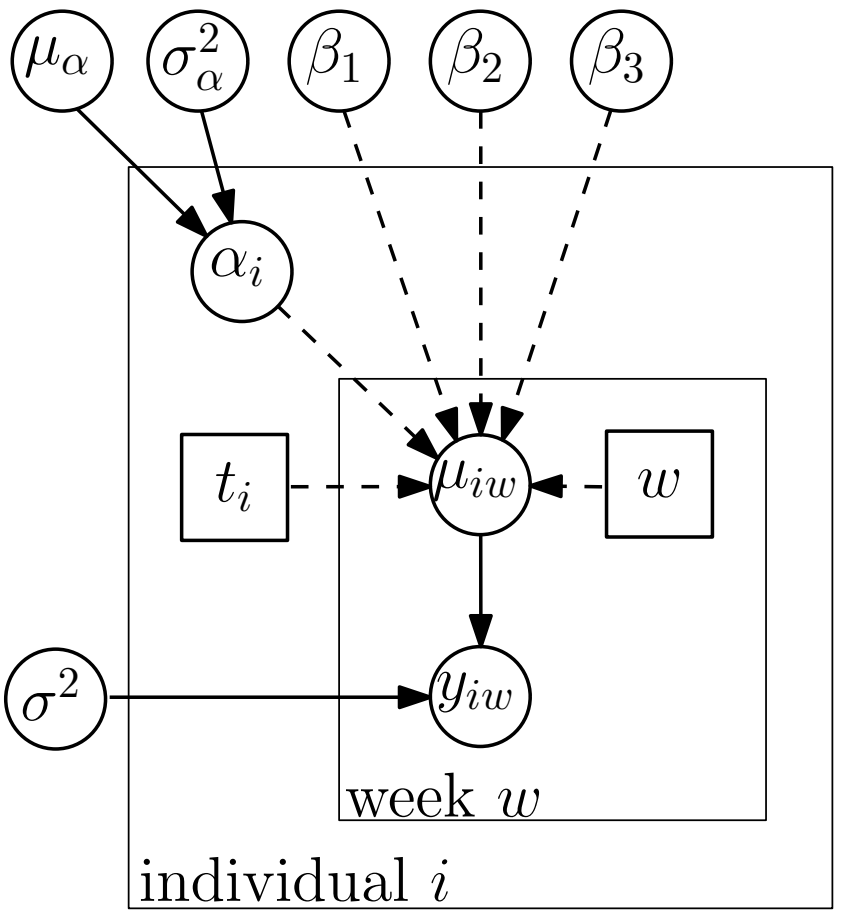
- This is an example of a *Hierarchical LM* or *Linear Mixed Model (LMM)* or *Random Coefficients* model

HAMD Example: DAGs for LM and LMM

non-hierarchical model (LM)



hierarchical model (LMM)



t_i represents the treatment indicator of individual i

HAMD Example: WinBUGS code for LM and LMM

WinBUGS code for non-hierarchical model:

```
model {
  for (i in 1:N) { # N individuals
    for (w in 1:W) { # W weeks
      hamd[i,w]~dnorm(mu[i,w],tau)
      mu[i,w]<-alpha+beta[treat[i]]*(w-1)
    }
  }
  # specification of priors ....
```

WinBUGS code for hierarchical model:

```
model {
  for (i in 1:N) { # N individuals
    for (w in 1:W) { # W weeks
      hamd[i,w]~dnorm(mu[i,w],tau)
      mu[i,w]<-alpha[i]+beta[treat[i]]*(w-1)
    }
    alpha[i]~dnorm(alpha.mu,alpha.tau) # random effects
  }
  # specification of priors ....
```

HAMD Example: WinBUGS code for priors

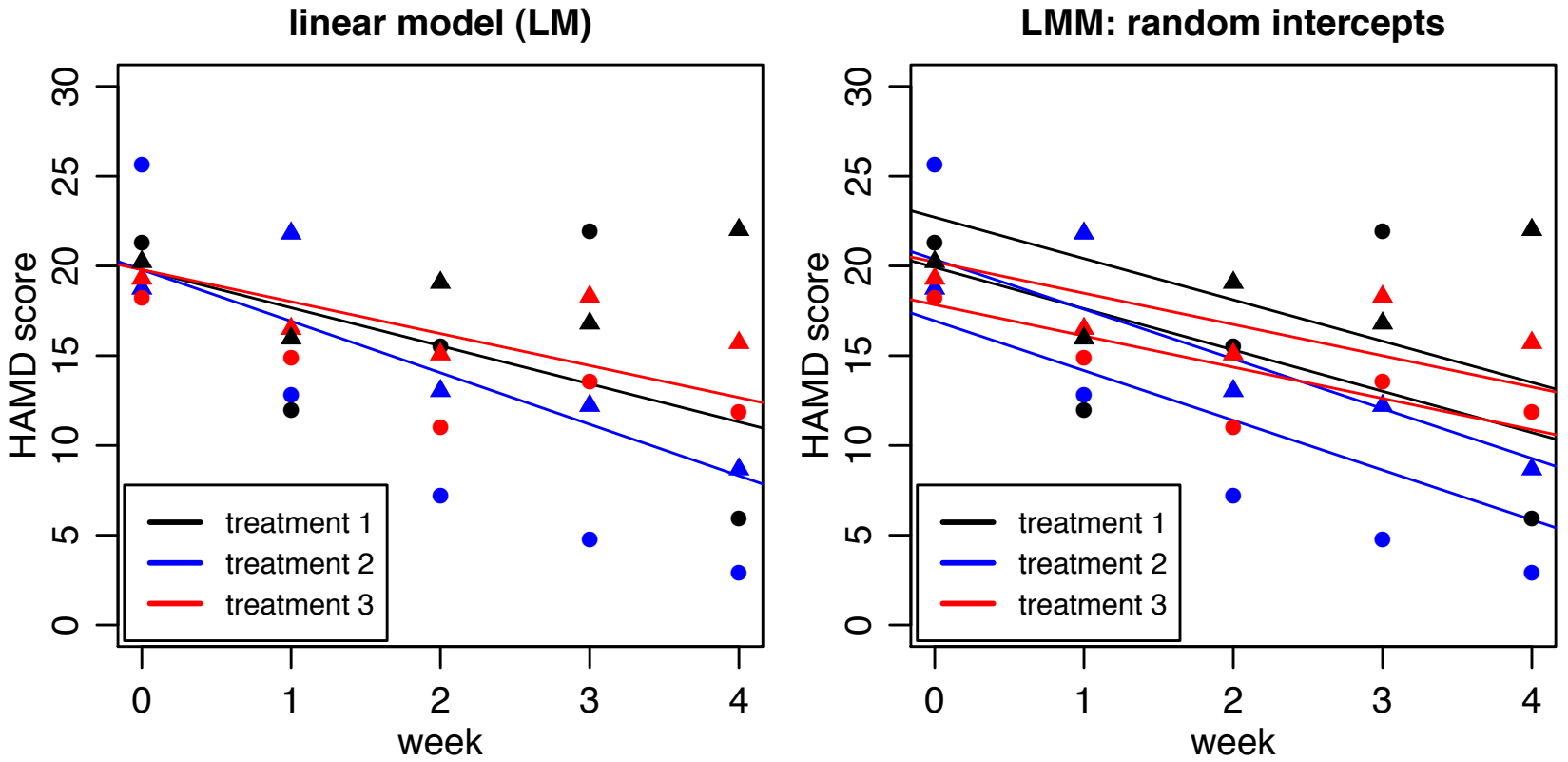
Prior specification for non-hierarchical model:

```
alpha~dnorm(0,0.00001)
for (t in 1:T){  # T treatments
  beta[t]~dnorm(0,0.00001)
  }
tau~dgamma(0.001,0.001)
sigma.sq<-1/tau  # Normal errors
```

Prior specification for hierarchical model:

```
alpha.mu~dnorm(0,0.00001)
alpha.sigma~dunif(0,100) # prior for random effects variances
alpha.sigma.sq<-pow(alpha.sigma,2)
for (t in 1:T){  # T treatments
  beta[t]~dnorm(0,0.00001)
  }
tau~dgamma(0.001,0.001)
sigma.sq<-1/tau  # Normal errors
```

HAMD Example: LM and LMM fitted lines



circles and triangles represent scores for 6 individuals (2 for each treatment)

- LM:
  - 3 regressions lines fitted, 1 for each treatment
  - each treatment has the same intercept, but a different slope
- LMM:
  - each individual has a different regression line
  - but for each treatment, individuals have the same slope

HAMD Example: results for LM and LMM

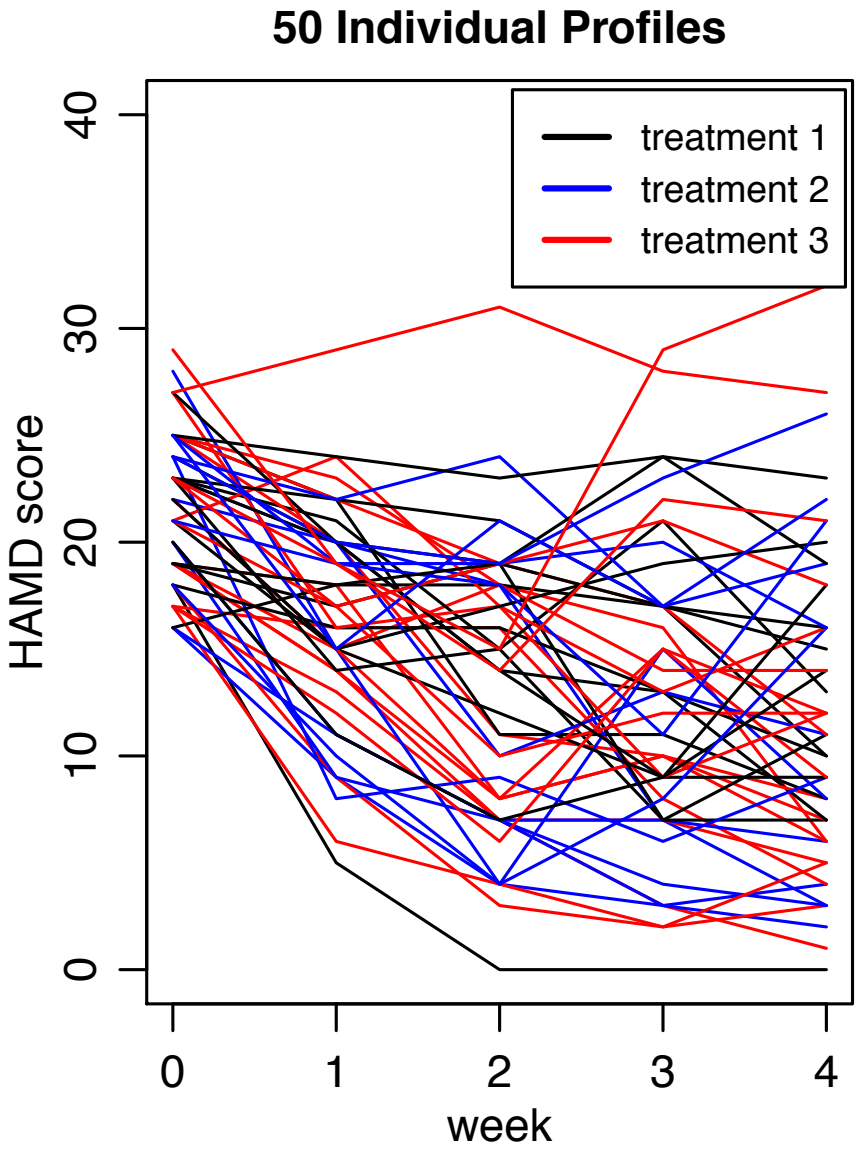
Table: posterior mean (95% credible interval) for the non-hierarchical and hierarchical models fitted to the HAMD data

	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)

Note

- the variability in the intercept in the hierarchical model
- how the residual variance (σ^2) is reduced when random effects are incorporated

HAMD Example: revisiting the data



The plot of the raw data

- indicates that separate intercepts are appropriate
- also suggests including separate slopes

So we add random slopes to the hierarchical model

HAMD Example: adding random slopes

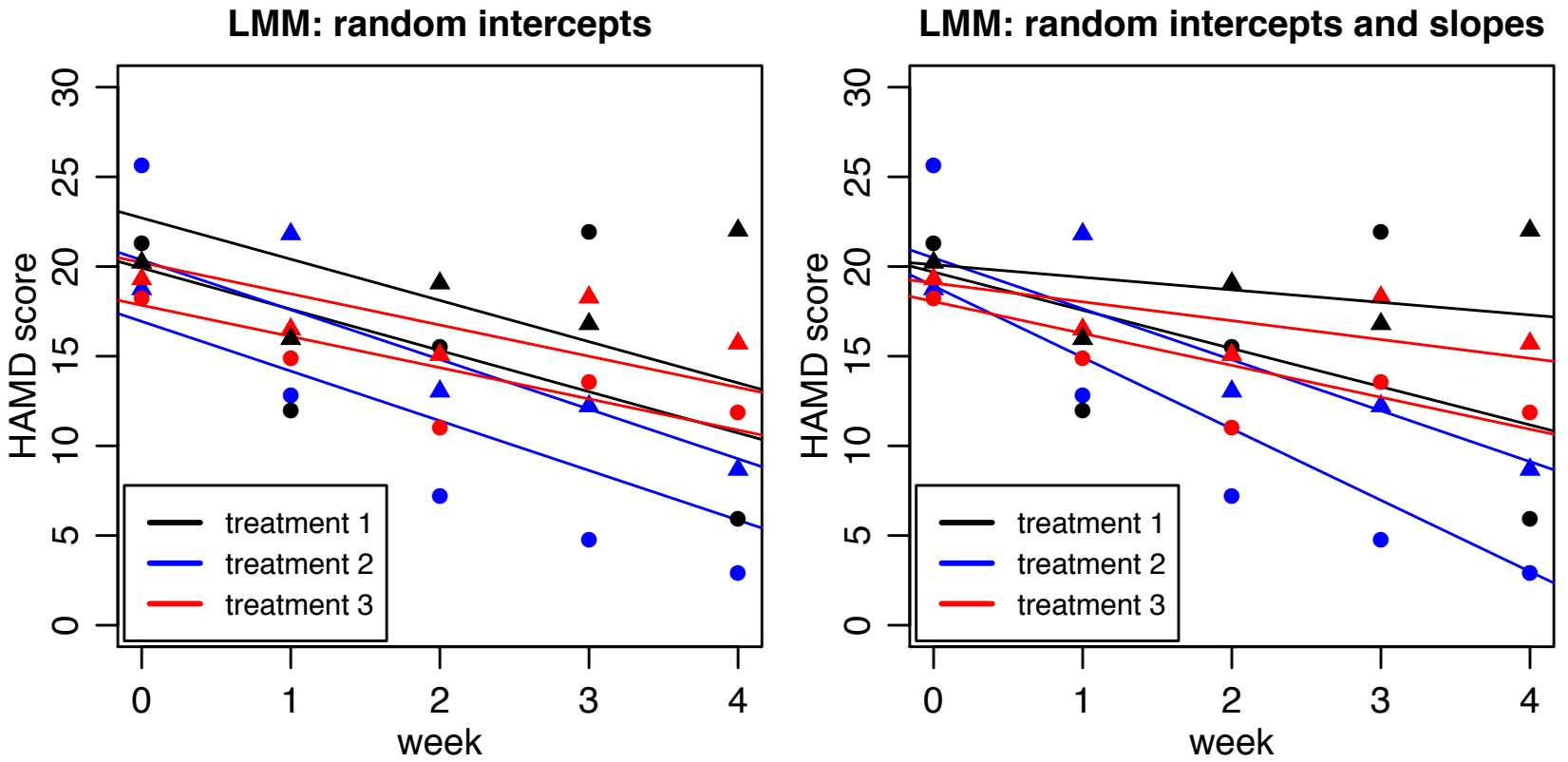
- Modify LMM to allow a separate slope for each individual:

$$\begin{aligned} y_{iw} &\sim \text{Normal}(\mu_{iw}, \sigma^2) \\ \mu_{iw} &= \alpha_i + \beta_{(treat(i),i)} w \end{aligned}$$

- As for the $\{\alpha_i\}$, assume that the $\{\beta_{(1,i)}\}$,$\{\beta_{(2,i)}\}$ & $\{\beta_{(3,i)}\}$ follow *common* prior distributions with vague priors on their *hyperparameters*

```
model {
  for (i in 1:N) {  # N individuals
    for (w in 1:W) {  # W weeks
      hamd[i,w]~dnorm(mu[i,w],tau)
      mu[i,w]<-alpha[i]+beta[treat[i],i]*(w-1)
    }
    alpha[i]~dnorm(alpha.mu,alpha.tau)
    for (t in 1:T){beta[t,i]~dnorm(beta.mu[t],beta.tau[t])}
  }
  # Priors
  for (t in 1:T){  # T treatments
    beta.mu[t]~dnorm(0,0.00001)
    beta.sigma[t]~dunif(0,100)
    beta.sigma.sq[t]<-pow(beta.sigma[t],2)
  }
  # specification of other priors as before ....
```

