Lecture 1.
Introduction to Bayesian Monte Carlo methods in WINBUGS

Summary

1. Probability as a means of representing uncertainty
2. Bayesian direct probability statements about parameters
3. Probability distributions
4. Monte Carlo simulation
5. Implementation in WinBUGS (and DoodleBUGS) - Demo
6. Directed graphs for representing probability models
7. Examples

How did it all start?

In 1763, Reverend Thomas Bayes of Tunbridge Wells wrote

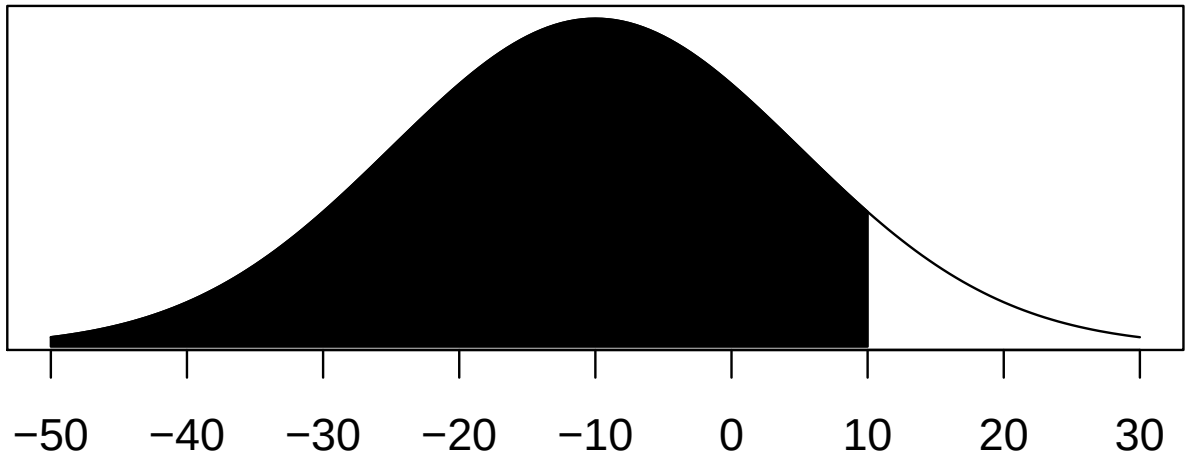
PROBLEM.

Given the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a ſingle trial lies ſomewhere between any two degrees of probability that can be named.

In modern language, given $r \sim \text{Binomial}(\theta, n)$, what is $\Pr(\theta_1 < \theta < \theta_2 | r, n)$?

Basic idea: Direct expression of uncertainty about unknown parameters

eg "There is an 89% probability that the absolute increase in major bleeds is less than 10 percent with low-dose PLT transfusions" (Tinmouth et al, Transfusion, 2004)



Why a direct probability distribution?

1. Tells us what we want: what are plausible values for the parameter of interest?

2. No *P-values*: just calculate relevant tail areas

3. No (difficult to interpret) *confidence intervals*: just report, say, central area that contains 95% of distribution

4. Easy to make predictions (see later)

5. Fits naturally into decision analysis / cost-effectiveness analysis / project prioritisation

6. There is a procedure for adapting the distribution in the light of additional evidence: i.e. *Bayes theorem* allows us to learn from experience

Inference on proportions

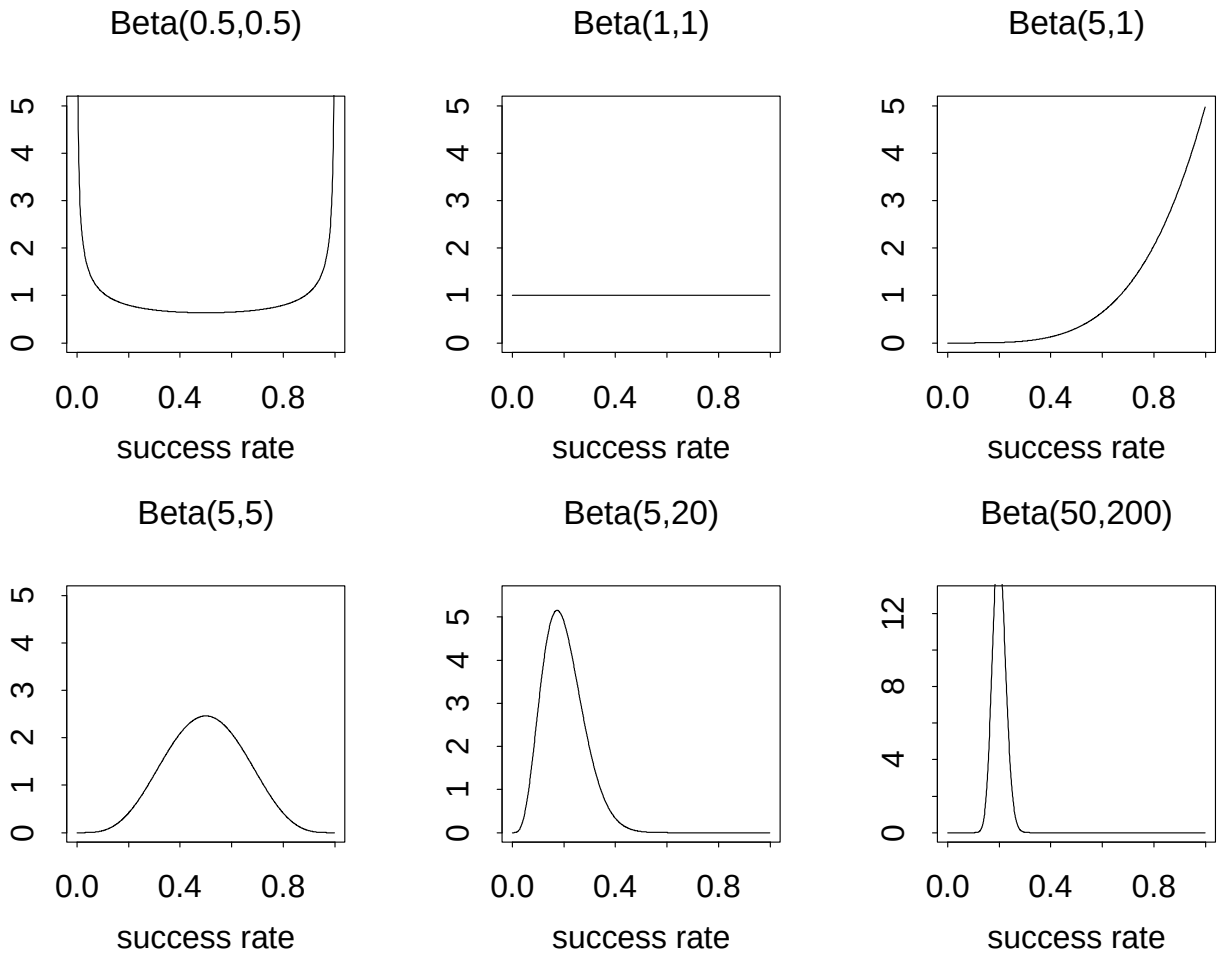
What is a reasonable form for a prior distribution for a proportion?

$\theta \sim$ Beta$[a, b]$ represents a beta distribution with properties:

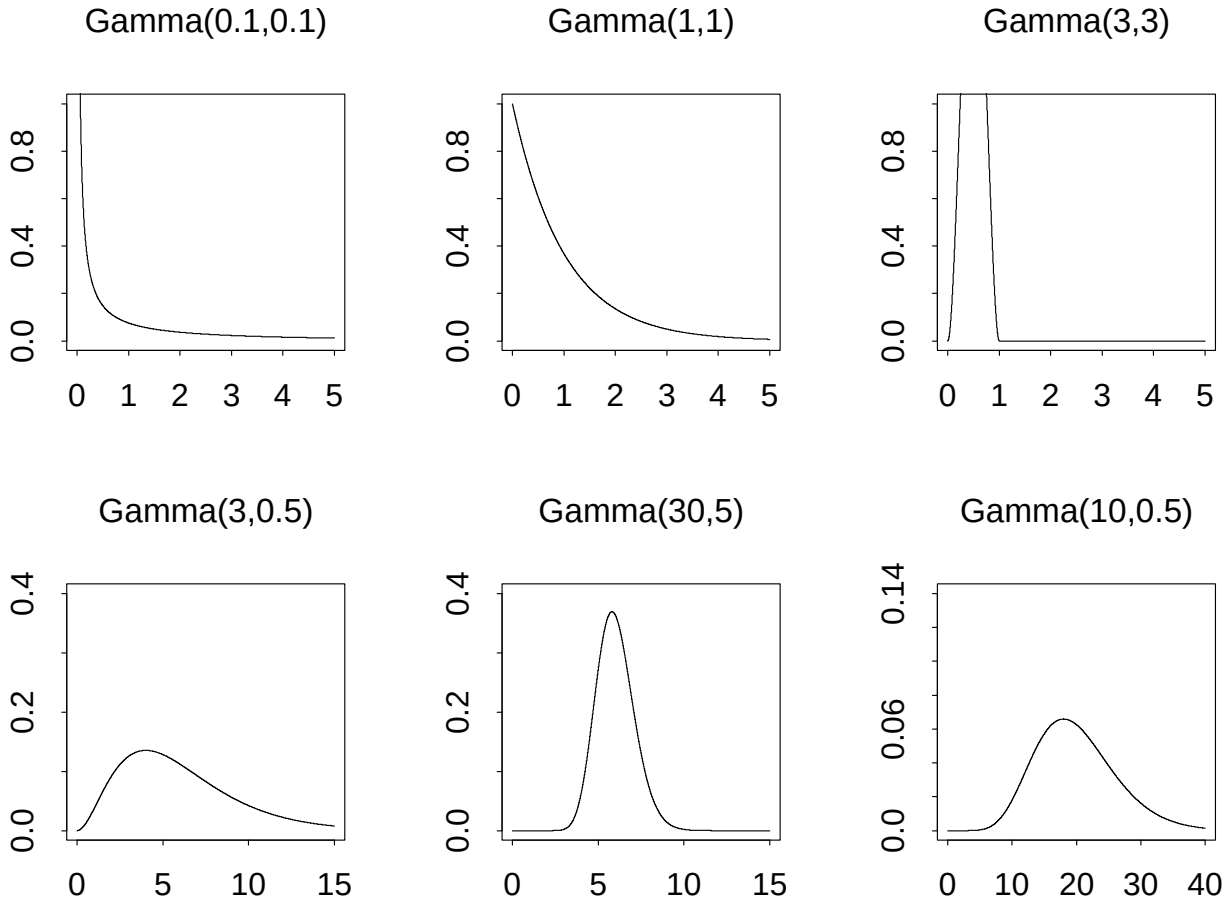
$$\begin{aligned} p(\theta|a,b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\,\theta^{a-1}\,(1-\theta)^{b-1}; \qquad \theta \in (0,1) \\ \mathsf{E}(\theta|a,b) &= \frac{a}{a+b} \\ \mathsf{V}(\theta|a,b) &= \frac{ab}{(a+b)^2(a+b+1)}: \end{aligned}$$

WinBUGS notation: `theta ~ dbeta(a,b)`

Beta distribution



Gamma distribution



The Gamma distribution

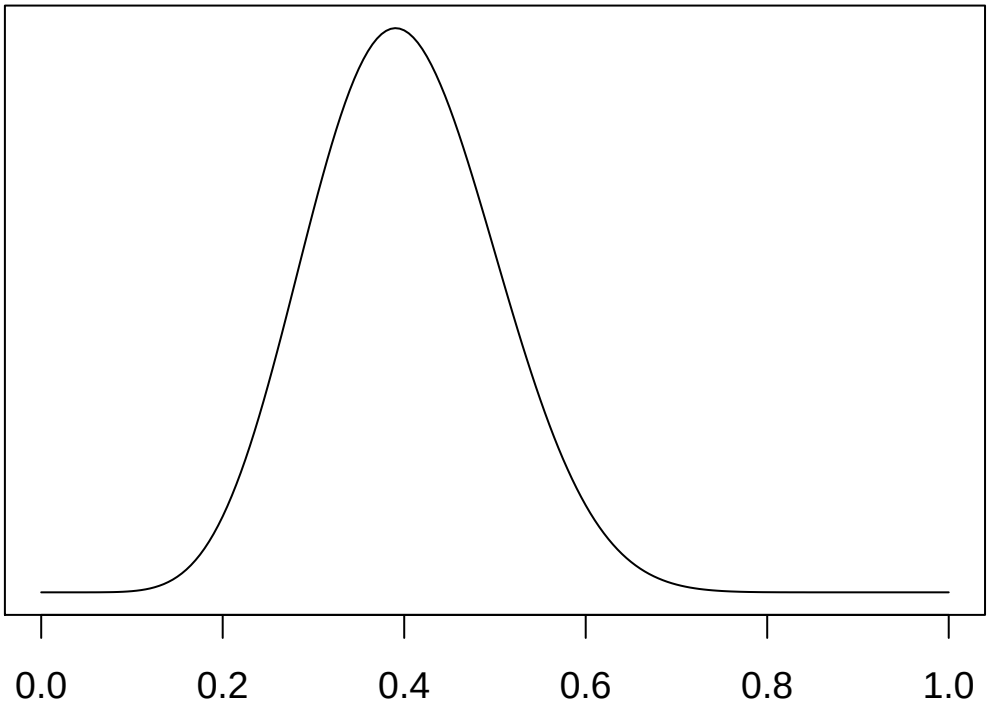
Flexible distribution for positive quantities. If $Y \sim$ Gamma$[a, b]$

$$\begin{aligned} p(y|a,b) &= \frac{b^a}{\Gamma(a)} y^{a-1}\, e^{-by}; \qquad y \in (0,\infty) \\ \mathsf{E}(Y|a,b) &= \frac{a}{b} \\ \mathsf{V}(Y|a,b) &= \frac{a}{b^2}. \end{aligned}$$

- Gamma[1,b] distribution is exponential with mean $1/b$
- Gamma$[\frac{v}{2}, \frac{1}{2}]$ is a Chi-squared χ^2_v distribution on v degrees of freedom
- $Y \sim$ Gamma[0.001,0.001] means that $p(y) \propto 1/y$, or that $\log Y \approx$ Uniform
- Used as conjugate prior distribution for inverse variances (precisions)
- Used as sampling distribution for skewed positive valued quantities (alternative to log normal likelihood) — MLE of mean is sample mean
- WinBUGS notation: `y ~ dgamma(a,b)`

Example: Drug

- Consider a drug to be given for relief of chronic pain
- Experience with similar compounds has suggested that annual response rates between 0.2 and 0.6 could be feasible
- Interpret this as a distribution with mean = 0.4, standard deviation 0.1
- A Beta[9.2,13.8] distribution has these properties



Beta[9.2, 13.8] prior distribution supporting response rates between 0.2 and 0.6,

Making predictions

Before observing a quantity Y, can provide its predictive distribution by integrating out unknown parameter

$$p(Y) = \int p(Y|\theta)p(\theta)d\theta.$$

Predictions are useful in e.g. cost-effectiveness models, design of studies, checking whether observed data is compatible with expectations, and so on.

If

$$\begin{aligned} \theta &\sim \text{Beta}[a, b] \\ Y_n &\sim \text{Binomial}(\theta, n), \end{aligned}$$

the exact predictive distribution for Y_n is known as the **Beta-Binomial**. It has the complex form

$$p(y_n) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \begin{pmatrix} n \\ y_n \end{pmatrix} \frac{\Gamma(a+y_n)\Gamma(b+n-y_n)}{\Gamma(a+b+n)}.$$

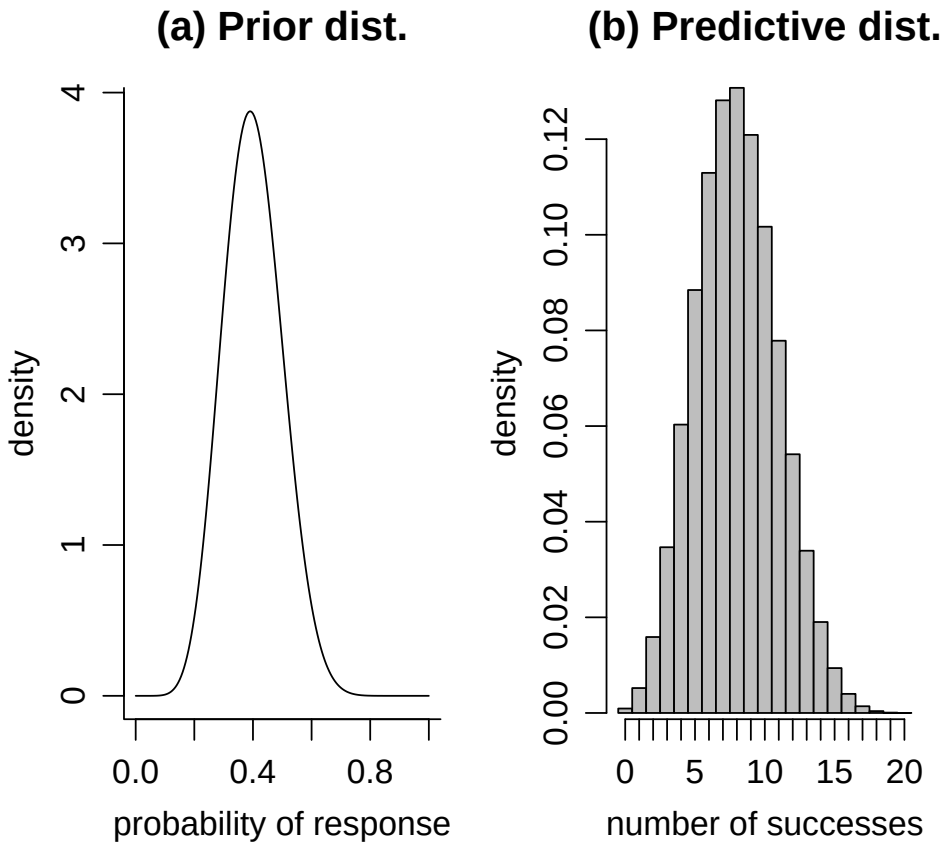
$$\text{E}(Y_n) = n\frac{a}{a+b}$$

If $a = b = 1$ (Uniform distribution), $p(y_n)$is uniform over 0,1,...,n.

But in WinBUGS we can just write

```
theta ~ dbeta(a,b)
Y     ~ dbin(theta,n)
```

and the integration is automatically carried out and does not require algebraic cleverness.



(a) is the Beta prior distribution

(b) is the predictive Beta-Binomial distribution of the number of successes Y in the next 20 trials

From Beta-binomial distribution, can calculate $P(Y_n \geq 15) = 0.015$.

Example: a Monte Carlo approach to estimating tail-areas of distributions

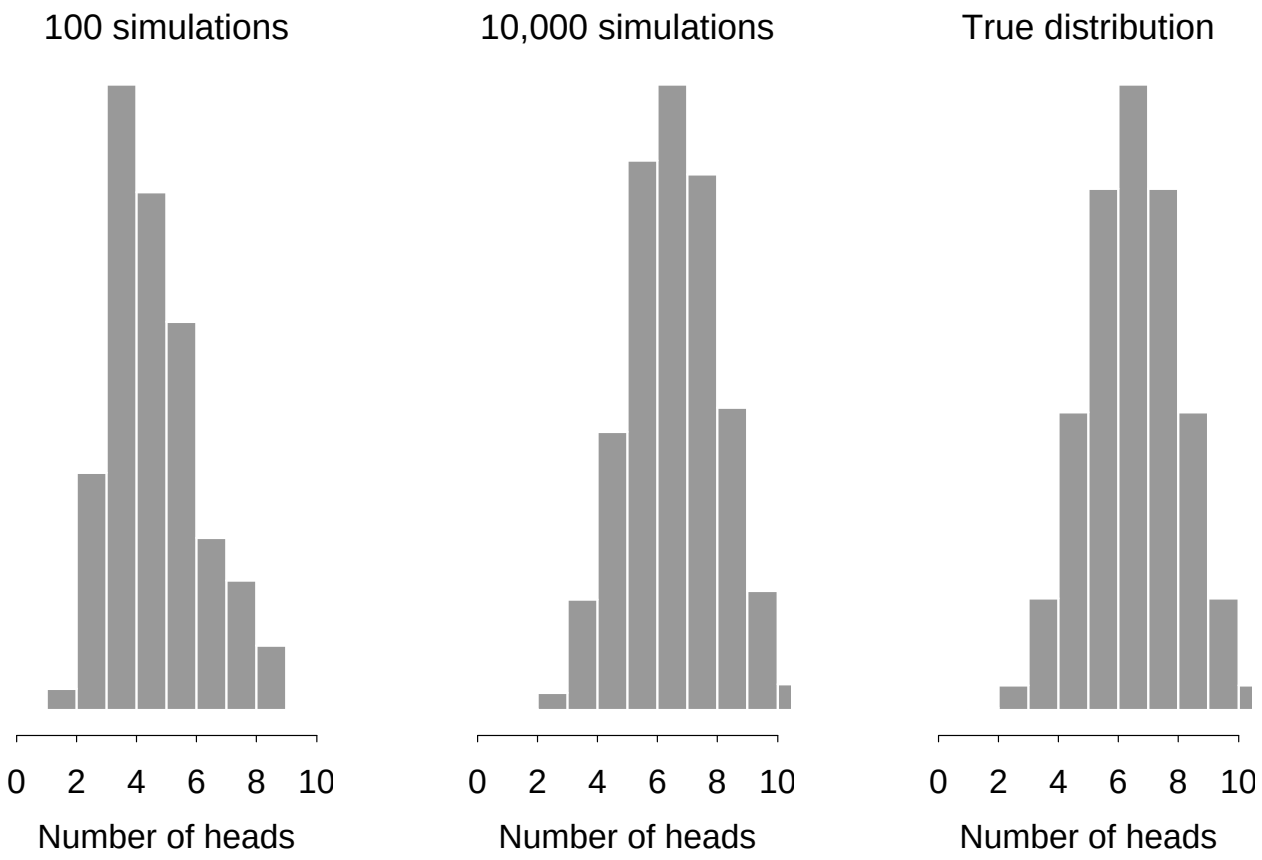
Suppose we want to know the probability of getting 8 or more heads when we toss a fair coin 10 times.

An *algebraic* approach:

$$\begin{aligned}
\Pr(\geq 8 \text{ heads}) &= \sum_{z=8}^{10} p\left(z|\pi = \frac{1}{2}, n = 10\right) \\
&= \binom{10}{8}\left(\frac{1}{2}\right)^8\left(\frac{1}{2}\right)^2 + \binom{10}{9}\left(\frac{1}{2}\right)^9\left(\frac{1}{2}\right)^1 + \binom{10}{10}\left(\frac{1}{2}\right)^{10}\left(\frac{1}{2}\right)^0 \\
&= 0.0547.
\end{aligned}$$

A *physical* approach would be to repeatedly throw a set of 10 coins and count the proportion of throws that there were 8 or more heads.

A *simulation* approach uses a computer to toss the coins!



Proportion with 8 or more 'heads' in 10 tosses:

(a) After 100 'throws' (0.02); (b) after 10,000 throws (0.0577); (c) the true Binomial distribution (0.0547)

General Monte Carlo analysis - 'forward sampling'

Used extensively in risk modelling - can think of as 'adding uncertainty' to a spreadsheet

- Suppose have logical function f containing uncertain parameters
- Can express our uncertainty as a prior distribution
- Simulate many values from this prior distribution
- Calculate f at the simulated values ('iterations')
- Obtain an empirical predictive distribution for f
- Sometimes termed *probabilistic sensitivity analysis*
- Can do in Excel add-ons such as `@RISK` or `Crystal Ball`.

The BUGS program

Bayesian inference Using Gibbs Sampling

- Language for specifying complex Bayesian models
- Constructs object-oriented internal representation of the model
- Simulation from full conditionals using Gibbs sampling
- Current version (WinBUGS 1.4) runs in Windows
- 'Classic' BUGS available for UNIX but this is an old version

WinBUGS is freely available from `http://www.mrc-bsu.cam.ac.uk/bugs`

- Scripts enable WinBUGS 1.4 to run in batch mode or be called from other software
- Interfaces developed for R, Excel, Splus, SAS, Matlab
- OpenBUGS site `http://www.rni.helsinki.fi/openbugs` provides an open source version

Running WinBUGS for Monte Carlo analysis (no data)

1. Open *Specification tool* from *Model* menu.
2. Program responses are shown on bottom-left of screen.
3. Highlight `model` by double-click. Click on *Check model*.
4. Click on *Compile*.
5. Click on *Gen Inits*.
6. Open *Update* from *Model* menu, and *Samples* from *Inference* menu.
7. Type nodes to be monitored into *Sample Monitor*, and click *set* after each.
8. Type * into *Sample Monitor*, and click *trace* to see sampled values.
9. Click on *Update* to generate samples.
10. Type * into *Sample Monitor*, and click *stats* etc to see results on all monitored nodes.

Using WinBUGS for Monte Carlo

The model for the 'coin' example is

$$Y \sim \text{Binomial}(0.5, 10)$$

and we want to know $P(Y \geq 8)$.

This model is represented in the BUGS language as

```
model{
 Y     ~  dbin(0.5,10)
 P8   <-  step(Y-7.5)
 }
```

`P8` is a step function which will take on the value 1 if `Y -7.5` is ≥ 0, *i.e.* `Y` is 8 or more, 0 if 7 or less.

Running this simulation for 100, 10000 and 1000000 iterations, and then taking the empirical mean of `P8`, provided the previous estimated probabilities that Y will be 8 or more.

Some aspects of the BUGS language

- `<-` represents logical dependence, *e.g.* `m <- a + b*x`
- `~` represents stochastic dependence, *e.g.* `r ~ dunif(a,b)`
- Can use arrays and loops

```
for (i in 1:n){
  r[i] ~ dbin(p[i],n[i])
  p[i] ~ dunif(0,1)
  }
```

- Some functions can appear on left-hand-side of an expression, *e.g.*

```
logit(p[i])<- a + b*x[i]
log(m[i])  <- c + d*y[i]
```

- `mean(p[])` to take mean of whole array, `mean(p[m:n])` to take mean of elements m to n. Also for `sum(p[])`.
- `dnorm(0,1)I(0,)` means the prior will be restricted to the range $(0, \infty)$.

Functions in the BUGS language

- `p <- step(x-.7)` $= 1$ if `x` ≥ 0.7, 0 otherwise. Hence monitoring `p` and recording its mean will give the probability that `x`≥ 0.7.
- `p <- equals(x,.7)` $= 1$ if `x` $= 0.7$, 0 otherwise.
- `tau <- 1/pow(s,2)` sets $\tau = 1/s^2$.
- `s <- 1/ sqrt(tau)` sets $s = 1/\sqrt{\tau}$.
- `p[i,k] <- inprod(pi[], Lambda[i,,k])` sets $p_{ik} = \sum_j \pi_j \Lambda_{ijk}$. `inprod2` may be faster.
- See 'Model Specification/Logical nodes' in manual for full syntax.

Some common Distributions

Expression	Distribution	Usage
`dbin`	binomial	`r ~ dbin(p,n)`
`dnorm`	normal	`x ~ dnorm(mu,tau)`
`dpois`	Poisson	`r ~ dpois(lambda)`
`dunif`	uniform	`x ~ dunif(a,b)`
`dgamma`	gamma	`x ~ dgamma(a,b)`

NB. The normal is parameterised in terms of its mean and *precision* = 1/ variance = $1/\text{sd}^2$.

See 'Model Specification/The BUGS language: stochastic nodes/Distributions' in manual for full syntax.

Functions cannot be used as arguments in distributions (you need to create new nodes).

Drug example: Monte Carlo predictions

Our prior distribution for proportion of responders in one year θ was Beta[9.2, 13.8].

Consider situation *before* giving 20 patients the treatment. What is the chance if getting 15 or more responders?

$$\begin{aligned} \theta &\sim \text{Beta}[9.2, 13.8] \quad \text{prior distribution} \\ y &\sim \text{Binomial}[\theta, 20] \quad \text{sampling distribution} \\ P_{\text{crit}} &= P(y \geq 15) \quad \text{Probability of exceeding critical threshold} \end{aligned}$$

```
# In BUGS syntax:
 model{
 theta     ~ dbeta(9.2,13.8)        # prior distribution
 y         ~ dbin(theta,20)         # sampling distribution
 P.crit   <- step(y-14.5)           # =1 if y >= 15, 0 otherwise

 }
```

WinBUGS output and exact answers

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
theta	0.4008	0.09999	9.415E-4	0.2174	0.3981	0.6044	1	10000
y	8.058	2.917	0.03035	3.0	8.0	14.0	1	10000
P.crit	0.0151	0.122	0.001275	0.0	0.0	0.0	1	10000

Note that the mean of the 0-1 indicator `P.crit` provides the estimated tail-area probability.

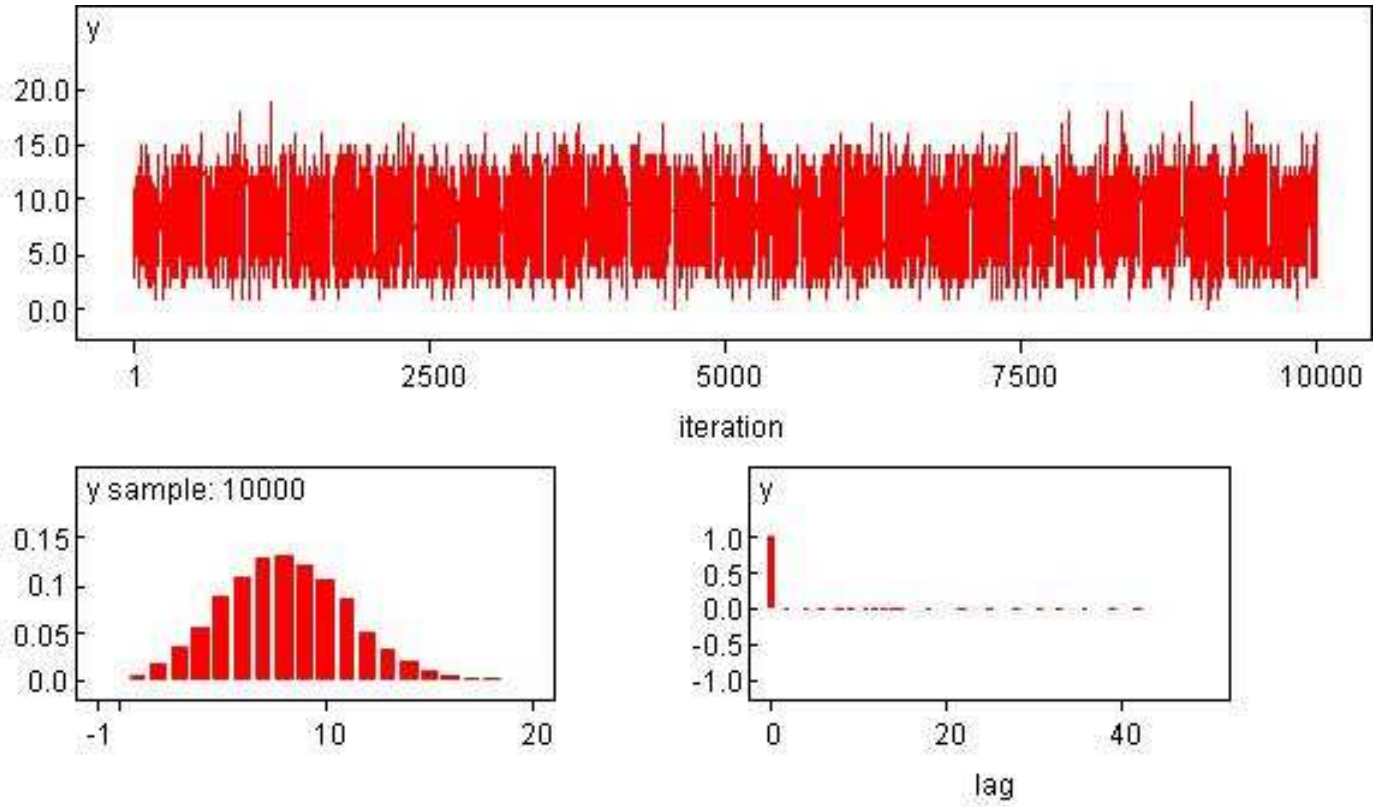
Exact answers from closed-form analysis:

- θ: mean 0.4 and standard deviation 0.1
- y: mean 8 and standard deviation 2.93.
- Probability of at least 15: 0.015

These are independent samples, and so MC error = SD/$\sqrt{\text{No.iterations}}$.