HAMD Example: random intercepts and slopes



circles and triangles represent scores for 6 individuals (2 for each treatment)

- LMM with random intercepts only:
  - each individual has a different regression line
  - but for each treatment, only intercept varies by individual
- LMM with random intercepts and random slopes:
  - now intercepts and slopes both vary
  - better fit for each individual

HAMD Example: results comparison

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

Alternative to random effects models

- In the linear model for the HAMD example, all the observations are assumed independent
  - clearly unrealistic
- We have dealt with this by introducing random effects to model the within person dependence
  - assumes the observations are independent conditional on the random intercepts and slopes
- Alternatively, we can explicitly model the autocorrelation between weekly visits for each individual using an autoregressive (or transition) model

Autoregressive models - without covariates

- Let $\boldsymbol{y} = (y_1, \ldots, y_T)$ be a time ordered sequence of observations
- Then an autoregressive Gaussian model for $\boldsymbol{y}$ is defined by
  - a time lag p
  - a set of coefficients $\{\gamma_1, \ldots \gamma_p\}$

  such that

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

- Simple case:

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

    or equivalently
  - $y_t \sim N(\gamma_1 y_{t-1}, \sigma_\epsilon^2)$

Autoregressive models - incorporating covariates

- We can incorporate explanatory variables by assuming that the residuals (rather than the observations themselves) are serially correlated
- Example: consider an autoregressive model of order 1, with a single explanatory variable, x, $t = 1, \ldots, n$
  - let $\mu_t = \beta x_t$
  - then $y_t | y_{t-1} = \mu_t + \gamma(y_{t-1} - \mu_{t-1}) + \epsilon_t; \quad \epsilon_t \sim \mathrm{N}(0, \sigma^2)$

    which can be written as

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

HAMD Example: an autoregressive model

- For the HAMD data, we can model the structure using an AR(1)
- Specify:

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

where $\mathcal{R}_{iw}$ follow a first-order autoregressive process defined by

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

- An equivalent form for implementation in WinBUGS is:

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

HAMD Example: AR(1) formulations

- Form implemented in WinBUGS is

$$\begin{aligned} y_{iw} &\sim \text{Normal}(\theta_{iw}, \sigma^2) \\ \theta_{i0} &= \mu_{i0} \\ \theta_{iw} &= \mu_{iw} + \gamma(y_{i(w-1)} - \mu_{i(w-1)}) \quad w \geq 1 \end{aligned}$$

- which can be written as

$$\begin{aligned} y_{i0} &= \mu_{i0} + \epsilon_{i0} \\ y_{iw} &= \mu_{iw} + \gamma(y_{i(w-1)} - \mu_{i(w-1)}) + \epsilon_{iw} \quad w \geq 1 \\ \epsilon_{iw} &\sim \text{Normal}(0, \sigma^2) \quad w = 0, \ldots, 4 \end{aligned}$$

- letting $\mathcal{R}_{iw} = y_{iw} - \mu_{iw}$, rewrite as

$$\begin{aligned} y_{iw} &= \mu_{iw} + \mathcal{R}_{iw} \\ \mathcal{R}_{i0} &= \epsilon_{i0} \\ \mathcal{R}_{iw} &= \gamma\mathcal{R}_{i(w-1)} + \epsilon_{iw} \quad w \geq 1 \\ \epsilon_{iw} &\sim \text{Normal}(0, \sigma^2) \quad w = 0, \ldots, 4 \end{aligned}$$

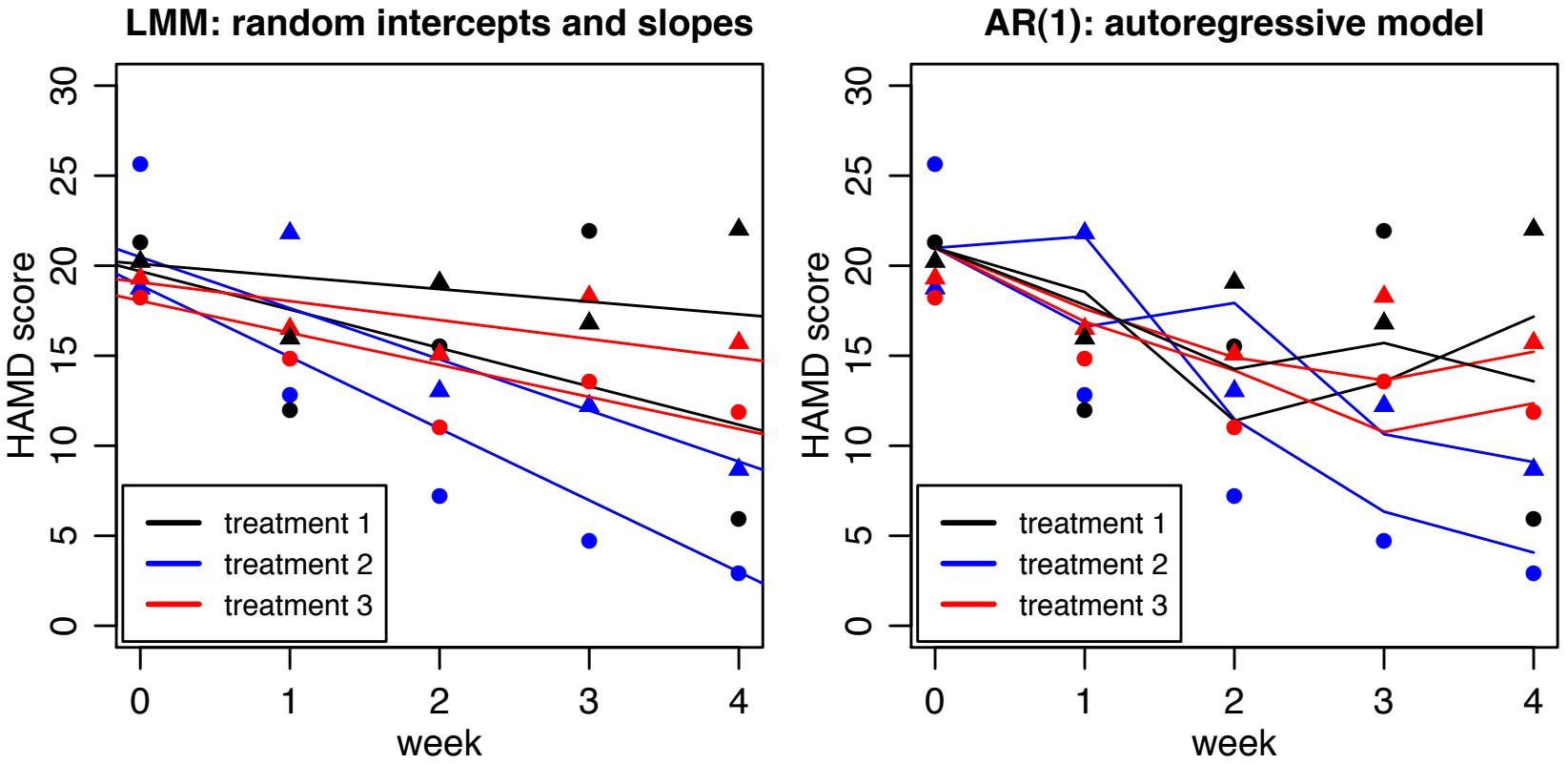
- So the two forms shown on the previous slide are equivalent

HAMD Example: WinBUGS code for AR(1) model

WinBUGS code for autoregressive model:

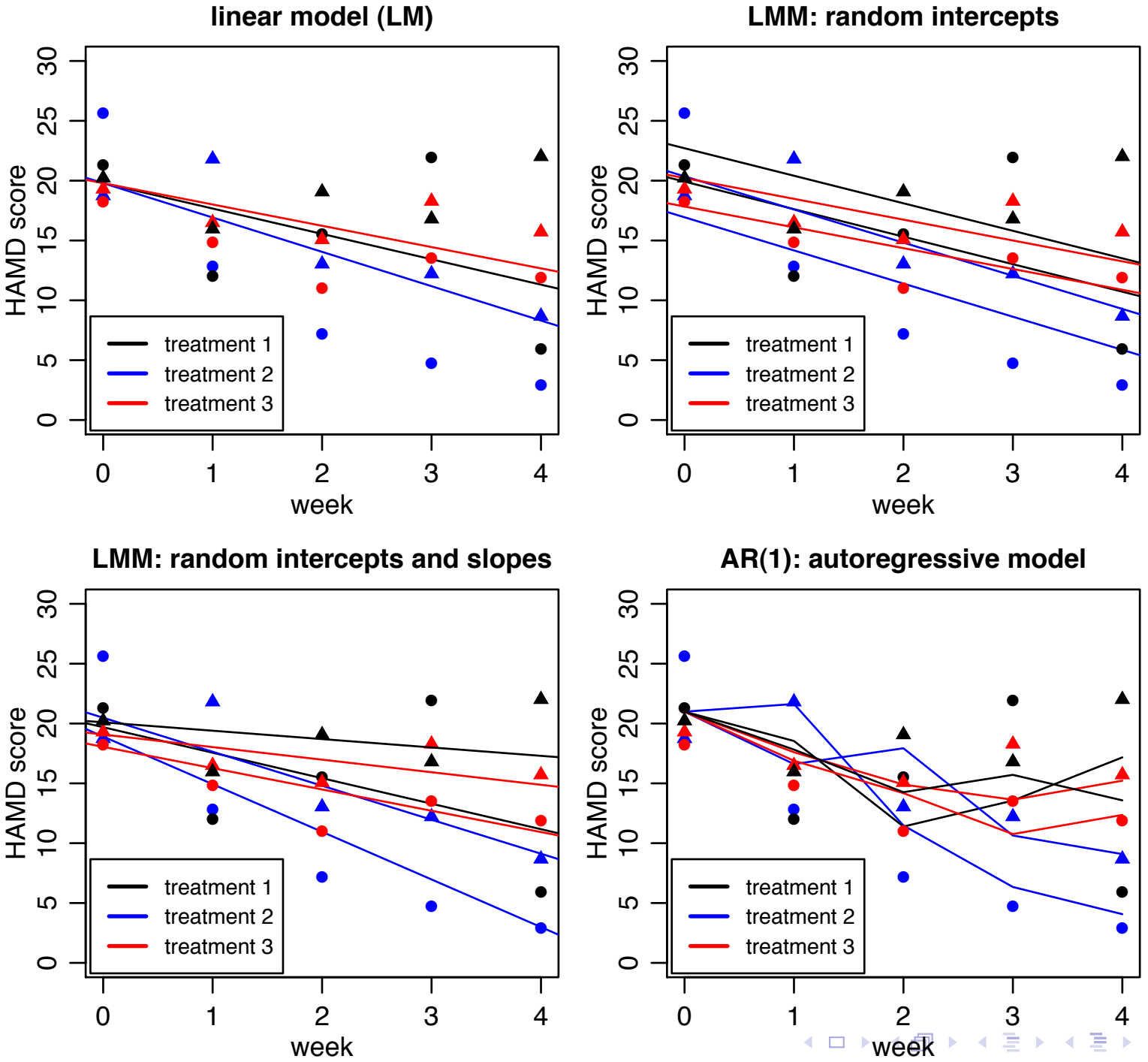
```
model {
  for (i in 1:N) {  # N individuals
    for (w in 1:W) {  # W weeks
      hamd[i,w]~dnorm(theta[i,w],tau)
      mu[i,w]<-alpha+beta[treat[i]]*(w-1)
    }
  theta[i,1]<-mu[i,1]  # week 0
  for (w in 2:W) {
    theta[i,w]<-mu[i,w]+gamma*(hamd[i,w-1]-mu[i,w-1])
    }
  }
  # Priors
  alpha~dnorm(0,0.00001)
  for (t in 1:T){beta[t]~dnorm(0,0.00001)}
  gamma~dnorm(0,0.00001)
  tau~dgamma(0.001,0.001)
  sigma.sq<-1/tau  # Normal errors
}
```

HAMD Example: LMM and AR(1) fitted lines



circles and triangles represent scores for 6 individuals (2 for each treatment)

- LMM with random intercepts and random slopes:
  - each individual has a linear line
- AR(1):
  - all lines start from a common intercept
  - paths vary according to residual at previous time point


linear model (LM)
LMM: random intercepts
LMM: random intercepts and slopes
AR(1): autoregressive model
HAMD score
week
treatment 1
treatment 2
treatment 3

HAMD Example: results comparison

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 Example: interpretation of results

- Study objective: are there any differences in the effects of the 3 treatments on the change in HAMD score over time?
- So we are particularly interested in the differences in the slope parameters, i.e.
  - $\beta_1 - \beta_2$, $\beta_1 - \beta_3$ and $\beta_2 - \beta_3$ or
  - $\mu_{\beta_1} - \mu_{\beta_2}$, $\mu_{\beta_1} - \mu_{\beta_3}$ and $\mu_{\beta_2} - \mu_{\beta_3}$ for models with random slopes
- To monitor these contrasts, add the following lines of WinBUGS code

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

or

```
contrasts[1]<-beta.mu[1]-beta.mu[2] ...
```

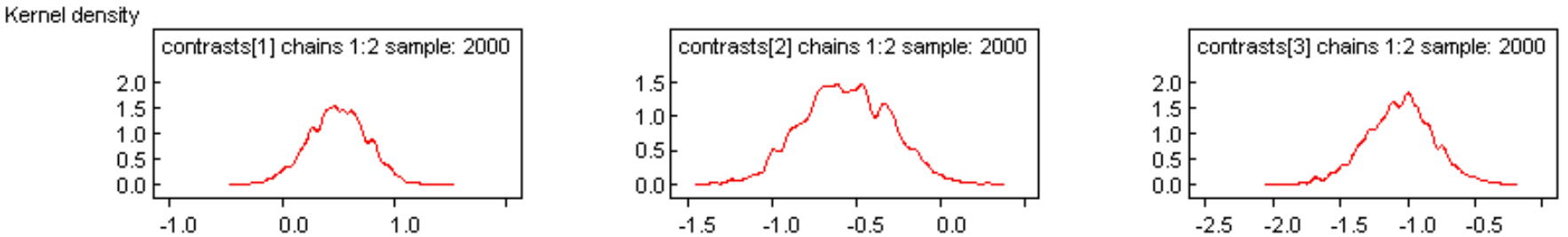
HAMD Example: contrasts

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 Example: model comparison

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

- Models which take account of the structure in the HAMD data (either by incorporating random effects or allowing for autocorrelation) fit better than a linear model
- The random effects models fit better than the autoregressive model of order 1

HAMD Example: model extensions (1)

It is straightforward to:

- allow for non-linearity by including a quadratic term

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

- include centre effects by allowing a different intercept for each centre
  - either as a fixed effect

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

  - or as random effects

$$\begin{aligned} \mu_{iw} &= \alpha_i + \beta_{treat(i)} w \\ \alpha_i &\sim \text{Normal}(\mu_{\alpha_{centre(i)}}, \sigma^2_{\alpha_{centre(i)}}) \end{aligned}$$

HAMD Example: model extensions (2)

We can also:

- allow a more complex covariance structure for the residuals by fitting an autoregressive model of order 2, AR(2)

$$
\begin{aligned}
\mathcal{R}_{i0} &= \epsilon_{i0} \\
\mathcal{R}_{i1} &= \gamma_1 \mathcal{R}_{i0} + \epsilon_{i1} \\
\mathcal{R}_{iw} &= \gamma_1 \mathcal{R}_{i(w-1)} + \gamma_2 \mathcal{R}_{i(w-2)} + \epsilon_{iw} \quad w \geq 2 \\
\epsilon_{iw} &\sim \text{Normal}(0, \sigma^2)
\end{aligned}
$$

All these complexities were included in the original data analysis

random effects v autoregressive models

- Random effects (hierarchical) linear and generalised linear models
  - suitable for repeated measurements
  - typically used for large number of individuals (n), small number of time points (T)
  - allow irregularly spaced observations
  - neglect time ordering and values
- Autoregressive models
  - suitable for temporally correlated responses
  - typically used for small n ($= 1$), large (T)
  - model temporal dependence of data or parameters explicitly
  - require equispaced data

Further points

- To summarise models:
  - LM: single level regression; errors assumed i.i.d. conditional on covariates, x
  - LMM: two level regression; errors assumed i.i.d. conditional on x and random effects
  - AR: single level regression; errors assumed i.i.d. conditional on x and past responses, y
- AR models with covariates:
  - AR(1) in the HAMD example was formulated so the residuals are serially correlated
  - AR models with covariates can also be set up so the mean responses are serially correlated

$$y_t|y_{t-1} = \mu_t + \gamma^\star y_{t-1} + \epsilon_t; \quad \epsilon_t \sim \mathrm{N}(0, \sigma^2)$$

  - the two forms give identical predictions, but the interpretation of the parameters is different