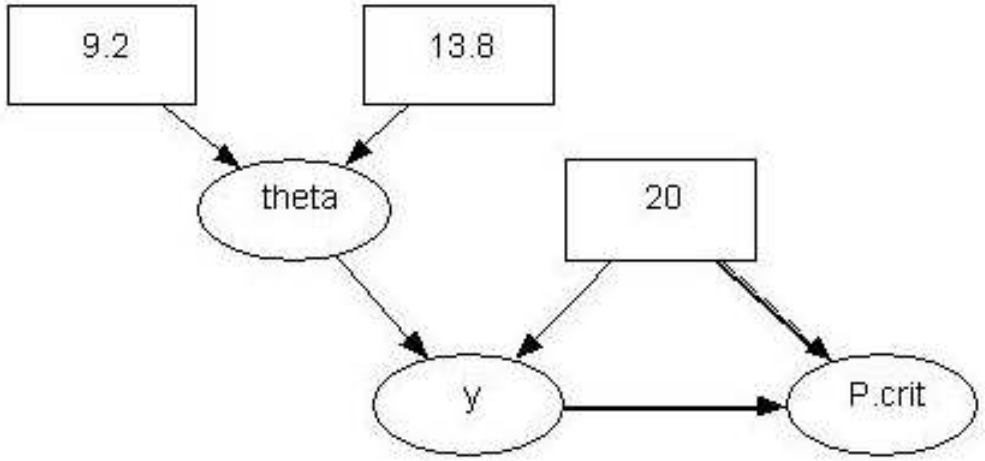
Can achieve arbitrary accuracy by running the simulation for longer.

WinBUGS output



Independent samples, and so no auto-correlation and no concern with convergence.

Graphical representation of models



- *Doodle* represents each quantity as a node in directed acyclic graph (DAG).
- Constants are placed in rectangles, random quantities in ovals
- Stochastic dependence is represented by a single arrow, and logical function as double arrow
- WinBUGS allows models to be specified graphically and run directly from the graphical interface
- Can write code from Doodles
- Good for explanation, but can be tricky to set up

Script for running Drug Monte Carlo example

Run from `Model/Script` menu

```
display('log')          # set up log file
check('c:/bugscourse/drug-MC')          # check syntax of model
#  data('c:/bugscourse/drug-data')      # load data file if there is one
compile(1)              # generate code for 1 simulations
# inits(1,'c:/bugscourse/drug-in1')     # load initial values if necessary
gen.inits()             # generate initial values for all unknown quantities
                        # not given initial values

set(theta)              # monitor the true response rate
set(y)                  # monitor the predicted number of successes
set(P.crit)             # monitor whether a critical number of successes occur
trace(*)                # watch some simulated values (although slows down simulation)
update(10000)           # perform 10000 simulations
history(theta)          # Trace plot of samples for theta
stats(*)                # Calculate summary statistics for all monitored quantities
density(theta)          # Plot distribution of theta
density(y)              # Plot distribution of y
```

Example: Power — uncertainty in a power calculation

- a randomised trial planned with n patients in each of two arms
- response with standard deviation $\sigma = 1$
- aimed to have Type 1 error 5% and 80% power
- to detect a true difference of $\theta = 0.5$ in mean response between the groups

Necessary sample size per group is

$$n = \frac{2\sigma^2}{\theta^2}(0.84 + 1.96)^2 = 63$$

Alternatively, for fixed n, the power is

$$\text{Power} = \Phi\left(\sqrt{\frac{n\theta^2}{2\sigma^2}} - 1.96\right).$$

Suppose we wish to express uncertainty concerning both θ and σ, e.g.

$$\theta \sim \mathrm{N}[0.5, 0.1^2], \quad \sigma \sim \mathrm{N}[1, 0.3^2].$$

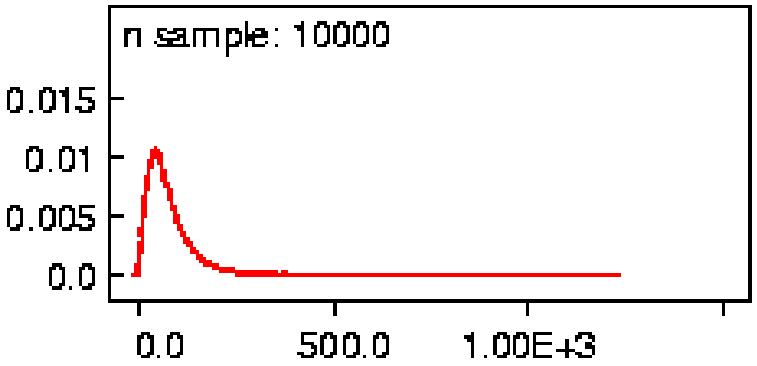
1. Simulate values of θ and σ from their prior distributions

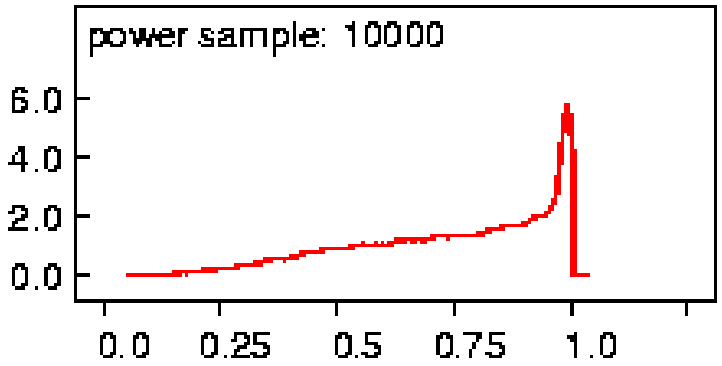
2. Substitute them in the formulae

3. Obtain a predictive distribution over n or Power

```
prec.sigma  <- 1/(0.3*0.3)      # transform sd to precision=1/sd2
prec.theta  <- 1/(0.1*0.1)
sigma      ~ dnorm(1, prec.sigma)I(0,)
theta      ~ dnorm(.5, prec.theta)I(0,)
n      <- 2 * pow(    (.84 +1.96) * sigma / theta ,  2 )
power  <- phi(  sqrt(63/2)* theta /sigma  -1.96  )
prob70 <-step(power-.7)
```

	Median	95% interval
n	62.5	9.3 to 247.2
Power (%)	80	29 to 100





For n= 63, the median power is 80%, and a trial of 63 patients per group could be seriously underpowered. There is a 37% chance that the power is less than 70%.

Lecture 2.
Introduction to conjugate Bayesian inference

What are Bayesian methods?

- Bayesian methods have been widely applied in many areas:
  - medicine / epidemiology
  - genetics
  - ecology
  - environmental sciences
  - social and political sciences
  - finance
  - archaeology
  - .....

- Motivations for adopting Bayesian approach vary:
  - natural and coherent way of thinking about science and learning
  - pragmatic choice that is suitable for the problem in hand

Spiegelhalter et al (2004) define a Bayesian approach as

> 'the explicit use of external evidence in the design, monitoring, analysis, interpretation and reporting of a [scientific investigation]'

They argue that a Bayesian approach is:

- more flexible in adapting to each unique situation
- more efficient in using all available evidence
- more useful in providing relevant quantitative summaries

than traditional methods

Example

A clinical trial is carried out to collect evidence about an unknown 'treatment effect'

Conventional analysis

- p-value for H_0: treatment effect is zero
- Point estimate and CI as summaries of size of treatment effect

Aim is to learn what this trial tells us about the treatment effect

Bayesian analysis

- Asks: 'how should this trial change our opinion about the treatment effect?'

The Bayesian analyst needs to explicitly state

- a reasonable opinion concerning the plausibility of different values of the treatment effect *excluding* the evidence from the trial (the **prior distribution**)
- the support for different values of the treatment effect based *solely* on data from the trial (the **likelihood**),

and to combine these two sources to produce

- a final opinion about the treatment effect (the **posterior distribution**)

The final combination is done using Bayes theorem, which essentially weights the likelihood from the trial with the relative plausibilities defined by the prior distribution

One can view the Bayesian approach as a formalisation of the process of learning from experience

Posterior distribution forms basis for all inference — can be summarised to provide

- point and interval estimates of treatment effect
- point and interval estimates of any function of the parameters
- probability that treatment effect exceeds a clinically relevant value
- prediction of treatment effect in a new patient
- prior information for future trials
- inputs for decision making
- ....

Bayes theorem and its link with Bayesian inference

Bayes' theorem Provable from probability axioms

Let A and B be events, then

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}.$$

If A_i is a set of mutually exclusive and exhaustive events (*i.e.* $p(\bigcup_i A_i) = \sum_i p(A_i) = 1$), then

$$p(A_i|B) = \frac{p(B|A_i)p(A_i)}{\sum_j p(B|A_j)p(A_j)}.$$

Example: use of Bayes theorem in diagnostic testing

- A new HIV test is claimed to have "95% sensitivity and 98% specificity"
- In a population with an HIV prevalence of 1/1000, what is the chance that patient testing positive actually has HIV?

Let A be the event that patient is truly HIV positive, $\overline{A}$ be the event that they are truly HIV negative.

Let B be the event that they test positive.

We want $p(A|B)$.

"95% sensitivity" means that $p(B|A) = .95$.

"98% specificity" means that $p(B|\overline{A}) = .02$.

Now Bayes theorem says

$$p(A|B) = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|\overline{A})p(\overline{A})}.$$

Hence $p(A|B) = \frac{.95 \times .001}{.95 \times .001 + .02 \times .999} = .045$.

Thus over 95% of those testing positive will, in fact, not have HIV.

- Our intuition is poor when processing probabilistic evidence
- The vital issue is *how should this test result change our belief that patient is HIV positive?*
- The disease prevalence can be thought of as a *'prior'* probability ($p = 0.001$)
- Observing a positive result causes us to modify this probability to $p = 0.045$. This is our *'posterior'* probability that patient is HIV positive.
- Bayes theorem applied to *observables* (as in diagnostic testing) is uncontroversial and established
- More controversial is the use of Bayes theorem in general statistical analyses, where *parameters* are the unknown quantities, and their prior distribution needs to be specified — this is **Bayesian inference**

Bayesian inference

Makes fundamental distinction between

- Observable quantities x, i.e. the data
- Unknown quantities θ

θ can be statistical parameters, missing data, mismeasured data ...

$\rightarrow$ parameters are treated as random variables

$\rightarrow$ in the Bayesian framework, we make probability statements about model parameters

! in the frequentist framework, parameters are fixed non-random quantities and the probability statements concern the data

As with any statistical analysis, we start by positing a model which specifies

$$p(x \mid \theta)$$

This is the **likelihood**, which relates all variables into a **'full probability model'**

From a Bayesian point of view

- θ is unknown so should have a **probability distribution** reflecting our uncertainty about it before seeing the data
  $\rightarrow$ need to specify a **prior distribution** $p(\theta)$

- x is known so we should condition on it
  $\rightarrow$ use Bayes theorem to obtain conditional probability distribution for unobserved quantities of interest given the data:

  $$p(\theta \mid x) = \frac{p(\theta)\, p(x \mid \theta)}{\int p(\theta)\, p(x \mid \theta)\, d\theta} \propto p(\theta)\, p(x \mid \theta)$$

  This is the **posterior distribution**

The prior distribution $p(\theta)$, expresses our uncertainty about θ **before** seeing the data.

The posterior distribution $p(\theta \mid x)$, expresses our uncertainty about θ **after** seeing the data.

Inference on proportions using a continuous prior

Suppose we now observe r positive responses out of n patients.

Assuming patients are independent, with common unknown response rate θ, leads to a binomial likelihood

$$p(r|n,\theta) \;=\; \left(\begin{array}{c} n \\ r \end{array} \right) \theta^r (1-\theta)^{n-r} \;\propto\; \theta^r (1-\theta)^{n-r}$$

θ needs to be given a continuous prior distribution.

Suppose that, before taking account of the evidence from our trial, we believe all values for θ are equally likely (is this plausible?) $\Rightarrow \theta \sim$ Unif(0, 1) i.e. $p(\theta) = \frac{1}{1-0} = 1$

Posterior is then

$$p(\theta|r,n) \propto \theta^r (1-\theta)^{(n-r)} \times 1$$

This has form of the *kernel* of a Beta(r+1, n-r+1) distribution (see lect 1), where

$$\theta \;\sim\; \text{Beta}(a,b) \;\equiv\; \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, \theta^{a-1}(1-\theta)^{b-1}$$

To represent external evidence that some response rates are more plausible than others, it is mathematically convenient to use a Beta(a, b) prior distribution for θ

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$$

Combining this with the binomial likelihood gives a posterior distribution

$$\begin{aligned} p(\theta \mid r, n) &\propto p(r \mid \theta, n)p(\theta) \\ &\propto \theta^{r}(1-\theta)^{n-r}\theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{r+a-1}(1-\theta)^{n-r+b-1} \\ &\propto \text{Beta}(r+a,\ n-r+b) \end{aligned}$$

Comments

- When the prior and posterior come from the same family of distributions the prior is said to be **conjugate** to the likelihood
  - Occurs when prior and likelihood have the same 'kernel'

- Recall from lecture 1 that a Beta(a, b) distribution has

$$\begin{aligned} \text{mean} &= a/(a+b), \\ \text{variance} &= ab/\left[(a+b)^2(a+b+1)\right] \end{aligned}$$

  Hence posterior mean is $E(\theta|r,n) = (r+a)/(n+a+b)$

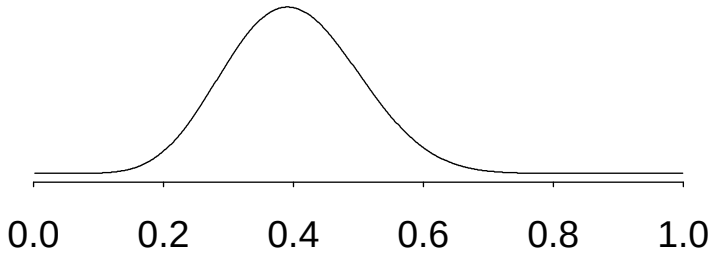
- a and b are equivalent to observing a priori $a-1$ successes in $a+b-2$ trials $\rightarrow$ can be elicited

- With fixed a and b, as r and n increase, $E(\theta|r,n) \rightarrow r/n$ (the MLE), and the variance tends to zero
  - This is a general phenomenon: as n increases, posterior distribution gets more concentrated and the likelihood dominates the prior

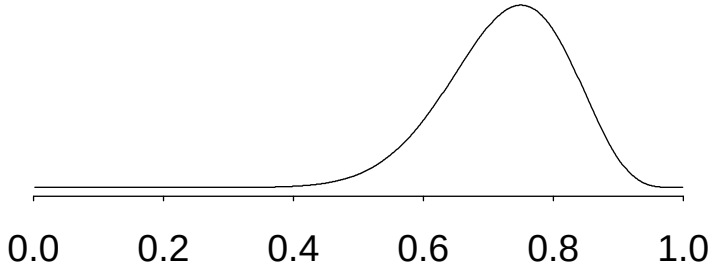
- A Beta(1, 1) is equivalent to Uniform(0, 1)

Example: Drug

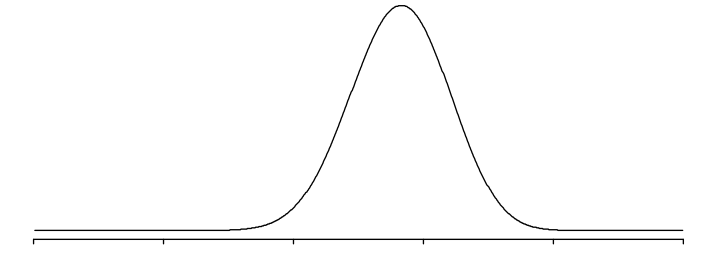
- Recall example from lecture 1, where we consider early investigation of a new drug
- Experience with similar compounds has suggested that response rates between 0.2 and 0.6 could be feasible
- We interpreted this as a distribution with mean = 0.4, standard deviation 0.1 and showed that a Beta(9.2,13.8) distribution has these properties
- Suppose we now treat $n = 20$ volunteers with the compound and observe $y = 15$ positive responses



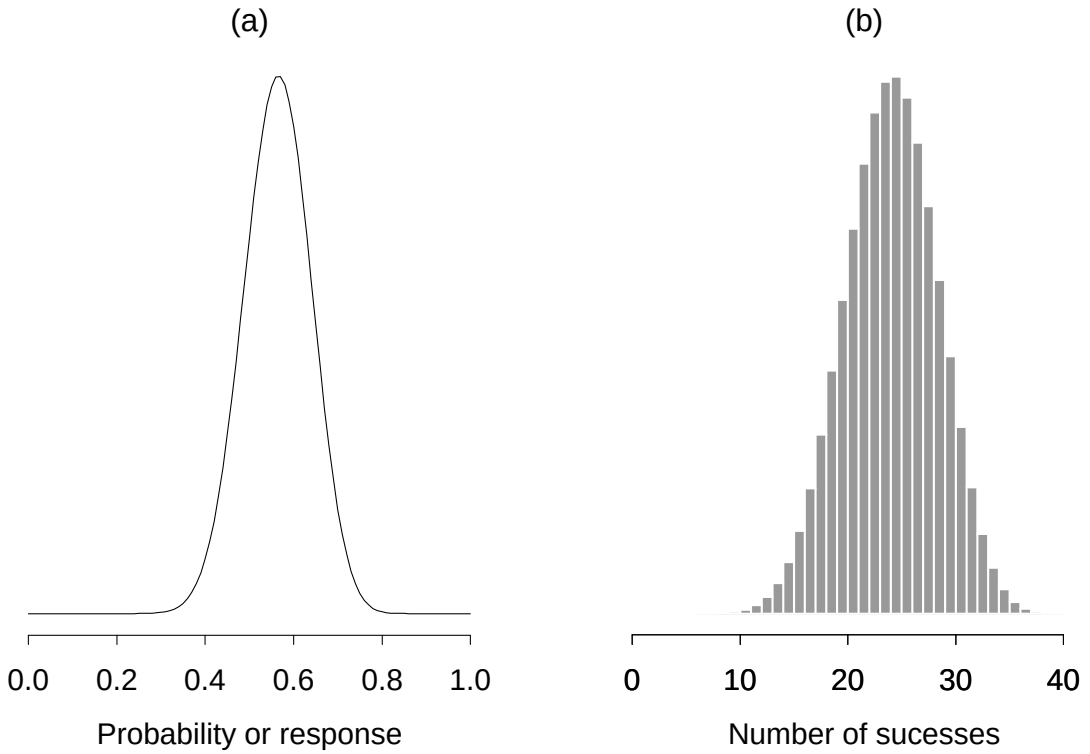
Beta(9.2, 13.8) prior distribution supporting response rates between 0.2 and 0.6



Likelihood arising from a Binomial observation of 15 successes out of 20 cases



Parameters of the Beta distribution are updated to (a+15, b+20-15) = (24.2, 18.8): mean 24.2/(24.2+18.8) = 0.56



(a) Beta posterior distribution after having observed 15 successes in 20 trials

(b) predictive Beta-Binomial distribution of the number of successes $\tilde{y}_{40}$ in the next 40 trials with mean 22.5 and standard deviation 4.3

Suppose we would consider continuing a development program if the drug managed to achieve at least a further 25 successes out of these 40 future trials

From Beta-binomial distribution, can calculate $P(\tilde{y}_{40} \geq 25) = 0.329$

Drug (continued): learning about parameters from data using Markov chain Monte-Carlo (MCMC) methods in WInBUGS

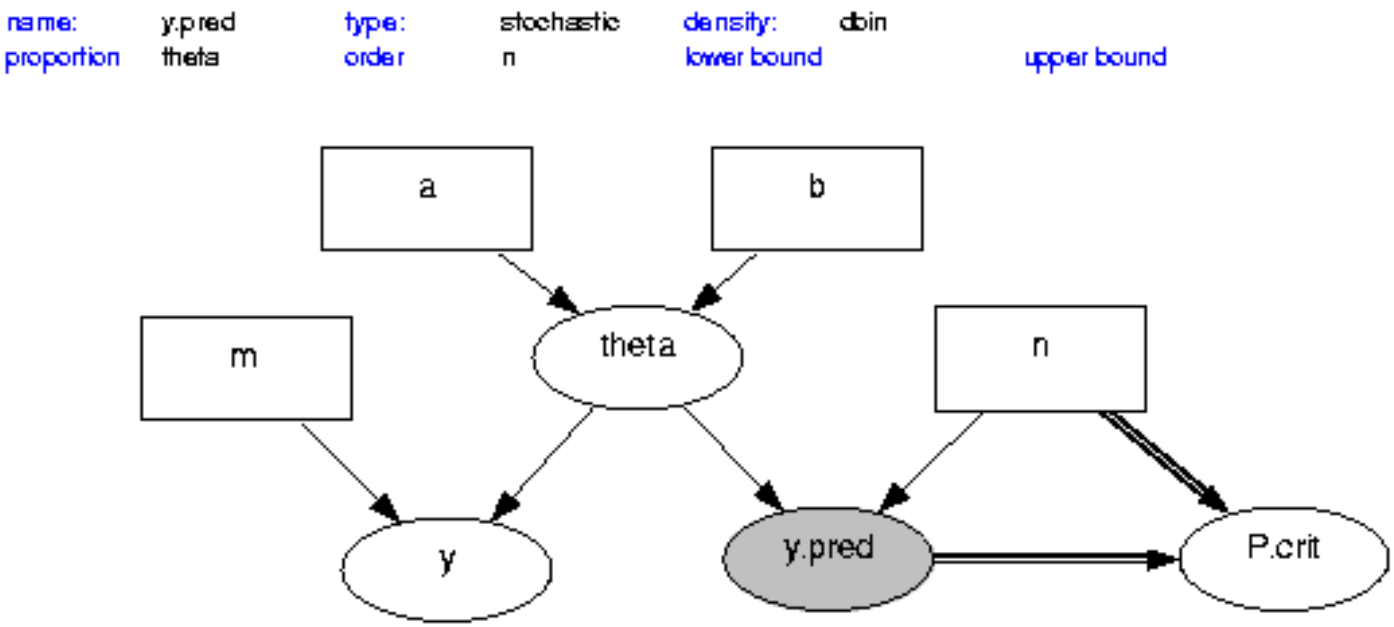
- Using MCMC (e.g. in WinBUGS), no need to explicitly specify posterior
- Can just specify the prior and likelihood separately
- WinBUGS contains algorithms to evaluate the posterior given (almost) arbitrary specification of prior and likelihood
  - posterior doesn't need to be closed form
  - but can (usually) recognise conjugacy when it exists

The drug model can be written

$$\begin{aligned}
\theta &\sim \text{Beta}[a, b] && \text{prior distribution} \\
y &\sim \text{Binomial}[\theta, m] && \text{sampling distribution} \\
y_{\text{pred}} &\sim \text{Binomial}[\theta, n] && \text{predictive distribution} \\
P_{\text{crit}} &= P(y_{\text{pred}} \geq n_{\text{crit}}) && \text{Probability of exceeding critical threshold}
\end{aligned}$$

```
# In BUGS syntax:

# Model description '
model {
  theta     ~ dbeta(a,b)              # prior distribution
  y         ~ dbin(theta,m)           # sampling distribution
  y.pred    ~ dbin(theta,n)           # predictive distribution
  P.crit   <- step(y.pred-ncrit+0.5)  # =1 if y.pred >= ncrit, 0 otherwise
}
```

Graphical representation of models



Note that adding data to a model is simply extending the graph.

Data files

Data can be written after the model description, or held in a separate .txt or .odc file

```
 list( a = 9.2,    # parameters of prior distribution
  b = 13.8,
  y = 15,      # number of successes
  m = 20,      # number of trials
  n = 40,      # future number of trials
 ncrit = 25)  # critical value of future successes
```

Alternatively, in this simple example, we could have put all data and constants into model description:

```
model{
  theta     ~ dbeta(9.2,13.8)      # prior distribution
  y         ~ dbin(theta,20)       # sampling distribution
  y.pred    ~ dbin(theta,40)       # predictive distribution
  P.crit   <- step(y.pred-24.5)    # =1 if y.pred >= ncrit, 0 otherwise

  y         <- 15
}
```

The WinBUGS data formats

WinBUGS accepts data files in:

1. Rectangular format (easy to cut and paste from spreadsheets)

```
 n[] r[]
 47  0
148 18
 ...
360 24
END
```

2. S-Plus format:

```
list(N=12,n = c(47,148,119,810,211,196,
          148,215,207,97,256,360),
     r = c(0,18,8,46,8,13,9,31,14,8,29,24))
```

Generally need a 'list' to give size of datasets etc.

Initial values

- WinBUGS can automatically generate initial values for the MCMC analysis using *gen inits*
- Fine if have informative prior information
- If have fairly 'vague' priors, better to provide reasonable values in an initial-values list

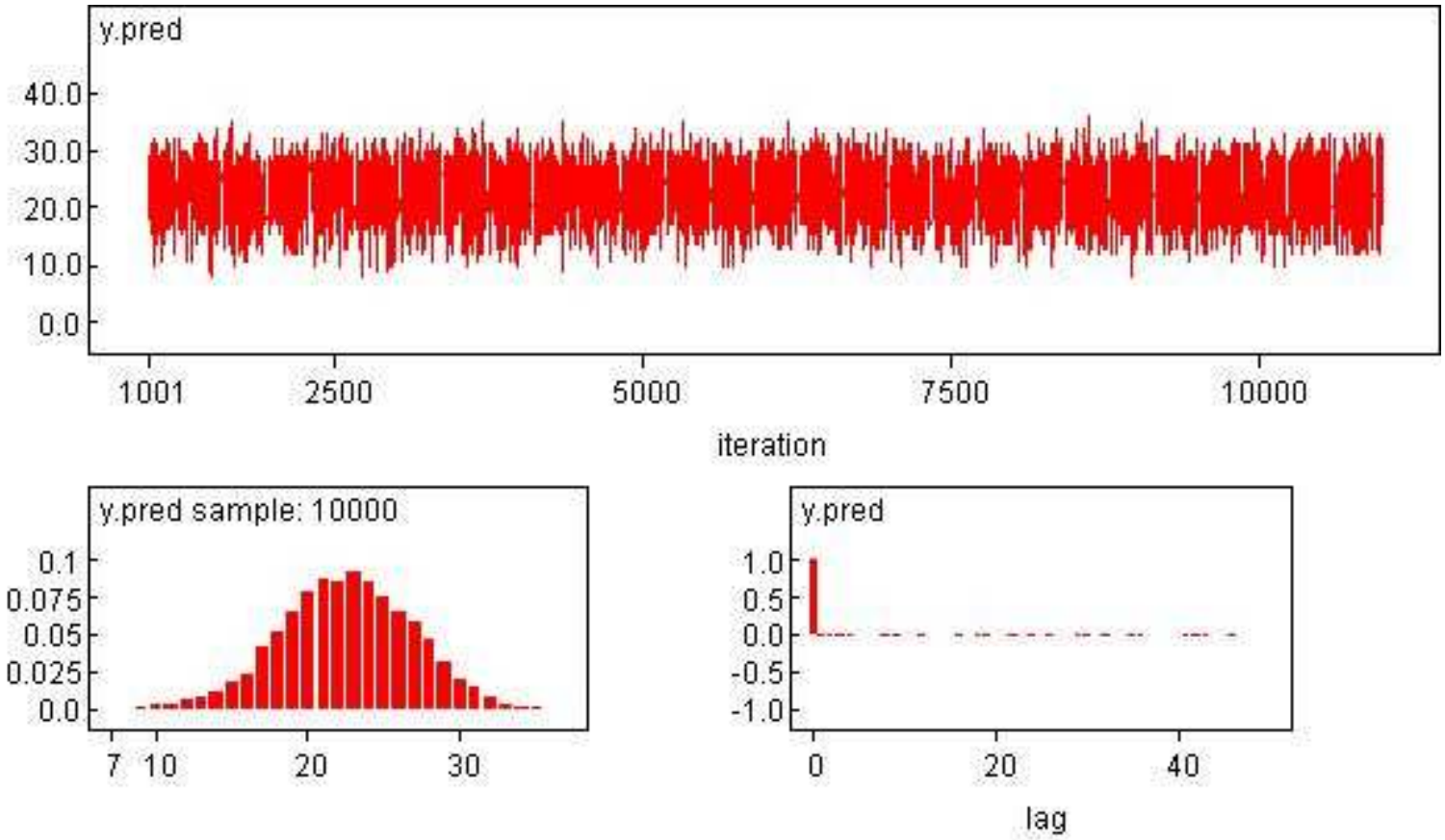
Initial values list can be after model description or in a separate file

```
list(theta=0.1)
```

Running WinBUGS for MCMC analysis (single chain)

1. Open *Specification tool* from *Model* menu.
2. Program responses are shown on bottom-left of screen.
3. Highlight `model` by double-click. Click on *Check model*.
4. Highlight start of data. Click on *Load data*.
5. Click on *Compile*.
6. Highlight start of initial values. Click on *Load inits*.
7. Click on *Gen Inits* if more initial values needed.
8. Open *Update* from *Model* menu.
9. Click on *Update* to burn in.
10. Open *Samples* from *Inference* menu.
11. Type nodes to be monitored into *Sample Monitor*, and click *set* after each.
12. Perform more updates.
13. Type * into *Sample Monitor*, and click *stats* etc to see results on all monitored nodes.

WinBUGS output



WinBUGS output and exact answers

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
theta	0.5633	0.07458	4.292E-4	0.4139	0.5647	0.7051	1001	30000
y.pred	22.52	4.278	0.02356	14.0	23.0	31.0	1001	30000
P.crit	0.3273	0.4692	0.002631	0.0	0.0	1.0	1001	30000

Exact answers from conjugate analysis

- θ: mean 0.563 and standard deviation 0.075
- Y^{pred}: mean 22.51 and standard deviation 4.31.
- Probability of at least 25: 0.329

MCMC results are within Monte Carlo error of the true values

Bayesian inference using the Normal distribution

Known variance, unknown mean

Suppose we have a sample of Normal data $x_i \sim \mathrm{N}(\theta, \sigma^2)$ $(i = 1, ..., n)$.

For now assume σ^2 is known and θ has a Normal prior $\theta \sim \mathrm{N}(\mu, \sigma^2/n_0)$

Same standard deviation σ is used in the likelihood and the prior. Prior variance is based on an 'implicit' sample size n_0

Then straightforward to show that the posterior distribution is

$$\theta|\boldsymbol{x} \sim \mathrm{N}\left(\frac{n_0\mu + n\overline{x}}{n_0 + n}, \frac{\sigma^2}{n_0 + n}\right)$$

- As n_0 tends to 0, the prior variance becomes larger and the distribution becomes 'flatter', and in the limit the prior distribution becomes essentially uniform over $-\infty, \infty$

- Posterior mean $(n_0\mu + n\overline{x})/(n_0 + n)$ is a weighted average of the prior mean μ and parameter estimate $\overline{x}$, weighted by their precisions (relative 'sample sizes'), and so is always a compromise between the two

- Posterior variance is based on an implicit sample size equivalent to the sum of the prior 'sample size' n_0 and the sample size of the data n

- As $n \to \infty$, $p(\theta|\boldsymbol{x}) \to \mathrm{N}(\overline{x}, \sigma^2/n)$ which does not depend on the prior

- Compare with frequentist setting, the MLE is $\widehat{\theta} = \overline{x}$ with $\mathrm{SE}(\widehat{\theta}) = \sigma/\sqrt{n}$, and sampling distribution

$$p(\widehat{\theta} \mid \theta) = p(\overline{x}|\theta) = \mathrm{N}(\theta, \sigma^2/n)$$

Example: 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 are tested throughout the year in each water supply zone
- Suppose we want to estimate the average THM concentration in a particular water zone, z
- Two independent measurements, x_{z1} and x_{z2} are taken and their mean, $\overline{x}_z$ is 130 $\mu g/l$
- Suppose we know that the assay measurement error has a standard deviation $\sigma_{[e]} = 5\mu g/l$
- What should we estimate the mean THM concentration to be in this water zone?