Case study: The ageing project

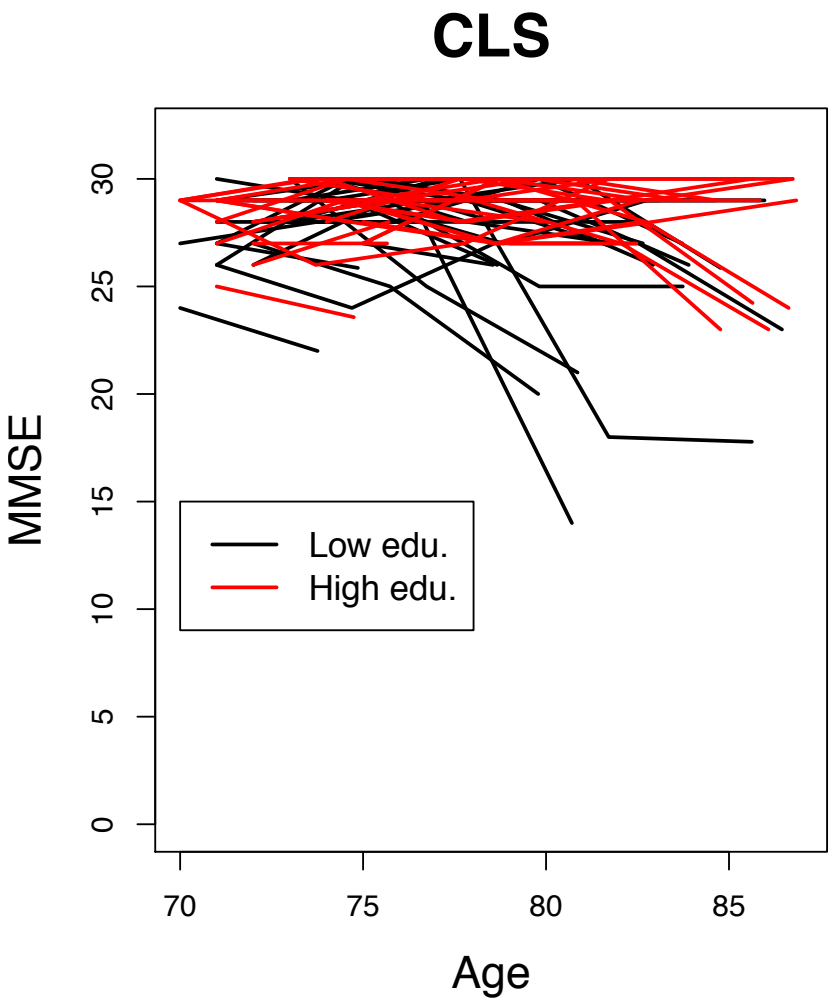
- Two objectives of this study:
  1. To evaluate effect of ageing on cognitive decline;
  2. To evaluate how different levels of education affect the rate of cognitive change over time (cognitive reserve).
- Cognitive reserve describes the mind's resilience to neuropathological damage of the brain. In the ageing study, psychologists hypothesise that people with higher education tend to have better cognitive performance and it slows down brain damage as age increases.
- In this study, we take the scores from Mini-Mental State Examination (MMSE) as a measure of the cognitive function.

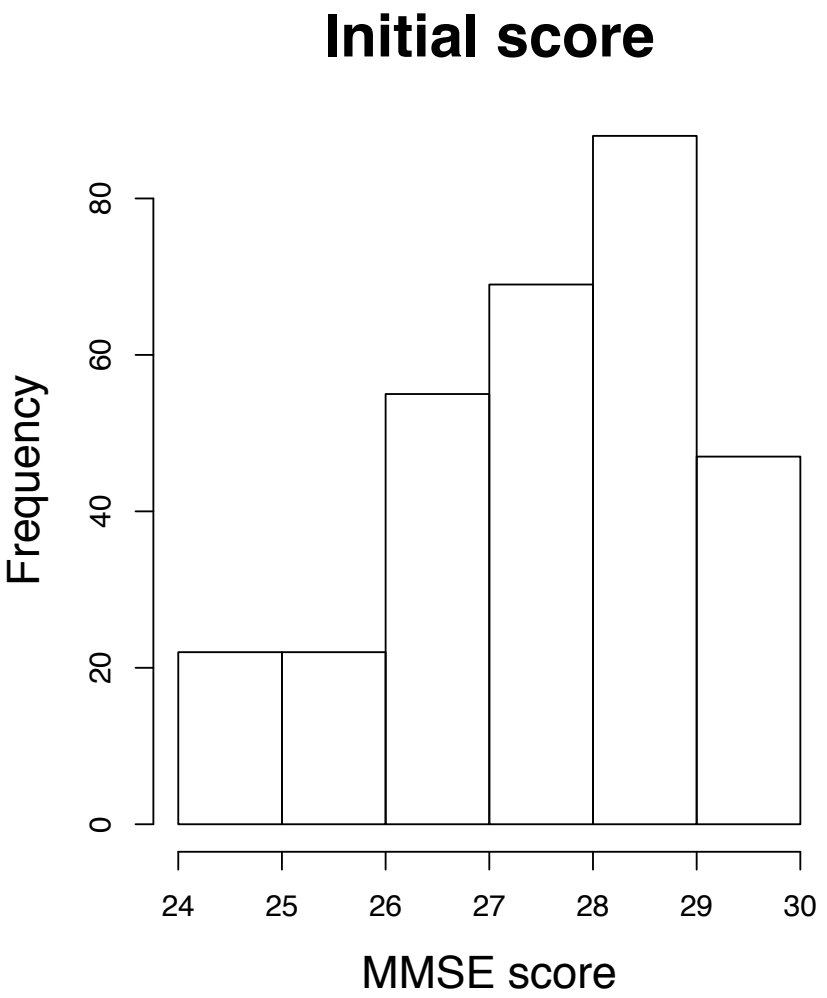
Ageing Example: Data

- We used data from the Canberra Longitudinal Study (CLS, Australia), which consists of 586 subjects.
- The study started in 1991.
- People were aged between 70 and 93 on entry.
- During the 14-years of follow-up, 4 exams were carried out.
- Outcome: MMSE score, takes values 0-30
  - the lower the score, the poorer the cognitive function
- Other variables
  - age at exams
  - gender (only looking at the male population in this lecture)
  - education (0=low and 1=high)

Ageing Example: Data

- Restrict to subjects whose first MMSE score was ≥ 24 to avoid potential dementia.





Ageing Example: Linear latent growth curve model

- To assess the ageing effect on cognitive performance, we model the MMSE scores, y_{ij}, as

$$\begin{aligned} y_{ij} &\sim \text{Normal}(\mu_{ij}, \sigma^2) \\ \mu_{ij} &= \alpha_i + \beta_i \cdot \text{age}^*_{ij} \end{aligned}$$

where

i: subjects;

j: occasions/MMSE exams;

age^*_{ij}: $\text{age}_{ij} - 75$;

α_i: the expected MMSE score at age 75 for subject i;

β_i: the rate of change in cognitive function per year for subject i;

σ^2: residual variance.

Ageing Example: Random effects

- We model the intercepts (α_i) and slopes (β_i) as random effects.
- Independent random effects priors can be assigned:

$$\begin{aligned} \alpha_i &\sim \text{Normal}(\mu_\alpha, \sigma^2_\alpha) \\ \beta_i &\sim \text{Normal}(\mu_\beta, \sigma^2_\beta) \\ \sigma_\alpha &\sim \text{Uniform}(0, 30) \\ \sigma_\beta &\sim \text{Uniform}(0, 10) \end{aligned} \tag{1}$$

- Alternatively, a joint prior allows for possible correlation between the two parameters for each subject *i*:

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim \text{multivariate Normal} \left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \Sigma \right) \tag{2}$$

- In both settings, μ_α and μ_β are assigned 'non-informative' hyperpriors, Normal(0,10000).

Ageing Example: Hyperprior on Σ^{-1}

- The conjugate prior for the inverse covariance matrix is the Wishart distribution,

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

where

R: a symmetric positive definite p x p matrix.

k: degrees of freedom and $k > p - 1$,

- with *smaller* values representing *higher* uncertainty.
- When $p = 2$, a common choice is to take $k = 2$.

- The expectation of a Wishart is $k \cdot \mathbf{R}^{-1}$. So we usually set $(1/k) \cdot \mathbf{R}$ to be a prior guess at the unknown true covariance matrix.
- Here we choose $\mathbf{R}=\begin{pmatrix} 20 & 0 \\ 0 & 2 \end{pmatrix}$.
- The Wishart distribution is a multivariate generalization of the Gamma distribution.

Ageing Example: Effect of education

- A natural way to examine the effect of education levels (and possibly other covariates) on the MMSE score at age 75 and the cognitive rate of change is to regress the two random effect terms against education, i.e.,

$$\begin{aligned} \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} &\sim \text{multivariate Normal}\left(\begin{pmatrix} \mu_{\alpha,i} \\ \mu_{\beta,i} \end{pmatrix}, \Sigma \right) \\ \mu_{\alpha,i} &= \eta_0 + \eta_1 \cdot \text{edu}_i \\ \mu_{\beta,i} &= \gamma_0 + \gamma_1 \cdot \text{edu}_i \end{aligned} \tag{3}$$

  - η_0: the average MMSE score at 75 for low educated person;
  - γ_0: the average rate of change in cognitive function for low educated person;
  - η_1: effect of high education on the average MMSE score at age 75;
  - γ_1: effect of high education on the average rate of change in cognitive function.
- 'Non-informative' priors are assigned to η_0, η_1, γ_0 and γ_1.

Ageing Example: Alternative parameterisation

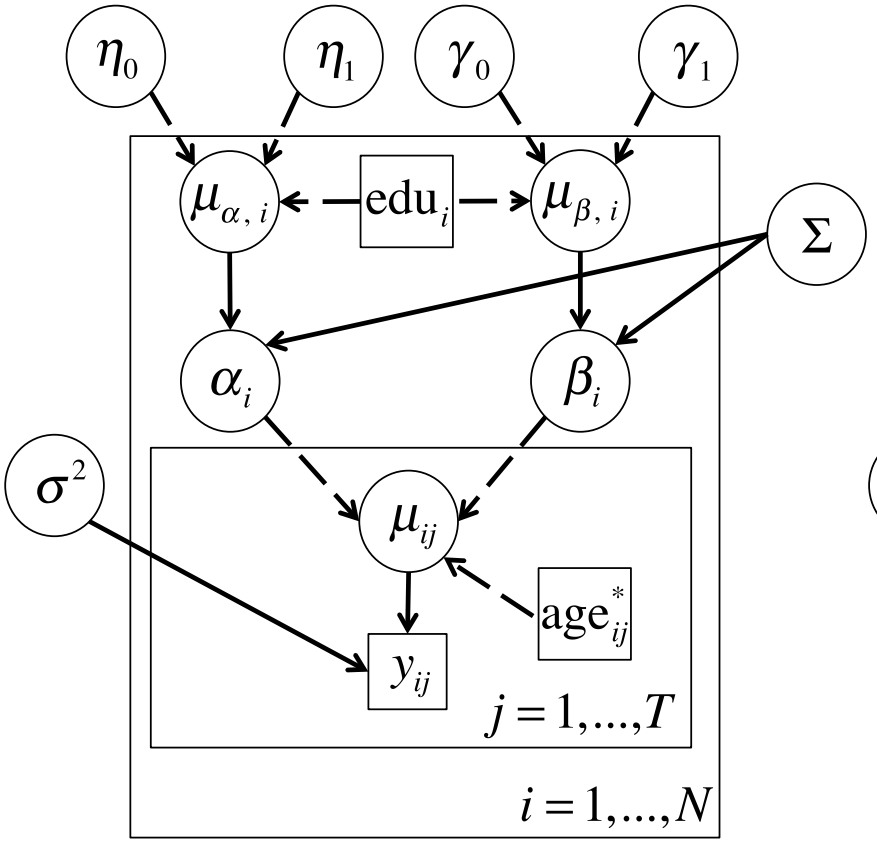
- An alternative modelling approach is to include education as an explanatory variable directly in the level 1 regression model:

$$\begin{aligned} \mu_{ij} &= (\eta_0 + \alpha_i) + (\gamma_0 + \beta_i) \cdot age^*_{ij} \\ &\quad + \eta_1 \cdot edu_i + \gamma_1 \cdot age^*_{ij} \cdot edu_i \\ \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} &\sim \text{multivariate Normal} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma \right) \end{aligned}$$

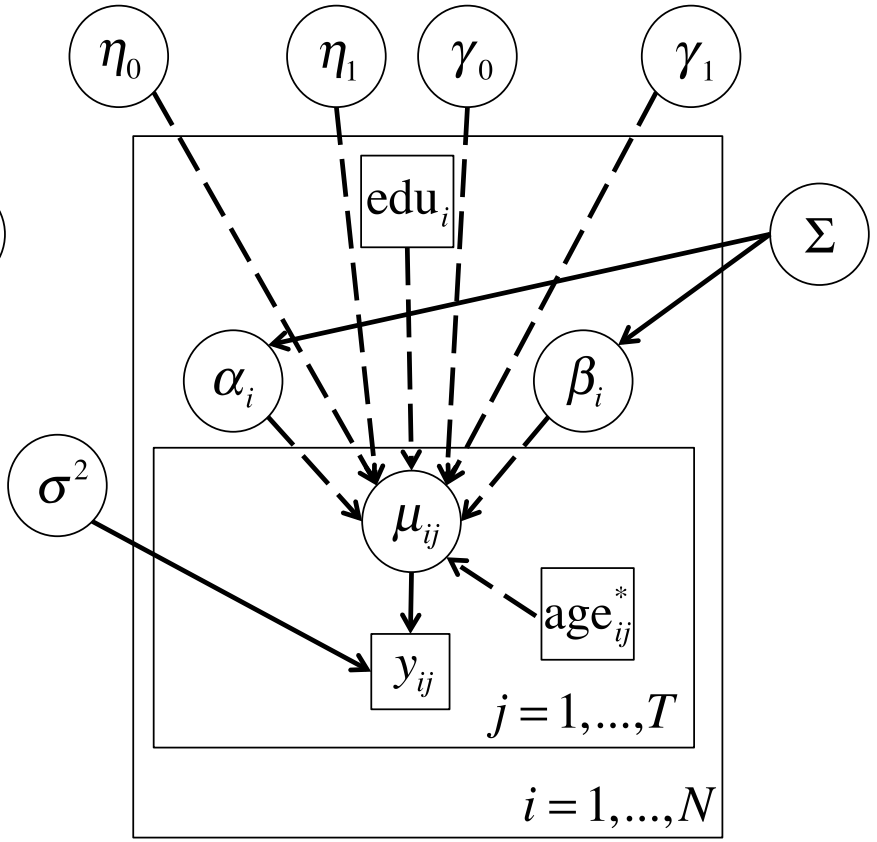
- The effect of education on the rate of change in cognitive function is quantified through the age $\times$ education interaction term, γ_1.
- This formulation is a reparameterisation (or, 'non-centred' version) of Model (3).