Let the mean THM conc. be denoted θ_z.

A standard analysis would use the sample mean $\overline{x}_z = 130\mu g/l$ as an estimate of θ_z, with standard error $\sigma_{[e]}/\sqrt{n} = 5/\sqrt{2} = 3.5\mu g/l$

A 95% confidence interval is $\overline{x}_z \pm 1.96 \times \sigma_{[e]}/\sqrt{n}$, i.e. 123.1 to 136.9 $\mu g/l$.

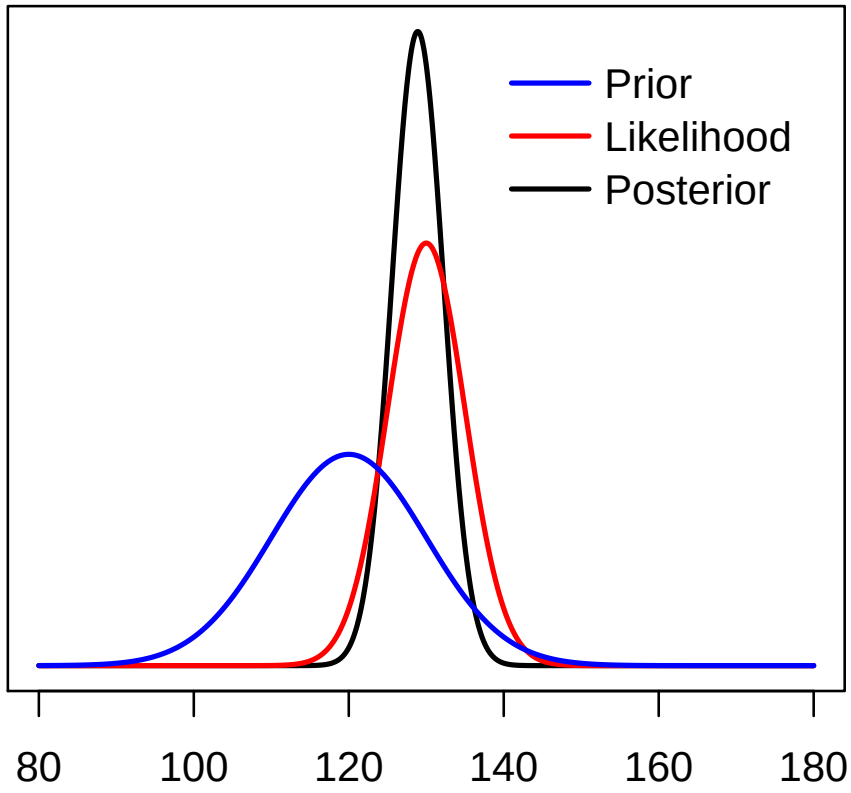
Suppose historical data on THM levels in other zones supplied from the same source showed that the mean THM concentration was 120 $\mu g/l$ with standard deviation 10 $\mu g/l$

- suggests Normal$(120, 10^2)$ prior for θ_z
- if we express the prior standard deviation as $\sigma_{[e]}/\sqrt{n_0}$, we can solve to find $n_0 = (\sigma_{[e]}/10)^2 = 0.25$
- so our prior can be written as $\theta_z \sim$ Normal$(120, \sigma^2_{[e]}/0.25)$

Posterior for θ_z is then

$$\begin{aligned} p(\theta_z|\boldsymbol{x}) &= \text{Normal}\left(\frac{0.25 \times 120 + 2 \times 130}{0.25 + 2}, \ \frac{5^2}{0.25 + 2}\right) \\ &= \text{Normal}(128.9, \ 3.33^2) \end{aligned}$$

giving 95% interval for θ_z of 122.4 to 135.4$\mu g/l$


Prior
Likelihood
Posterior
80
100
120
140
160
180
mean THM concentration, ug/l (theta)

Prediction

Denoting the posterior mean and variance as $\mu_n = (n_0\mu + n\bar{x})/(n_0 + n)$ and $\sigma_n^2 = \sigma^2/(n_0 + n)$, the *predictive distribution* for a new observation $\tilde{x}$ is

$$p(\tilde{x}|\boldsymbol{x}) = \int p(\tilde{x}|\boldsymbol{x}, \theta)p(\theta|\boldsymbol{x})d\theta$$

which generally simplifies to

$$p(\tilde{x}|\boldsymbol{x}) = \int p(\tilde{x}|\theta)p(\theta|\boldsymbol{x})d\theta$$

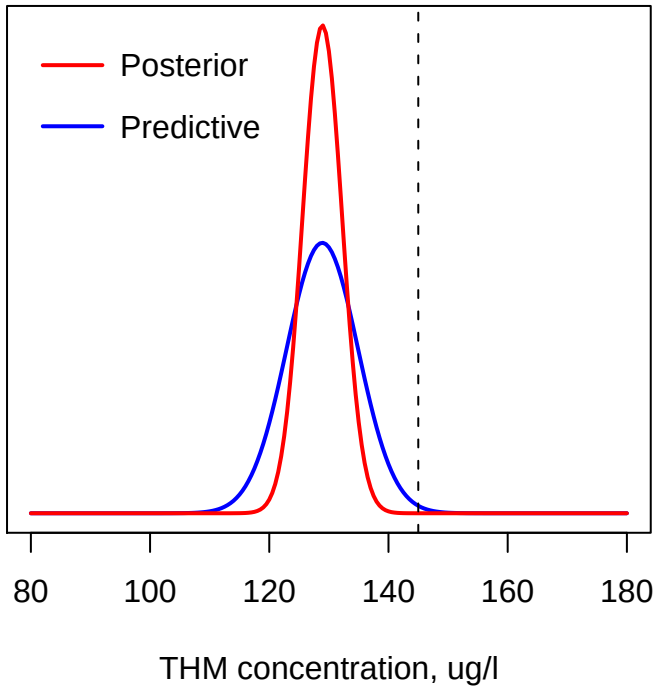
which can be shown to give

$$p(\tilde{x}|\boldsymbol{x}) \sim \mathsf{N}\left(\mu_n, \sigma_n^2 + \sigma^2\right)$$

So the predictive distribution is centred around the posterior mean with variance equal to sum of the posterior variance and the sample variance of $\tilde{x}$

Example: THM concentration (continued)

- Suppose the water company will be fined if THM levels in the water supply exceed $145\mu g/l$
- Predictive distribution for THM concentration in a future sample taken from the water zone is $\mathrm{N}(128.9, 3.33^2 + 5^2) = \mathrm{N}(128.9, 36.1)$
- Probability that THM concentration in future sample exceeds $145\mu g/l$ is $1 - \Phi[(145 - 128.9)/\sqrt{(36.1)}] = 0.004$



Bayesian inference using count data

Suppose we have an independent sample of counts $x_1, ..., x_n$ which can be assumed to follow a Poisson distribution with unknown mean μ:

$$p(\boldsymbol{x}|\mu) = \prod_i \frac{\mu^{x_i} e^{-\mu}}{x_i!}$$

The kernel of the Poisson likelihood (as a function of μ) has the same form as that of a Gamma(a, b) prior for μ:

$$p(\mu) = \frac{b^a}{\Gamma(a)} \mu^{a-1} e^{-b\mu}$$

Note: A Gamma(a, b) density has mean a/b and variance a/b^2

This implies the following posterior

$$\begin{aligned} p(\mu \mid \boldsymbol{x}) &\propto p(\mu)\, p(\boldsymbol{x} \mid \mu) \\ &= \frac{b^a}{\Gamma(a)} \mu^{a-1} e^{-b\mu} \prod_{i=1}^{n} e^{-\mu} \frac{\mu^{x_i}}{x_i!} \\ &\propto \mu^{a+n\overline{x}-1}\, e^{-(b+n)\mu} \\ &= \text{Gamma}(a + n\overline{x},\, b + n). \end{aligned}$$

The posterior is another (different) Gamma distribution.

The Gamma distribution is said to be the *conjugate* prior.

$$E(\mu \mid \boldsymbol{x}) = \frac{a + n\overline{x}}{b + n} = \overline{x} \left(\frac{n}{n+b} \right) + \frac{a}{b} \left(1 - \frac{n}{n+b} \right)$$

So posterior mean is a compromise between the prior mean a/b and the MLE $\overline{x}$

Example: London bombings during WWII

- Data below are the number of flying bomb hits on London during World War II in a 36 km^2 area of South London

- Area was partitioned into 0.25 km^2 grid squares and number of bombs falling in each grid was counted

Hits, x	0	1	2	3	4	7
Number of areas, n	229	211	93	35	7	1

Total hits, $\sum_i n_i x_i = 537$
Total number of areas, $\sum_i n_i = 576$

- If the hits are random, a Poisson distribution with constant hit rate θ should fit the data

- Can think of $n = 576$ observations from a Poisson distribution, with $\overline{x} = 537/576 = 0.93$

The 'invariant' Jeffreys prior (see later) for the mean θ of a Poisson distribution is $p(\theta) \propto 1/\sqrt{\theta}$, which is equivalent to an (improper) Gamma(0.5,0) distribution. Therefore

$$\begin{aligned} p(\theta|\boldsymbol{y}) &= \text{Gamma}(a + n\overline{x}, b + n) = \text{Gamma}(537.5, 576) \\ \mathbb{E}(\theta|\boldsymbol{y}) &= \frac{537.5}{576} = 0.933; \qquad \text{Var}(\theta|\boldsymbol{y}) = \frac{537.5}{576^2} = 0.0016 \end{aligned}$$

Note that these are almost exactly the same as the MLE and the square of the SE(MLE)

Summary

For all these examples, we see that

- the posterior mean is a compromise between the prior mean and the MLE
- the posterior s.d. is less than each of the prior s.d. and the s.e.(MLE)

> *'A Bayesian is one who, vaguely expecting a horse and catching a glimpse of a donkey, strongly concludes he has seen a mule'* (Senn, 1997)

As $n \to \infty$,

- the posterior mean $\to$ the MLE
- the posterior s.d. $\to$ the s.e.(MLE)
- the posterior does not depend on the prior.

These observations are generally true, when the MLE exists and is unique

Choosing prior distributions

When the posterior is in the same family as the prior then we have what is known as *conjugacy*. This has the advantage that prior parameters can usually be interpreted as a *prior sample*. Examples include:

Likelihood	Parameter	Prior	Posterior
Normal	mean	Normal	Normal
Normal	precision	Gamma	Gamma
Binomial	success prob.	Beta	Beta
Poisson	rate or mean	Gamma	Gamma

- Conjugate prior distributions are mathematically convenient, but do not exist for all likelihoods, and can be restrictive

- Computations for non-conjugate priors are harder, but possible using MCMC (see next lecture)

Calling WinBUGS from other software

- Scripts enable WinBUGS 1.4 to be called from other software
- Interfaces developed for R, Splus, SAS, Matlab
- See `www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml`
- Andrew Gelman's `bugs` function for R is most developed - reads in data, writes script, monitors output etc. Now packaged as `R2WinBUGS`.
- OpenBUGS site http://mathstat.helsinki.fi/openbugs/ provides an open source version, including `BRugs` package which works from within R

Further reading

Berry (1996) (Introductory text on Bayesian methods, with medical slant)

Lee (2004) (Good intro to Bayesian inference; more mathematical than Berry; 3rd edition contains WinBUGS examples)

Bernardo and Smith (1994) (Advanced text on Bayesian theory)

Lecture 3.
Introduction to MCMC

Why is computation important?

- Bayesian inference centres around the posterior distribution

$$p(\boldsymbol{\theta}|x) \propto p(x|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})$$

  where $\boldsymbol{\theta}$ is typically a large vector of parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2, ..., \theta_k\}$

- $p(x|\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ will often be available in closed form, but $p(\boldsymbol{\theta}|x)$ is usually not analytically tractable, and we want to
  - obtain marginal posterior $p(\theta_i|x) = \int\int ... \int p(\boldsymbol{\theta}|x)\, d\boldsymbol{\theta}_{(-i)}$ where $\boldsymbol{\theta}_{(-i)}$ denotes the vector of θ's excluding θ_i
  - calculate properties of $p(\theta_i|x)$, such as mean $(= \int \theta_i p(\theta_i|x) d\theta_i)$, tail areas $(= \int_T^\infty p(\theta_i|x) d\theta_i)$ etc.

$\rightarrow$ numerical integration becomes vital

Monte Carlo integration

We have already seen that Monte Carlo methods can be used to simulate values from prior distributions and from **closed form** posterior distributions

If we had algorithms for sampling from arbitrary (typically high-dimensional) posterior distributions, we could use Monte Carlo methods for Bayesian estimation:

- Suppose we can draw samples from the joint posterior distribution for $\boldsymbol{\theta}$, *i.e.*

$$(\theta_1^{(1)}, ..., \theta_k^{(1)}), (\theta_1^{(2)}, ..., \theta_k^{(2)}), ..., (\theta_1^{(N)}, ..., \theta_k^{(N)}) \quad \sim \quad p(\boldsymbol{\theta}|x)$$

- Then
  - $\theta_1^{(1)}, ..., \theta_1^{(N)}$ are a sample from the marginal posterior $p(\theta_1|x)$
  - $E(g(\theta_1)) = \int g(\theta_1)p(\theta_1|x)d\theta_1 \approx \frac{1}{N}\sum_{i=1}^{N} g(\theta_1^{(i)})$

$\rightarrow$ this is Monte Carlo integration

$\rightarrow$ theorems exist which prove convergence in limit as $N \rightarrow \infty$ even if the sample is dependent (crucial to the success of MCMC)

How do we sample from the posterior?

- We want samples from joint posterior distribution $p(\boldsymbol{\theta}|x)$
- *Independent* sampling from $p(\boldsymbol{\theta}|x)$ may be difficult
- **BUT** *dependent* sampling from a *Markov chain* with $p(\boldsymbol{\theta}|x)$ as its stationary (equilibrium) distribution is easier
- A sequence of random variables $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, ...$ forms a Markov chain if $\theta^{(i+1)} \sim p(\theta|\theta^{(i)})$ *i.e.* conditional on the value of $\theta^{(i)}$, $\theta^{(i+1)}$ is independent of $\theta^{(i-1)}, ..., \theta^{(0)}$
- Several standard 'recipes' available for designing Markov chains with required stationary distribution $p(\boldsymbol{\theta}|x)$
  - Metropolis *et al.* (1953); generalised by Hastings (1970)
  - **Gibbs Sampling** (see Geman and Geman (1984), Gelfand and Smith (1990), Casella and George (1992)) is a special case of the Metropolis-Hastings algorithm which generates a Markov chain by sampling from **full conditional distributions**
  - See Gilks, Richardson and Spiegelhalter (1996) for a full introduction and many worked examples.

Gibbs sampling

Let our vector of unknowns $\boldsymbol{\theta}$ consist of k sub-components $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)$

1) Choose starting values $\theta_1^{(0)}$, $\theta_2^{(0)}$, ..., , $\theta_k^{(0)}$

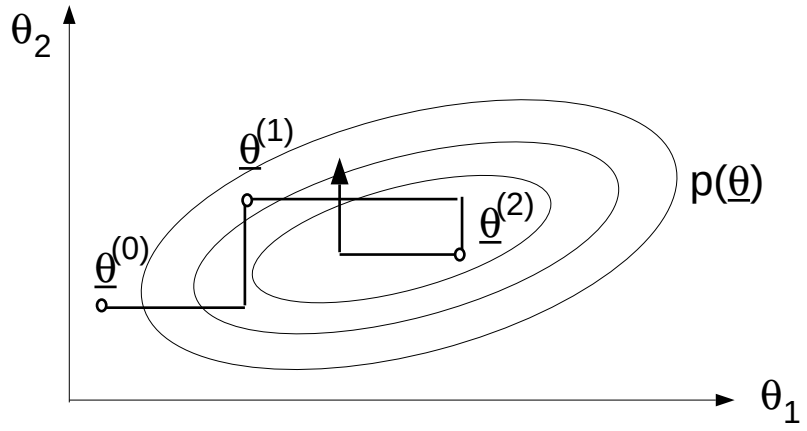
2) Sample $\theta_1^{(1)}$ from $p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, ..., , \theta_k^{(0)}, x)$
   Sample $\theta_2^{(1)}$ from $p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, ..., , \theta_k^{(0)}, x)$
   .....
   Sample $\theta_k^{(1)}$ from $p(\theta_k|\theta_1^{(1)}, \theta_2^{(1)}, ..., , \theta_{k-1}^{(1)}, x)$

3) Repeat step 2 many 1000s of times
   – eventually obtain sample from $p(\boldsymbol{\theta}|x)$

The conditional distributions are called 'full conditionals' as they condition on all other parameters

Gibbs sampling ctd.

Example with $k = 2$



- Sample $\theta_1^{(1)}$ from $p(\theta_1|\theta_2^{(0)}, x)$
- Sample $\theta_2^{(1)}$ from $p(\theta_2|\theta_1^{(1)}, x)$
- Sample $\theta_1^{(2)}$ from $p(\theta_1|\theta_2^{(1)}, x)$
- .....

$\boldsymbol{\theta}^{(n)}$ forms a Markov chain with (*eventually*) a stationary distribution $p(\boldsymbol{\theta}|x)$.

Using MCMC methods

There are two main issues to consider

- Convergence (how quickly does the distribution of $\boldsymbol{\theta}^{(t)}$ approach $p(\boldsymbol{\theta}|x)$?)
- Efficiency (how well are functionals of $p(\boldsymbol{\theta}|x)$ estimated from $\{\boldsymbol{\theta}^{(t)}\}$?)

Checking convergence

This is the users responsibility!

- Note: Convergence is to target **distribution** (the required posterior), not to a single value.

- Once convergence reached, samples should look like a random scatter about a stable mean value

Convergence diagnosis

- How do we know we have reached convergence?
- *i.e.* How do we the know number of 'burn-in' iterations?
- Many 'convergence diagnostics' exist, but none foolproof
- CODA and BOA software contain large number of diagnostics

Gelman-Rubin-Brooks diagnostic

- A number of runs
- Widely differing starting points
- Convergence assessed by quantifying whether sequences are much further apart than expected based on their internal variability
- Diagnostic uses components of variance of the multiple sequences

Example: A dose-response model

Consider the following response rates for different doses of a drug

dose x_i	No. subjects n_i	No. responses r_i
1.69	59	6
1.72	60	13
1.75	62	18
1.78	56	28
1.81	63	52
1.83	59	53
1.86	62	61
1.88	60	60

Fit a logistic curve with 'centred' covariate $(x_i - \overline{x})$:

$$\begin{aligned} r_i &\sim \text{Bin}(p_i, n_i) \\ \text{logit } p_i &= \alpha + \beta(x_i - \overline{x}) \\ \alpha &\sim \text{N}(0, 10000) \\ \beta &\sim \text{N}(0, 10000) \end{aligned}$$

Checking convergence with multiple runs

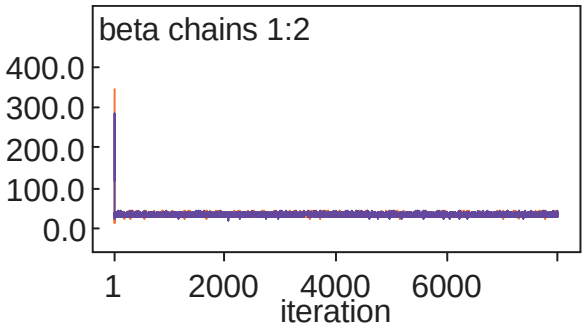
- Set up multiple initial value lists, *e.g.*
  ```
  list(alpha=-100, beta=100)
  list(alpha=100, beta=-100)
  ```
- Before clicking *compile*, set *num of chains* to 2
- Load both sets of initial values
- Monitor from the start of sampling
- Assess how much burn-in needed using the *bgr* statistic

Using the *bgr* statistic

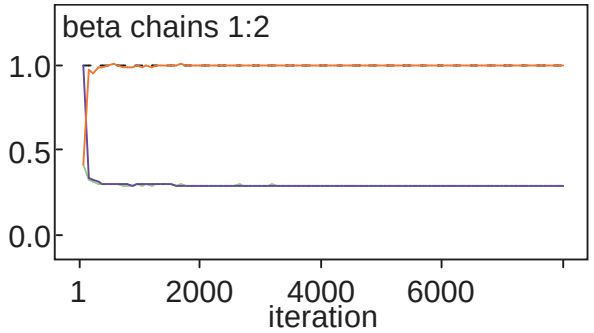
- *Green*: width of 80% intervals of pooled chains: should be stable
- *Blue*: average width of 80% intervals for chains: should be stable
- *Red*: ratio of pooled/within: should be near 1
- Double-click on plot, then *cntl* + right click gives statistics

Output for 'centred' analysis

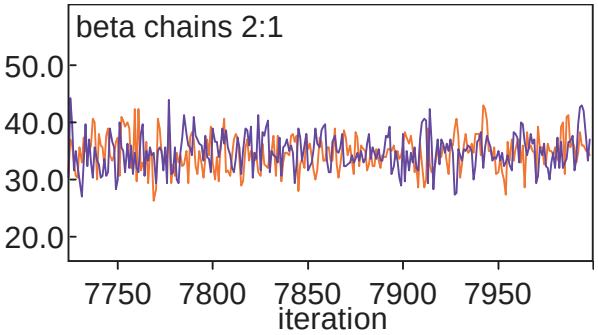
history



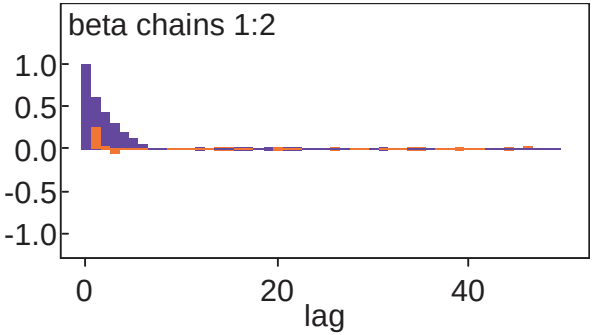
bgr diagnostic



trace



autocorrelation



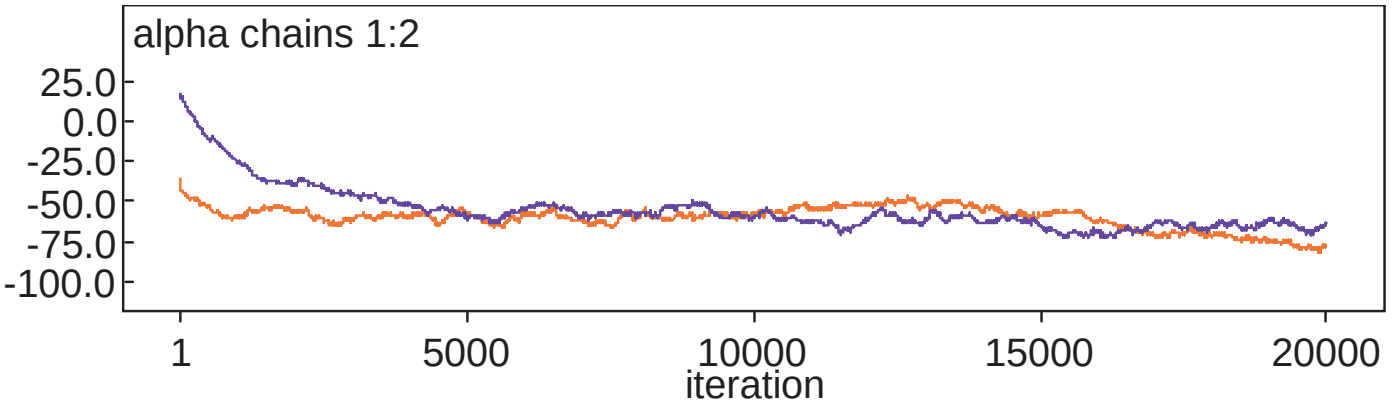
node	mean	sd	MC error	2.5%	median	97.5%	start	sample
alpha	0.7489	0.139	0.00138	0.4816	0.7468	1.026	1001	14000
beta	34.6	2.929	0.02639	29.11	34.53	40.51	1001	14000

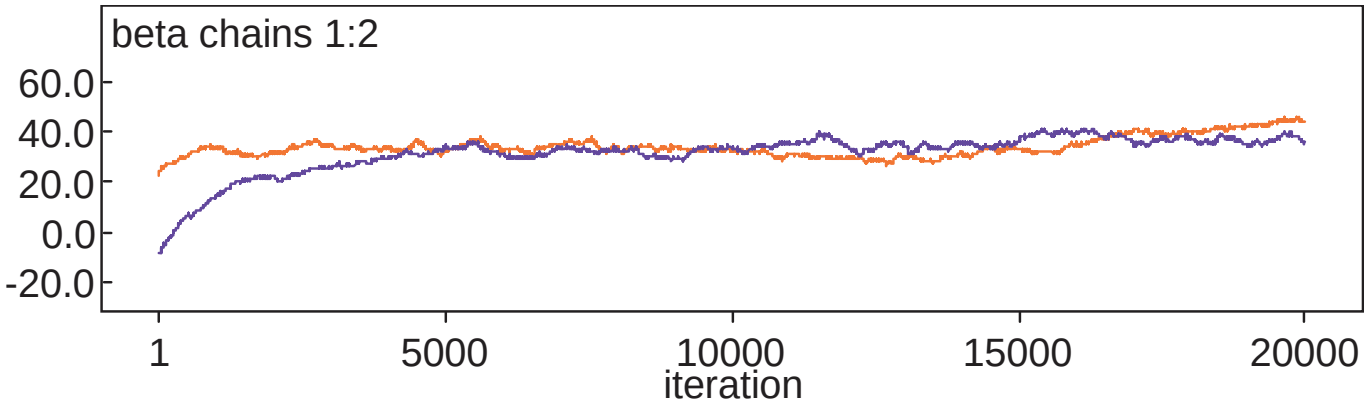
Problems with convergence

Fit a logistic curve with 'un-centred' covariate x:

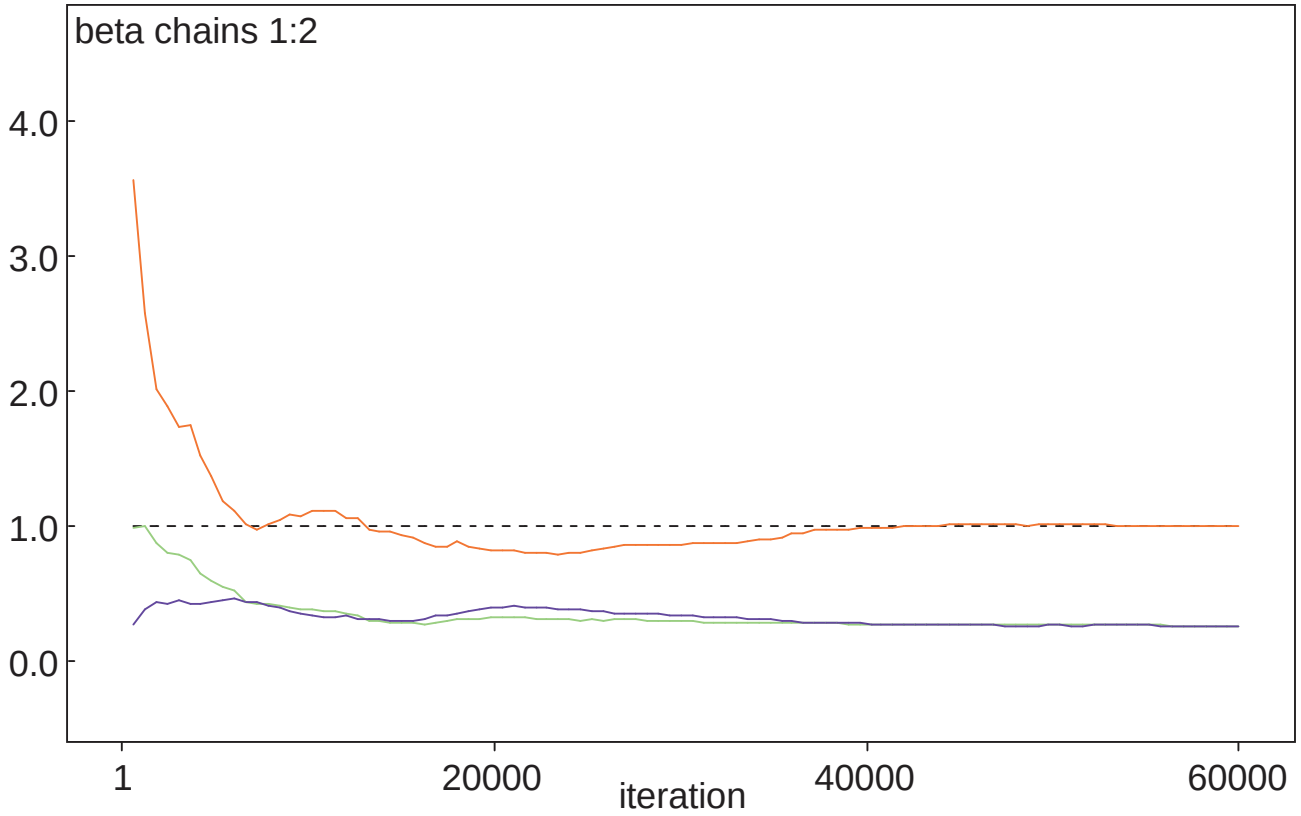
$$\begin{aligned} r_i &\sim \text{Bin}(p_i, n_i) \\ \text{logit } p_i &= \alpha + \beta x_i \\ \alpha &\sim \text{N}(0, 10000) \\ \beta &\sim \text{N}(0, 10000) \end{aligned}$$

History plots for ‘un-centred’ analysis





bgr output for 'un-centred' analysis

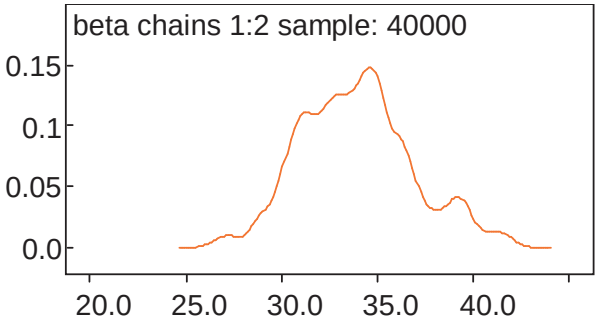


Drop first 40,000 iterations as burn-in

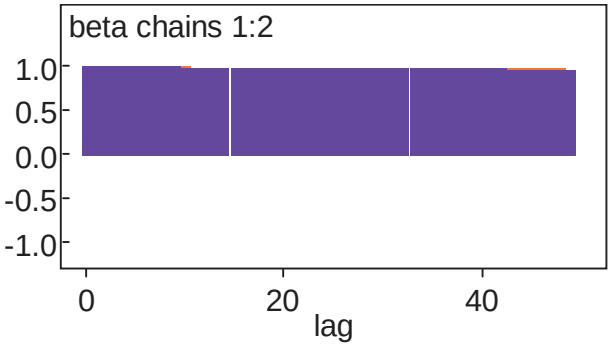
node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta	33.97	2.955	0.1734	28.7	33.89	40.3	40001	40000

Output for 'un-centred' analysis

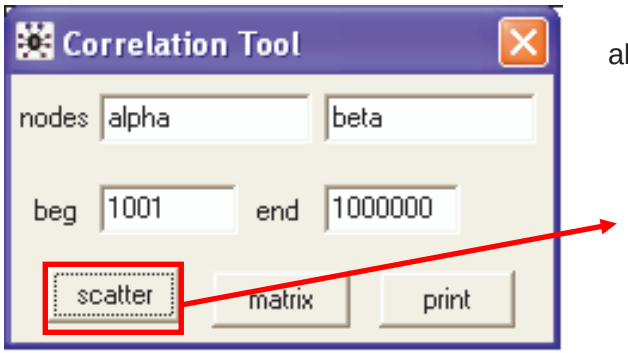
posterior density



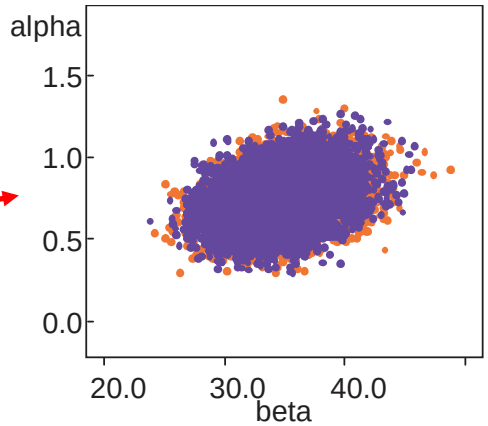
autocorrelation



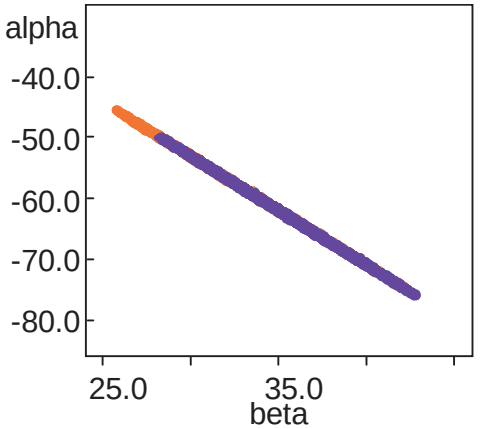
bivariate posteriors



centred



un-centred



How many iterations after convergence?

- After convergence, further iterations are needed to obtain samples for posterior inference.
- More iterations = more accurate posterior estimates.
- Efficiency of sample mean of θ as estimate of theoretical posterior expectation $E(\theta)$ usually assessed by calculating Monte Carlo standard error (MC error)
- MC error = standard error of posterior sample mean as estimate of theoretical expectation for given parameter
- MC error depends on
  - true variance of posterior distribution
  - posterior sample size (number of MCMC iterations)
  - autocorrelation in MCMC sample
- Rule of thumb: want MC error $< 1 - 5\%$ of posterior SD

Inference using posterior samples from MCMC runs

A powerful feature of the Bayesian approach is that all inference is based on the joint posterior distribution

$\Rightarrow$ can address wide range of substantive questions by appropriate summaries of the posterior

- Typically report either mean or median of the posterior samples for each parameter of interest as a point estimate

- 2.5% and 97.5% percentiles of the posterior samples for each parameter give a 95% posterior credible interval (interval within which the parameter lies with probability 0.95)

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta	34.60	2.929	0.0239	29.11	34.53	40.51	1001	14000

So point estimate of `beta` would be 34.60, with 95% credible interval (29.11, 40.51)

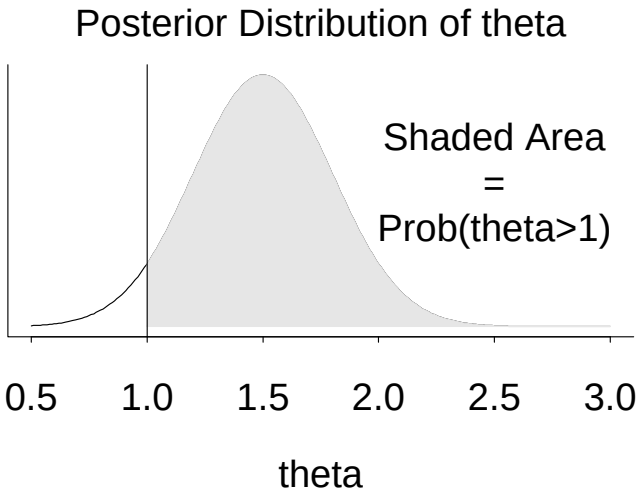
Probability statements about parameters

- Classical inference cannot provide probability statements about parameters (e.g. p-value is not Pr(H_0 true), but probability of observing data as or more extreme than we obtained, given that H_0 is true)

- In Bayesian inference, it is simple to calculate e.g. Pr($\theta > 1$):

  = Area under posterior distribution curve to the right of 1

  = Proportion of values in posterior sample of θ which are > 1



- In `WinBUGS` use the `step` function:
  `p.theta <- step(theta - 1)`

- For discrete parameters, may also be interested in Pr($\delta = \delta_0$):
  `p.delta <- equals(delta, delta0)`

- Posterior means of `p.theta` and `p.delta` give the required probabilities

Complex functions of parameters

- Classical inference about a function of the parameters $g(\theta)$ requires construction of a specific estimator of $g(\theta)$. Obtaining appropriate error can be difficult.

- Easy using MCMC: just calculate required function $g(\theta)$ as a logical node at each iteration and summarise posterior samples of $g(\theta)$

In dose-response example, suppose we want to estimate the $ED95$: that is the dose that will provide 95% of maximum efficacy.

$$\begin{aligned} \text{logit}\,0.95 &= \alpha + \beta(ED95 - \overline{x}) \\ ED95 &= (\text{logit}\,0.95 - \alpha)/\beta + \overline{x} \end{aligned}$$

Simply add into model

```
ED95 <- (logit(0.95) - alpha)/beta + mean(x[])
```

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
ED95	1.857	0.007716	8.514E-5	1.843	1.857	1.874	1001	10000

How to rank if you must

- Recent trend in UK towards ranking 'institutional' performance e.g. schools, hospitals
- Might also want to rank treatments, answer 'which is the best' etc
- Rank of a point estimate is a highly unreliable summary statistic

$\Rightarrow$ Would like measure of uncertainty about rank

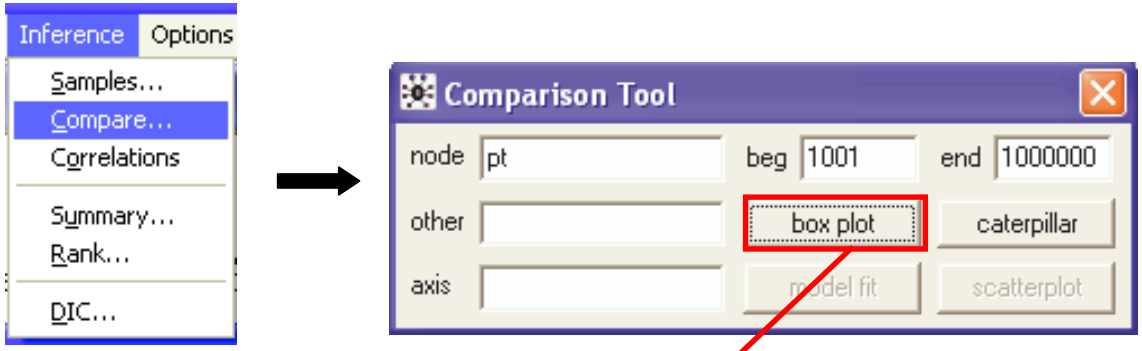
- Bayesian methods provide *posterior interval estimates* for ranks
- WinBUGS contains 'built-in' options for ranks:
  - `Rank` option of `Inference` menu monitors the rank of the elements of a specified vector
  - `rank(x[], i)` returns the rank of the i^{th} element of `x`
  - `equals(rank(x[],i),1)` =1 if i^{th} element is ranked lowest, 0 otherwise. Mean is probability that i^{th} element is 'best' (if counting adverse events)
  - `ranked(x[], i)` returns the value of the i^{th}-ranked element of `x`

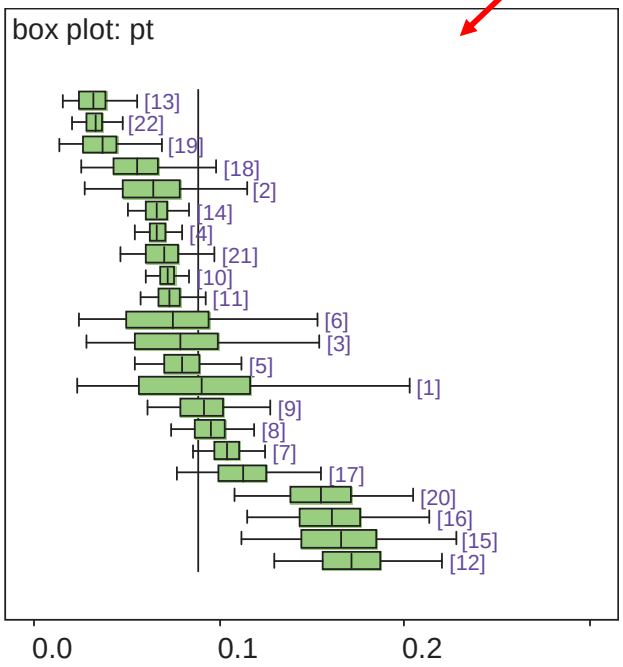
Example of ranking: 'Blocker' trials

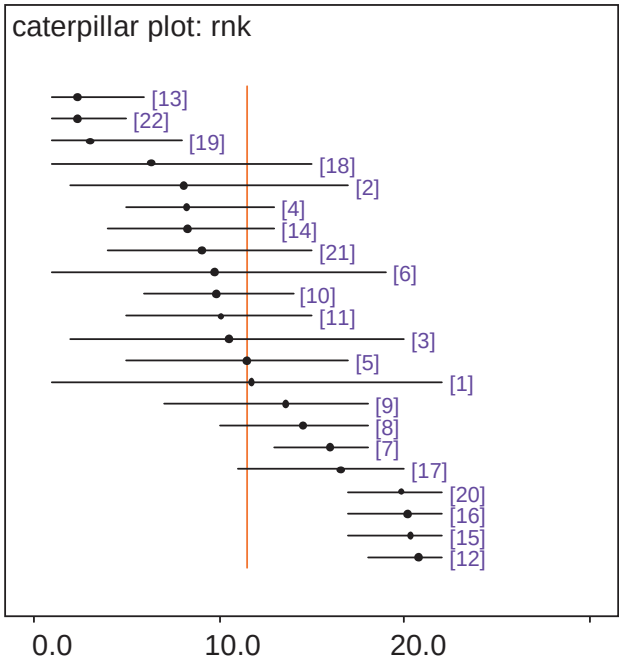
- 22 trials of beta-blockers used in WinBUGS manual to illustrate random-effects meta-analysis.
- Consider just treatment arms: which trial has the lowest mortality rate?
- Assume independent 'Jeffreys' beta[0.5, 0.5] prior for each response rate.

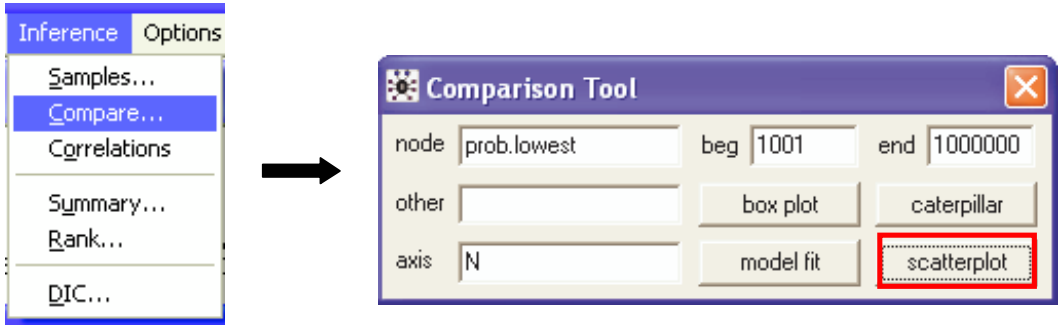
```
for( i in 1 : Num) {
   rt[i] ~ dbin(pt[i],nt[i])
   pt[i] ~ dbeta(0.5,0.5)                # Jeffreys prior
   rnk[i] <-  rank(pt[], i)              # rank of i'th trial
   prob.lowest[i]  <- equals(rnk[i],1)   # prob that i'th trial lowest
   N[i]<-i                               # used for indexing plot
}
```

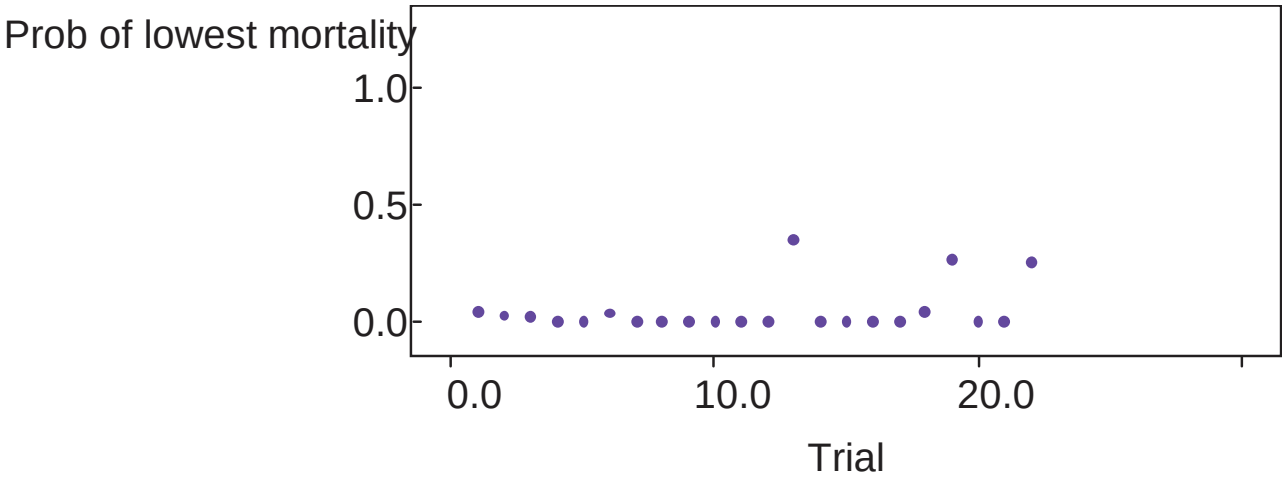
Mortality rates and ranks











Ranking methods may be useful when

- comparing alternative treatments
- comparing subsets
- comparing response-rates, cost-effectiveness or any summary measure

Further reading

Gelfand and Smith (1990) (key reference to use of Gibbs sampling for Bayesian calculations)

Casella and George (1992) (Explanation of Gibbs sampling)

Brooks (1998) (tutorial paper on MCMC)

Spiegelhalter et al (1996) (Comprehensive coverage of practical aspects of MCMC)

Lecture 4.

Bayesian linear regression models

Bayesian regression models

Standard (and non standard) regression models can be easily formulated within a Bayesian framework.

- Specify probability distribution (likelihood) for the data
- Specify form of relationship between response and explanatory variables
- Specify prior distributions for regression coefficients and any other unknown (nuisance) parameters

Some advantages of a Bayesian formulation in regression modelling include:

- Easy to include parameter restrictions and other relevant prior knowledge
- Easily extended to non-linear regression
- Easily 'robustified'
- Easy to make inference about functions of regression parameters and/or predictions
- Easily extended to handle missing data and covariate measurement error

Linear regression

Consider a simple linear regression with univariate Normal outcome y_i and a vector of covariates $x_{1i}, ..., x_{pi}$, $i = 1, ..., n$

$$\begin{aligned} y_i &= \beta_0 + \sum_{k=1}^{p} \beta_k x_{ki} + \epsilon_i \\ \epsilon_i &\sim \text{Normal}(0, \sigma^2) \end{aligned}$$

An equivalent Bayesian formulation would typically specify

$$\begin{aligned} y_i &\sim \text{Normal}(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \sum_{k=1}^{p} \beta_k x_{ki} \\ (\beta_0, \beta_1, ..., \beta_p, \sigma^2) &\sim \text{Prior distributions} \end{aligned}$$

A typical choice of 'vague' prior distribution (see later for more details) that will give numerical results similar to OLS or MLE is:

$$\begin{aligned} \beta_k &\sim \text{N}(0, 100000) \quad k = 0, ..., p \\ 1/\sigma^2 &\sim \text{Gamma}(0.001, 0.001) \end{aligned}$$

Example: New York Crime data

- 23 Precincts in New York City
- Response = THEFT: seasonally adjusted changes in larcenies (thefts) from a 27-week base period in 1966 to a 58-week experimental period in 1966-1967
- Predictors = MAN: % change in police manpower; DIST: district indicator (1 Downtown, 2 Mid-town, 3 Up-town)
- Model specification:

$$
\begin{aligned}
\text{THEFT}_i &\sim \text{Normal}(\mu_i, \sigma^2) \quad i = 1, ..., 21 \\
\mu_i &= \alpha + \beta \times \text{MAN}_i + \text{<effect of DIST>} \\
1/\sigma^2 &\sim \text{Gamma}(0.001, 0.001) \\
\alpha &\sim \text{N}(0, 100000) \\
\beta &\sim \text{N}(0, 100000) \\
\text{Prior} &\ \text{on} \ \text{coefficients for DIST effect}
\end{aligned}
$$

Specifying categorical covariates using the BUGS language

DIST$_i$ is a 3-level categorical explanatory variable

Two alternative ways of specifying model in BUGS language

1. Create usual 'design matrix' in data file:

```
  MAN[] THEFT[] DIST2[] DIST3[]
-15.76    3.19        0       0  # district 1
  0.98   -3.45        0       0
  3.71    0.04        0       0
.......
 -9.56    3.68        0       0
 -2.06    8.63        1       0  # district 2
 -0.76   10.82        1       0
 -6.30   -0.50        1       0
.......
 -2.82   -2.02        1       0
-16.19    0.94        0       1  # district 3
-11.00    4.42        0       1
......
-10.77    1.58        0       1
END
```

BUGS model code is then

```
for (i in 1:N) {
  THEFT[i] ~ dnorm(mu[i], tau)
  mu[i] <- alpha + beta*MAN[i] + delta2*DIST2[i] + delta3*DIST3[i]
}
alpha ~ dnorm(0, 0.00001)
beta ~ dnorm(0, 0.00001)
delta2 ~ dnorm(0, 0.00001)
delta3 ~ dnorm(0, 0.00001)
tau ~ dgamma(0.001, 0.001)
sigma2 <- 1/tau
```

Note: BUGS parameterises normal in terms of mean and **precision** (1/variance)!!

Initial values file would be something like

```
list(alpha = 1, beta = -2, delta2 = -2, delta3 = 4, tau = 2)
```

2. Alternatively, input explanatory variable as single vector coded by its level:

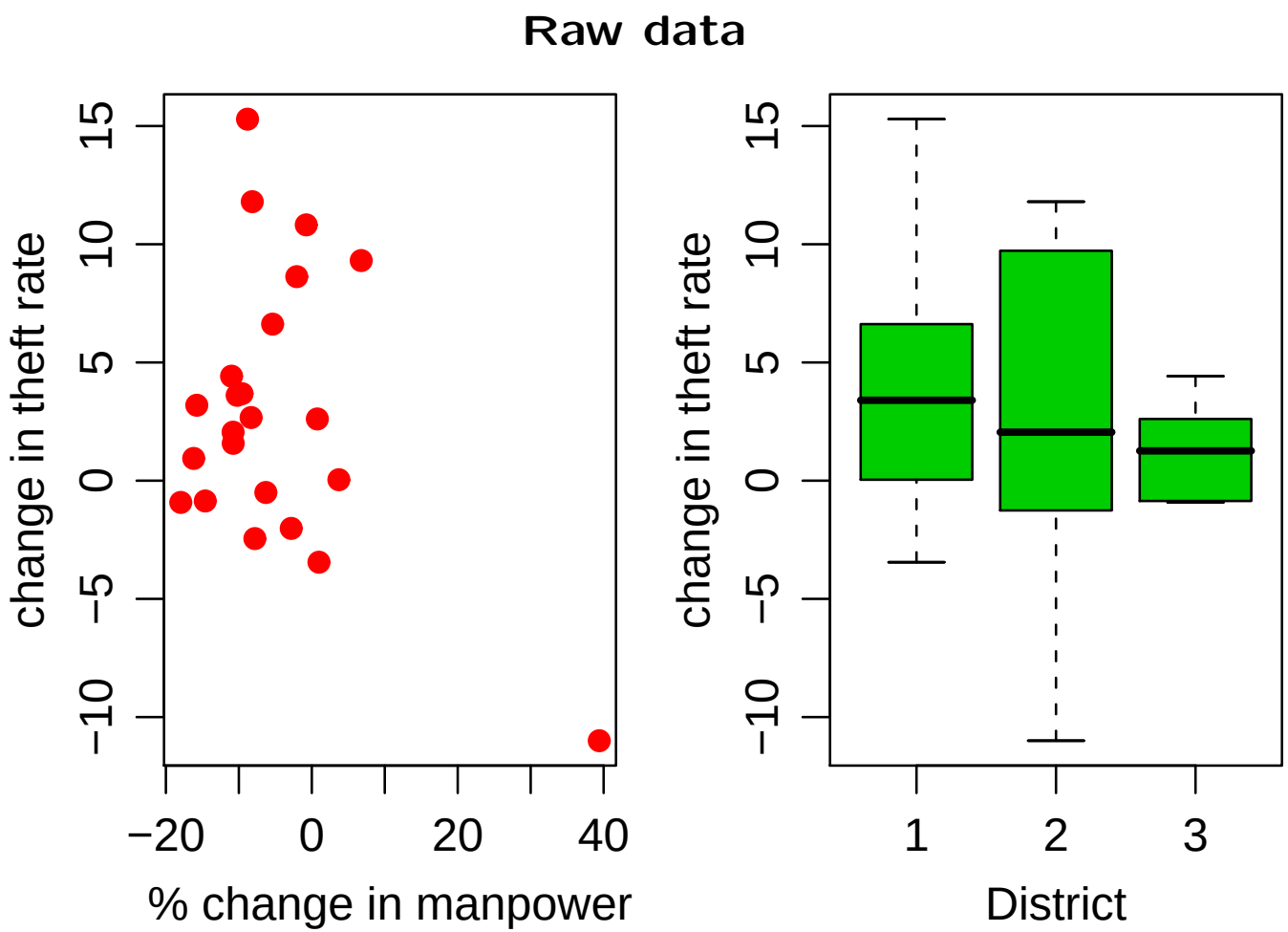
```
  MAN[]   THEFT[]  DIST[]
-15.76       3.19       1
  0.98      -3.45       1
  3.71       0.04       1
.....
 -9.56       3.68       1
 -2.06       8.63       2
 -0.76      10.82       2
 -6.30      -0.50       2
.....
 -2.82      -2.02       2
-16.19       0.94       3
-11.00       4.42       3
.....
-10.77       1.58       3
END
```

Then use 'double indexing' feature of BUGS language

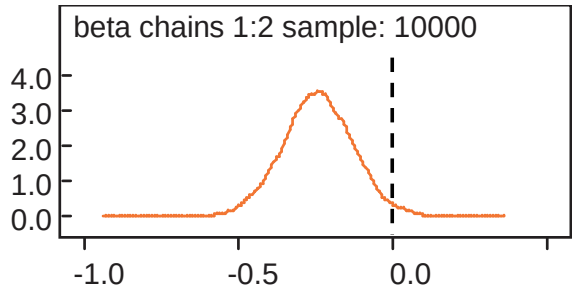
```
for (i in 1:23) {
  THEFT[i] ~ dnorm(mu[i], tau)
  mu[i] <- alpha + beta*MAN[i] + delta[DIST[i]]
}
alpha ~ dnorm(0, 0.00001)
beta ~ dnorm(0, 0.00001)
delta[1] <- 0  # set coefficient for reference category to zero
delta[2] ~ dnorm(0, 0.00001)
delta[3] ~ dnorm(0, 0.00001)
tau ~ dgamma(0.001, 0.001)
sigma2 <- 1/tau
```

In initial values file, need to specify initial values for `delta[2]` and `delta[3]` but not `delta[1]`. Use following syntax:

```
list(alpha = 1, beta = -2, delta = c(NA, -2, 4), tau = 2)
```


Raw data
change in theft rate
15
10
5
0
−5
−10
−20
0
20
40
% change in manpower
change in theft rate
15
10
5
0
−5
−10
1
2
3
District

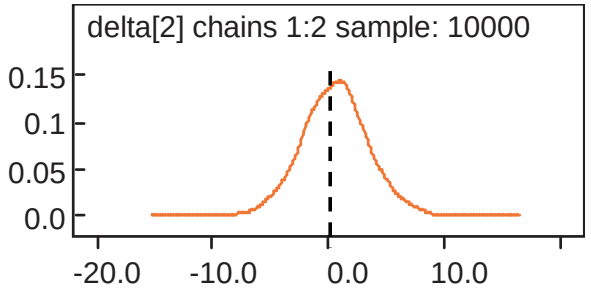
Change in theft rate per 1% increase in police manpower



Posterior mean -0.24

95% interval (-0.47, -0.01)

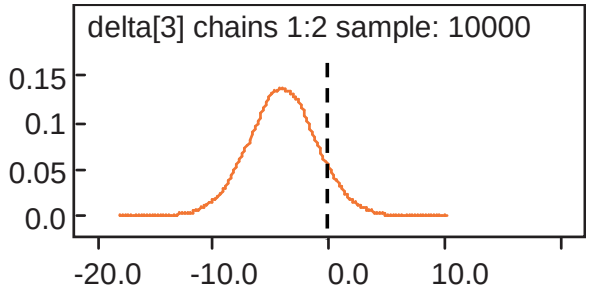
Change in theft rate in Midtown relative to Downtown



Posterior mean 0.6

95% interval (-5.1, 6.6)

Change in theft rate in Uptown relative to Downtown

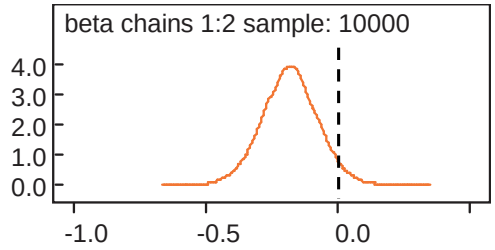


Posterior mean -4.0

95% interval (-10.0, 2.1)

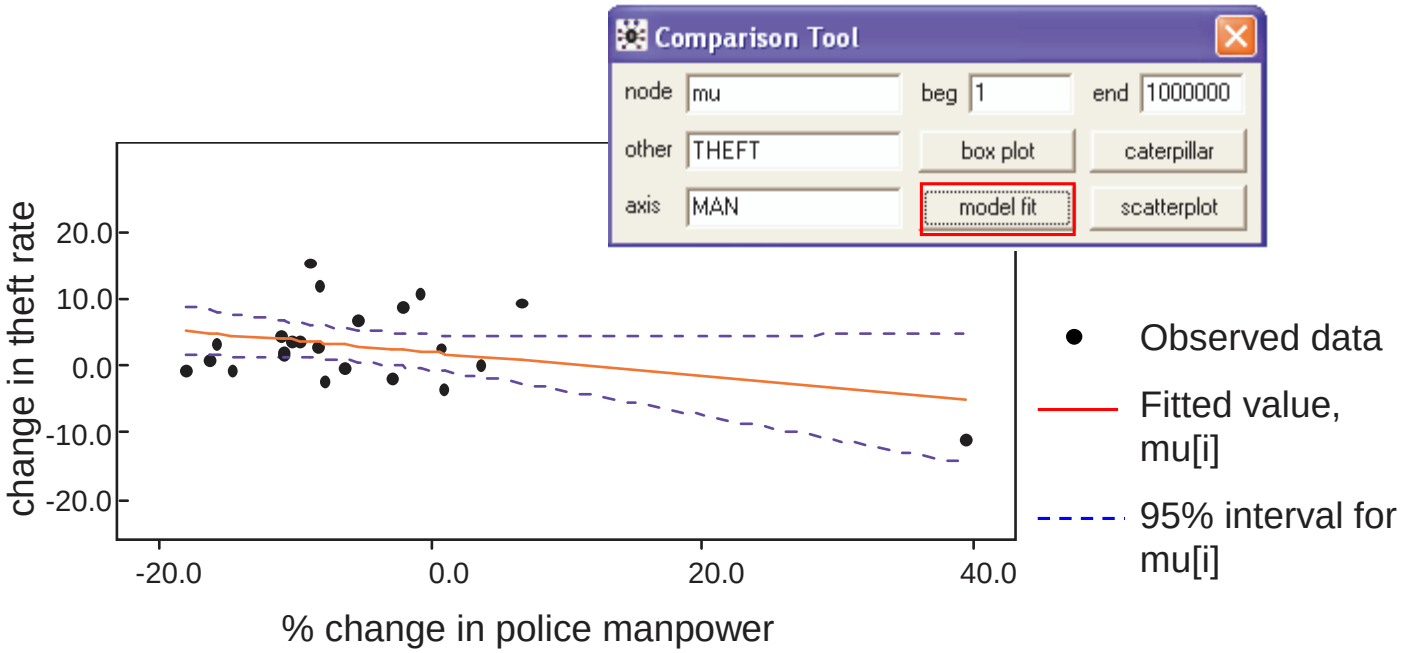
- 95% intervals for DIST effect both include zero
  → drop DIST from model (see later for Bayesian model comparison criteria)

Change in theft rate per 1% increase in police manpower



Posterior mean -0.18

95% interval (-0.39, 0.04)

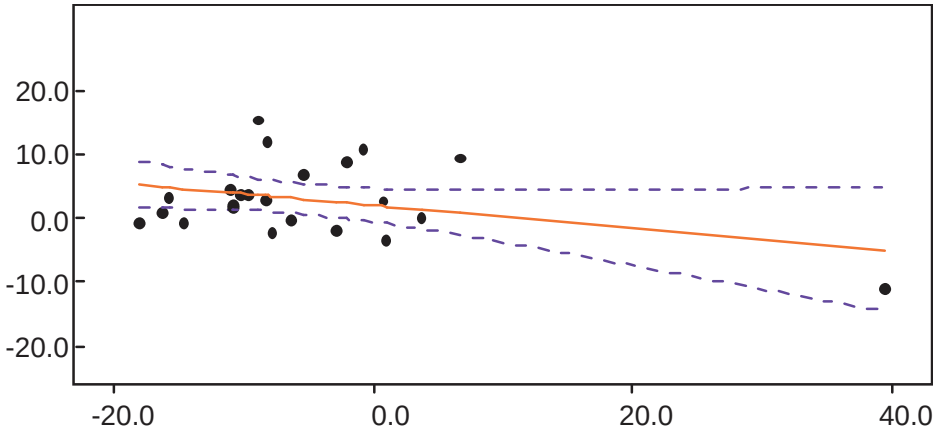


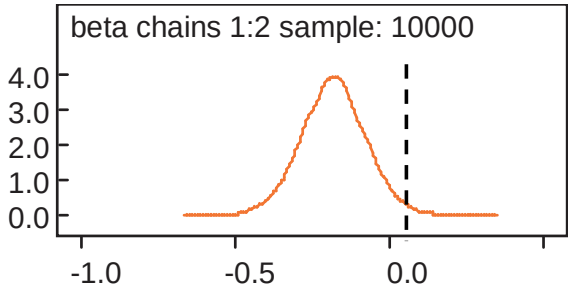
- Influential point corresponds to 20^{th} Precinct
- During 2nd period, manpower assigned to this Precinct was experimentally increased by about 40%
- No experimental increases in any other Precinct

  $\rightarrow$ Robustify model assuming t-distributed errors

```
for (i in 1:23) {
  THEFT[i] ~ dt(mu[i], tau, 4)   # robust likelihood (t on 4 df)
  mu[i] <- alpha + beta*MAN[i]
}
alpha ~ dnorm(0, 0.00001)
beta ~ dnorm(0, 0.00001)
tau ~ dgamma(0.001, 0.001)
sigma2 <- 1/tau

dummy <- DIST[1]  # ensures all variables in data file appear in model code
```