Ageing Example: DAGs

Hierarchically centred



Non-centred



Ageing Example: WinBUGS code

- WinBUGS code to implement the hierarchically centred version of the model

```
model {
  for (i in 1:N) {
    for (j in 1:T) {
      y[i,j] ~ dnorm(mu.y[i,j], tau.y)
      mu.y[i,j] <- alpha[i] + beta[i] * age.star[i,j]
    }

  #    MVN for joint modelling intercepts and slopes
    alpha[i] <- inter.slope[i,1]
    beta[i] <- inter.slope[i,2]
    inter.slope[i,1:2] ~ dmnorm(mu[i,1:2],tau.mu[1:2,1:2])

  #    modelling the random effects with education
    mu[i,1] <- eta0 + eta1 * edu[i]
    mu[i,2] <- gamma0 + gamma1 * edu[i]
  }
```

$$
\begin{aligned}
y_{ij} &\sim \text{Normal}(\mu_{ij}, \sigma^2) \\
\mu_{ij} &= \alpha_i + \beta_i \cdot \text{age}^*_{ij} \\
\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} &\sim \text{MVN}\left(\begin{pmatrix} \mu_{\alpha,i} \\ \mu_{\beta,i} \end{pmatrix}, \Sigma \right) \\
\mu_{\alpha,i} &= \eta_0 + \eta_1 \cdot \text{edu}_i \\
\mu_{\beta,i} &= \gamma_0 + \gamma_1 \cdot \text{edu}_i
\end{aligned}
$$

Ageing Example: WinBUGS code (cnt.)

```
#    priors
  eta0 ~ dnorm(0,0.0001)
  eta1 ~ dnorm(0,0.0001)
  gamma0 ~ dnorm(0,0.0001)
  gamma1 ~ dnorm(0,0.0001)

#    residual variance
  tau.y ~ dgamma(0.5,0.005)
  sigma2.y <- pow(tau.y,-1)

#    Wishart prior for the inverse covariance matrix
  tau.mu[1:2,1:2] ~ dwish(R[1:2,1:2],2)
  R[1,1] <- 20
  R[1,2] <- 0
  R[2,1] <- 0
  R[2,2] <- 2

#    covariance matrix
  Sigma[1:2,1:2] <- inverse(tau.mu[1:2,1:2])

#    variances of the random effects
  sigma2.alpha <- Sigma[1,1]
  sigma2.beta  <- Sigma[2,2]

#    correlation between the intercepts and slopes
  cor.alpha.beta <- Sigma[1,2]/(pow(sigma2.alpha,0.5)*pow(sigma2.beta,0.5))

#    average MMSE score of men at 75 with high education
  mmse.high.edu <- eta0 + eta1
#    average rate of cognitive change of men with high education
  rate.high.edu <- gamma0 + gamma1
}
```

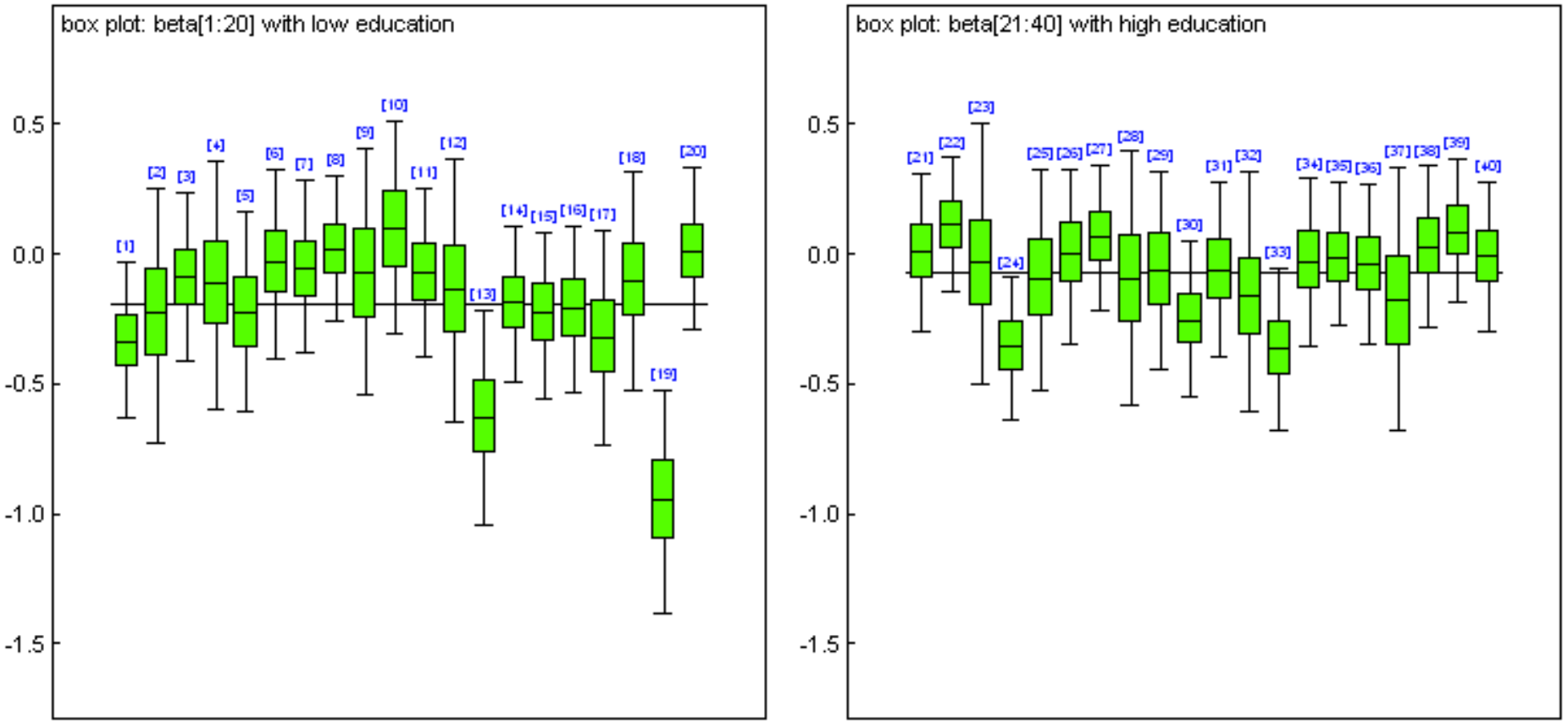
Ageing Example: Data structure

- Some subjects have missing responses at some exams
- Here we assume missing at random (See Lecture 6)
  - just include missing value indicator (`NA`) for any missing responses in the data file
  - covariates (age, education) fully observed
  - No need to change model code to deal with the missing responses in this case since assuming MAR
  - WinBUGS will automatically impute (predict) values for missing responses from the fitted model at each iteration

y[,1]	y[,2]	y[,3]	y[,4]	age[,1]	age[,2]	age[,3]	age[,4]
24	28	NA	NA	81	85	89	93
27	NA	27	30	71	75	79	83
29	30	30	NA	72	76	80	84
...				...			

Ageing Example: Results

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



Ageing Example: Results

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)

Note:

- There was a significant decline in MMSE performance (hence cognitive function) over time;
- Men with low education showed a faster decline in cognitive function than those who had higher education. However, such a difference was not statistically significant.

Ageing Example: Results

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

Ageing Example: Ceiling effect

- Because MMSE is designed to screen for dementia, not quite a test of cognitive ability, a healthy individual should be able to achieve the maximum score quite easily.
- A score of 30 suggests that this person could at least achieve the maximum score at that particular time $\rightarrow$ right censoring.
- We can model these censored data as follows:

$$y_{ij}^{*} \sim \text{Normal}(\mu_{ij}, \sigma^2) \cdot \text{I}(\text{lower}_{ij}, +\infty)$$

where

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

and I(lower,upper) is the censoring function in WinBUGS.

Ageing Example: WinBUGS code

- WinBUGS code ignoring censoring

```
model {
  for (i in 1:N) {
    for (j in 1:T) {
      y[i,j] ~ dnorm(mu.y[i,j], tau.y)
      mu.y[i,j] <- alpha[i] + beta[i] * age.star[i,j]
    }
  }
  ...
```

- WinBUGS code allowing for censoring

```
model {
  for (i in 1:N) {
    for (j in 1:T) {
      y.s[i,j] ~ dnorm(mu.y[i,j], tau.y)I(lower.lim[i,j],)
      mu.y[i,j] <- alpha[i] + beta[i] * age.star[i,j]
    }
  }
  ...
```

Ageing Example: Data structure

- For model ignoring censoring

```
y[,1]   y[,2]   y[,3]   y[,4]
24      28      NA      NA
27      NA      27      30
29      30      30      NA
...
```

- For model allowing for censoring

```
y.s[,1] y.s[,2] y.s[,3] y.s[,4]
24      28      NA      NA
27      NA      27      NA
29      NA      NA      NA
...

lower.lim[,1] lower.lim[,2] lower.lim[,3] lower.lim[,4]
0             0             0             0
0             0             0             30
0             30            30            0
...
```

Ageing Example: Results

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)

- Conclusions remain the same.
- Uncertainty increased slightly due to loss of information from allowing for censoring.

Lecture 5. Hierarchical models for complex patterns of variation

Outline

There is huge scope for elaborating the basic hierarchical models discussed in lecture 1 to reflect additional complexity in the data, e.g

- Adding further levels to the hierarchy (patients within wards within hospitals, pupils within schools within local authorities,...)
- Adding non-nested (cross-classified) levels (patients within GPs crossed with hospitals,...)
- Repeated observations on some/all units
- Modelling temporal or spatial structure in data
- Non-normal random effects distributions, mixture models
- Hierarchical models for variance components, ...

Models for repeated (longitudinal) observerations were covered in previous lecture

In this lecture, we will discuss some of the other examples:

- Hierarchical models with $>$ 2 levels
- Cross-classified models
- Hierarchical models for variances

3-level hierarchical models

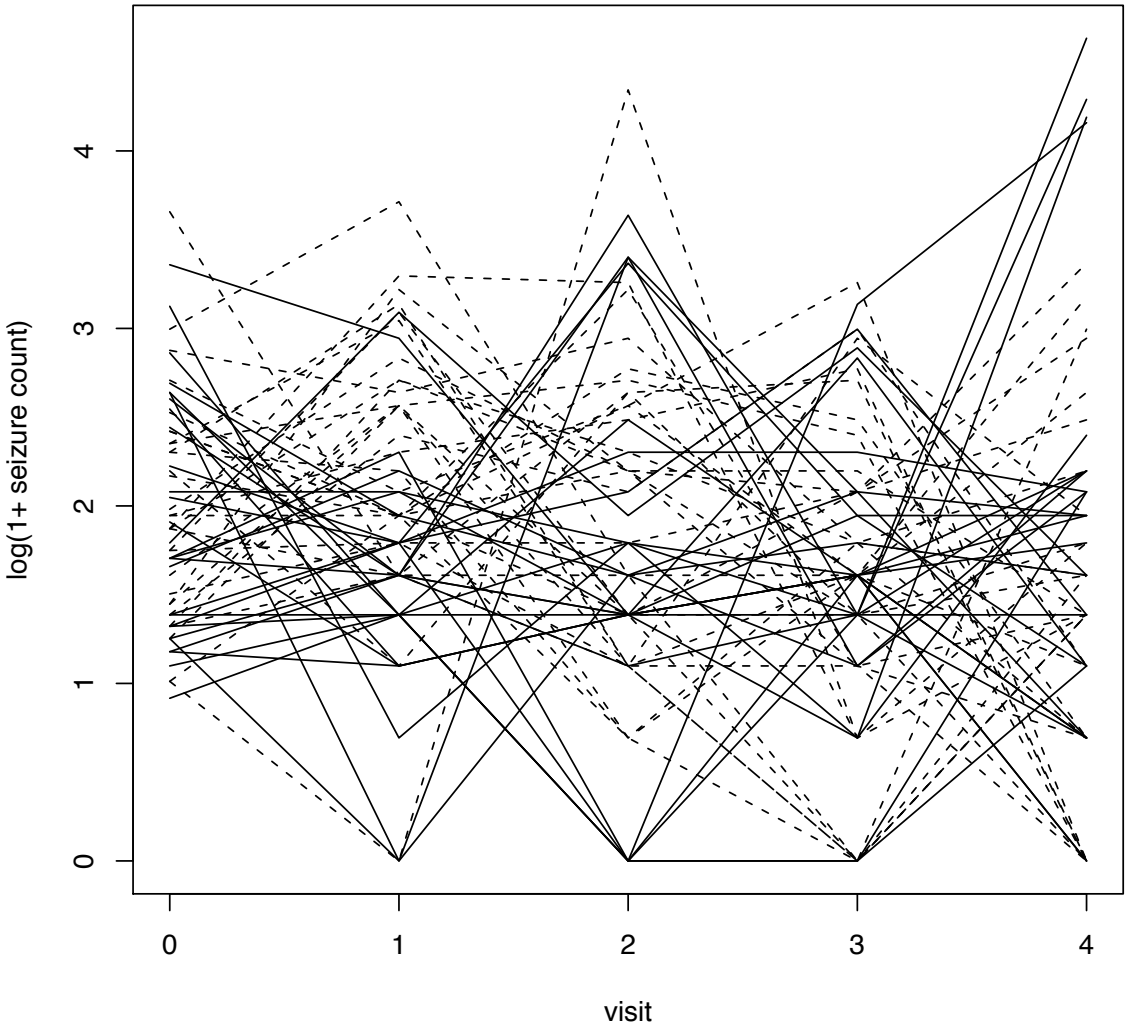
Example: Epilepsy

- Data from randomized trial of anti-convulsant therapy in 59 epilepsy patients followed-up on 4 separate occasions. (See WinBUGS Examples Volume I.)
- The variables are:

T	Treatment: Progabide versus Placebo
A	Age at randomization (years)
B	Seizure count in 8-week baseline period
V_1	Seizure count in 2 weeks before first follow-up visit
V_2	Seizure count in 2 weeks before second follow-up visit
V_3	Seizure count in 2 weeks before third follow-up visit
V_4	Seizure count in 2 weeks before fourth follow-up visit.

- Study objective: to estimate the treatment effect and to investigate heterogeneity in patient responses

Raw data



How should we model these data ?

- We typically use the *Poisson* distribution to describe the sampling variation for count data
  - Could fit a simple Poisson regression model to estimate treatment effect
- However, this is unlikely to adequately model the variation in the data for 2 reasons:
  1. Observations from the same patient are likely to be correlated, even after accounting for available covariates
     - → We should include patient-specific random effects
     - → Counts for each patient assumed to be conditionally independent Poisson random variables given the patient random effect
  2. Count data are often 'overdispersed', i.e. have larger variance than assumed by Poisson distribution
     - → Could use negative-binomial (2 parameter) sampling distribution instead
     - → Or include observation-level random effect to account for overdispersion

Model for Epilepsy example

V_{ij} is the seizure count for the ith patient at the jth visit

$$\begin{aligned} V_{ij} &\sim \text{Poisson}(\mu_{ij}) \\ \log \mu_{ij} &= \beta_{0i} + \beta_1 T_i + \beta_2 A_i + \beta_3 B_i + \phi_{ij} \end{aligned}$$

Observation-level random effects:

$$\phi_{ij} \sim \text{Normal}(\delta, \phi^2)$$

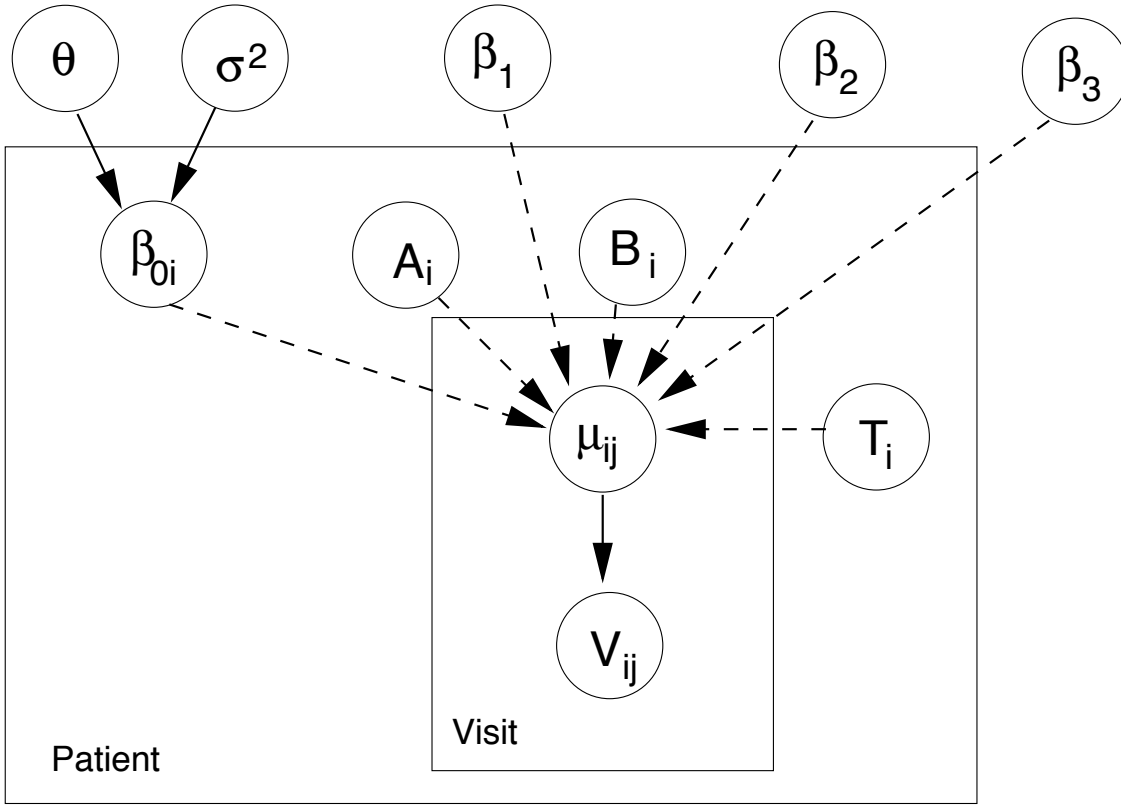
Patient-level random intercepts:

$$\beta_{0i} \sim \text{Normal}(\theta, \sigma^2)$$

Assume vague priors for the hyperparameters of the random effects distributions and the regression coefficients:

$$\begin{aligned} \theta, \delta, \beta_1, \beta_2, \beta_3 &\sim \text{Normal}(0, 100000) \\ \sigma, \phi &\sim \text{Unif}(0, 10) \end{aligned}$$

Graphical model for Epilepsy example



Note that V_{ij} are *conditionally independent* given the random effect β_{0i} for each patient, and the other parameters, but are *marginally dependent*

Epilepsy: Some alternative models

Besides the model discussed (Model 2), we also fitted a simple (non-hierarchical) GLM model (Model 0) and a model (Model 1) with just patient random effects

$$
\begin{aligned}
\textit{Model 0:} \quad & \log \mu_{ij} &&= \beta_0 + \beta_1 T_i + \beta_2 A_i + \beta_3 B_i \\
\textit{Model 1:} \quad & \log \mu_{ij} &&= \beta_{0i} + \beta_1 T_i + \beta_2 A_i + \beta_3 B_i \\
\textit{Model 2:} \quad & \log \mu_{ij} &&= \beta_{0i} + \beta_1 T_i + \beta_2 A_i + \beta_3 B_i + \phi_{ij}
\end{aligned}
$$

Results for different models for the epilepsy data

Parameters	Model 0	Model 1	Model 2
Treatment	-0.017 (0.05)	-0.33 (0.17)	-0.30 (0.16)
Age	0.58 (0.11)	0.30 (0.36)	0.35 (0.37)
Baseline	1.22 (0.03)	1.03 (0.11)	1.04 (0.11)
σ^2_β	–	0.31 (0.08)	0.27 (0.08)
σ^2_ϕ	–	–	0.14 (0.03)
DIC	1728	1279	1156
p_D	3.9	49.5	120.7

- Marked change in regression coefficients and increase in uncertainty (posterior SD) after inclusion of random effects (Models 1, 2 versus 0)
- Substantial heterogeneity between patients (σ^2_β) even after accounting for covariates (age, treatment and baseline seizure count)
- Could explore heterogeneity further by extending model to include random slopes for each subject

Cross-classified random effects models

- Straightforward to extend basic 2-level hierarchical model to include non-nested random effects structures, e.g.
  - THM measurements cross-classified within zones and years
  - pupils cross-classified within primary and secondary schools
- Easiest to formulate cross-classified models in WinBUGS using nested index notation (see example)

Example: Schools – exam scores cross-classified by primary and secondary school

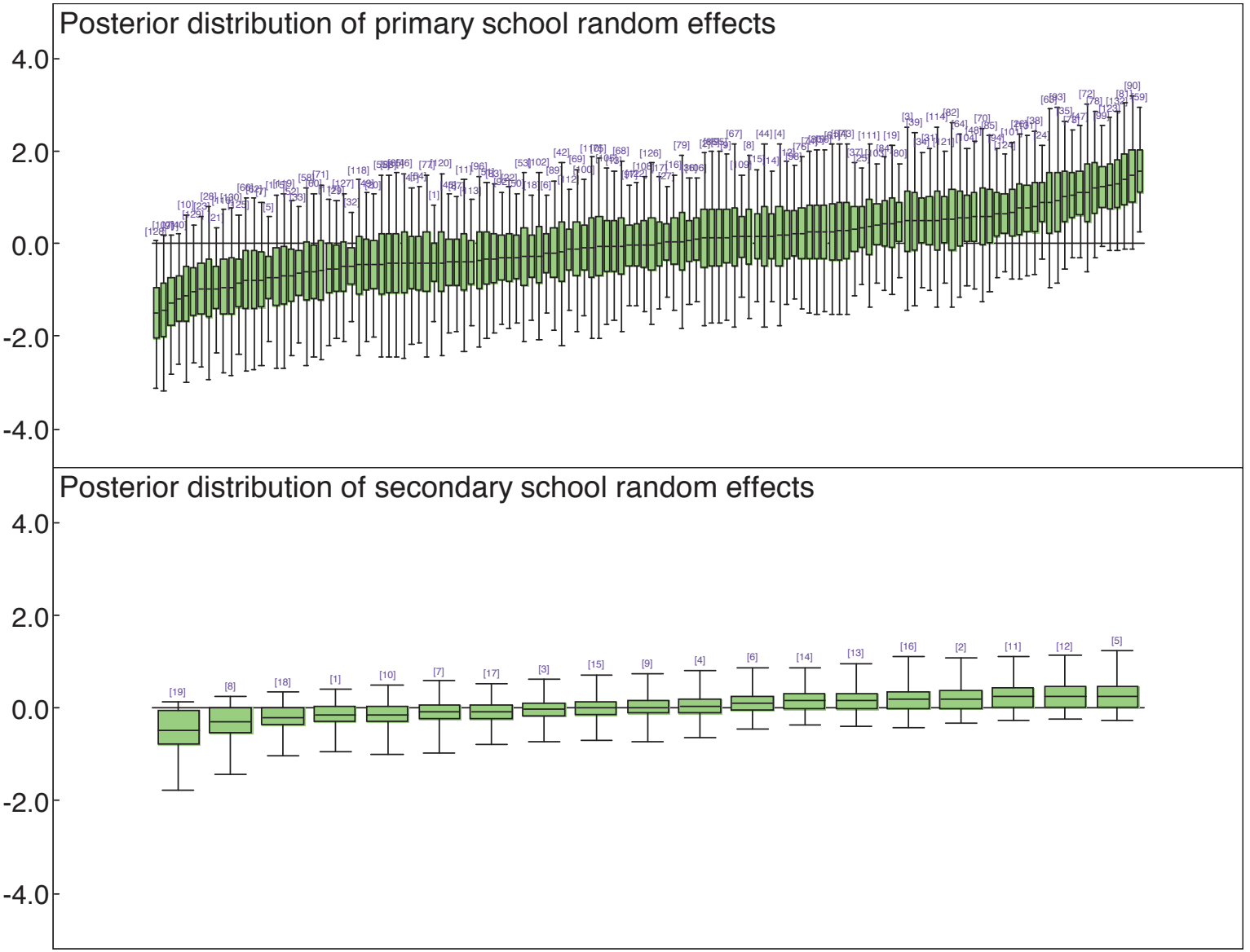
- These data were obtained from the MLwiN website `www.mlwin.com/softrev/2lev-xc.html`
- We use a random sample of 800 children who attended 132 primary schools and 19 secondary schools in Scotland
- The following variables were used
  - `Y`: Exam attainment score of pupils at age 16
  - `VRQ`: verbal reasoning score taken on secondary school entry
  - `SEX`: Pupil's gender (0 = boy, 1 = girl)
  - `PID`: Primary school identifying code
  - `SID`: Secondary school identifying code
- A normal hierarchical model is fitted, with independent random effects for primary school and secondary school
- Verbal reasoning score and gender are included as 'fixed' covariate effects (but note that in Bayesian framework, 'fixed' effect coefficients are still assigned prior distributions)

BUGS model code

```
for(i in 1:Nobs) {
 Y[i] ~ dnorm(mu[i], tau.e)
 mu[i] <- alpha + beta[1]*SEX[i] + beta[2]*VRQ[i] +
                  theta.ps[PID[i]] + theta.ss[SID[i]]
}
### random effects distributions
 for(j in 1:Nprim) { theta.ps[j] ~ dnorm(0, tau.ps) } # primary
 for(k in 1:Nsec) { theta.ss[k] ~ dnorm(0, tau.ss) } # secondary
### priors on regression coefficients and variances
 tau.e ~ dgamma(0.001, 0.001)
 sigma2.e <- 1/tau.e              # residual error variance
 tau.ps ~ dgamma(0.001, 0.001)
 sigma2.ps <- 1/tau.ps            # between primary school var.
 tau.ss ~ dgamma(0.001, 0.001)
 sigma2.ss <- 1/tau.ss            # between secondary school var.
 alpha ~ dnorm(0, 0.000001)       # intercept
 for(q in 1:2) {beta[q] ~ dnorm(0, 0.000001)} # regression coeff.
### percentage of total variance explained
 VPC.ps <- sigma2.ps/(sigma2.e+sigma2.ps+sigma2.ss) # primary
 VPC.ss <- sigma2.ss/(sigma2.e+sigma2.ps+sigma2.ss) # secondary
```

Results

Parameters	Model 1		Model 2	
α	5.53	(5.17, 5.88)	5.85	(5.59, 6.10)
β_1 (sex)	–	–	0.23	(-0.08, 0.53)
β_2 (VRQ)	–	–	0.16	(0.15, 0.17)
$\sigma^2_{[e]}$	8.18	(7.35, 9.10)	4.49	(4.03, 5.00)
$\sigma^2_{[ps]}$	1.12	(0.43, 1.98)	0.36	(0.08, 0.70)
$\sigma^2_{[ss]}$	0.19	(0.10, 0.82)	0.02	(0.0007, 0.12)
VPC_{ps}	11.8%	(4.7%, 19.8%)	7.4%	(1.5%, 13.8%)
VPC_{ss}	2.0%	(0.1%, 8.3%)	0.4%	(0.01%, 2.4%)
DIC	4008		3514	
p_D	58.0		43.8	


Posterior distribution of primary school random effects
4.0
2.0
0.0
-2.0
-4.0
Posterior distribution of secondary school random effects
4.0
2.0
0.0
-2.0
-4.0
[19]
[8]
[18]
[1]
[10]
[7]
[17]
[3]
[15]
[9]
[4]
[6]
[14]
[13]
[16]
[2]
[11]
[12]
[5]

Heteroscedasticity

- Heteroscedasticity $\rightarrow$ non constant variance
- Can occur at any level of hierarchical model
- Easily handled in MCMC framework by modelling variance as a specified function of other variables

Example: complex level 1 variation in Schools

Original model:

$$\begin{aligned} Y_i &\sim \text{Normal}(\mu_i, \sigma^2_{[e]}) \\ \mu_i &= \alpha + \beta_1 \text{SEX}_i + \beta_2 \text{VRQ}_i + \theta_{[ps]\text{PID}_i} + \theta_{[ss]\text{SID}_i} \\ &.... \end{aligned}$$

Complex level 1 variation depending on VRQ:

$$\begin{aligned} Y_i &\sim \text{Normal}(\mu_i, \sigma^2_{[e]i}) \\ \log \sigma^2_{[e]i} &= \gamma_1 + \gamma_2 \text{VRQ}_i \\ \mu_i &= ..... \end{aligned}$$

Along with priors on α, β_k and random effects variances, also need priors on coefficients of variance model:

$$\gamma_k \sim \text{Normal}(0, 0.000001); \quad k = 1, 2$$

BUGS model code

```
for(i in 1:Nobs) {
 Y[i] ~ dnorm(mu[i], tau.e[i])
 mu[i] <- alpha + beta[1]*SEX[i] + beta[2]*VRQ[i] +
                  theta.ps[PID[i]] + theta.ss[SID[i]]

 # complex level 1 variance
 logsigma2.e[i] <- gamma[1] + gamma[2]*VRQ[i]
 tau.e[i] <- 1/exp(logsigma2.e[i])
}

# remaining code is same as before
......
......
# except no longer need prior on residual error variance
##tau.e ~ dgamma(0.001, 0.001)
##sigma2.e <- 1/tau.e              # residual error variance

# instead need to include priors
# on coefficients of the variance model
for(k in 1:2) { gamma[k] ~ dnorm(0, 0.000001)    }
```

BUGS model code (continued)....

```
## VPC will now depend on value of VRQ

# level 1 variance for child with VRQ in lowest 10th percentile
sigma2.e.lowVRQ <- exp(gamma[1] + gamma[2] * (-19))

# level 1 variance for child with VRQ in highest 10th percentile
sigma2.e.hiVRQ <- exp(gamma[1] + gamma[2] * 15)

## percentage of total variance explained
## by primary school effects....

# .....for pupils with low VRQ
VPC.ps.lowVRQ <- sigma2.ps /
                 (sigma2.e.lowVRQ + sigma2.ps + sigma2.ss)

# .....for pupils with hi VRQ
VPC.ps.hiVRQ <- sigma2.ps /
                (sigma2.e.hiVRQ + sigma2.ps + sigma2.ss)
```

Initial values

- Remember to edit initial values from previous model to:
  - remove initial values for `tau.e`
  - add initial values for `gamma` vector
- Some care needed when specifying initial values for `gamma[2]` to avoid numerical problems in BUGS
  - `gamma[2]` measures effect of unit change in VRQ (which ranges from -30 to 40) on log residual variance
  - Residual variance was around 5 from previous analysis, so expect values of log variance around log 5 = 1.6

  $\Rightarrow$ `gamma[2]` should be quite small ($<<1$)

e.g.

```
list(alpha = 0, tau.ps = 1, tau.ss = 1,
     beta=c(0,0), gamma=c(1, 0.001))
```

Results

Parameter	Posterior mean	95% CI
γ_2	0.019	(0.008, 0.029)
VPC_{ps} (low VRQ)	9.0%	(2.0%, 18.0%)
VPC_{ps} (hi VRQ)	5.0%	(1.0%, 10.4%)
VPC_{ss} (low VRQ)	0.6%	(0.01%, 3.3%)
VPC_{ss} (hi VRQ)	0.4%	(0.01%, 1.9%)
DIC	3503	
p_D	43.3	

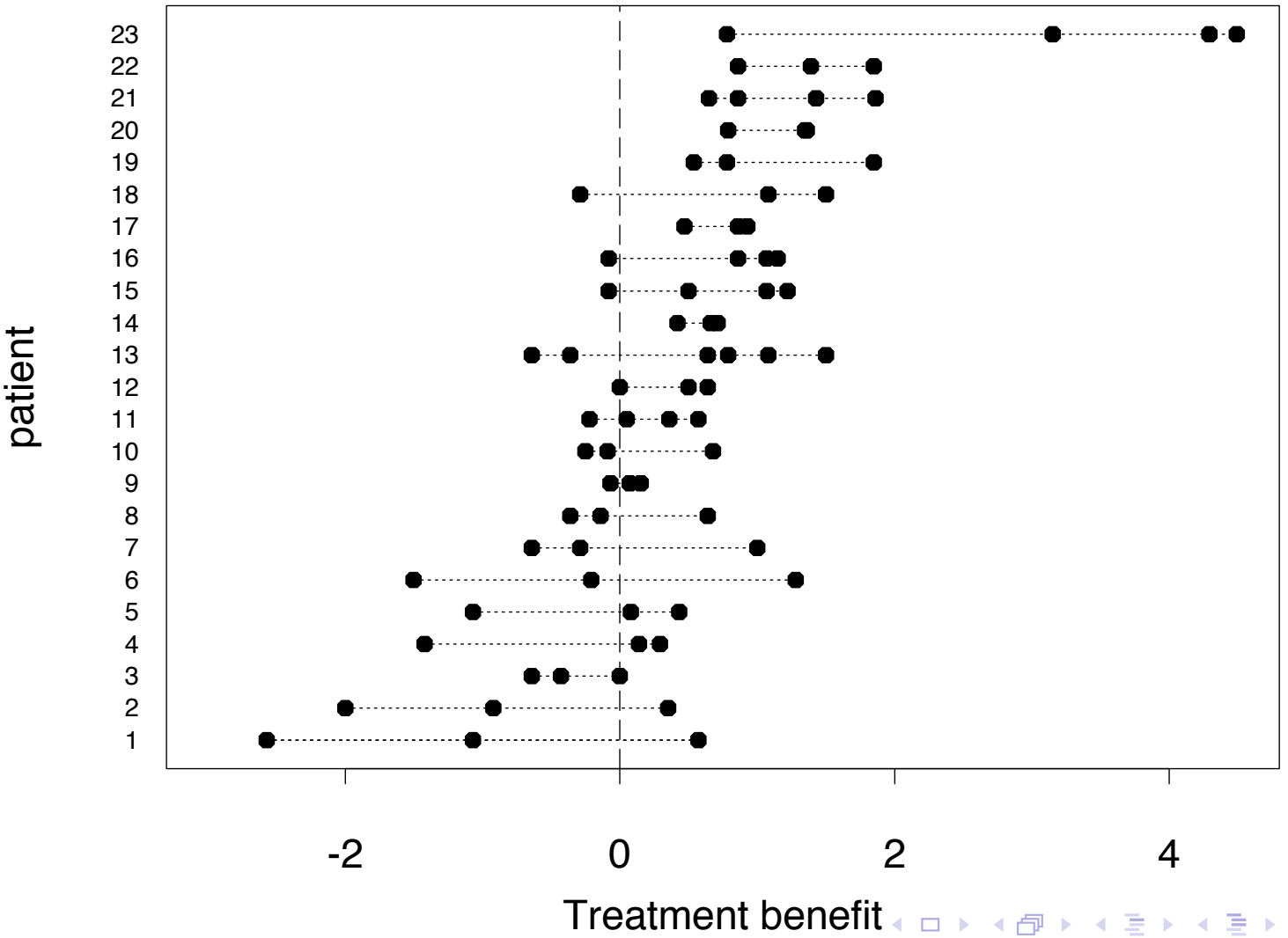
Recall model with homoscedastic level 1 variance had DIC = 3514, p_D = 43.8, so heteroscedastic model preferred

Hierarchical models for variances

Example: N-of-1 trials

- N-of-1 trials $\rightarrow$ repeated within-person crossover trials
- Often suitable for investigating short-term symptom relief in chronic conditions
- Example:
  - **Intervention:** Amitriptyline for treatment of fibromyalgia to be compared with placebo.
  - **Study design:** 23 N-of-1 studies - each patient treated for a number of periods (3 to 6 per patient), and in each period both amitriptyline and placebo were administered in random order
  - **Outcome measure:** Difference in response to a symptom questionnaire in each paired crossover period. A positive difference indicates Amitriptyline is superior
  - **Evidence from study:** 7/23 experienced benefit from the new treatments in all their periods