Model with Normal errors

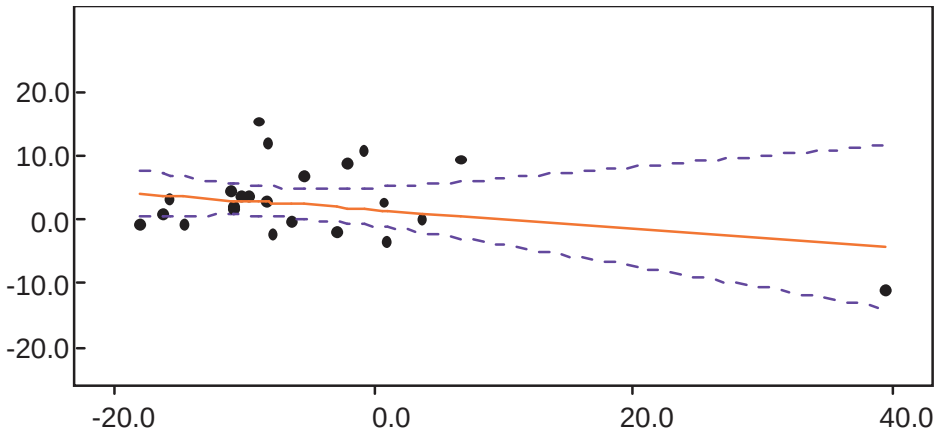


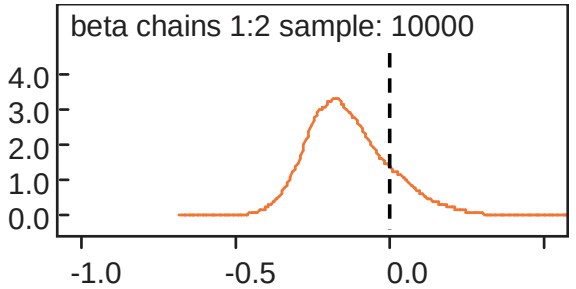


Posterior mean -0.18

95% interval (-0.39, 0.04)

Model with Student t errors



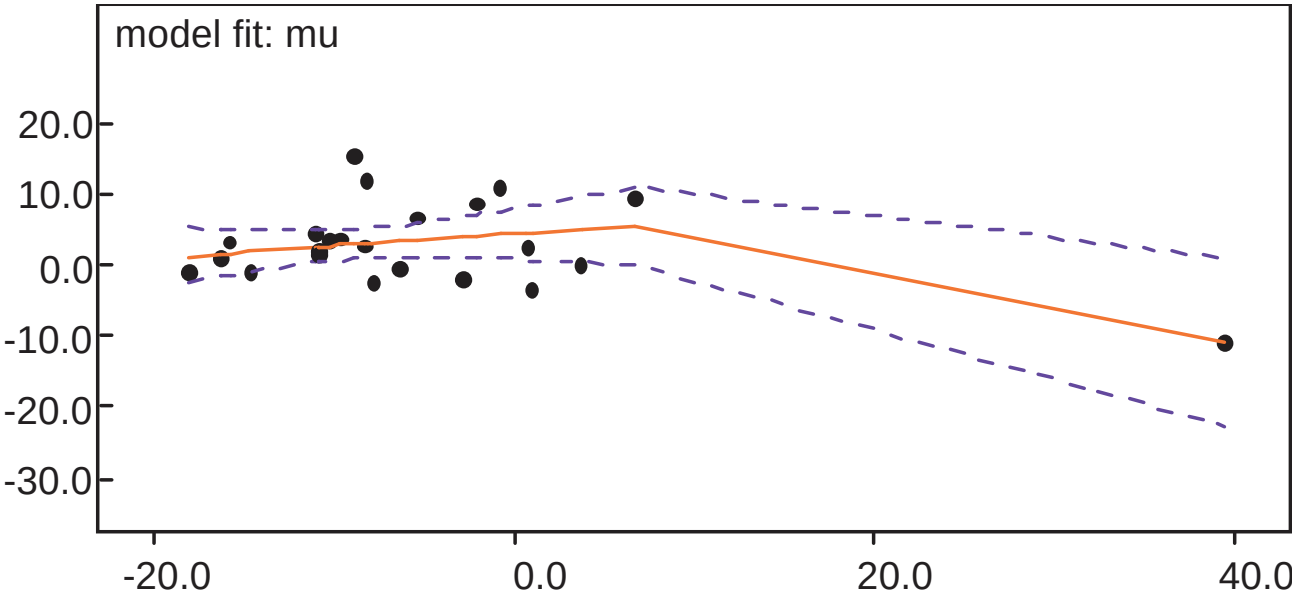


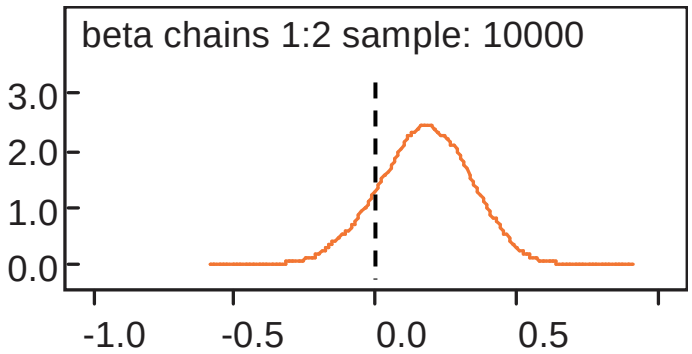
Posterior mean -0.13

95% interval (-0.36, 0.18)

- Precinct 20 still quite influential
- Add additional covariate corresponding to a binary indicator for Precinct 20
  - Equivalent to fitting separate (saturated) model to Precinct 20

```
for(i in 1:23) {
  THEFT[i] ~ dt(mu[i], tau, 4) # robust likelihood (t on 4 df)
  mu[i] <- alpha + beta*MAN[i] + delta*PREC20[i]   # separate term for precinct 20
}
alpha ~ dnorm(0, 0.000001)
beta ~ dnorm(0, 0.000001)
delta ~ dnorm(0, 0.000001)
tau ~ dgamma(0.001, 0.001)
sigma2 <- 1/tau          # residual error variance

dummy <- DIST[1] # ensures all variables in data file appear in model code

# Create indicator variable for precinct 20
# (alternatively, could add this variable to data file)
for(i in 1:13) {   PREC20[i] <- 0   }
PREC20[14] <- 1
for(i in 15:23) {   PREC20[i] <- 0   }
```





Posterior mean 0.18

95% interval (-0.16, 0.49)

Specifying prior distributions

Why did we choose a $N(0, 100000)$ prior for each regression coefficient and a $Gamma(0.001, 0.001)$ prior for the inverse of the error variance?

Choice of prior is, in principle, subjective

- it might be elicited from experts (see Spiegelhalter et al (2004), sections 5.2, 5.3)

- it might be more convincing to be based on historical data, *e.g.* a previous study

  $\rightarrow$ assumed relevance is still a subjective judgement (see Spiegelhalter at al (2004), section 5.4)

- there has been a long and complex search for various 'non-informative', 'reference' or 'objective' priors (Kass and Wasserman, 1996)

'Non-informative' priors

- Better to refer to as 'vague', 'diffuse' or 'minimally informative' priors
- Prior is vague with respect to the likelihood
  - prior mass is diffusely spread over range of parameter values that are plausible, i.e. supported by the data (likelihood)

Uniform priors (Bayes 1763; Laplace, 1776)

Set $p(\theta) \propto 1$

- This is improper ($\int p(\theta)d\theta \neq 1$)
- The posterior will still usually be proper
- Inference is based on the likelihood $p(x \mid \theta)$
- It is not really objective, since a flat prior $p(\theta) \propto 1$ on θ does not correspond to a flat prior on $\phi = g(\theta)$, but to $p(\phi) \propto \left|\frac{d\theta}{d\phi}\right|$ where $\left|\frac{d\theta}{d\phi}\right|$ is the Jacobian
  - Note: Jacobian ensures area under curve (probability) in a specified interval (θ_1, θ_2) is preserved under the transformation $\rightarrow$ same area in interval $(\phi_1 = g(\theta_1), \phi_2 = g(\theta_2))$

Proper approximations to Uniform$(-\infty, \infty)$ prior:

- $p(\theta)$ = Uniform(a, b) where a and b specify an appropriately wide range, e.g. Uniform$(-1000, 1000)$

- $p(\theta)$ = N$(0, V)$ where V is an appropriately large value for the variance, e.g. N$(0, 100000)$

- Recall that WinBUGS parameterises Normal in terms of mean and precision, so vague normal prior will be, e.g. `theta ~ dnorm(0, 0.00001)`

Jeffreys' invariance priors

Consider 1-to-1 transformation of $\theta : \phi = g(\theta)$, e.g. $\phi = 1 + \theta^3$

Transformation of variables: $p(\theta)$ is equivalent to $p(\phi) = p(\theta = g^{-1}(\phi)) \left|\frac{d\theta}{d\phi}\right|$

Jeffreys proposed defining a non-informative prior for θ as $p(\theta) \propto I(\theta)^{1/2}$ where $I(\theta)$ is Fisher information for θ

$$I(\theta) = -\mathbb{E}_{X|\theta}\left[\frac{\partial^2 \log p(X|\theta)}{\partial\theta^2}\right] = \mathbb{E}_{X|\theta}\left[\left(\frac{\partial \log p(X|\theta)}{\partial\theta}\right)^2\right]$$

- Fisher Information measures curvature of log likelihood
- High curvature occurs wherever small changes in parameter values are associated with large changes in the likelihood
  - Jeffreys' prior gives more weight to these parameter values
  - data provide strong information about parameter values in this region
  - ensures data dominate prior everywhere
- Jeffreys' prior is invariant to reparameterisation because

$$I(\phi)^{1/2} = I(\theta)^{1/2}\left|\frac{d\theta}{d\phi}\right|$$

Examples of Jeffreys' priors

- Normal case: unknown mean m, known variance v
  Sample $x_1, \ldots, x_n$ from $N(m, v)$

$$\log p(x|m) = -\sum \frac{(x_i - m)^2}{2v} + C \qquad \Rightarrow I(m) = n/v$$

  So Jeffreys' prior for m is $\propto 1$, i.e. the Uniform distribution
- Normal case: known mean m, unknown variance v, with $s = \sum (x_i - m)^2$

$$\log p(x|v) = -n/2 \log v - \frac{s}{2v} \qquad \Rightarrow I(v) = \frac{n}{2v^2}$$

  So Jeffreys' prior for v is $\propto v^{-1}$
  This improper distribution is approximated by a Gamma(ϵ, ϵ) distribution with $\epsilon \longrightarrow 0$
  Note: $p(v) \propto v^{-1}$ is equivalent to a uniform prior on $\log v$

Some recommendations

Distinguish

- *primary* parameters of interest in which one may want minimal influence of priors

- *secondary* structure used for smoothing *etc.* in which informative priors may be more acceptable

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)

Location parameters (e.g. 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

Scale parameters

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

$$\begin{aligned} p(\sigma^2) &\propto \frac{1}{\sigma^2} \propto \text{Gamma}(0,0) \\ p(\log(\sigma)) &\propto \text{Uniform}(-\infty, \infty) \end{aligned}$$

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

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

  which may be approximated by a 'just proper' prior

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

  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)`
  or alternatively `tau <- 1/exp(logsigma2);  logsigma2 ~ dunif(-100, 100)`

Sensitivity analysis plays a crucial role in assessing the impact of particular prior distributions, whether elicited, derived from evidence, or reference, on the conclusions of an analysis.

Informative priors

- An informative prior expresses specific, definite information about a variable
- Example: a prior distribution for the temperature at noon tomorrow
  - A reasonable approach is to make the prior a normal distribution with mean equal to today's noontime temperature, with variance equal to the day-to-day variance of atmospheric temperature
- Posterior from one problem (today's temperature) becomes the prior for another problem (tomorrow's temperature)
- Priors elicited from experts can be used to take account of domain-specific knowledge, judgement, experience
- Priors can also be used to impose constraints on variables (e.g. based on physical or assumed properties) and bound variables to plausible ranges

Example: Trade union density

(Western and Jackman, 1994)

- Example of regression analysis in comparative research
- What explains cross-national variation in union density?
- Union density is defined as the percentage of the work force who belongs to a trade union
- Competing theories:
  - Wallerstein: union density depends on the size of the civilian labour force (LabF)
  - Stephens: union density depends on industrial concentration (IndC)
  - Note: These two predictors correlate at -0.92.

- Data: n = 20 countries with a continuous history of democracy since World War II

- Variables: Union density (Uden), (log) labour force size (LabF), industrial concentration (IndC), left wing government (LeftG), measured in late 1970s

- Fit linear regression model to compare theories

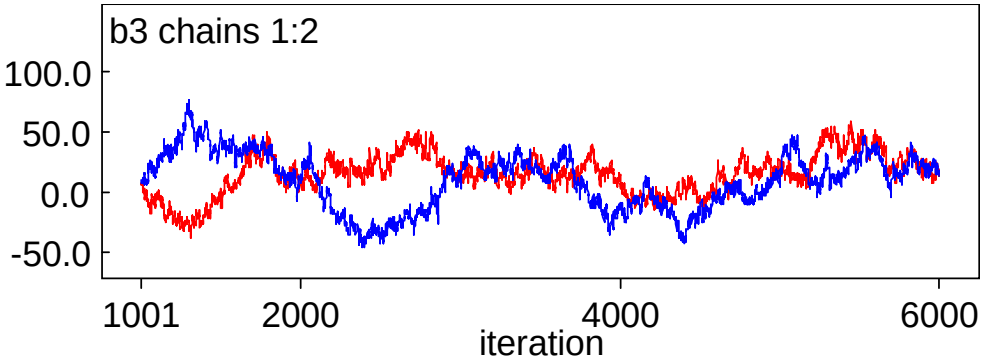
$$
\begin{aligned}
\text{Uden}_i &\sim \text{N}(\mu_i, \sigma^2) \\
\mu_i &= b_0 + b_1(\text{LeftG}_i - \overline{\text{LeftG}}) + b_2(\text{LabF}_i - \overline{\text{LabF}}) + b_3(\text{IndC}_i - \overline{\text{IndC}})
\end{aligned}
$$

Vague priors:

$$
\begin{aligned}
1/\sigma^2 &\sim \text{Gamma}(0.001, 0.001) \\
b_0 &\sim \text{N}(0, 100000) \\
b_1 &\sim \text{N}(0, 100000) \\
b_2 &\sim \text{N}(0, 100000) \\
b_3 &\sim \text{N}(0, 100000)
\end{aligned}
$$

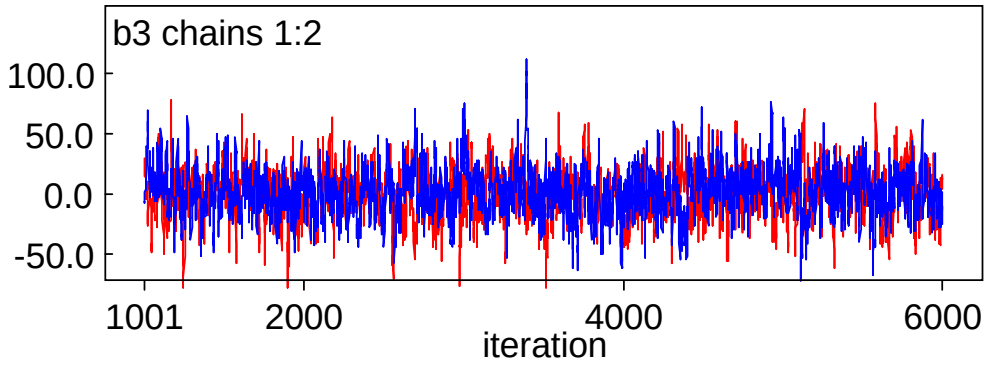
Trace plots, posterior estimates and MC error for regression coefficients

Without centering covariates



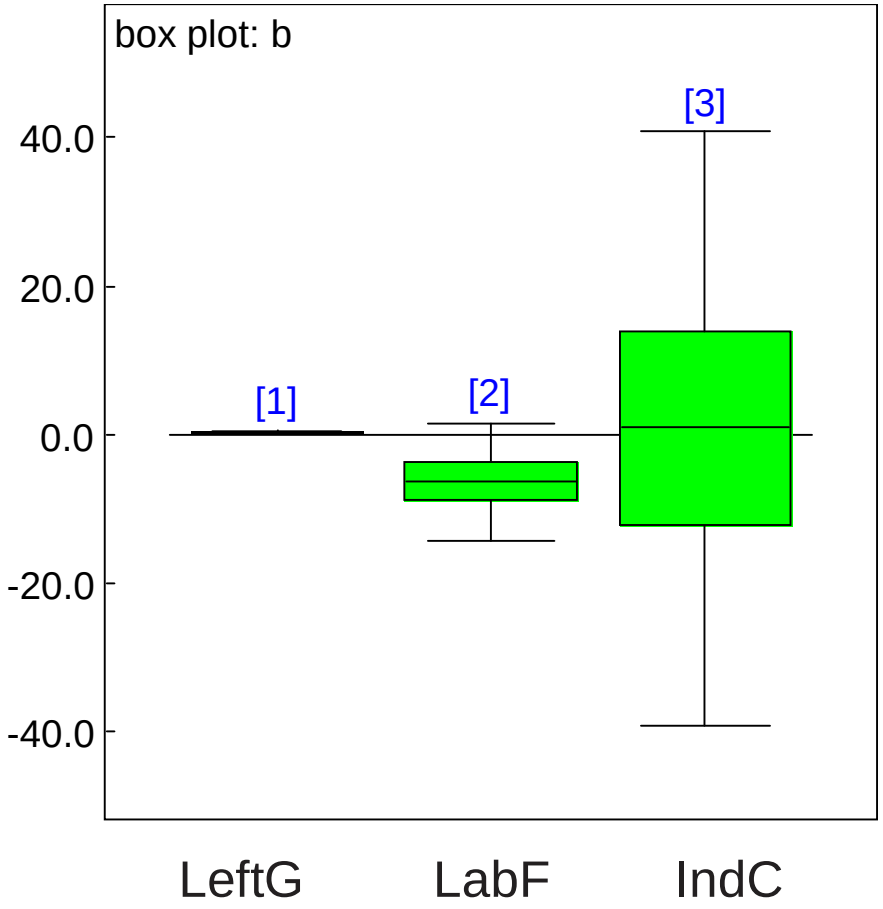
	mean	sd	MC error
b0	61.7	62.8	5.19
b1	0.27	0.08	.002
b2	-4.14	4.18	0.34
b3	12.1	20.6	1.67

With centered covariates



	mean	sd	MC error
b0	54.0	2.48	0.02
b1	0.27	0.08	.001
b2	-6.33	3.96	0.13
b3	0.98	20.2	0.67

Posterior distribution of regression coefficients



Motivation for Bayesian approach with informative priors

- Because of small sample size and multicollinear variables, not able to adjudicate between theories
- Data tend to favour Wallerstein (union density depends on labour force size), but neither coefficient estimated very precisely
- Other historical data are available that could provide further relevant information
- Incorporation of prior information provides additional structure to the data, which helps to uniquely identify the two coefficients

Wallerstein informative prior

- Believes in negative labour force effect
- Comparison of Sweden and Norway in 1950:
  - doubling of labour force corresponds to 3.5-4% drop in union density
  - on log scale, labour force effect size $\approx -3.5/\log(2) \approx -5$
- Confidence in direction of effect represented by prior SD giving 95% interval that excludes 0

$$b_2 \sim \mathrm{N}(-5, 2.5^2)$$

- Vague prior assumed for IndC effect, $b_3 \sim \mathrm{N}(0, 100000)$

Stephens informative prior

- Believes in positive industrial concentration effect
- Decline in industrial concentration in UK in 1980s:
  - drop of 0.3 in industrial concentration corresponds to about 3% drop in union density
  - industrial concentration effect size $\approx 3/0.3 = 10$
- Confidence in direction of effect represented by prior SD giving 95% interval that excludes 0

$$b_3 \sim \mathrm{N}(10, 5^2)$$

- Vague prior assumed for IndC effect, $b_3 \sim \mathrm{N}(0, 100000)$

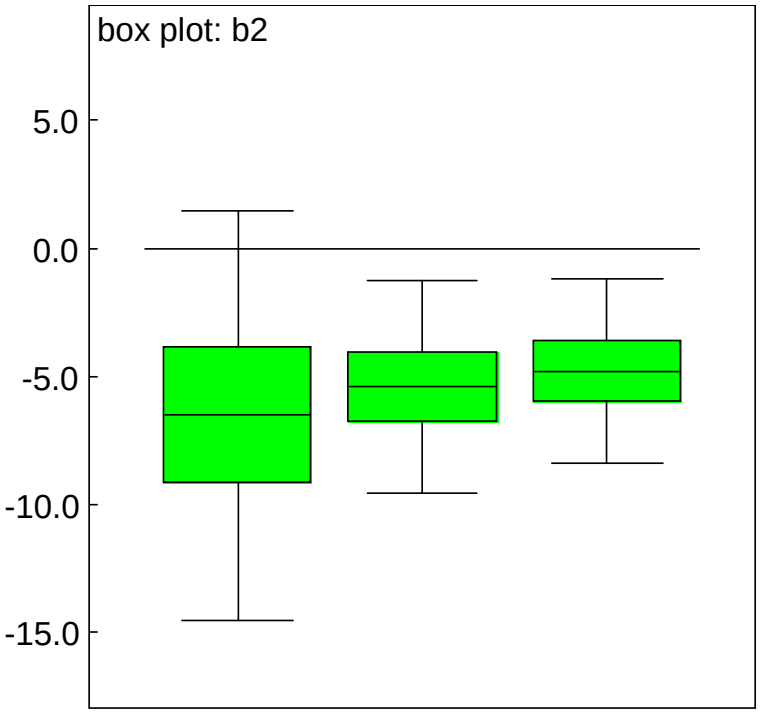
Both Wallerstein and Stephens priors

- Both believe left-wing governments assist union growth
- Assuming 1 year of left-wing government increases union density by about 1% translates to effect size of 0.3
- Confidence in direction of effect represented by prior SD giving 95% interval that excludes 0

$$b_1 \sim \mathrm{N}(0.3, 0.15^2)$$

- Vague prior $b_0 \sim \mathrm{N}(0, 100000)$ assumed for intercept

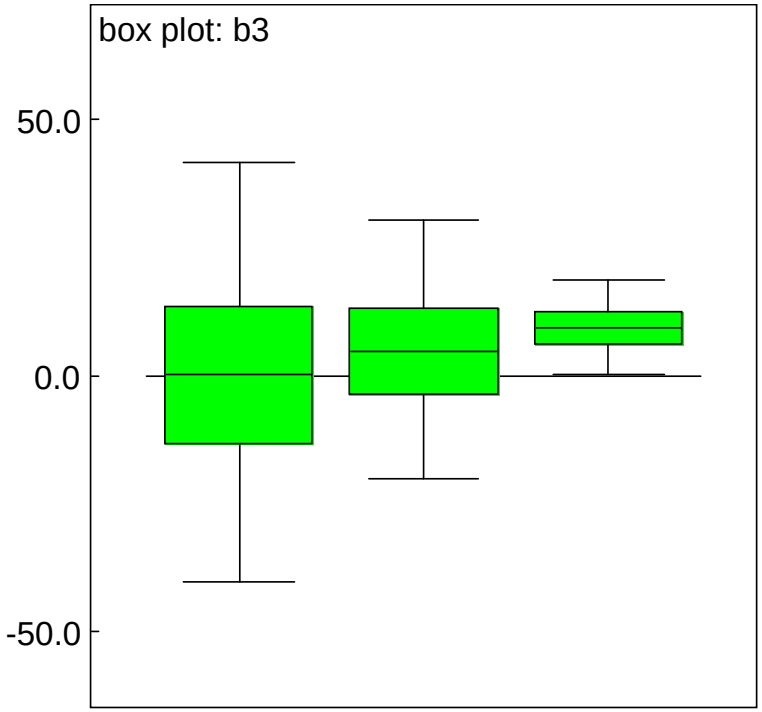
Effect of Labour Force Size (Wallerstein hypothesis)



b2 (LabF): Vague Info Vague

b3 (IndC): Vague Vague Info

Effect of Industrial Concentration (Stephens hypothesis)



b2 (LabF): Vague Info Vague

b3 (IndC): Vague Vague Info

Comments

- Effects of LabF and IndC estimated more precisely
- Both sets of prior beliefs support inference that labour-force size decreases union density
- Only Stephens prior supports conclusion that industrial concentration increases union density
- Choice of prior is subjective – if no consensus, can we be satisfied that data have been interpreted fairly?
  - Sensitivity to priors (e.g. repeat analysis using priors with increasing variance) — see Practical exercises
  - Sensitivity to data (e.g. residuals, influence diagnostics) — see later lecture

Multivariate responses

- In many applications, it is common to collect data on a number of different outcomes measured on the same units, e.g.
  - sample survey, where respondents asked several different questions
  - experiment with several different outcomes measured on each unit

- May wish to fit regression model to each response
  - May have different covariates in each regression
  - But errors may be correlated
  - Might also wish to impose cross-equation parameter restrictions

  $\rightarrow$ Seemingly Unrelated Regressions (SUR) (Zellner, 1962)

Bayesian approach to SUR models $\rightarrow$ model vector of response for each unit as multivariate normal (could also have robust version using multivariate t)

Possible to extend Bayesian SUR models to binary, categorical, count responses using multivariate latent variable approach.

Example: Analysis of compositional data

- Compositional data are vectors of proportions $\boldsymbol{p}_i = (p_{i1}, ..., p_{iJ})$ representing relative contributions of each of J categories to the whole, e.g.
  - Proportion of income spent on different categories of expenditure
  - Proportion of electorate voting for different political parties
  - Relative abundance of different species in a habitat
  - Chemical composition of a rock or soil sample
  - Proportion of deaths from different causes in a population
- Regression models for compositional data must satisfy two constraints

$$0 \leq p_{ij} \leq 1$$
$$\sum_j p_{ij} = 1$$

- Two main modelling strategies, treating vector of proportions as the data (sufficient statistics)
  - Model $\boldsymbol{p}_i$ using a Dirichlet likelihood (multivariate generalisation of a beta distribution)
    - assumes the ratios of "compositions" (i.e. proportions) are independent
  - Multivariate logistic normal model (Aitchison, 1986) — apply *additive log ratio* (alr) transformation, $y_{ij} = \log(p_{ij}/p_{iJ})$ and model $\boldsymbol{y}_i$ as multivariate normal
    - allows dependence between ratios of proportions
    - this can be thought of as a type of SUR model
- To allow for sampling variability in the observed counts (including zero counts), model counts (rather than proportions) as multinomial (see later)

Multivariate logistic normal model

Define

$$y_{ij} = \log(p_{ij}/p_{iJ})$$

the log ratios of proportions in each category relative to a reference category J

Note that $p_{iJ} = 1 - \sum_{j \neq J} p_{ij}$, so

$$p_{ij} = \frac{\exp y_{ij}}{1 + \sum_{j \neq J} \exp y_{ij}}$$

Since y_{ij} are unconstrained, can model vector $\boldsymbol{y}_i = \{y_{ij}, j \neq J\}$ as multivariate normal

Example: British General Election 1992

- Data originally analysed by Katz and King (1999), and formulated as BUGS example by Simon Jackman

- Data consist of vote proportions for Conservative (j=1), Labour (j=2) and Lib-Dem (j=3) parties from 1992 General Election for each of 521 constituencies

- Additive log ratio transformation applied to proportions, taking Lib-Dem vote as reference category

- Covariates include lagged values of the log ratios from previous election, and indicators of the incumbency status of each party's candidate

BUGS model code

```
for(i in 1:521){
  y[i,1:2] ~ dmnorm(mu[i,], prec[ , ])
  for(j in 1:2){
    mu[i,j] <- beta[j,1]*x[i,1] + beta[j,2]*x[i,2] + beta[j,3]*x[i,3] +
               beta[j,4]*x[i,4] + beta[j,5]*x[i,5] + beta[j,6]*x[i,6]
  }
}

## priors for elements of precision matrix
prec[1:2,1:2] ~ dwish(R[,],k)
R[1,1] <- .01; R[1,2] <- 0; R[2,1] <- 0; R[2,2] <- .01;  k <- 2

# convert precision to covariance matrix
sigma[1:2,1:2] <- inverse(prec[ , ])
rho <- sigma[1,2]/sqrt(sigma[1,1]*sigma[2,2])  # correlation

## Priors for regression coefficients
for(j in 1:2){
  for(k in 1:6) {
    beta[j,k] ~ dnorm(0, 0.000001)
  }
}
```

Priors on precision matrix of multivariate normal

The multivariate generalisation of the Gamma (or χ^2) distribution is the Wishart distribution, which arises in classical statistics as the distribution of the sum-of-squares-and-products matrix in multivariate normal sampling.

The Wishart distribution $\mathrm{W}_p(k, R)$ for a symmetric positive definite $p \times p$ matrix Ω has joint density function proportional to

$$|R|^{k/2}|\Omega|^{(k-p-1)/2} \exp\left(-(1/2)\mathrm{tr}(R\Omega)\right)$$

in terms of two parameters: a real scalar $k > p-1$ and a symmetric positive definite matrix R. The expectation of this distribution is

$$E[\Omega] = kR^{-1}$$

When the dimension p is 1, i.e. reverting back to univariate case, it is easy to show that the Wishart distribution becomes the more familiar:

$$\mathrm{W}_1(k, R) \equiv \mathrm{Gamma}(k/2, R/2) \equiv (\chi^2_k)/R$$

If we use the Wishart distrubution as a prior distribution for a precision matrix Ω in sampling from $N_p(\mu, \Omega^{-1})$, we find, generalising the univariate case above, that we get the same form for the posterior for Ω – another Wishart distribution.

In view of the result above for the expectation of the Wishart distribution, we usually set $(1/k)R$ to be a prior guess at the unknown true variance matrix. A common choice is to take $k = p$.