Raw data for each patient



Statistical model

- If y_{kj} is the j^{th} measurement on the k^{th} individual, we assume

$$y_{kj} \sim \mathrm{N}(\theta_k, \sigma_k^2)$$

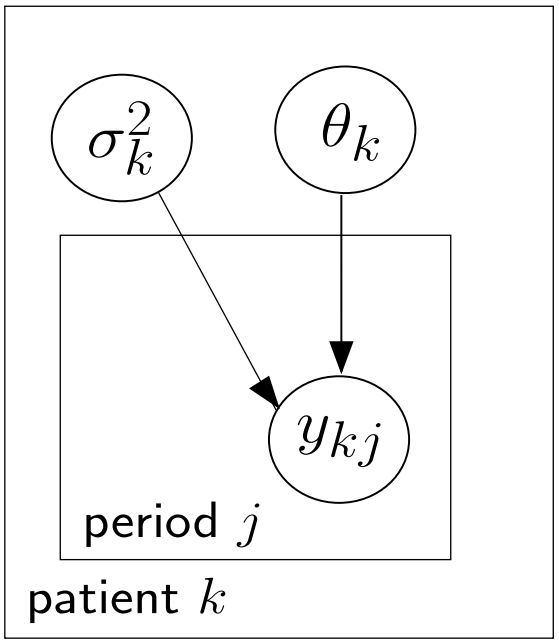
- Assume both θ_k's and σ_k^2's are *exchangeable*, in the sense there is no reason to expect systematic differences and we act as if they are drawn from some common prior distribution.
- We make the specific distributional assumption that

$$\begin{aligned} \theta_k &\sim \mathrm{N}(\mu_\theta, \phi_\theta^2) \\ \log(\sigma_k^2) &\sim \mathrm{N}(\mu_\sigma, \phi_\sigma^2) \end{aligned}$$

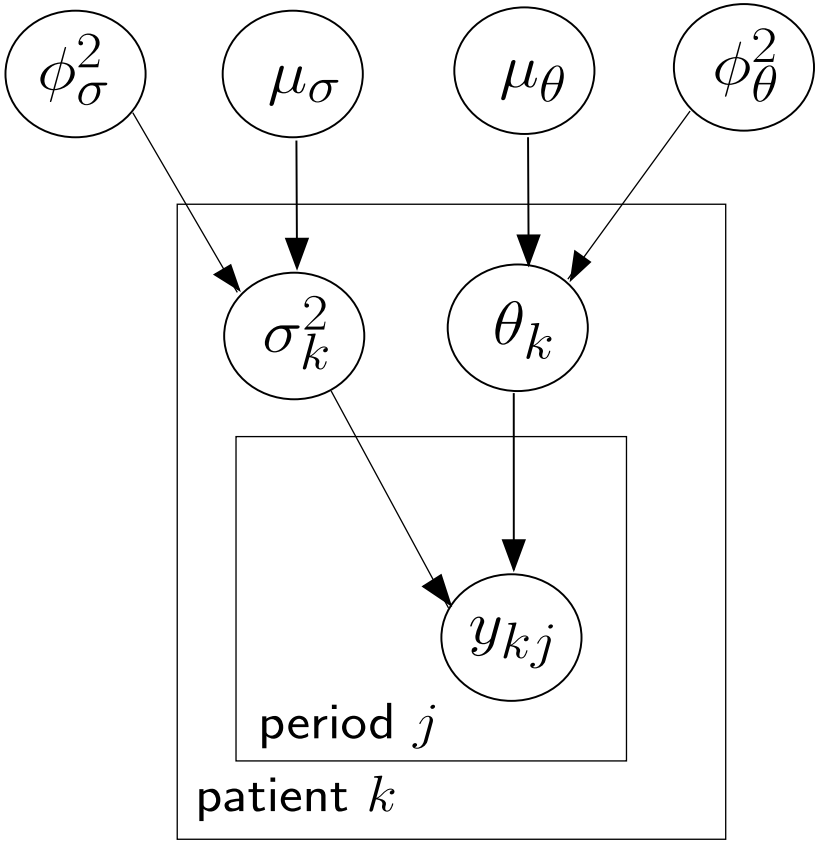
  A normal distribution for the log-variances is equivalent to a log-normal distribution for the variances
- Uniform priors adopted for $\mu_\theta, \phi_\theta, \mu_\sigma$ and ϕ_σ.

Graphical model for n-of-1 example

Independent effect

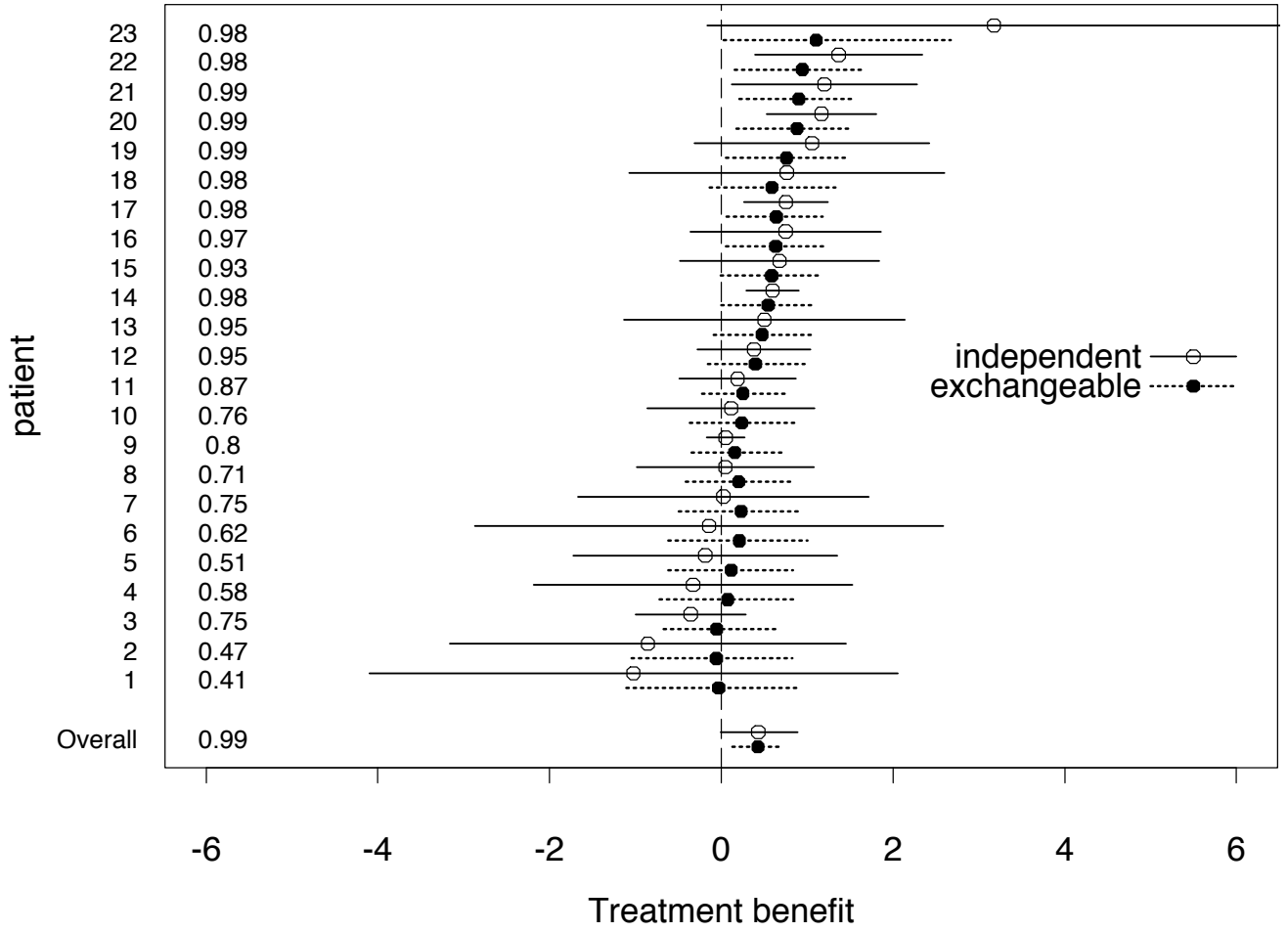


Exchangeable means and variances



Results

Estimates and 95% intervals for treatment effect, and posterior probability that effect > 0



Interpretation

- Exchangeable model shrinks in the extreme patients, reflecting the limited information from each individual (see patient 23)
- It might be felt the model is exercising undue influence in this situation
- Despite shrinkage, narrower intervals mean that 9 patients have 95% intervals excluding 0 compared to 6 with the independent analysis
- One consequence of allowing exchangeable variances is that patient 9 has a *wider* interval under the exchangeable model
  - patient 9's observations were very close together $\rightarrow$ very narrow interval under independence model
- Straightforward to include patient-level covariates
- Sensitivity analysis to the shape of both the sampling and the random-effects distribution: say assuming t-distributions.

Lecture 6.
Missing data

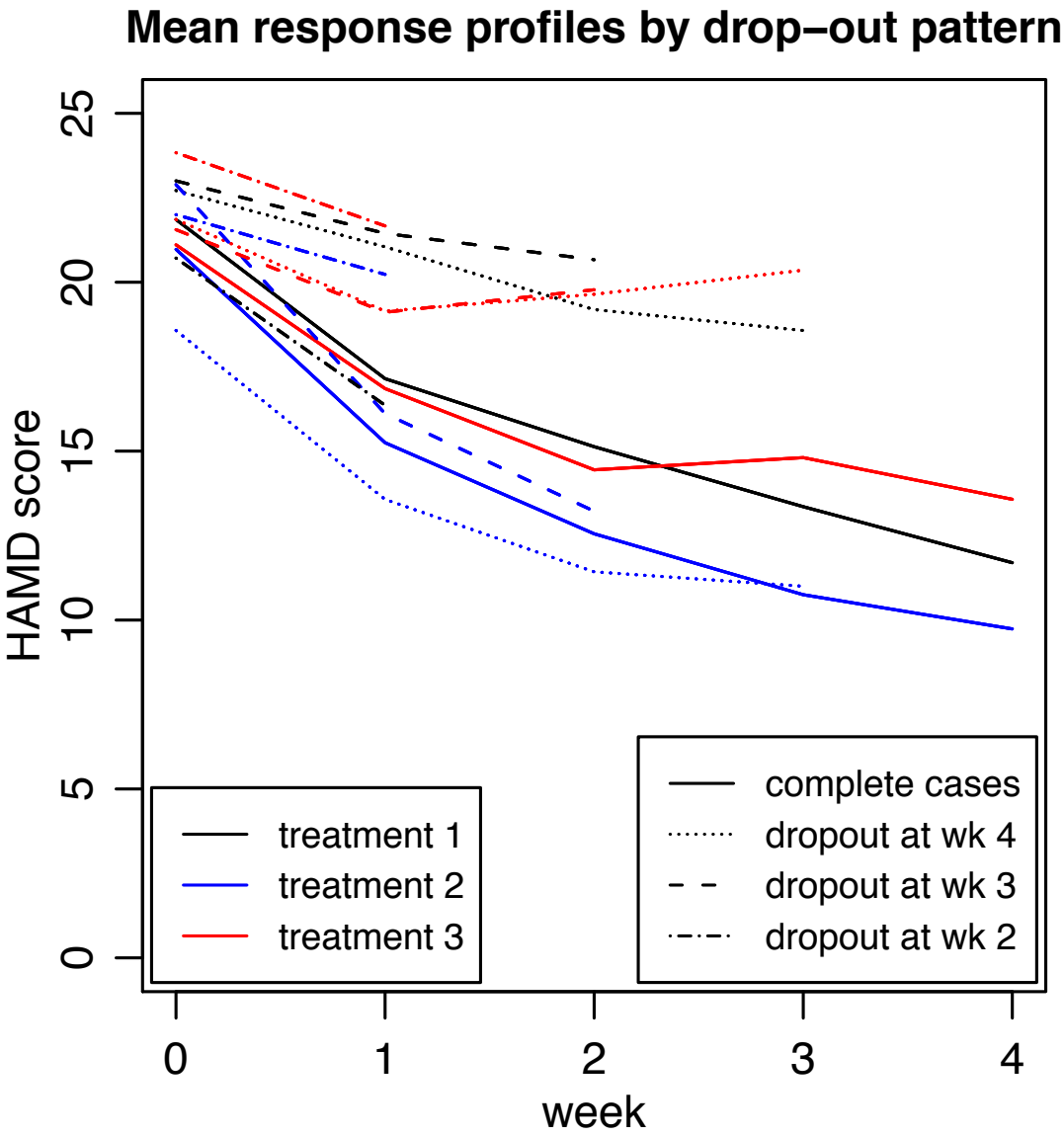
Missing data

- Missing data are common!
- Usually handled inadequately
- Typically, missing data are ignored and only complete cases analysed - known as a complete case analysis (CC)
- Advantage of CC
  - simple
- Disadvantages of CC
  - often introduces bias
  - inefficient as data from incomplete cases discarded
- Bayesian hierarchical models can easily be adapted to
  - include partially observed cases
  - incorporate realistic assumptions about the reasons for the missingness

HAMD example: revisited

- In Lecture 4 we analysed antidepressant clinical trial data using complete case analysis
- In so doing, we
  - discarded partial data from 121 out of 367 subjects
  - did not consider why subjects dropped out of the study (although we were implicitly making assumptions)
- In this lecture, we revisit the HAMD example and perform a more 'principled' analysis
- We also look at how to deal with missing covariates using a low birth weight example

HAMD example: missing scores



- Individuals allocated treatments 1 and 3 generally have higher profiles if they dropped out rather than remained in the study
- But the drop-out and cc profiles are similar for treatment 2

HAMD example: model of missingness

- The full data for the HAMD example includes:
  - y, HAMD score (observed and missing values)
  - t, treatment
  - and a missing data indicator for y, m, s.t.

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

- So we can model
  - y (random effects model)
  - and the probability that y is missing using:

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

We now consider different possibilities for p_{iw}, which depend on the assumptions we make about the missing data mechanism

Types of missing data

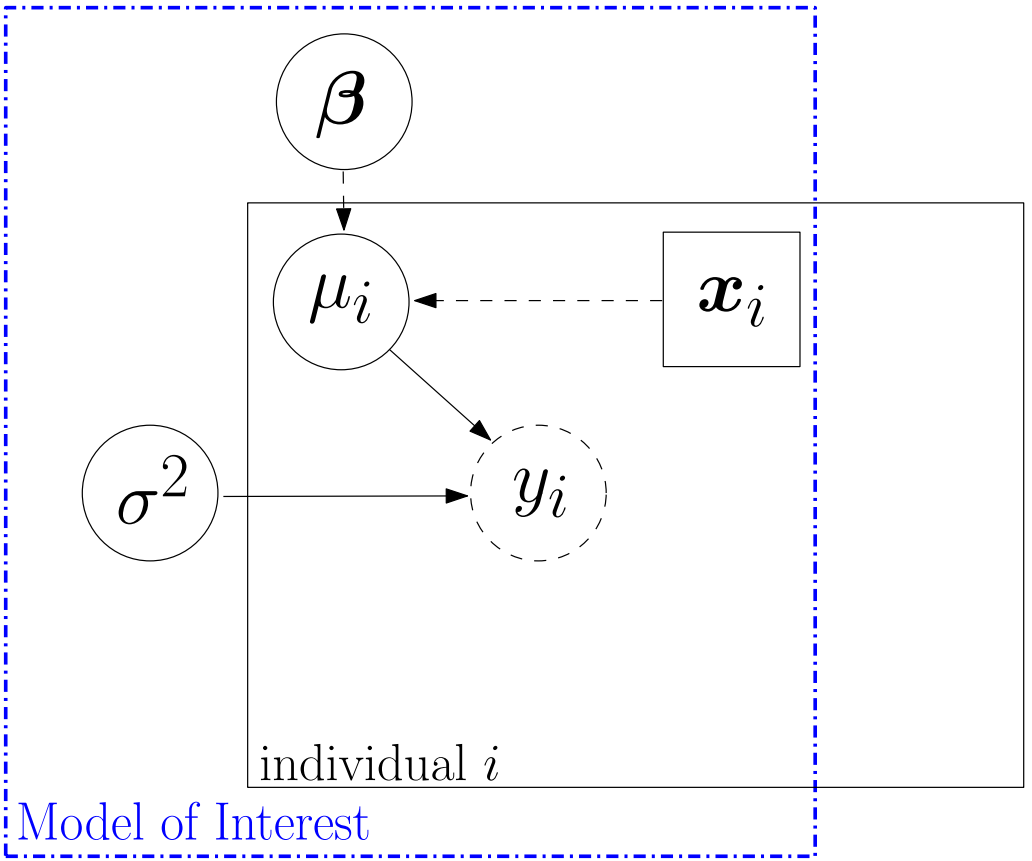
Following Rubin, missing data are generally classified into 3 types

Consider the mechanism that led to the missing HAMD scores (y)
recall: p_{iw} is the probability y_{iw} is missing for individual i in week w

- Missing Completely At Random (MCAR)
  - missingness does not depend on observed or unobserved data
  - e.g. $logit(p_{iw}) = \theta_0$
- Missing At Random (MAR)
  - missingness depends only on observed data
  - e.g. $logit(p_{iw}) = \theta_0 + \theta_1 t_i$ or $logit(p_{iw}) = \theta_0 + \theta_2 y_{i0}$
    note: drop-out from week 2; weeks 0 and 1 completely observed
- Missing Not At Random (MNAR)
  - neither MCAR or MAR hold
  - e.g. $logit(p_{iw}) = \theta_0 + \theta_3 y_{iw}$

DAG: Model of Interest (Analysis Model)