Note

- In BUGS language, you must specify the dimension of vectors or arrays on the left hand side of multivariate distributions

- In above example, each row (observation) of the 521×2 matrix $\boldsymbol{y}$ is a vector of length 2, hence `y[i,1:2] ~ dmnorm.....`

- Likewise, `prec` is a 2×2 matrix, hence `prec[1:2, 1:2] ~ dwish.....`

- You **cannot** specify the dimension to be a parameter (even if the value of the parameter is specified elsewhere in the code or data file), e.g.

```
J <- 2
prec[1:J, 1:J] ~ dwish(R[,], k)
```

  will give an error at compilation

- You do not need to specify the dimension of vectors or arrays on the right hand side of distribution statements, e.g. dimension of `R[,]` is not specified above (the dimension is implicit from dimension of left hand side)

Interpretation of model parameters

- Interpretation of parameter estimates on a multivariate log-odds scale is difficult
- Using the inverse alr transformation, easy to recover estimates of expected proportions or predicted counts in different categories
- Effect of covariates can be examined by calculating difference or ratio of expected proportions for different values of the covariate
  - Using MCMC, easy to obtain uncertainty intervals for such contrasts
- Example: effect of incumbency on expected proportion of votes for each party
  - For party j, calculate expected alr-transformed proportion for two values of incumbency: (1) party j's candidate is incumbent; (2) open seat (no candidate is incumbent)
  - Hold values of all other covariates constant, e.g. at their means
  - Use inverse alr transformation to obtain expected proportions under each incumbency value, and take difference

BUGS code for calculating incumbency effects

```
for(j in 1:2){
 # value of mu with Conservative incumbent and average values of other variables
 mu.con[j] <- beta[j,1]*mean(x[,1]) + beta[j,2]*mean(x[,2]) +
              beta[j,3]*mean(x[,3]) + beta[j,4]*1

 # value of mu with Labour incumbent and average values of other variables
 mu.lab[j] <- beta[j,1]*mean(x[,1]) + beta[j,2]*mean(x[,2]) +
              beta[j,3]*mean(x[,3]) + beta[j,5]*1

 # value of mu with open seat and average values of other variables
 mu.open[j] <- beta[j,1]*mean(x[,1]) + beta[j,2]*mean(x[,2]) + beta[j,3]*mean(x[,3])

 # expected proportions
 exp.mu.con[j] <- exp(mu.con[j]); p.con[j] <- exp.mu.con[j]/(1 + sum(exp.mu.con[]))

 exp.mu.lab[j] <- exp(mu.lab[j]); p.lab[j] <- exp.mu.lab[j]/(1 + sum(exp.mu.lab[]))

 exp.mu.open[j] <- exp(mu.open[j]); p.open[j] <- exp.mu.open[j]/(1 + sum(exp.mu.open[]))
}

# difference in expected proportions due to incumbency
incumbency.con <- p.con[1] - p.open[1]
incumbency.lab <- p.lab[2] - p.open[2]
```

Results

	Posterior mean	95% CI
Expected vote (Cons), area 1	44.8%	(44.2%, 45.4%)
Expected vote (Lab), area 1	46.1%	(45.3%, 46.8%)
Expected vote (LibDem), area 1	9.1%	(8.7%, 9.5%)
incumbency advantage (Cons)	−0.06%	(−0.75%, 0.70%)
incumbency advantage (Lab)	−0.30%	(−1.6%, 1.0%)
incumbency advantage (LibDem)	8.6%	(5.1%, 12.3%)
ρ (correlation between log ratio for Cons:LibDem and Lab:LibDem)	0.87	(0.85, 0.89)

Note: results for incumbency advantage agree with those in Tomz et al (2002) but not with Katz and King, who analysed data from 10 consecutive elections and used empirical Bayes shrinkage priors on the β coefficients across years

Multivariate t likelihood

- Multivariate logistic normal for compositional data relies on assumption that the log ratios are approximately multivariate normal
- Katz and King (1999) argue that this assumption is not appropriate for British election data
  - majority of constituencies tend to be more clustered, and a minority more widely dispersed, than the multivariate normal implies
- K&K propose replacing multivariate normal by a heavier-tailed multivariate student t distribution
- Multivariate t has 3 parameters: p-dimensional mean vector, $p \times p$ inverse scale (precision) matrix and a scalar degrees of freedom parameter
- A Wishart prior can be used for the inverse scale matrix
- Degrees of freedom parameter can either be fixed, or assigned a prior distribution
- Note: as degrees of freedom $\rightarrow \infty$, t $\rightarrow$ Normal

BUGS code for multivariate t likelihood

- Only need to change 2 lines of code
  1. Likelihood:

```
##  y[i,1:2] ~ dmnorm(mu[i,1:2], prec[ , ])
   y[i,1:2] ~ dmt(mu[i,1:2], prec[ , ], nu)
```

  2. Specify either fixed value for degrees of freedom, `nu`, or a suitable prior

```
## nu <- 4
  nu ~ dunif(2, 250)
```

Results

	Normal	**Student t**
Expected vote (Cons), area 1	44.8% (44.2%, 45.4%)	44.7% (44.2%, 45.2%)
Expected vote (Lab), area 1	46.1% (45.3%, 46.8%)	46.4% (45.7%, 47.0%)
Expected vote (LibDem), area 1	9.1% (8.7%, 9.5%)	8.8% (8.6%, 9.3%)
incumbency advantage (Cons)	−0.06% (−0.75%, 0.70%)	0.00% (-0.65%, 0.60%)
incumbency advantage (Lab)	−0.30% (−1.6%, 1.0%)	−0.50% (−1.7%, 0.70%)
incumbency advantage (LibDem)	8.6% (5.1%, 12.3%)	2.3% (−2.0%, 8.3%)
ρ	0.87 (0.85, 0.89)	0.87 (0.85, 0.90)
ν	–	4.5 (3.4, 5.9)
DIC	−480	−1220
p_D	14.5	11.8

Lecture 5.

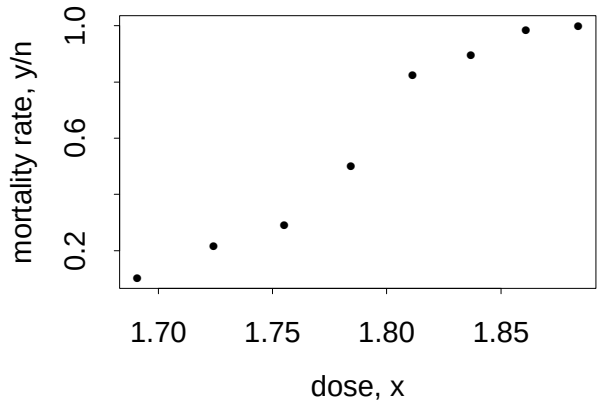
Further Bayesian regression models

Generalised Linear regression Models

- Specification of Bayesian GLMs follows straightforwardly from previous discussion of linear models
- No closed form solution available, but straightforward to obtain samples from posterior using MCMC

Example: Beetles

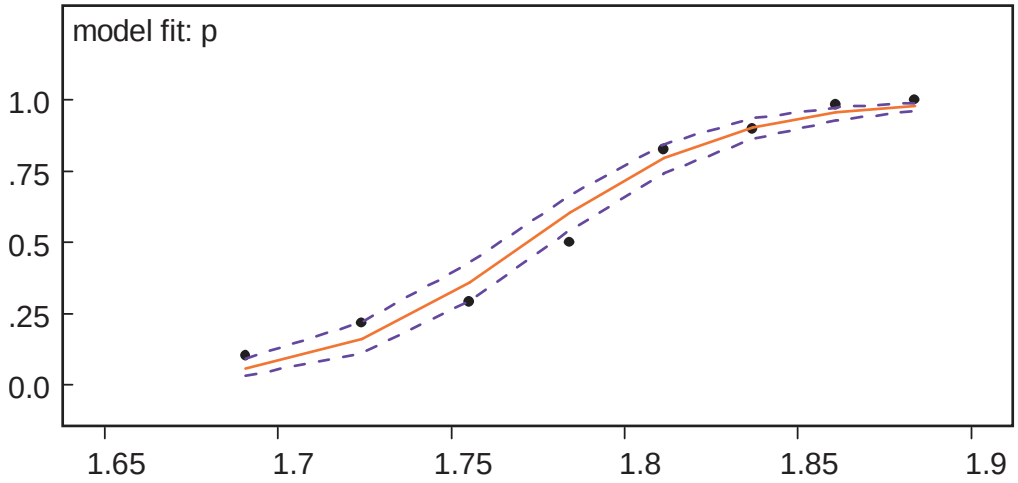
Dobson (1983) analyses binary dose-response data from a bioassay experiment in which the numbers of beetles killed after 5 hour exposure to carbon disulphide at N=8 different concentrations are recorded.



We start by fitting a logistic regression model

$$\begin{aligned} y_i &\sim \text{Binomial}(p_i, n_i) \\ \text{logit} p_i &= \alpha + \beta(x_i - \overline{x}) \\ \alpha &\sim \text{Normal}(0, 10000) \\ \beta &\sim \text{Normal}(0, 10000) \end{aligned}$$

Beetles: logistic regression model fit (red = posterior mean of p_i; blue = 95% interval; black dots = observed rate y_i/n_i)

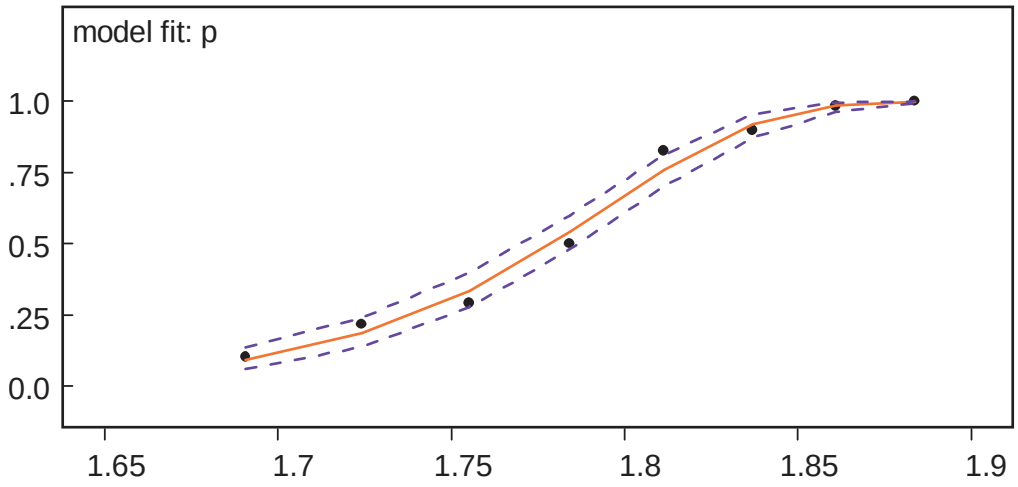


dose level i	obs. rate y_i/n_i	posterior mean of p_i	95% interval
1	0.10	0.06	(0.03, 0.09)
2	0.22	0.16	(0.11, 0.22)
3	0.29	0.36	(0.29, 0.43)
4	0.50	0.61	(0.54, 0.67)
5	0.83	0.80	(0.74, 0.85)
6	0.90	0.90	(0.86, 0.94)
7	0.98	0.96	(0.93, 0.97)
8	1.00	0.98	(0.96, 0.99)

Some evidence of lack of fit at extremes, so try alternative complementary log-log link function

$$\begin{aligned} y_i &\sim \text{Binomial}(p_i, n_i) \\ \text{cloglog} p_i &= \alpha + \beta(x_i - \overline{x}) \\ \alpha &\sim \text{Normal}(0, 10000) \\ \beta &\sim \text{Normal}(0, 10000) \end{aligned}$$

Beetles: cloglog regression model fit (red = posterior mean of p_i; blue = 95% interval; black dots = observed rate y_i/n_i)



dose level i	obs. rate y_i/n_i	posterior mean of p_i	95% interval
1	0.10	0.09	(0.06, 0.14)
2	0.22	0.19	(0.14, 0.24)
3	0.29	0.34	(0.28, 0.40)
4	0.50	0.54	(0.48, 0.60)
5	0.83	0.76	(0.70, 0.81)
6	0.90	0.92	(0.87, 0.95)
7	0.98	0.98	(0.96, 0.99)
8	1.00	1.00	(0.99, 1.00)

Can write probit model in two different ways

$$\text{probit}p_i \quad = \quad \alpha + \beta(x_i - \overline{x})$$

or

$$p_i \quad = \quad \Phi(\alpha + \beta(x_i - \overline{x}))$$

In WinBUGS , either

```
probit(p[i]) <- alpha + beta*(x[i]-mean(x[]))
```

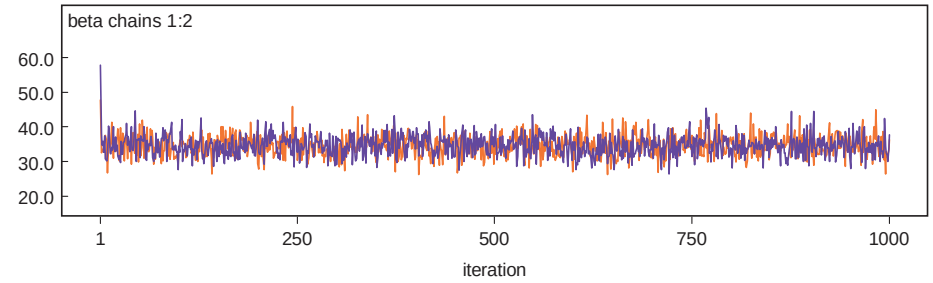
or

```
p[i] <- phi(alpha + beta*(x[i]-mean(x[])))
```

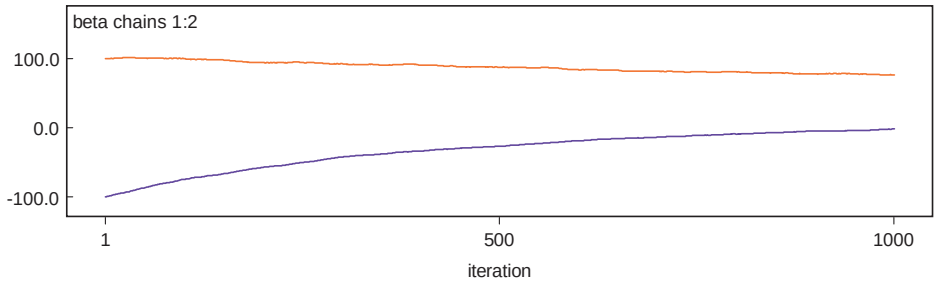
The second way is *slower*, but can be *more robust* to numerical problems.

Note the importance of centering the covariate (dose) in this example to reduce correlations between the parameters

History plot for slope, β: Centred covariate

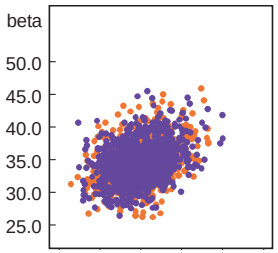


History plot for slope, β: Uncentred covariate

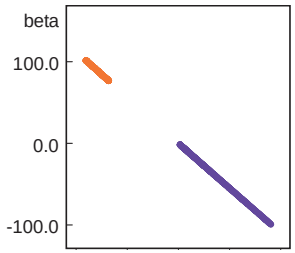


Bivariate scatter plot showing correlation between sampled values of α and β

Centered covariate



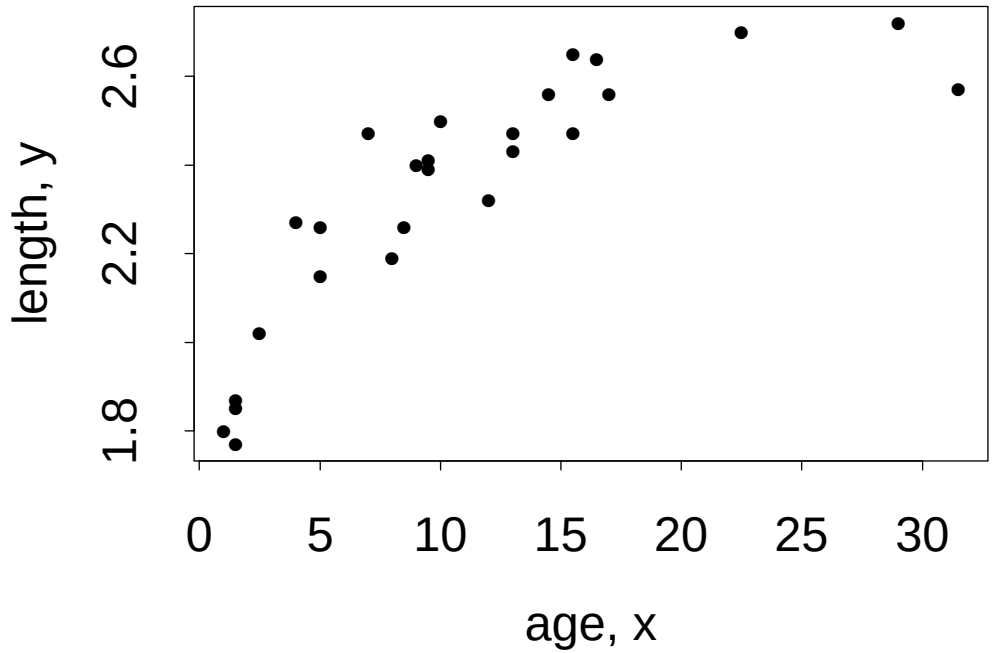
Uncentred covariate



Non linear regression models

Example: Dugongs

Carlin and Gelfand (1991) consider data on length (y_i) and age (x_i) measurements for 27 dugongs (sea cows) captured off the coast of Queensland



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 as e.g.

$$\begin{aligned} \alpha &\sim \text{Uniform}(0, 100) \\ \beta &\sim \text{Uniform}(0, 100) \\ \gamma &\sim \text{Uniform}(0, 1) \end{aligned}$$

Alternatively, vague Normal priors with appropriate bounds could be specified for α and β, e.g.

$$\begin{aligned} \alpha &\sim \text{Normal}(0, 10000)I(0,) \\ \beta &\sim \text{Uniform}(0, 10000)I(0,) \end{aligned}$$

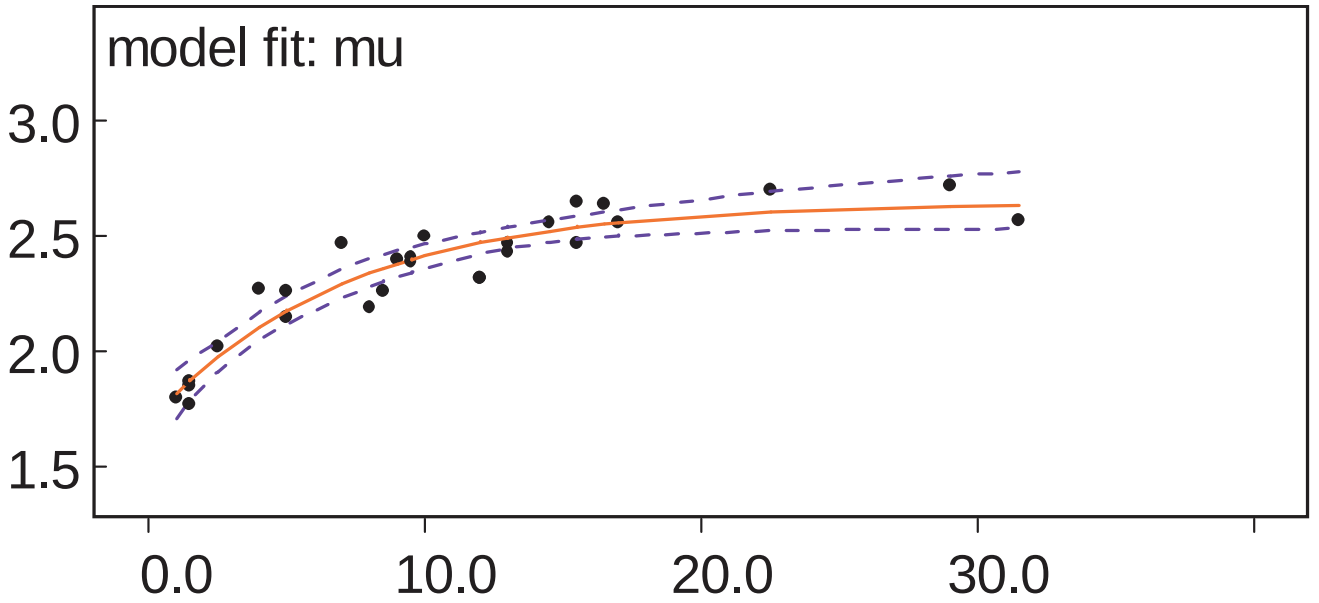
For the sampling variance, could specify uniform prior log variance or log sd scale

$$\log \sigma \sim \text{Uniform}(-10, 10)$$

or gamma prior on precision scale

$$1/\sigma^2 \sim \text{Gamma}(0.001, 0.001)$$

Dugongs: model fit (red = posterior mean of μ_i; blue = 95% interval)



Categorical data

1. Data recorded as $y_i = 0$ or 1

```
y[i] ~ dbern(p[i])
probit(p[i]) <- beta0 + beta1*x1[i] .....
```

NB `probit` can be rather fragile:

```
p[i] <- phi(beta0 + beta1*x1[i] .....)
```

is slower but may be more robust.

2. A single categorical variable y_i= 1,2,3

```
y[i] ~ dcat(p[])
```

where `p[]` is an array of probabilities

Modelling unknown denominators

Suppose we are told that a fair coin has come up heads 10 times - how many times (n) has it been tossed?

We want to specify a uniform prior distribution for n

1. Could give a continuous prior distribution for n and use 'round' function

```
model {
   r <- 10
   q <- 0.5
   r ~ dbin(q, n)
   n.cont ~ dunif(1, 100)
   n <- round(n.cont)
}
```

BUGS output:

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
n	21.08	4.794	0.07906	13.0	21.0	32.0	1001	5000
n.cont	21.08	4.804	0.07932	13.31	20.6	32.0	1001	5000

We can be 95% sure that the coin has been tossed between 13 and 32 times

2. Or a discrete uniform prior on 10 to 100

```
model {
  r <- 10
  q <- 0.5
  r ~ dbin(q, n)
  # discrete prior on 10 to 100
  for(j in 1:9)   {  p[j]<-0   }
  for(j in 10:100){  p[j]<-1/91}
  n ~ dcat(p[])
}
```

BUGS output:

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
n	21.07	4.761	0.03929	13.0	21.0	32.0	1001	10000

We obtain a similar answer to before, i.e. we can be 95% sure that the coin has been tossed between 13 and 32 times

Multinomial data

Suppose observed data are arrays of counts in K categories,
e.g. $\boldsymbol{y}$ = (1,2,4), (3,3,3) etc

```
for(i in 1:N) {
  y[i,1:3] ~ dmulti(q[], n[i])
  n[i] <- sum(y[i,])
}
```

If have no covariates, can use Dirichlet prior for `q[]`, so that $p(\boldsymbol{q}) \propto q_k^{\alpha_k}$

```
  q[1:3] ~ ddirch(alpha[])
```

$\boldsymbol{\alpha}$ needs to be fixed (i.e. can't learn about $\boldsymbol{\alpha}$), *e.g.*

```
for(k in 1:3) {
   alpha[k]<-1
}
```

gives uniform prior on $\boldsymbol{q}$

Remember that you need to specify the dimension of vectors or arrays on the left hand side of distributions in the BUGS languange

If have covariates, can use multinomial-logistic model

$$\eta_{ik} = \log \frac{q_{ik}}{q_{i1}} = \alpha_k + \sum_p \beta_{pk} x_{pi}, \quad k = 2, ..., K; i = 1, ..., N$$

- Conceptually, this is equivalent to $K - 1$ logistic regressions comparing category $k > 1$ with category 1

- *cf* log ratio model for compositional data — here we are modelling the observed counts in each category (rather than the observed log ratios), and treating the underlying probabilities as unknown parameters

- Can also re-write* model in terms of the original probabilities rather than the log ratios (log odds):

$$q_{ik} = \frac{\phi_{ik}}{\sum_k \phi_{ik}} \qquad \text{where } \phi_{ik} = e^{\eta_{ik}} = e^{\alpha_k + \sum_p \beta_{pk} x_{pi}}$$

  with constraint that $\phi_{i1} = 1$ (i.e. $\eta_{i1} = 0$)

*To verify this result exponentiate the first equation for the log ratio to obtain $q_{ik} = q_{i1} e^{\eta_{ik}}$, and note that the convention $\eta_{i1} = 0$ makes this formula valid for all k. Next sum over k and use the fact that $\sum_k q_{ik} = 1$ to obtain $q_{i1} = 1/\sum_k e^{\eta_{ik}}$.

Example: Car choice

Foster et al (1998) present data on choice of car (family, sporty or work) for 263 customers by gender, age and marital status.

Age	**Sex**	**Marital**	**Car type (y)**		
			Family	**Sporty**	**Work**
< 25	F	Married	6	6	3
$25 - 35$	F	Married	32	8	8
> 35	F	Married	13	2	4
< 25	M	Married	5	4	1
$25 - 35$	M	Married	37	12	11
> 35	M	Married	13	2	3
< 25	F	Single	4	11	0
$25 - 35$	F	Single	9	5	2
> 35	F	Single	4	0	2
< 25	M	Single	4	11	3
$25 - 35$	M	Single	10	14	6
> 35	M	Single	3	4	1

Here we will view car type as the response and age, sex and marital status as predictors

BUGS code

```
for(i in 1:12){
  y[i,1:3] ~ dmulti(q[i,], n[i])
  n[i] <- sum(y[i,])
  for(k in 1:3) {
    q[i,k] <- phi[i,k] / sum(phi[i,])
    log(phi[i,k]) <- alpha[k] + beta[k, age[i]] + delta[k]*sex[i] + gamma[k]*marital[i]
  }
}
# constraints
alpha[1]<-0; beta[1,1]<-0; beta[1,2]<-0; beta[1,3]<-0; delta[1]<-0; gamma[1]<-0
beta[2,1] <- 0; beta[3,1] <- 0   # further constraints (baseline age categories)
# priors
for(k in 2:3) {
  alpha[k] ~ dnorm(0, 0.0001)
  beta[k, 2] ~ dnorm(0, 0.0001);   beta[k, 3] ~ dnorm(0, 0.0001)
  delta[k] ~ dnorm(0, 0.0001)
  gamma[k] ~ dnorm(0, 0.0001)

  # odds ratios of choosing car type k (2=sporty, 3=work) versus type 1 (family)
  OR.age25[k] <- exp(beta[k,2])    # odds ratio for age 25-35 vs <25
  OR.age35[k] <- exp(beta[k,3])    # odds ratio for age >35 vs <25
  OR.male[k] <- exp(delta[k])      # odds ratio for males vs females
  OR.single[k] <- exp(gamma[k])    # odds ratio for single vs married
}
```

Data

```
age[]   sex[]   marital[]   y[,1]   y[,2]   y[,3]
1   1   1   6   6   3
2   1   1   32  8   8
3   1   1   13  2   4
1   2   1   5   4   1
2   2   1   37  12  11
3   2   1   13  2   3
1   1   2   4   11  0
2   1   2   9   5   2
3   1   2   4   0   2
1   2   2   4   11  3
2   2   2   10  14  6
3   2   2   3   4   1
END
```

Initial values

```
list(alpha = c(NA, 1, 1), delta = c(NA, 1, 1),  gamma = c(NA, 1, 1),
      beta = structure(.Data = c(NA, NA, NA,
                                 NA, 1, 1,
                                 NA, 1, 1), .Dim=c(3,3))
)
```

Note syntax for arrays (`beta`) — data read row by row, with `.Dim` statement specifying number of rows and number of columns.

Results

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
OR.age25[2]	0.3343	0.1262	0.00313	0.1513	0.3135	0.633	4001	52000
OR.age25[3]	1.036	0.6156	0.01894	0.3354	0.8885	2.603	4001	52000
OR.age35[2]	0.1905	0.09994	0.00190	0.0589	0.1695	0.440	4001	52000
OR.age35[3]	1.06	0.7099	0.01921	0.2781	0.8796	2.874	4001	52000
OR.male[2]	1.525	0.4869	0.01719	0.7882	1.457	2.681	4001	52000
OR.male[3]	1.324	0.482	0.01787	0.624	1.237	2.482	4001	52000
OR.single[2]	3.529	1.143	0.04155	1.809	3.354	6.247	4001	52000
OR.single[3]	1.502	0.6006	0.02165	0.6331	1.402	2.973	4001	52000

- So, for example, the odds of buying a sporty rather than family car is over 3.5 times higher for single customers compared to married customers (95% CI 1.8, 6.2) (`OR.single[2]`)

- May be more efficient to fit this using Poisson distributions: see `alli` example in manual

- Can also extend to ordered categorical data (see `Bones` and `inhalers` examples in manual)

Original data were in subject-specific format with age as a continuous variable:

```
age[] sex[] marital[] y[]
34  2   1   1
36  2   2   2
23  2   1   1
....
```

To fit model to individual level data, use **categorical** rather than **multinomial** likelihood

```
for(i in 1:263){
  y[i] ~ dcat(q[i,])
  for(k in 1:3) {
    q[i,k] <- phi[i,k] / sum(phi[i,])
    log(phi[i,k]) <- alpha + beta[k]*age[i] + delta[k]*sex[i] + gamma[k]*marital[i]
  }
}
# constraints
alpha[1]<-0; beta[1]<-0; delta[1]<-0; gamma[1]<-0

# priors
for(k in 2:3) {
  alpha[k] ~ dnorm(0, 0.0001)
  beta[k] ~ dnorm(0, 0.0001)
  delta[k] ~ dnorm(0, 0.0001)
  gamma[k] ~ dnorm(0, 0.0001)

  # odds ratios of choosing car type k (2=sporty, 3=work) versus type 1 (family)
  OR.age[k] <- exp(beta[k])       # odds ratio per year
  OR.male[k] <- exp(delta[k])     # odds ratio for males vs females
  OR.single[k] <- exp(gamma[k])   # odds ratio for single vs married
}
```

Results

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
OR.age[2]	0.8818	0.0267	4.753E-4	0.8292	0.8822	0.9326	4001	22500
OR.age[3]	0.9701	0.0293	4.388E-4	0.9112	0.9707	1.027	4001	22500
OR.male[2]	1.404	0.4324	0.02191	0.7466	1.333	2.375	4001	22500
OR.male[3]	1.32	0.5035	0.0258	0.6219	1.234	2.573	4001	22500
OR.single[2]	3.802	1.256	0.06471	1.986	3.573	6.814	4001	22500
OR.single[3]	1.532	0.5835	0.02781	0.7131	1.435	3.008	4001	22500

- Results for `OR.single` and `OR.male` very similar to previous model
- Odds ratios for age indicate
  - significant reduction in odds of buying sporty versus family car as age increases
  - slight but non-significant reduction in odds of buying work versus family car as age increases

Lecture 6.

Predictions, missing data, model checking, model comparison

Making predictions

- Important to be able to predict unobserved quantities for
  - 'filling-in' missing or censored data
  - model checking - are predictions 'similar' to observed data?
  - making predictions!