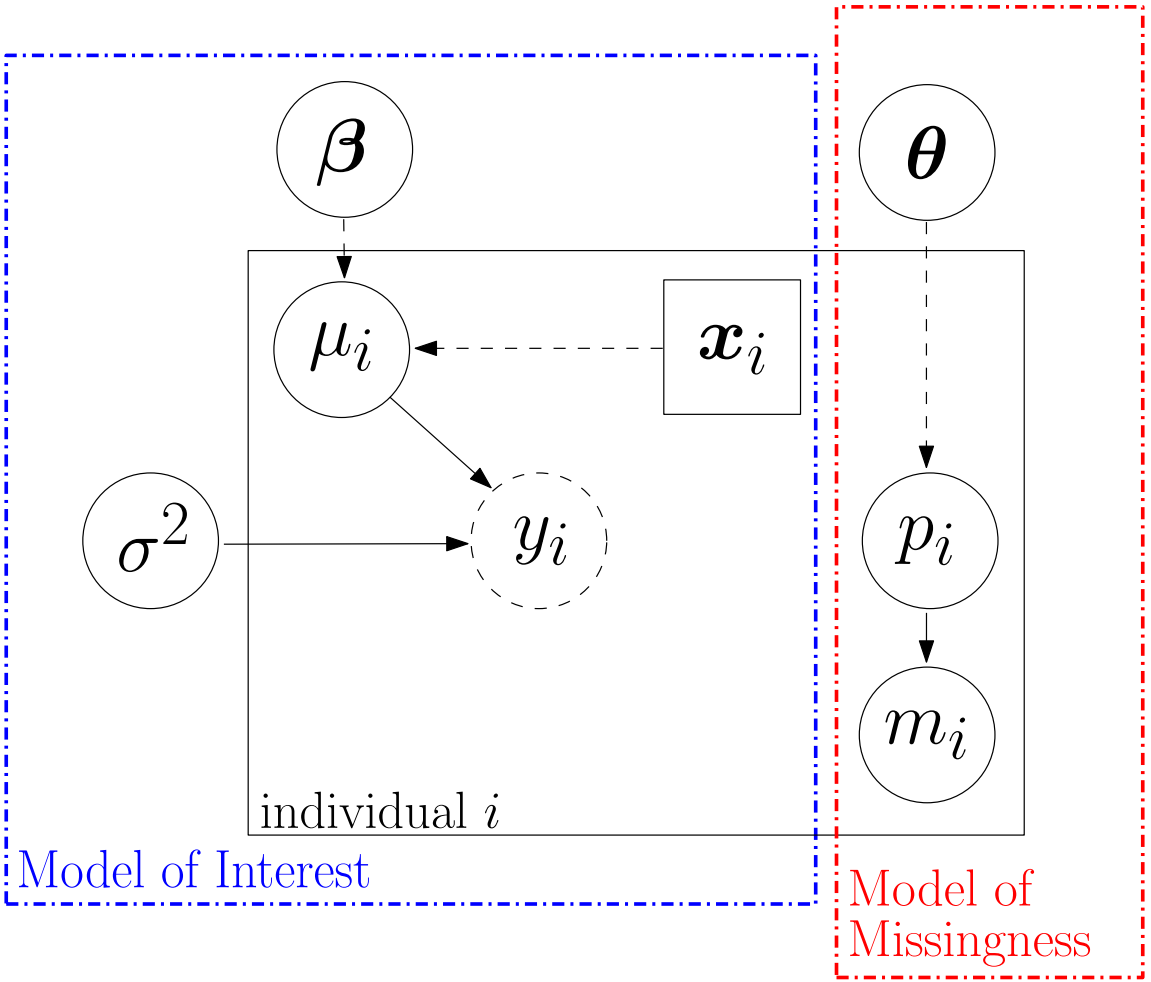
A typical regression model of interest



Note: $\boldsymbol{x}$ is completely observed, but y has missing values

DAG: Missing Completely At Random (MCAR)

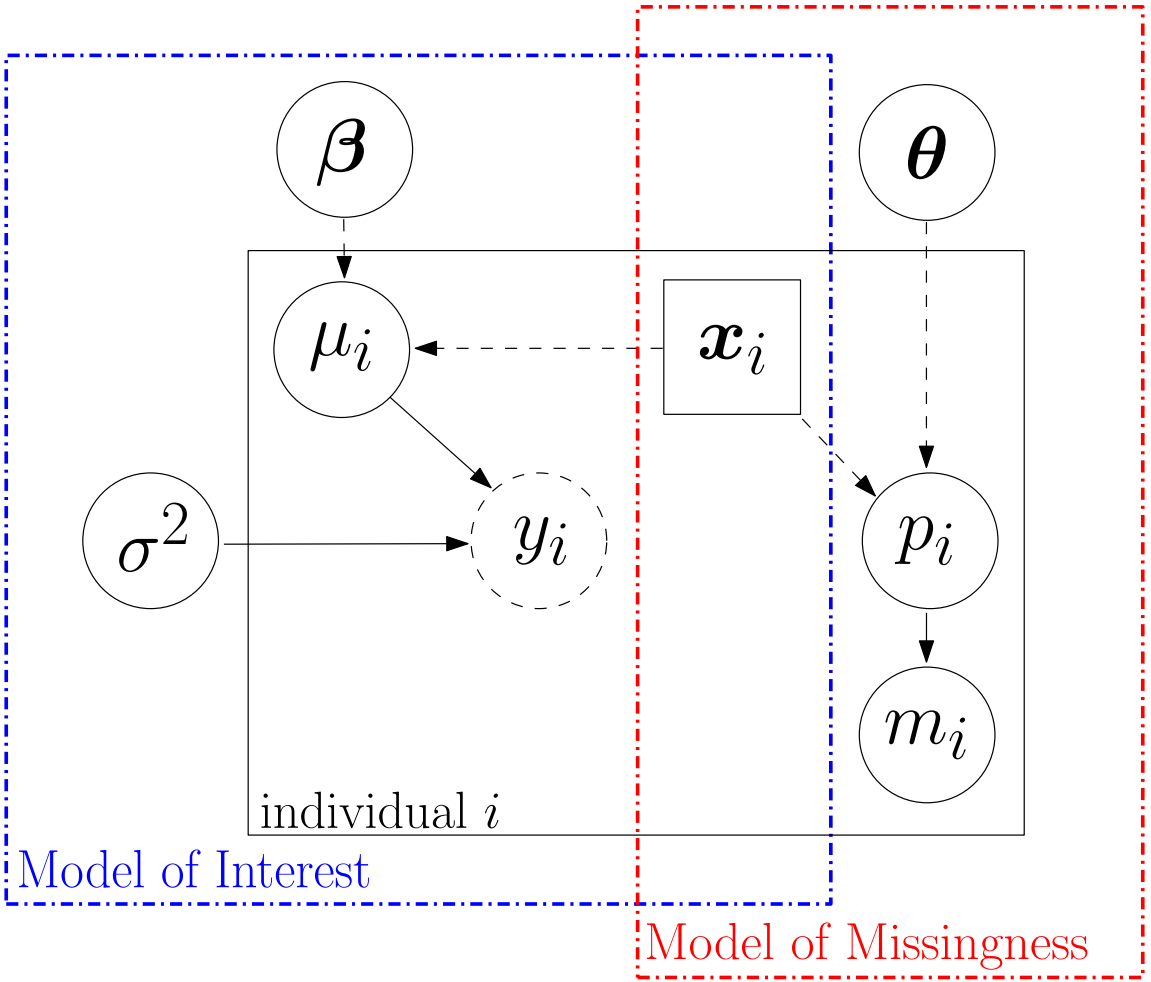
A regression model of interest + model for probability of y missing



Note: $\boldsymbol{x}$ is completely observed, but y has missing values

DAG: Missing At Random (MAR)

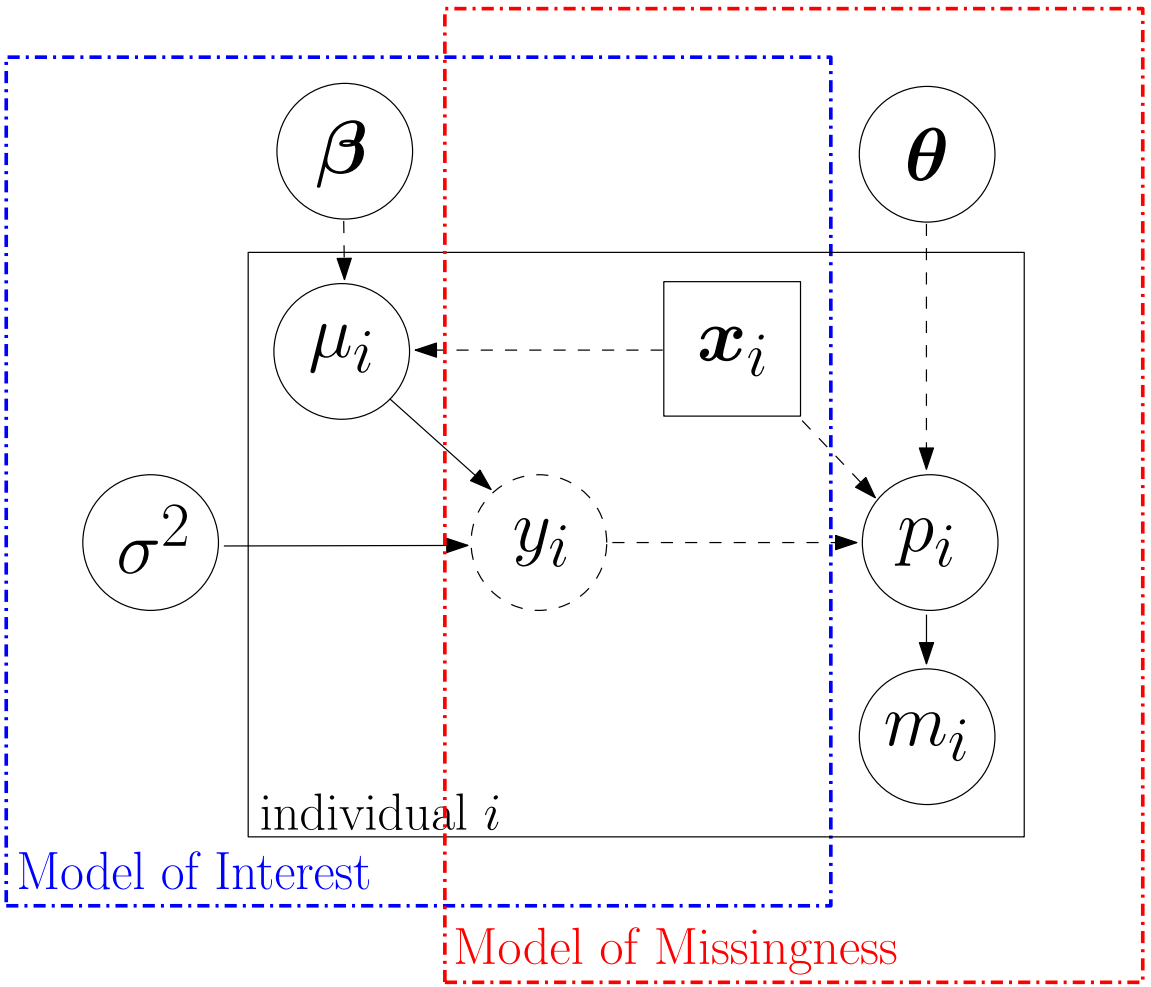
A regression model of interest + model for probability of y missing



Note: $\boldsymbol{x}$ is completely observed, but y has missing values

DAG: Missing Not At Random (MNAR)

A regression model of interest + model for probability of y missing



Note: $\boldsymbol{x}$ is completely observed, but y has missing values

Joint model: general notation

We now use general notation, but in the HAMD example $\boldsymbol{z} = (y, t)$

- Let $\boldsymbol{z} = (z_{ij})$ denote a rectangular data set
  $i = 1, ..., n$ individuals and $j = 1, ..., k$ variables
- Partition $\boldsymbol{z}$ into observed and missing values, $\boldsymbol{z} = (\boldsymbol{z}^{obs}, \boldsymbol{z}^{mis})$
- Let $\boldsymbol{m} = (m_{ij})$ be a binary indicator variable such that

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

- Let β and θ denote vectors of unknown parameters
- Then the joint model (likelihood) of the full data is

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

Joint model: integrating out the missingness

- The joint model, $f(\boldsymbol{z}^{obs}, \boldsymbol{z}^{mis}, \boldsymbol{m}|\beta, \theta)$, cannot be evaluated in the usual way because it depends on missing data
- However, the marginal distribution of the observed data can be obtained by integrating out the missing data,

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

We now look at a factorisation of the joint model

Joint model: selection model factorisation

- The joint model can be factorised as

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

  which simplifies with appropriate conditional independence assumptions

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

- This factorisation is known as a selection model
- $f(\boldsymbol{z}^{obs}, \boldsymbol{z}^{mis}|\beta)$ is the usual likelihood you would specify if all the data had been observed
- $f(\boldsymbol{m}|\boldsymbol{z}^{obs}, \boldsymbol{z}^{mis}, \theta)$ represents the missing data mechanism and describes the way in which the probability of an observation being missing depends on other variables (measured or not) and on its own value
- For some types of missing data, the form of the conditional distribution of $\boldsymbol{m}$ can be simplified

Simplifying the factorisation for MAR and MCAR

- Recall we wish to integrate out the missingness

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

- MAR missingness depends only on observed data, i.e.

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

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

- MCAR missingness is a special case of MAR that does not even depend on the observed data

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

$$\text{So, } f(\boldsymbol{z}^{obs}, \boldsymbol{m}|\beta, \theta) = f(\boldsymbol{m}|\theta) f(\boldsymbol{z}^{obs}|\beta)$$

Ignorable/Nonignorable missingness

The missing data mechanism is termed ignorable if

1. the missing data are MCAR or MAR and
2. the parameters β and θ are distinct

In the Bayesian setup, an additional condition is

3. the priors on β and θ are independent

'Ignorable' means we can ignore the model of missingness,
but does not necessarily mean we can ignore the missing data!

However if the data mechanism is nonignorable,
then we cannot ignore the model of missingness

Assumptions

- In contrast with the sampling process, which is often known, the missingness mechanism is usually unknown
- The data alone cannot usually definitively tell us the sampling process
  - But with fully observed data, we can usually check the plausibility of any assumptions about the sampling process e.g. using residuals and other diagnostics
- Likewise, the missingness pattern, and its relationship to the observations, cannot definitively identify the missingness mechanism
  - Unfortunately, the assumptions we make about the missingness mechanism cannot be definitively checked from the data at hand

Sensitivity analysis

- The issues surrounding the analysis of data sets with missing values therefore centre on assumptions
- We have to
  - decide which assumptions are reasonable and sensible in any given setting - contextual/subject matter information will be central to this
  - ensure that the assumptions are transparent
  - explore the sensitivity of inferences/conclusions to the assumptions

Bayesian methods for handling missing data

- Bayesian approach treats missing data as additional unknown quantities for which a posterior distribution can be estimated
  - no fundamental distinction between missing data and unknown parameters
- 'Just' need to specify appropriate joint model for observed and missing data and model parameters, and estimate in usual way using MCMC
- Fully model-based approach to missing data

In what follows, it is helpful to distinguish between

- missing response and missing covariate data (regression context) i.e. we let $\boldsymbol{z} = (y, \boldsymbol{x})$ where y is the response of interest and $\boldsymbol{x}$ is a set of covariates
- ignorable and non-ignorable missingness mechanisms, since when mechanism is ignorable, specifying the joint model reduces to specifying $f(\boldsymbol{z}^{obs}, \boldsymbol{z}^{mis}|\beta)$, and ignoring $f(\boldsymbol{m}|\boldsymbol{z}^{obs}, \theta)$

Missing response data
- assuming missing data mechanism is ignorable (I)

In this case, $z^{mis} = y^{mis}$ and $\boldsymbol{z}^{obs} = (y^{obs}, \boldsymbol{x})$

- Usually treat fully observed covariates as fixed constants rather than random variables with a distribution
- Joint model $f(\boldsymbol{z}^{obs}, \boldsymbol{z}^{mis}|\beta)$ reduces to specification of $f(\boldsymbol{y}^{obs}, \boldsymbol{y}^{mis}|\boldsymbol{x}, \beta)$
- $f(\boldsymbol{y}^{obs}, \boldsymbol{y}^{mis}|\boldsymbol{x}, \beta)$ is just the usual likelihood we would specify for fully observed response y
- Estimating the missing responses y^{mis} is equivalent to posterior prediction from the model fitted to the observed data

  $\Rightarrow$ Imputing missing response data under an ignorable mechanism will not affect estimates of model parameters

Missing response data
- assuming missing data mechanism is ignorable (II)

In WinBUGS

- denote missing response values by `NA` in the data file
- specify response distribution (likelihood) as you would for complete data
- missing data are treated as additional unknown parameters

  $\Rightarrow$ WinBUGS will automatically simulate values for the missing observations according to the specified likelihood distribution, conditional on the current values of all relevant unknown parameters

HAMD example: ignorable missing data mechanism

- Assume the missing data mechanism is ignorable for the HAMD example
  - the probability of the score being missing is not related to the current score (or change in score since the previous week)
- Use the same model and WinBUGS code as for the complete case analysis, but the data now includes incomplete records, with the missing values denoted by NA
- Extract of data file

```
list(N=367, W=5, T=3,
hamd=structure(.Data= c(...23, 15, 6, 9, 11,
  ...
  25, 15, 16, 10, NA,
  ...
  20, 29, 31, NA, NA,
  ...
  23, 19, NA, NA, NA, ...), .Dim=c(367, 5)) ...
```

HAMD example: impact on treatment comparisons

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

Including all the partially observed cases in the analysis provides stronger evidence that:

- treatment 2 is more effective than treatment 1
- treatment 2 is more effective than treatment 3

HAMD example: informative missing data mechanism

- However, if we think the probability of the score being missing might be related to the current score, then we must jointly fit
  1. an analysis model - already defined
  2. a model for the missing data mechanism
- Assume that the probability of drop-out is related to the current score, then model the missing response indicator

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

  where $\bar{y}$ is the mean score

- typically, very little information about θ_1 in data
- information depends on parametric model assumptions and error distribution
- advisable to use informative priors (Mason, 2009; Mason, 2008)

HAMD Example: code for model of missingness

WinBUGS code for the model of missingness

```
# Model of Missingness
for (w in 3:W) { # weeks with drop-out
  for (i in 1:N[w]) {
    # exclude individuals who have already dropped out
    hamd.ind[i,w]~dbern(p[i,w])
    # probability of drop-out depends on current score
    logit(p[i,w])<-theta0+theta1*(hamd[i,w]-hamd.mean)
  }
}
hamd.mean<-16
# mildly informative priors
theta0~dlogis(0,1)
theta1~dnorm(0,1.48)  # limit to 5 fold change
```

HAMD Example: MAR v MNAR

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

Allowing for informative missingness with dependence on the current HAMD score:

- has a slight impact on the treatment comparisons
- yields a 95% interval comparing treatments 1 & 3 that includes 0

HAMD Example: sensitivity analysis

- Since the true missingness mechanism is unknown and cannot be checked, sensitivity analysis is essential
- We have already assessed the results of
  - assuming the missingness mechanism is ignorable
  - using an informative missingness mechanism of the form

$$\text{logit}(p_i) = \theta_0 + \theta_1(y_{iw} - \bar{y})$$

- However, we should also look at alternative informative missingness mechanisms, e.g.
  - allow drop-out probability to be dependent on the *change* in score

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

  - allow different θ for each treatment
  - use different prior distributions

HAMD Example: sensitivity analysis results

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

- This is a sensitivity analysis, we do NOT choose the “best" model
- Model comparison with missing data is very tricky
  - we cannot use the DIC automatically generated by WinBUGS on its own (Mason et al., 2010)
- The range of results should be presented

Missing covariate data
- assuming missing data mechanism is ignorable (I)

In this case, $\boldsymbol{z}^{mis} = \boldsymbol{x}^{mis}$ and $\boldsymbol{z}^{obs} = (y, \boldsymbol{x}^{obs})$

- To include records with missing covariates (associated response and other covariates observed) we
  - now have to treat covariates as random variables rather than fixed constants
  - must build an imputation model to predict their missing values
- Typically, the joint model, $f(\boldsymbol{z}^{obs}, \boldsymbol{z}^{mis}|\beta) = f(y, \boldsymbol{x}^{obs}, \boldsymbol{x}^{mis}|\beta)$ is factorised as

$$f(y, \boldsymbol{x}^{obs}, \boldsymbol{x}^{mis}|\beta) = f(y|\boldsymbol{x}^{obs}, \boldsymbol{x}^{mis}, \beta_y) f(\boldsymbol{x}^{obs}, \boldsymbol{x}^{mis}|\beta_x)$$

where β is partitioned into conditionally independent subsets (β_y, β_x)

Missing covariate data
- assuming missing data mechanism is ignorable (II)

- The first term in the joint model factorisation, $f(y|\boldsymbol{x}^{obs}, \boldsymbol{x}^{mis}, \beta_y)$, is the usual likelihood for the response given fully observed covariates
- The second term, $f(\boldsymbol{x}^{obs}, \boldsymbol{x}^{mis}|\beta_x)$ can be thought of as a 'prior model' for the covariates (which are treated as random variables, not fixed constants), e.g.
  - joint prior distribution, say MVN
  - regression model for each variable with missing values
- It is not necessary to explicitly include response, y, as a predictor in the prior imputation model for the covariates, as its association with $\boldsymbol{x}$ is already accounted for by the first term in the joint model factorisation (unlike multiple imputation)

Missing covariate data
- assuming missing data mechanism is ignorable (III)

In WinBUGS

- Denote missing covariate values by `NA` in the data file
- Specify usual regression analysis model, which will depend on partially observed covariates
- In addition, specify prior distribution or regression model for the covariate(s) with missing values
- WinBUGS will automatically simulate values from the posterior distribution of the missing covariates (which will depend on the prior model for the covariates and the likelihood contribution from the corresponding response variable)
- Uncertainty about the covariate imputations is automatically accounted for in the analysis model

LBW Example: low birth weight data

- Study objective: is there an association between trihalomethane (THM) concentrations and the risk of full term low birth weight?
  - THM is a by-product of chlorine water disinfection potentially harmful for reproductive outcomes
- The variables we will use are:
  - Y: binary indicator of low birth weight (outcome)
  - X: binary indicator of THM concentrations (exposure of interest)
  - $\boldsymbol{C}$: mother's age, baby gender, deprivation index (vector of measured confounders)
  - U: smoking (a partially measured confounder)
- We have data for 8969 individuals, but only 931 have an observed value for smoking
  - 90% of individuals will be discarded if we use CC

LBW Example: missingness assumptions

- Assume that *smoking* is MAR
  - probability of smoking being missing does not depend on whether the individual smokes
  - this assumption is reasonable as the missingness is due to the sample design of the underlying datasets
- Also assume that the other assumptions for ignorable missingness hold (see slide 15), so we do not need to specify a model for the missingness mechanism
- However, since *smoking* is a covariate, we must specify an imputation model if we wish to include individuals with missing values of *smoking* in our dataset
  - the design depends on the deprivation index, so this should be included in the imputation model

LBW Example: specification of joint model

- Analysis model: logistic regression for outcome, low birth weight

$$Y_i \sim Bernoulli(p_i)$$
$$logit(p_i) = \beta_0 + \beta_X X_i + \boldsymbol{\beta}_C^T \boldsymbol{C}_i + \beta_U U_i$$
$$\beta_0, \beta_X, \ldots \sim \text{Normal}(0, 10000^2)$$

- Imputation model: logistic regression for missing covariate, smoking

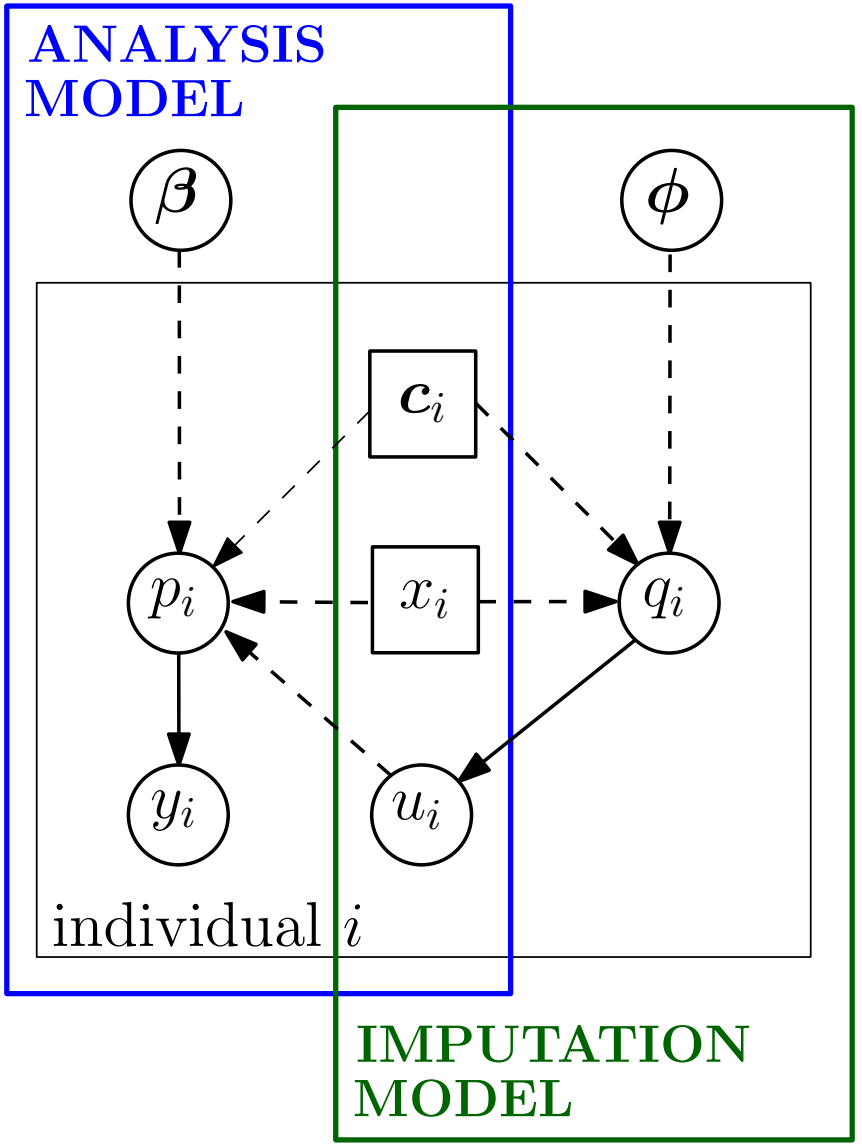
$$U_i \sim Bernoulli(q_i)$$
$$logit(q_i) = \phi_0 + \phi_X X_i + \boldsymbol{\phi}_C^T \boldsymbol{C}_i$$
$$\phi_0, \phi_X, \ldots \sim \text{Normal}(0, 10000^2)$$

- Unlike multiple imputation, we do not need to include Y as a predictor in the imputation model

LBW Example: WinBUGS code

```
### ANALYSIS MODEL ###
for (i in 1:N) { # N individuals
  lbw[i] ~ dbern(p[i])
  logit(p[i]) <- beta0+beta.X[thm[i]]+beta.C1[age[i]]
    +beta.C2[sex[i]]+beta.C3*dep[i]+beta.U*smoke[i]
}
### COVARIATE IMPUTATION MODEL ###
for (i in 1:N) {
  smoke[i] ~ dbern(q[i])
  logit(q[i]) <- phi0+beta.X[thm[i]]+phi.C1[age[i]]
    +phi.C2[sex[i]]+phi.C3*dep[i]
}
### PRIORS FOR OUTCOME MODEL ###
beta0 ~ dnorm(0,0.00000001)
beta.X[1] <- 0   # alias first level of thm beta
beta.X[2] ~ dnorm(0,0.00000001) ...
### PRIORS FOR COVARIATE IMPUTATION MODEL ###
phi0 ~ dnorm(0,0.00000001) ...
### CALCULATE THE ODDS RATIOS ###
thm.or <- exp(beta.thm[2]) ...
```

LBW Example: graphical representation



LBW Example: results

Table: Low Birth Weight Analysis Results

	Odds ratio (95% interval)			
	CC (N=931)		All (N=8969)	
Trihalomethanes				
$> 60 \mu g/L$	2.36	(0.96,4.92)	1.17	(1.01,1.37)
Mother's age				
≤ 25	0.89	(0.32,1.93)	1.05	(0.74,1.41)
$25 - 29^{\star}$		1		1
$30 - 34$	0.13	(0.00,0.51)	0.80	(0.55,1.14)
≥ 35	1.53	(0.39,3.80)	1.14	(0.73,1.69)
Male baby	0.84	(0.34,1.75)	0.76	(0.58,0.95)
Deprivation index	1.74	(1.05,2.90)	1.34	(1.17,1.53)
Smoking	1.86	(0.73,3.89)	1.92	(0.80,3.82)

* Reference group

- CC analysis is very uncertain
- Extra records shrink intervals substantially (Best, 2011)

Comments on covariate imputation models

- Covariate imputation model gets more complex if > 1 missing covariates
  - typically need to account for correlation between missing covariates
  - could assume multivariate normality if covariates all continuous
  - for mixed binary, categorical and continuous covariates, could fit latent variable (multivariate probit) model (Chib and Greenberg 1998; Best 2011)
- If we assume that *smoking* is MNAR, then we must add a third part to the model
  - a model of missingness with a missingness indicator variable for smoking as the response

Concluding remarks

- Bayesian hierarchical models with both missing responses and covariates have the potential to become quite complicated, particularly if we cannot assume ignorable missingness
- However, the Bayesian framework is well suited to the process of building complex models, linking smaller sub-models as a coherent joint model
- A typical model may consist of 3 parts:
  1. analysis model
  2. covariate imputation model
  3. model of missingness
- Typically need informative priors to help identify selection models for informative non-response
- Sensitivity analysis to examine impact of modelling assumptions for non-ignorable missing data mechanisms is essential (Mason et al., 2010a)

References and Further Reading

Best, NG, Spiegelhalter, DJ, Thomas, A and Brayne, CEG (1996). Bayesian analysis of realistically complex models. *J R Statist Soc A*, **159**, 323–342.

Best, NG (2011). Bayesian approaches for combining multiple data sources to adjust for missing confounders. Plenary Lecture, 4th International Joint Meeting of the Institute of Mathematical Statistics and the International Society for Bayesian Analysis, Utah, Jan 5-7 2011. Available at `www.bias-project.org.uk`.

Breslow, N (1990). Biostatistics and Bayes. *Statistical Science*, **5**, 269–298.

Congdon, P (2001) Bayesian statistical modelling. Wiley.

Diggle, P (1988). An approach to the analysis of repeated measurements. *Biometrics*, **44**, 959–971.

Diggle, P and Kenward MG (1994). Informative Drop-out in Longitudinal Data Analysis (with discussion). *Journal of the Royal Statistical Society C*, **43**, 49–93.

Dunson, D (2001). Commentary: Practical advantages of Bayesian analysis in epidemiologic data. *American Journal of Epidemiology*, **153**, 1222–1226.

Fisher, LD (1996). Comments on Bayesian and frequentist analysis and interpretation of clinical trials — comment. *Controlled Clinical Trials*, **17**, 423–34.

Gelman, A (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, 515–533.

Gelman, A, Carlin, JC, Stern, H and Rubin, DB (2004). *Bayesian Data Analysis*, 2nd edition, Chapman & Hall, New York.

Gustafson, P (2003). *Measurement Error and Misclasification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*, Chapman & Hall/CRC Press.

Kass, RE and Wasserman, L (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343–70.

Little RJA and Rubin DB (2002). *Statistical Analysis with Missing Data*, 2nd edition, Wiley, New Jersey.

Marshall, EC and Spiegelhalter, DJ (2003). Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine*, **22**, 1649–60.

Mason, A (2008). Methodological developments for combining data. Presentation at the 'ESRC Research Methods Festival'. Oxford, July 2008. Available at `www.bias-project.org.uk`.

Mason, A (2009). Bayesian methods for modelling non-random missing data mechanisms in longitudinal studies. PhD thesis, Imperial College London. Available at `www.bias-project.org.uk`.

Mason, A, Richardson, S and Best, N (2010). Using DIC to compare selection models with non-ignorable missing responses. Technical report, Imperial College London. Available at `www.bias-project.org.uk`.

Mason, A, Richardson, S, Plewis I, and Best, N (2010a). Strategy for modelling non-random missing data mechanisms in observational studies using Bayesian methods. Technical report, Imperial College London. Available at `www.bias-project.org.uk`.

O'Hagan, A (2003). HSSS Model Criticism. In *Highly Structured Stochastic Systems*, (eds. PJ Green, NL Hjort, and ST Richardson). Oxford University Press, Oxford.

Richardson, S (1996). Measurement error. In *Markov chain Monte Carlo in Practice*, (eds. DJ Spiegelhalter, WR Gilks, and S Richardson). Chapman & Hall, London, pp. 401-417.

Richardson, S and Best, NG (2003). Bayesian hierarchical models in ecological studies of health-environment effects, *Environmetrics*, **14**, 129-147.

Spiegelhalter, DJ (1998). Bayesian graphical modelling: a case-study in monitoring health outcomes. *Journal of the Royal Statistical Society, Series C*, **47**, 115–133.

Spiegelhalter, DJ, Gilks, WR and Richardson, S (1996). *Markov chain Monte Carlo in Practice*, Chapman & Hall, London.

Spiegelhalter, DJ, Abrams, K and Myles, JP (2004). *Bayesian Approaches to Clinical Trials and Health Care Evaluation*, Wiley, Chichester.

Spiegelhalter, DJ, Best, NG, Carlin, BP, and van der Linde, A (2002). Bayesian measures of model complexity and fit (with discussion). *J Roy Statist Soc B*, **64**, 583–639.

Vaida F and Blanchard S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, **92**, 351–70.