- Easy in MCMC/WinBUGS; just specify a stochastic node without a data-value - it will be automatically predicted

- Provides automatic imputation of missing data

- Easiest case is where there is no data at all: just 'forward sampling' from prior, *Monte Carlo* methods

Example: Dugongs — prediction

Suppose we want to project beyond current observations, eg at ages 35 and 40

Could explicitly set up predictions

```
for (i in 1:N){
    y[i]   ~   dnorm( mu[i], inv.sigma2 )
    mu[i] <-   alpha - beta * pow(gamma, x[i])
    }
mu35  <-   alpha - beta * pow(gamma, 35)
mu40  <-   alpha - beta * pow(gamma, 40)
y35    ~   dnorm( mu35, inv.sigma2 )
y40    ~   dnorm( mu40, inv.sigma2 )
```

Interval around μ_{40} will reflect uncertainty concerning fitted parameters

Interval around y_{40} will additionally reflect sampling error σ and uncertainty about σ

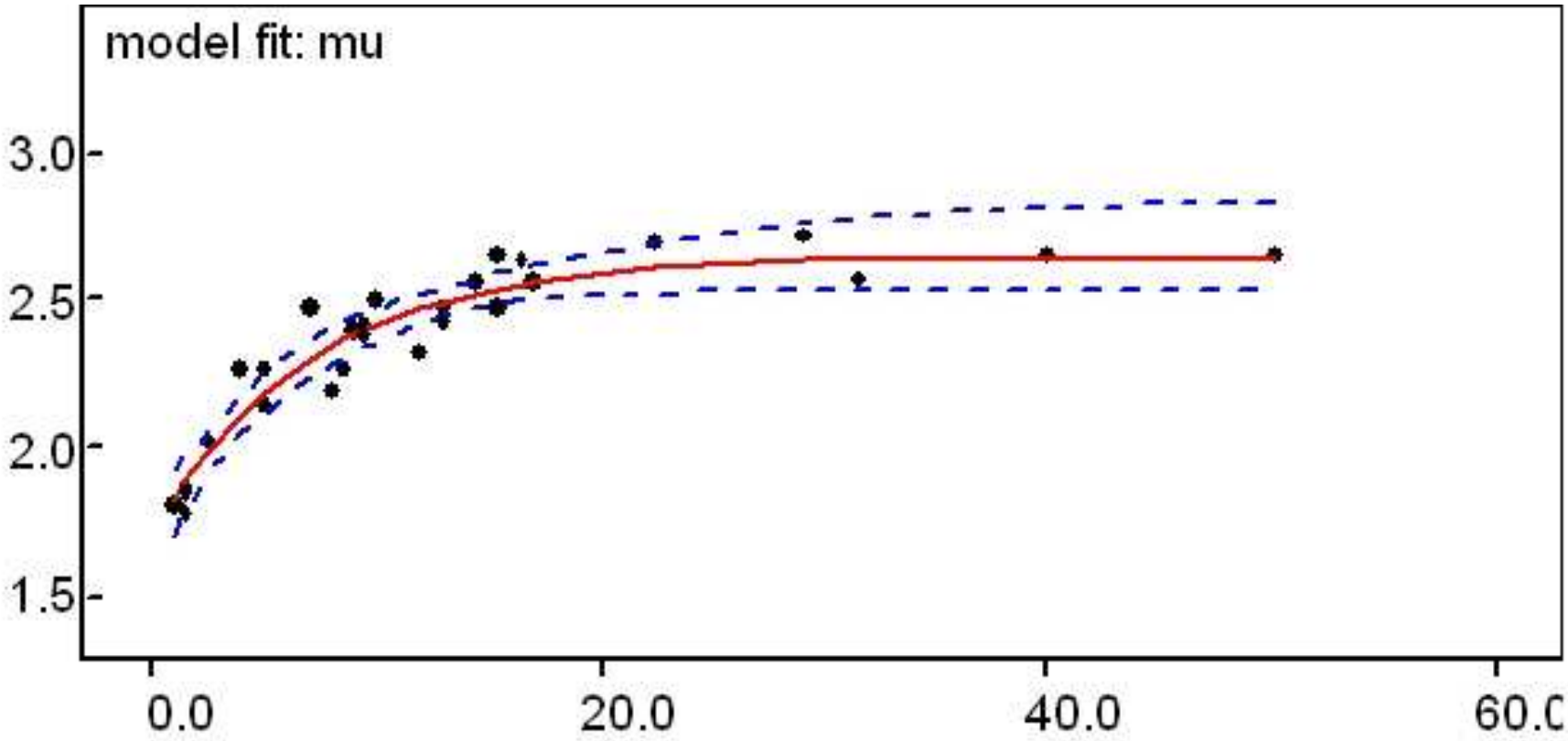
Dugongs: prediction as missing data

Easier to set up as missing data - WinBUGS automatically predicts it

```
list(x = c( 1.0,  1.5,  1.5,  1.5, 2.5,   4.0,  5.0,  5.0,  7.0,
              8.0,  8.5,  9.0,  9.5, 9.5,  10.0, 12.0, 12.0, 13.0,
             13.0, 14.5, 15.5, 15.5, 16.5, 17.0, 22.5, 29.0, 31.5, 35, 40),
       Y = c(1.80, 1.85, 1.87, 1.77, 2.02, 2.27, 2.15, 2.26, 2.47,
             2.19, 2.26, 2.40, 2.39, 2.41, 2.50, 2.32, 2.32, 2.43,
             2.47, 2.56, 2.65, 2.47, 2.64, 2.56, 2.70, NA, NA), N = 29)
```

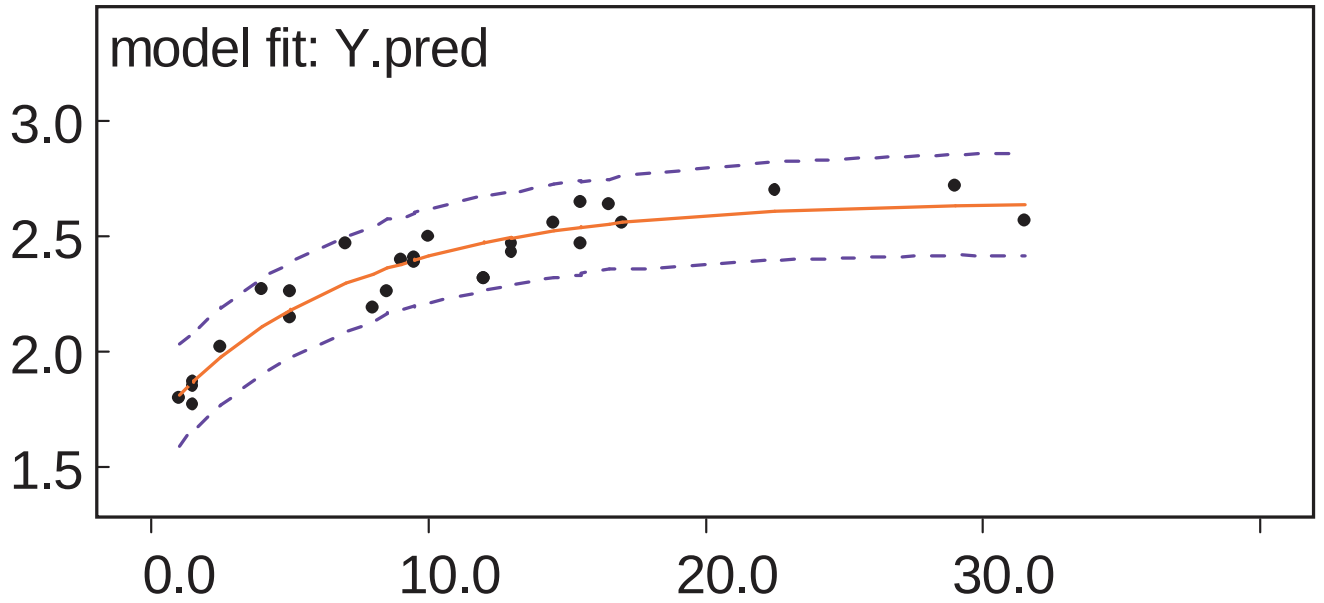
node	mean	sd	MC error	2.5%	median	97.5%	start	sample
mu[28]	2.651	0.07189	0.00423	2.533	2.642	2.815	1001	10000
Y[28]	2.651	0.1228	0.004537	2.415	2.648	2.902	1 001	10000
mu[29]	2.655	0.07825	0.004772	2.533	2.644	2.837	1001	10000
Y[29]	2.653	0.1275	0.005026	2.413	2.649	2.921	1001	10000

Dugongs: projections



Dugongs: prediction as model checking

```
y.pred[i]   ~   dnorm( mu[i], inv.sigma2 )
```



Criticism of 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 in non-hierarchical models

- Standardised Pearson residuals

$$(y - \theta)/\sigma$$

  where $\theta = \mathsf{E}[Y]$, $\sigma^2 = \mathsf{V}[Y]$

- In Bayesian analysis these are random quantities, with distributions

- If assuming Normality, then

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

  has a Uniform[0,1] distribution under true θ and σ

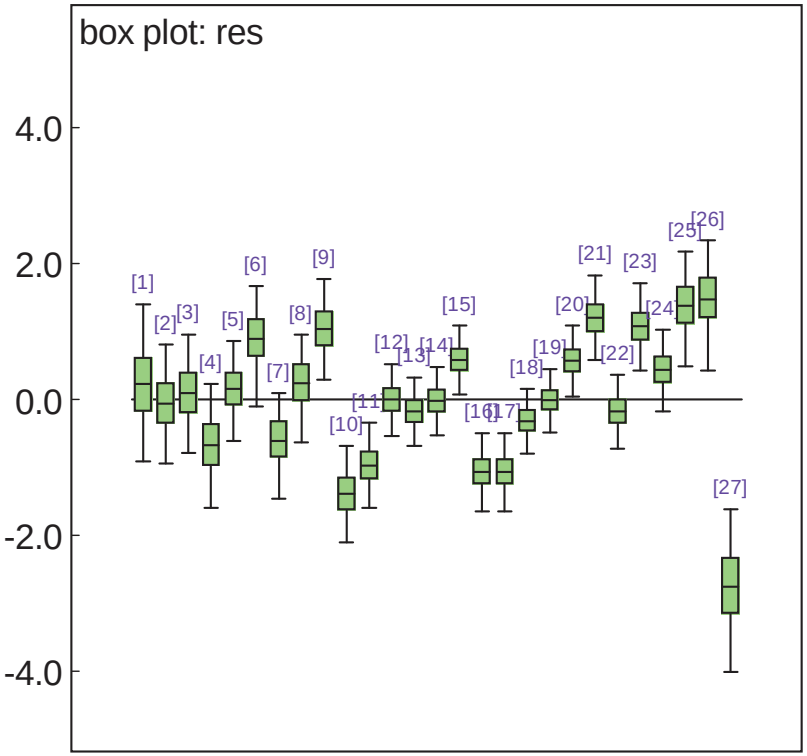
Example: Dugongs — non-linear models

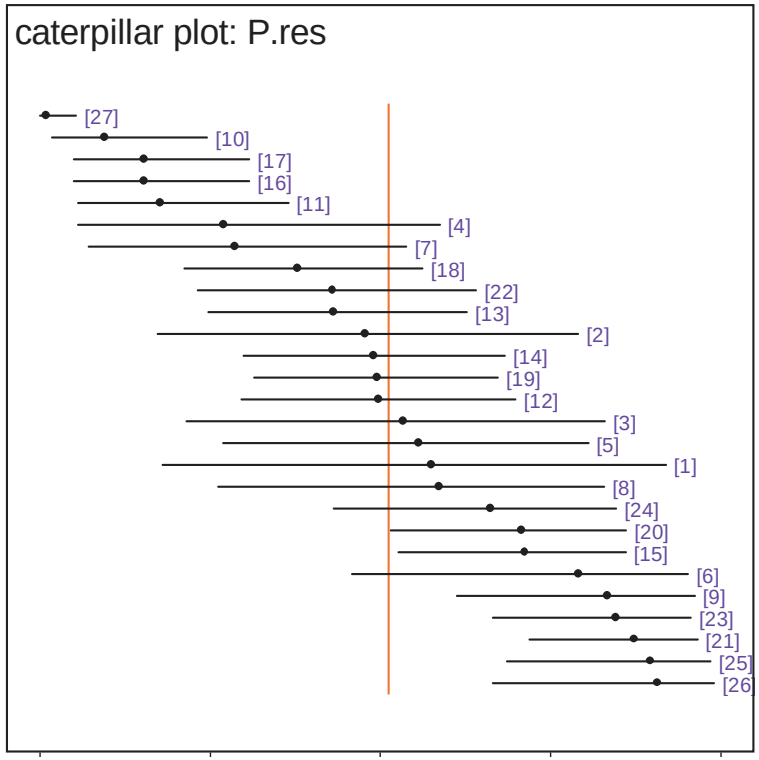
$$y_i \sim \mathsf{N}[\alpha - \beta\gamma^{x_i}, \sigma^2]$$

```
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
  p.res[i] <- phi(res[i])
}
alpha       ~ dunif(0,100)
beta        ~ dunif(0,100)
gamma       ~ dunif(0, 1)
log.sigma   ~ dunif(-10,10)   #  uniform prior on log(standard deviation)
log(sigma) <- log.sigma
inv.sigma2 <- 1/(sigma*sigma) # inv.sigma2 is precision
```

- Can monitor standardised residuals, `res`, and their P-values, `p.res`

Dugongs: residuals





Comments on residuals

- Could use $X^2 = \Sigma_i r_i^2$ as overall measure of residual variation
- Gelman et al suggest plotting single draw in Q-Q plot.
- Can also calculate deviance residuals.
- Note: not independent, so best used informally
- Multivariate version: Mahalanobis distance

$$M_i = (\boldsymbol{y}_i - \mathbb{E}(\boldsymbol{y}_i))' \mathsf{Var}(\boldsymbol{y}_i)^{-1} (\boldsymbol{y}_i - \mathbb{E}(\boldsymbol{y}_i))$$

- If want single value for testing distributional shape, could plug-in posterior means, or use residual posterior means, but better to use approximate **predictive** residuals calculated outside BUGS:

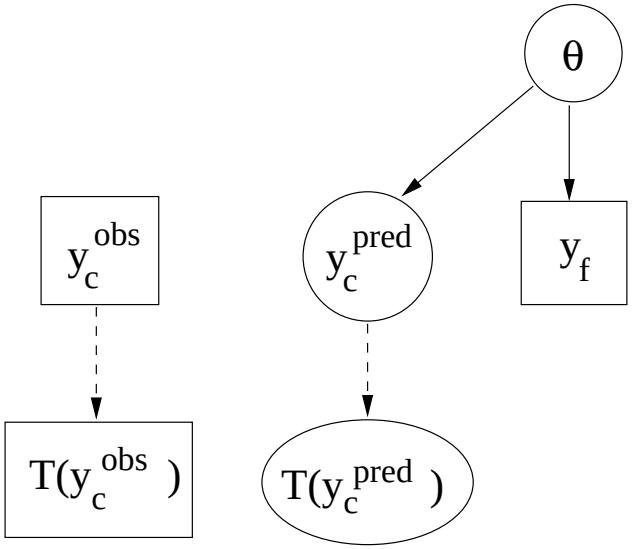
$$r_i = \frac{y_i - \mathsf{E}(Y_i^{pred})}{\sqrt{\mathsf{V}(Y_i^{pred})}}$$

  Multivariate version is

$$M_i = \left(\boldsymbol{y}_i - \mathbb{E}(\boldsymbol{y}_i^{pred})\right)' \mathsf{Var}(\boldsymbol{y}_i^{pred})^{-1} \left(\boldsymbol{y}_i - \mathbb{E}(\boldsymbol{y}_i^{pred})\right)$$

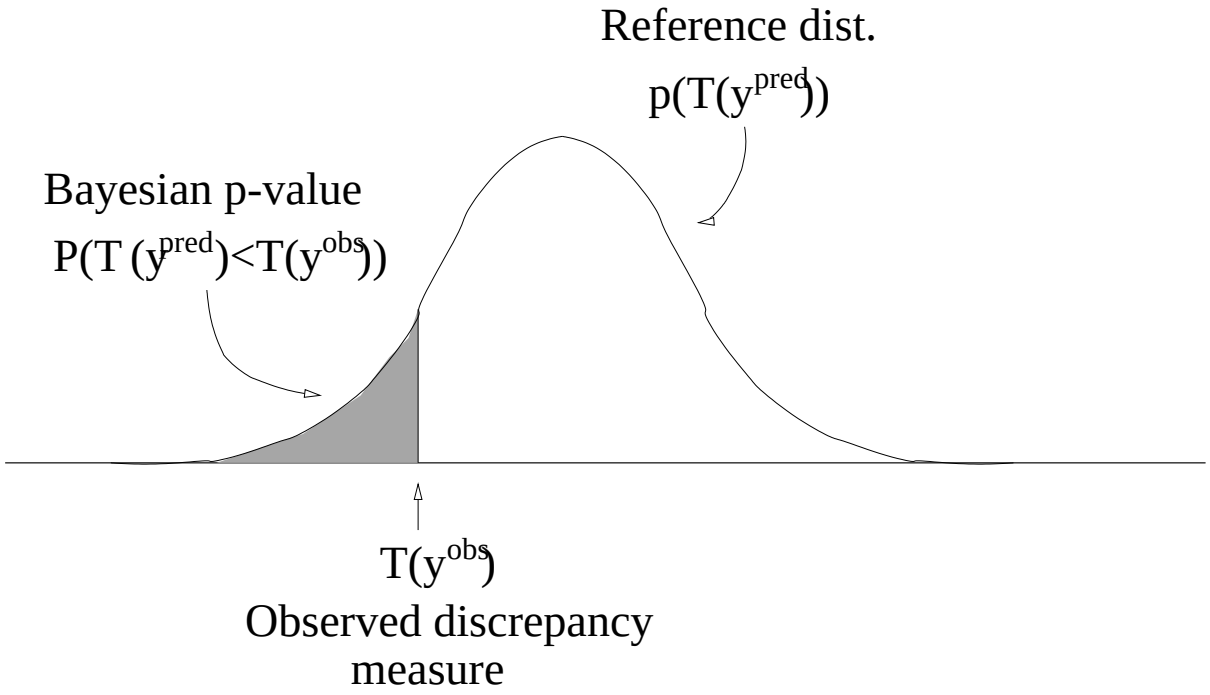
Cross-validation

- Compare observed data (or function of data) with predicted values under the H_0 model
- Assessing *conflict* between the observed data and their predictive distribution.
- Split data into y_f used to fit the model; y_c for criticism.



- $T(y_c)$ termed *discrepancy statistic* by Gelman et al (1995a,b) such that it has extreme value if data conflict with the model
- common choice is just $T(y_{ci}) = y_{ci}$ to check for individual outliers

Bayesian p-values



- (1-sided) probability that predicted data could be more extreme than observed, as measured by the discrepancy statistic $T(y_c)$
- Has Unif[0,1] distribution under H_0
- Suffers from usual problems of multiple testing

Comments on Cross-validation

- suffers from usual problems of masking
- does not target specific model assumptions
- tiresome to implement using MCMC

Alternatives to cross-validation

- importance resampling
- approximate $p(Y_i^{pred}|y_{\setminus i})$ by $p(Y_i^{pred}|y)$
  - just replicate data ('posterior predictive P-values')
  - replicate parameters *and* data ('mixed predictive P-values')
- can end up with many P-values - use ideas of False Discovery Rates

Model comparison

What is the 'deviance'?

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

$$D(\theta) = -2 \log p(y|\theta) \tag{1}$$

- In WinBUGS the quantity `deviance` is automatically calculated, where θ are the parameters that appear in the stated sampling distribution of y

- The full normalising constants for $p(y|\theta)$ are included in `deviance`

- *e.g.* for Binomial data `y[i] ~ dbin(theta[i],n[i])`, the `deviance` is

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

What is the 'standardised deviance'?

- In generalised linear models the saturated deviance is (loosely) defined as $D(y)$ - the deviance with the observations substituted for their expectations
- We define the standardised deviance as $D(\theta) - D(y)$
- *e.g.* for Binomial data, Bayesian standardised deviance is

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

- Just sum of deviance residuals
- This is a random quantity with a posterior distribution
- If model fits the data, expected to have χ^2_I distribution, where I is the dimensionality of θ.
- Can be used as absolute measure of fit
- In WinBUGS you currently need to calculate it yourself

Use of mean deviance as measure of fit

- Dempster (1974) suggested plotting posterior distribution of deviance
- Many authors suggested using posterior mean deviance $\overline{D} = \mathbb{E}[D]$ as a measure of fit
- Invariant to parameterisation of θ
- Robust, generally converges well
- But more complex models will fit the data better and so will have smaller $\overline{D}$
- Need to have some measure of 'model complexity' to trade off against $\overline{D}$

Bayesian model comparison using DIC

- Natural way to compare models is to use criterion based on trade-off between the fit of the data to the model and the corresponding complexity of the model
- Spiegelhalter et al (2002) proposed a Bayesian model comparison criterion based on this principle:

  Deviance Information Criterion, DIC = 'goodness of fit' + 'complexity'

- They measure fit via the deviance

$$D(\theta) = -2 \log L(\text{data}|\theta)$$

- Complexity measured by estimate of the 'effective number of parameters':

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

  i.e. posterior mean deviance minus deviance evaluated at the posterior mean of the parameters

- The DIC is then defined analagously to AIC as

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

  Models with smaller DIC are better supported by the data

- DIC can be monitored in `WinBUGS` from `Inference/DIC` menu

- 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.
- Not a function of the marginal likelihood of the data, so *not* aiming for Bayes factor behaviour.
- Do not believe there is any 'true' model.
- p_D is not invariant to reparameterisation.
- p_D can be be negative! (not desirable)
- Alternative to p_D suggested

Could DIC be improved?

- It would be better if WinBUGS used the posterior mean of the 'direct parameters' (eg those that appear in the WinBUGS distribution syntax) to give a 'plug-in' deviance, rather than the posterior means of the stochastic parents.

- Users are free to calculate this themselves: could dump out posterior means of 'direct' parameters in likelihood, then calculate deviance outside WinBUGS or by reading posterior means in as data and checking `deviance` in `node info`

- Lesson: need to be careful with highly non-linear models, where posterior means may not lead to good predictive estimates

- Same problem arises with mixture models

DIC is allowed to be negative - not a problem!

- A probability density $p(y|\theta)$ can be greater than 1 if has a small standard deviation
- Hence a deviance can be negative, and a DIC negative
- Only *differences* in DIC are important: its absolute size is irrelevant
- Suppose observe data (-0.01, 0.01)
- Unknown mean (uniform prior), want to choose between three models with $\sigma = 0.001, 0.01, 0.1$.

	Dbar	Dhat	pD	DIC
y1	177.005	176.046	0.959	177.964
y2	-11.780	-12.740	0.961	-10.819
y3	-4.423	-5.513	1.090	-3.332

- Each correctly estimates the number of unknown parameters.
- The middle model ($\sigma = 0.01$) has the smallest DIC, which is negative.

Why won't DIC work with mixture likelihoods?

- WinBUGS currently 'greys out' DIC if the likelihood depends on any discrete parameters
- So cannot be used for mixture likelihoods
- Not clear what estimate to plug in for class membership indicator – mode?
- If mixture is represented marginally (ie not using an explicit indicator for class membership), could use $\overline{\theta}$ but could be taking mean of bimodal distribution and get poor estimate
- Celeux et al (2003) have made many suggestions
- Can still be used if prior (random effects) is a mixture

More on missing data in WinBUGS

Missing response data, assuming missing data mechanism is ignorable

- denote missing observations 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

Ignorable missing response data is essentially a prediction problem — see earlier dugongs example

If missing data mechanism is informative

- need explicit model for missing data mechanism
- usually need informative priors on parameters of missing data model as no information in the data
- See Best *et al.* (1996) for one example

Missing covariate data

- denote missing observations by `NA` in the data file
- specify prior distribution for the covariate
  - e.g. if X is a continuous covariate containing some missing values, could specify $X_i \sim \text{Normal}(\mu, \sigma^2)$ or build regression model relating X_i to other observed covariates
  - can then assume vague priors for μ and σ^2; posterior distribution of μ and σ^2 will be informed by the observed part of the vector of X's
- WinBUGS will automatically simulate values from the posterior distribution of the missing covariates (which will depend on the prior for the X's and the likelihood contribution from the corresponding response variable)

Example: Childhood malaria in the Gambia

Diggle et al (2002)

Data:

- 2035 children in 65 villages in the Gambia
- Response: Binary indicator of presence of malarial parasites in blood sample taken from each child
- Covariates include: child's age and use of bed nets, inclusion/exclusion of village from primary health care system and greenness of surrounding vegetation (from satellite information)

Questions of interest include:

- Does sleeping under a bed net reduce risk of malaria?

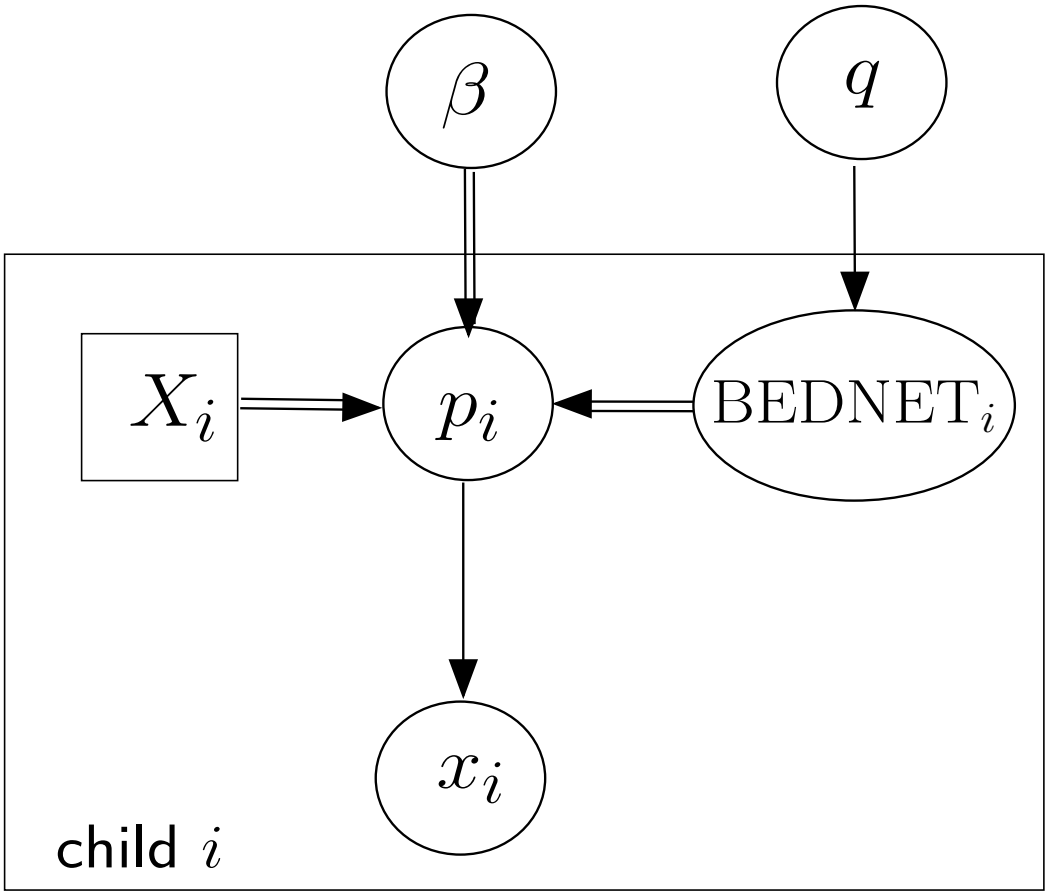
- Here we consider a slightly modified version of Diggle et al's dataset:
  - BEDNET = binary indicator of whether child sleeps under a (treated) bed net
  - Suppose the value of BEDNET is missing for 30% of children
- Consider 2 alternative models for the missing covariate:
  1. Probability of BEDNET = 1 is same for all children *a priori*

$$\begin{aligned} \text{BEDNET}_i &\sim \text{Bernoulli}(q) \\ q &\sim \text{Beta}(1,1) \end{aligned}$$

  2. Probability of BEDNET = 1 depends on whether or not village belongs to primary health care system (PHC)

$$\begin{aligned} \text{BEDNET}_i &\sim \text{Bernoulli}(q_i) \\ \text{logit}q_i &= \gamma_1 + \gamma_2 \text{PHC}_i; \quad (+ \text{ vague priors on } \gamma_1 \text{ and } \gamma_2) \end{aligned}$$

DAG for Model 1



WinBUGS code for model 1

```
model {
  for(i in 1:2035) {
    Y[i] ~ dbern(p[i])
    logit(p[i]) <- alpha + beta.age[AGE[i]] + beta.bednet*BEDNET[i] +
                   beta.green*(GREEN[i] - mean(GREEN[])) + beta.phc*PHC[i]
  }
  # model for missing exposure variable
  for(i in 1:2035) { BEDNET[i] ~ dbern(q)  } # prior model for whether or not child
                                             # i sleeps under treated bednet
  q ~ dbeta(1, 1) # vague prior (uniform) on prob of sleeping under treated bednet

  # vague priors on regression coefficients
  alpha ~ dflat()
  beta.bednet ~ dflat()
  ........etc.......

  # calculate odds ratios of interest
  OR.bednet <- exp(beta.bednet)      # odds ratio of malaria for children using
                                     # treated bednets
  PP.bednet <- step(0.8 - OR.bednet) # probability that using treated bed net
                                     # reduces risk of malaria by at least 20%
}
```

WinBUGS code for model 2

- Replace model for missing exposure variable by

```
# model for missing exposure variable
for(i in 1:2035) {
   BEDNET[i] ~ dbern(q[i])  # prior model for whether or not child i
                            # sleeps under treated bednet
   logit(q[i]) <- gamma[1] + gamma[2]*PHC[i] # allow prob of using treated
                                             # bednet to depend on whether
                                             # or not village belongs to
                                             # primary health care system
}
for(k in 1:2) {   gamma[k] ~ dflat()     }
OR.treated.phc <- exp(gamma[2])              # odds ratio of sleeping under
                                             # treated bednet for children
                                             # living in villages in the PHC
```

Results

	`OR.bednet`		`PP.bednet`	`OR.age2`	
	Mean	95% interval		Mean	95% interval
No missing data	0.57	(0.45, 0.72)	0.99	1.40	(1.06, 1.81)
Model 1	0.66	(0.49, 0.86)	0.93	1.39	(1.06, 1.80)
Model 2	0.64	(0.47, 0.83)	0.95	1.41	(1.06, 1.83)
Single imputation*	0.76	(0.61, 0.95)	0.68	1.40	(1.05, 1.80)
Complete case (exclude all cases with missing data)	0.63	(0.47, 0.83)	0.96	1.70	(1.20, 2.35)

*Imputed using observed proportion of bed net users

Lecture 7.

Introduction to hierarchical models

Often interested in making inferences on 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 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 (unless very large sample sizes)

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

- Recall conjugate normal-normal model for the THM example in lecture 2
- Full data includes THM measurements for 70 water supply zones
- We assume a normal likelihood for the data in each zone

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

- Notice that we now have 70 distinct mean parameters θ_z
- What prior should we specify for each θ_z?
- Note that we now also take $\sigma^2_{[e]}$ to be unknown and assume a vague prior $1/\sigma^2_{[e]} \sim \text{Gamma}(0.001, 0.001)$

Identical parameters

- Assume that the mean THM levels are the same in all zones, $\theta_z = \theta \; \forall z$

- 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)

  $\rightarrow$ conjugate normal-normal model discussed in Lecture 2

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

Independent parameters

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

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

- This will give posterior estimates $E(\theta_z|\boldsymbol{x}_z) \approx \overline{x}_z$ (the raw 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

  $\rightarrow$ 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:

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

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

$\Rightarrow$ assign prior distributions to μ and $\sigma^2_{[z]}$, e.g.

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

$\Rightarrow$ *joint prior distribution* for the entire set of parameters

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

Then apply Bayes theorem as usual to simultaneously estimate joint posterior distribution of all the unknown quantities:

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

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]

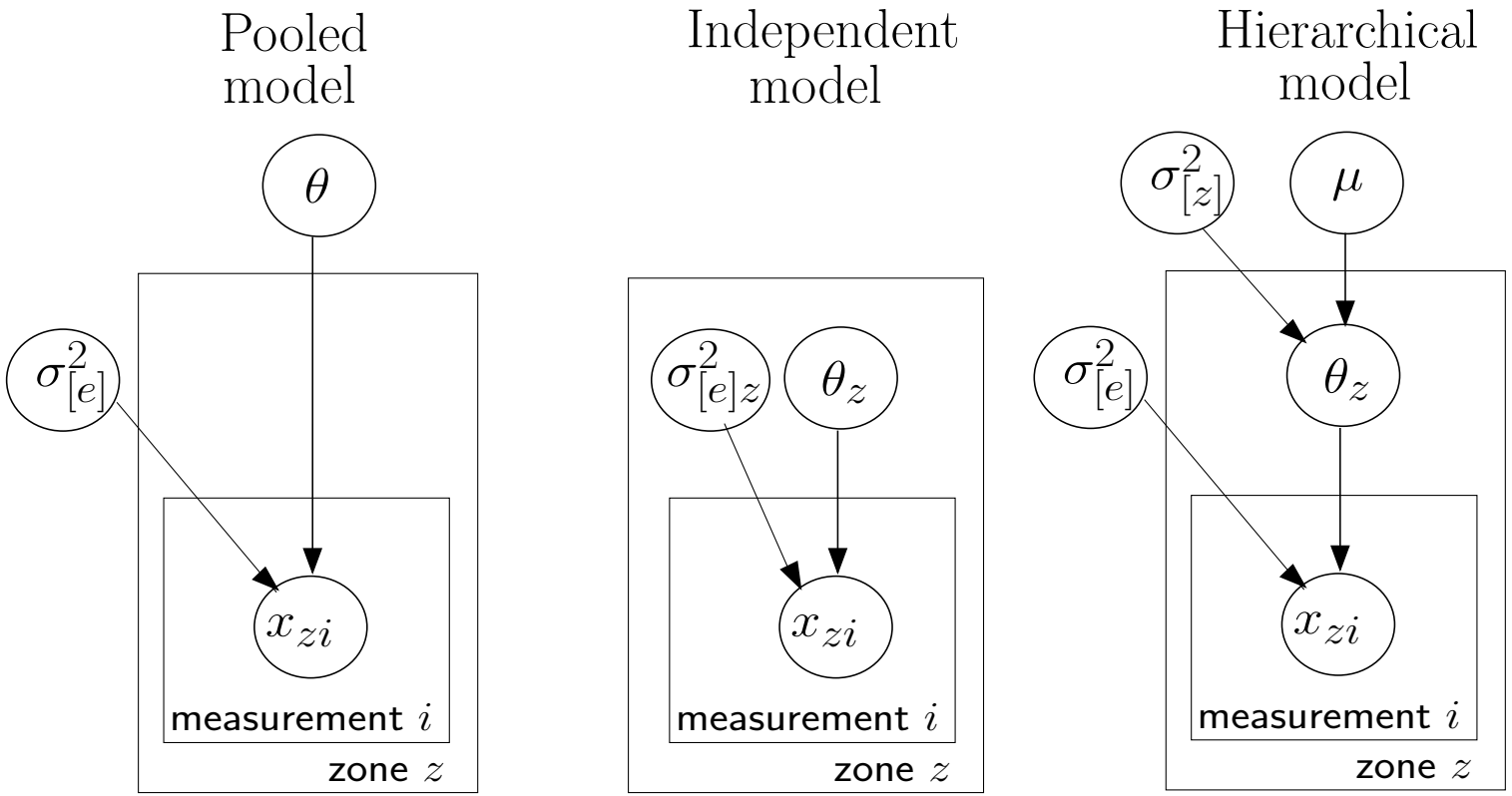
Advantages of this approach:

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 population (prior) parameters μ and $\sigma^2_{[z]}$
- leads to *global smoothing* of the zone mean THM levels
- reflects our full uncertainty about the true values of μ and $\sigma^2_{[z]}$

Such models are called *Hierarchical* or *Random effects* or *Multilevel* models

Graphical representation of THM models

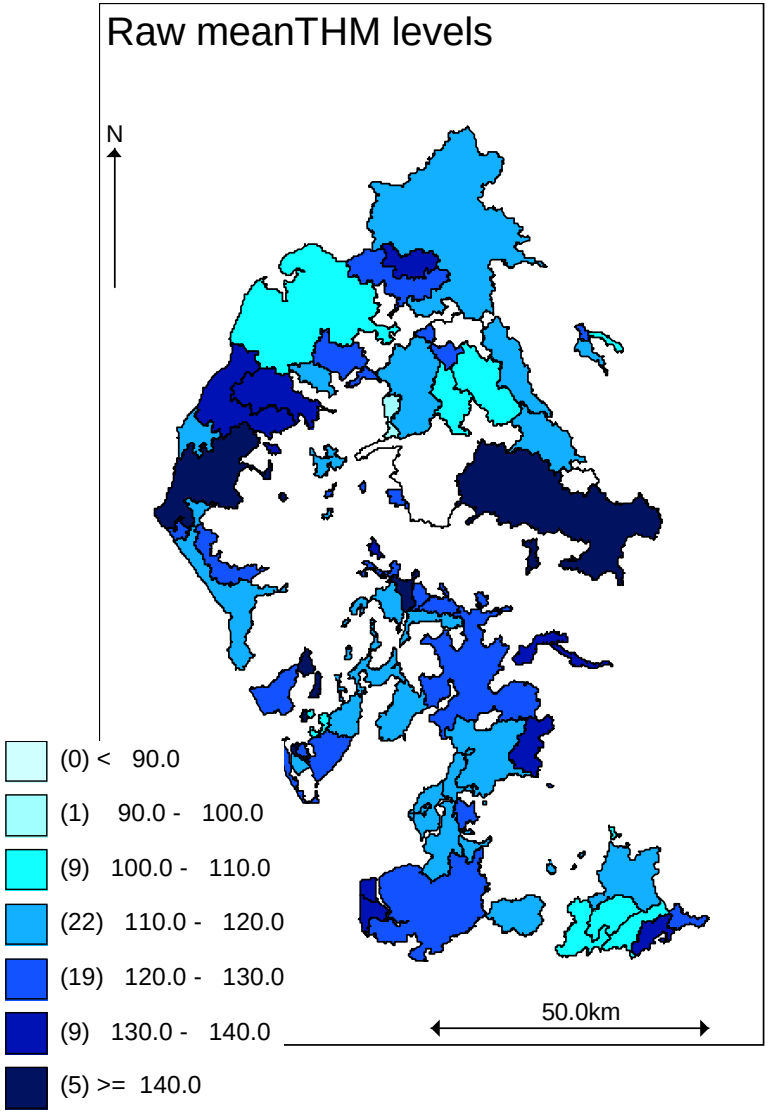


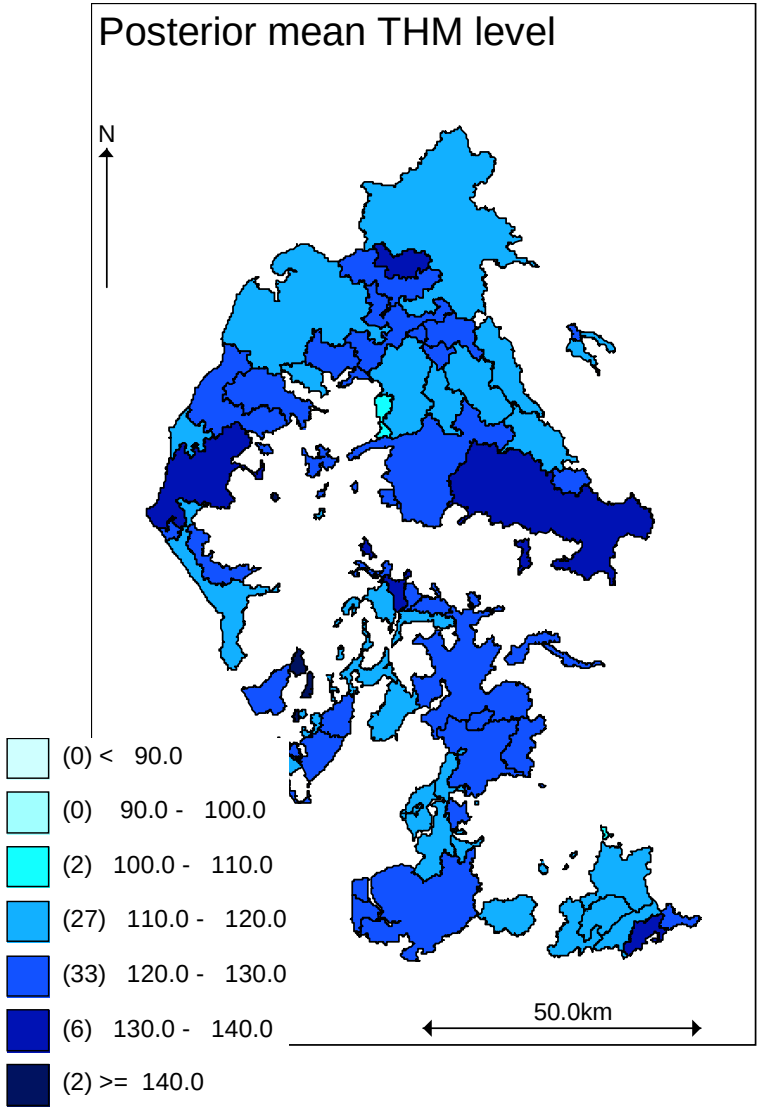
Parameter Interpretation for hierarchical model

- 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)

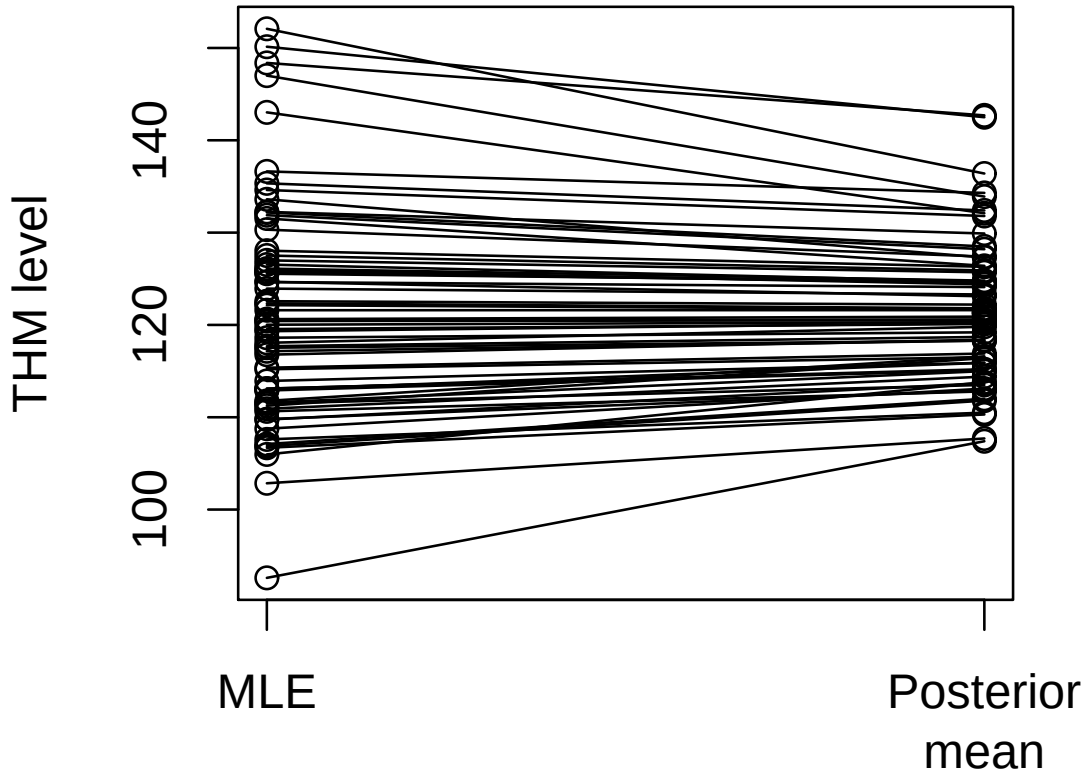
Note: could elaborate model to allow zone-specific residual error variance $\sigma^2_{[e]z}$ with a hierarchical prior:

$$\begin{aligned} x_{zi} &\sim \text{Normal}(\theta_z, \sigma^2_{[e]z}), \quad i = 1, ...n_z \\ \log\left(\sigma^2_{[e]z}\right) &\sim \text{Normal}(v, \phi^2) \\ v &\sim \text{Uniform}(-100, 100) \\ 1/\phi^2 &\sim \text{Gamma}(0.001, 0.001) \end{aligned}$$

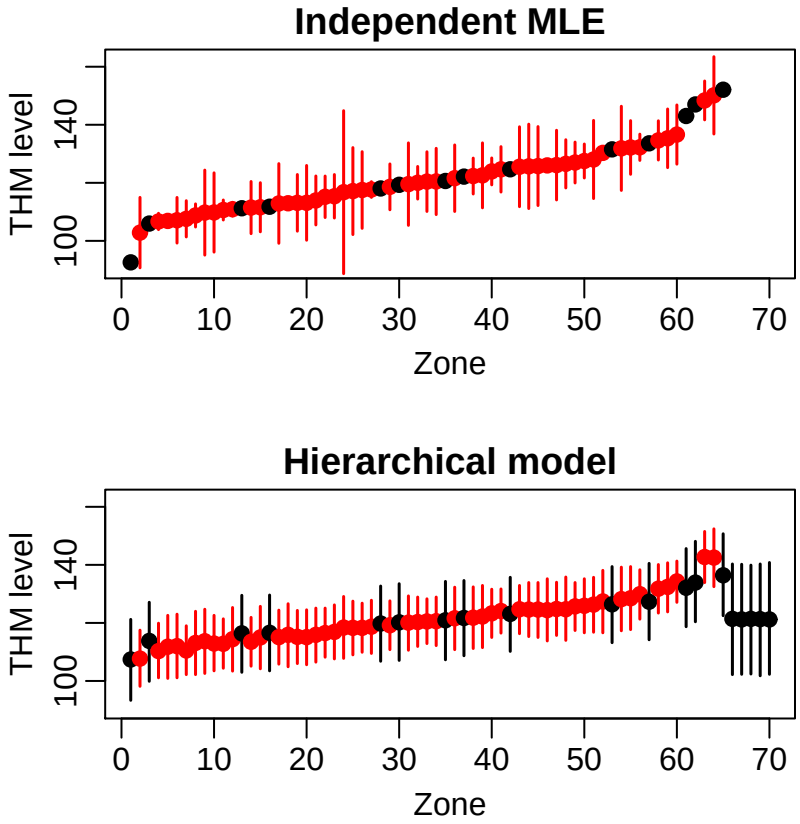

Raw meanTHM levels
N
(0) < 90.0
(1) 90.0 - 100.0
(9) 100.0 - 110.0
(22) 110.0 - 120.0
(19) 120.0 - 130.0
(9) 130.0 - 140.0
(5) >= 140.0
50.0km


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

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

General form for hierarchical model

In general, suppose we have data $\boldsymbol{x}$ and parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$

- Likelihood $p(\boldsymbol{x}|\boldsymbol{\theta})$ models structure of the observables
- Prior $p(\boldsymbol{\theta})$ is decomposed into conditional distributions $p(\boldsymbol{\theta}|\phi_2) \times p(\phi_2|\phi_3) \times \ldots \times p(\phi_m)$

The marginal prior distribution for $\boldsymbol{\theta}$ is then

$$p(\boldsymbol{\theta}) \;=\; \int p(\boldsymbol{\theta}|\phi_2) \;\times\; p(\phi_2|\phi_3) \;\times \times\, p(\phi_{m-1}|\phi_m) \;\times\; p(\phi_m)\, d\phi_2\, d\phi_3 \ldots d\phi_m$$

- ϕ_k are called *hyperparameters* of level k and are introduced to simplify prior specification.

- The conditional prior distributions $p(\phi_{k-1}|\phi_k)$ express structural judgements (e.g. exchangeability, spatial correlation...)

- Theoretically there can be as many levels as necessary, but in practice it is usually hard to interpret parameters of level 3 or higher

- A non-informative prior is usually specified for the marginal distribution of the top-level parameters

Implementation of THM hierarchical model in WinBUGS

Data contain between 0 and 6 observations per zone → '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)
 }

# priors 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
```

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, 107.7, 124.7, .....),
     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)
}

# priors 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
```

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, 107.7, 124.7,...),
     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
124.6    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)

}

# priors 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
```

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,
                      124.7,    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
  111.6    112.5     98.6    107.7       NA        NA
     NA       NA       NA       NA       NA        NA
  124.7       NA       NA       NA       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

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

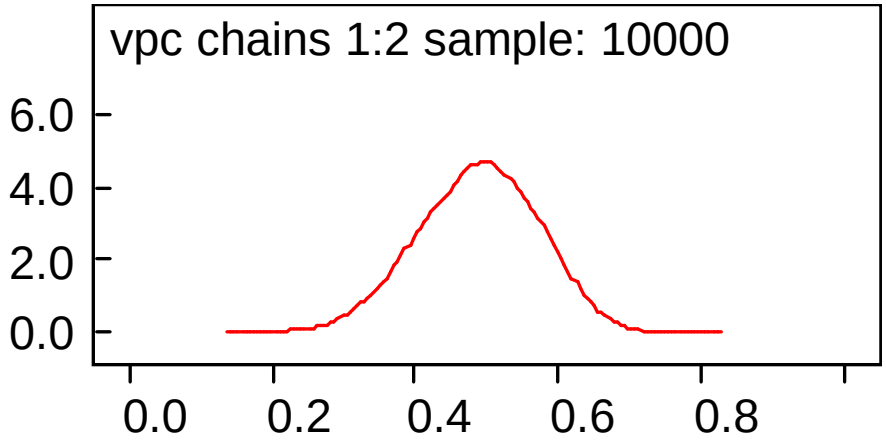
  - $\sigma^2_{[e]}$ is the 'level 1' variance (i.e. variance of Normal likelihood)
  - $\sigma^2_{[z]}$ is the 'level 2' variance (i.e. random effects variance)

- 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

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

Hierarchical centering

- Above formulation of THM model is **hierarchically centered**
  - random effect is centered around overall mean

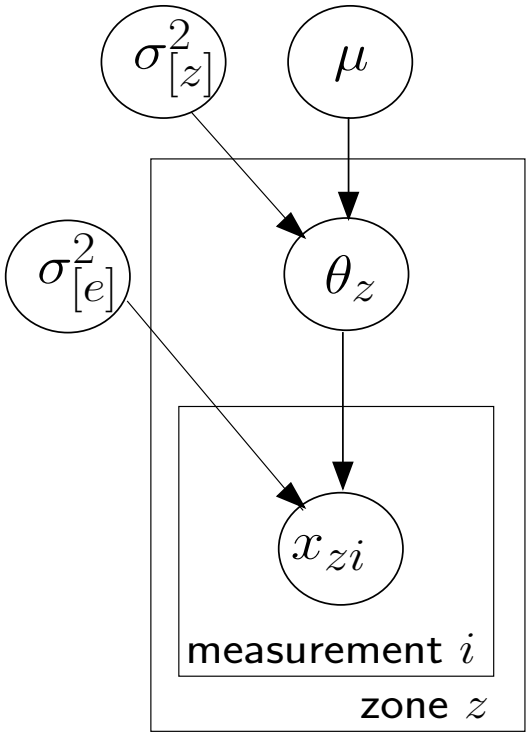
$$\theta_z \sim \text{Normal}(\mu, \sigma^2_{[z]})$$

- Alternatively, could use **non-centered** parameterisation
  - random effect is a priori independent of overall mean

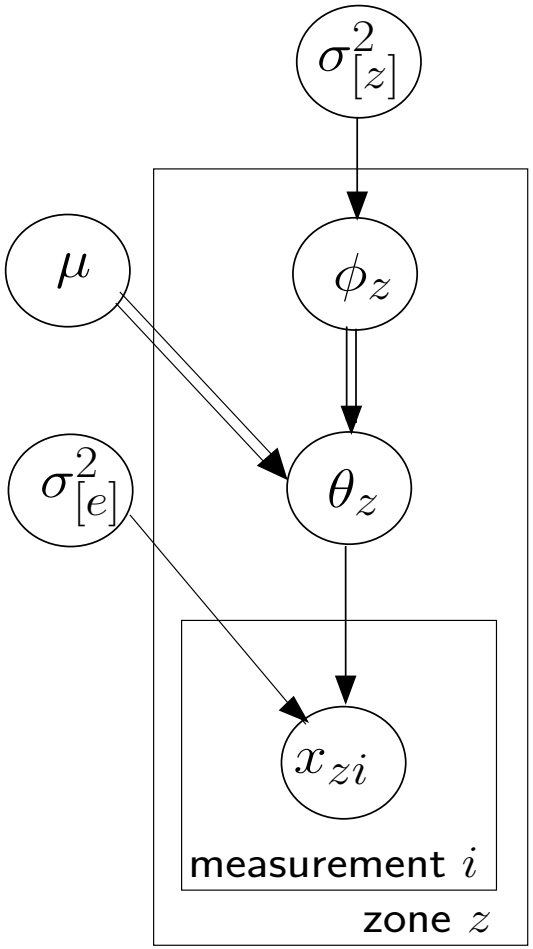
$$\begin{aligned} \theta_z &= \mu + \phi_z \\ \phi_z &\sim \text{Normal}(\mu, \sigma^2_{[z]}) \end{aligned}$$

- Choice of parameterisation can have big impact on mixing of MCMC chain
- Depends on relative size of variances at level 1 ($\sigma^2_{[e]}$) and level 2 ($\sigma^2_{[z]}$) — see Gelfand, Sahu and Carlin (1995)
  - If VPC large ($\sigma^2_{[z]} >> \sigma^2_{[e]}$), hierarchical centering more efficient
  - If VPC small ($\sigma^2_{[z]} << \sigma^2_{[e]}$), non-centered parameterisation more efficient
- The following results are based on simulating data for THM example with different VPCs

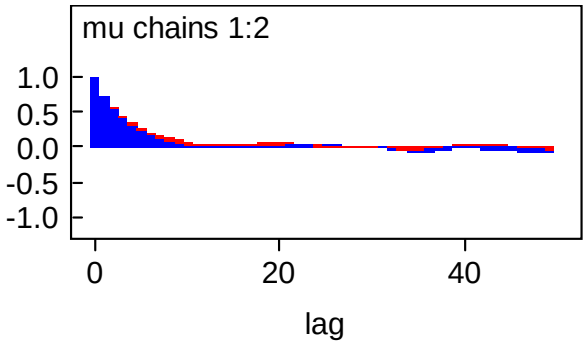
Hierarchically centered model



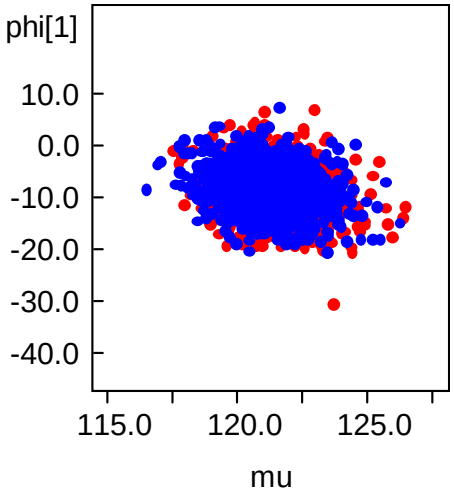
Non centered model



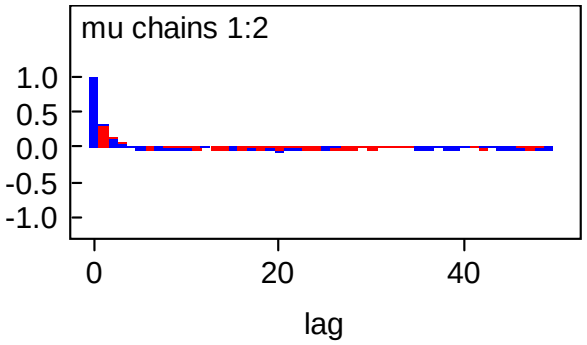
Non hierarchically centered



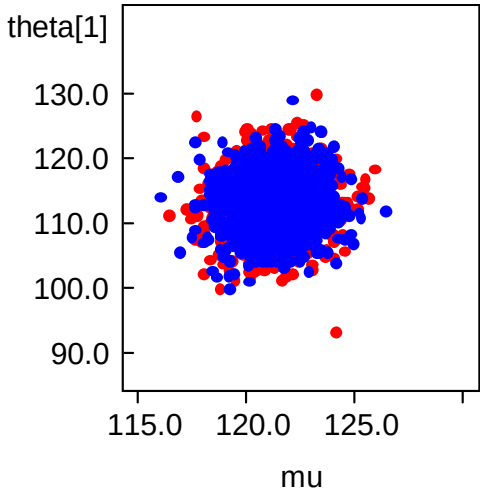
mean	sd	MC error
121.3	1.42	0.039



Hierarchically centered

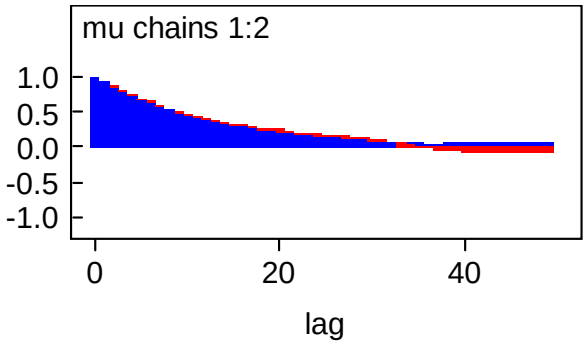


mean	sd	MC error
121.3	1.41	0.017

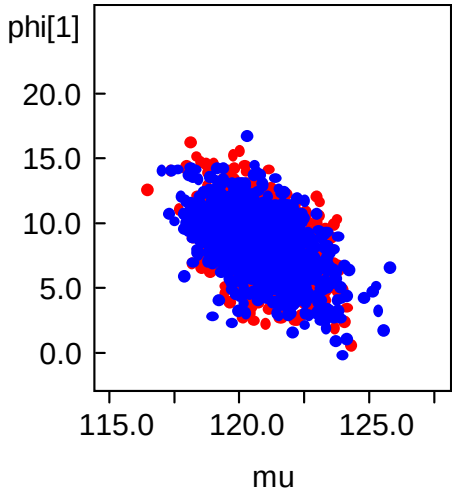


VPC = 0.49; 95% interval (0.32, 0.64)

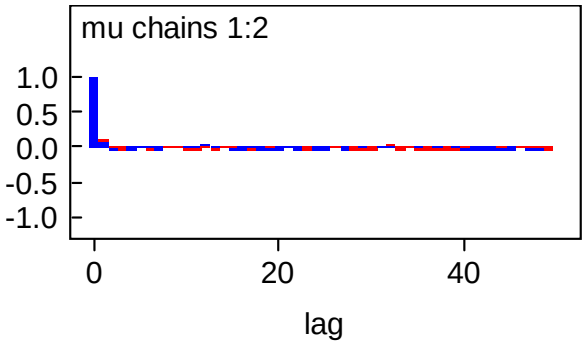
Non hierarchically centered



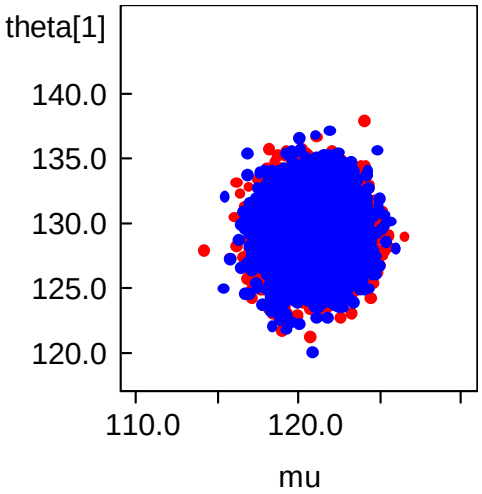
mean	sd	MC error
120.9	1.37	0.060



Hierarchically centered

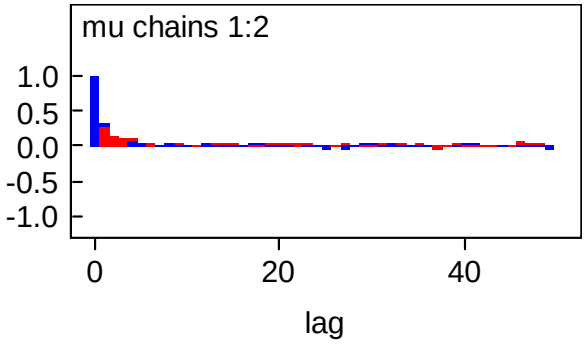


mean	sd	MC error
120.8	1.43	0.014

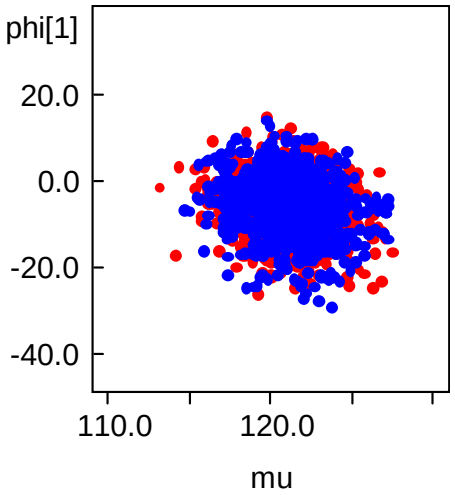


VPC = 0.83; 95% interval (0.75, 0.89)

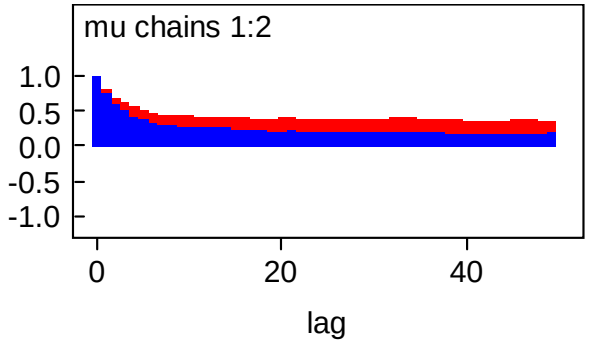
Non hierarchically centered



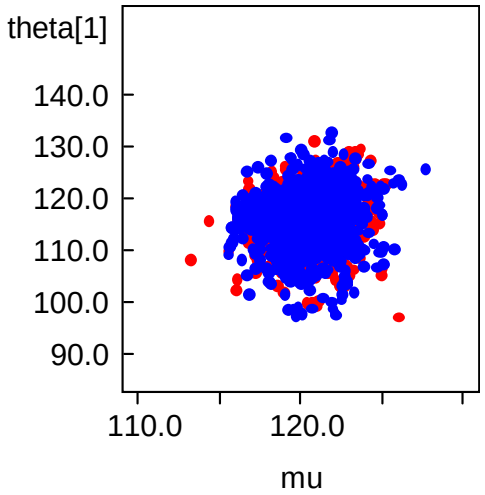
mean	sd	MC error
120.8	2.01	0.039



Hierarchically centered



mean	sd	MC error
120.3	2.06	0.102



VPC = 0.10; 95% interval (0.01, 0.30)

Comparison with alternative estimation methods

Consider a general 2-level model

- Likelihood $p(\boldsymbol{x}|\boldsymbol{\theta})$
- Prior $p(\boldsymbol{\theta})$ decomposed into
  - $p(\boldsymbol{\theta}|\phi)$ (random effects distribution)
  - $p(\phi)$ (prior on hyperparameters)

Empirical Bayes Estimation

- In full hierarchical Bayes, ϕ is treated as unknown and assigned a prior distribution
  - Posterior distribution of random effects is

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \int p(\boldsymbol{\theta}, \phi|\boldsymbol{x}) d\phi$$

- In Empirical Bayes, a plug-in estimate $\widehat{\phi}$ is used and inference proceeds as in the non-hierarchical case with ϕ known
  - Posterior distibution of random effects is

$$p(\boldsymbol{\theta}|\boldsymbol{x}, \widehat{\phi})$$

  - Ignores uncertainty in hyperparameters ϕ
  - Point estimates usually similar to full Bayes, but interval estimates too narrow

- $\widehat{\phi}$ typically estimated using the *marginal* likelihood $p(\boldsymbol{x}|\phi) = \int p(\boldsymbol{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\phi) d\boldsymbol{\theta}$, either by maximization or method of moments

- EB uses the data twice (once to estimate prior parameters, then to estimate posterior)

IGLS/RIGLS Estimation

- (R)IGLS = (Restricted) Iterative Generalised Least Squares
- Provides ML estimates of fixed effects and variance components in Gaussian multi-level models
- As with EB, no prior distribution is assumed for ϕ, and estimation is essentially based on marginal likelihood $p(\boldsymbol{x}|\phi)$
- As with EB, ignoring full uncertainty in hyperparamters leads to overly precise estimates

Example: Meta-analysis of Magnesium following MI

(See example 3.13 in Spiegelhalter et al (2004))

Intervention: Intravenous magnesium sulphate may have a protective effect after acute myocardial infarction (AMI).

Data: Series of 8 small randomised trials estimating odds ratio for in-hospital mortality in magnesium vs. control groups

Here we take the reported log odds ratio, x_i, for each trial i as the 'data', and assume these are normally distributed with known sampling variance, s_i^2 = reported SE^2 of the log odds ratio

Analysis: Random effects meta-analysis

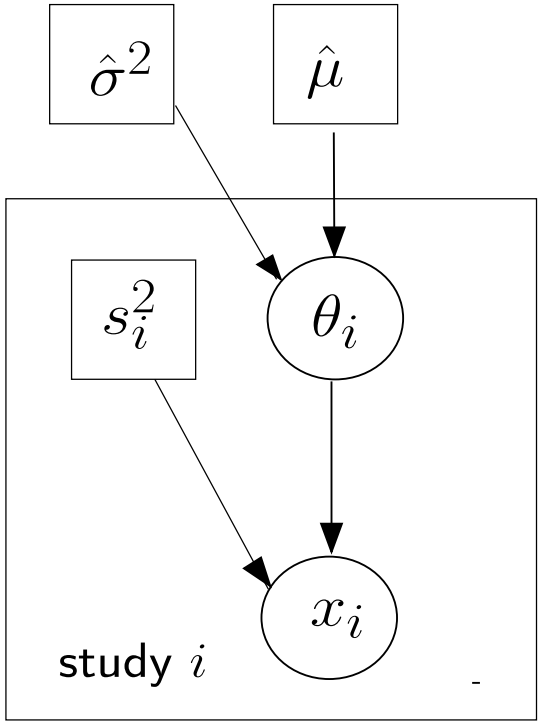
$$\begin{aligned} x_i &\sim \text{N}(\theta_i, s_i^2) \\ \theta_i &\sim \text{N}(\mu, \sigma^2) \end{aligned}$$

- Empirical Bayes: use 'plug-in' values of $\widehat{\mu} = \log(0.58)$ and $\widehat{\sigma} = 0.29$ based on method of moments estimator
- Full Bayes: assume diffuse priors

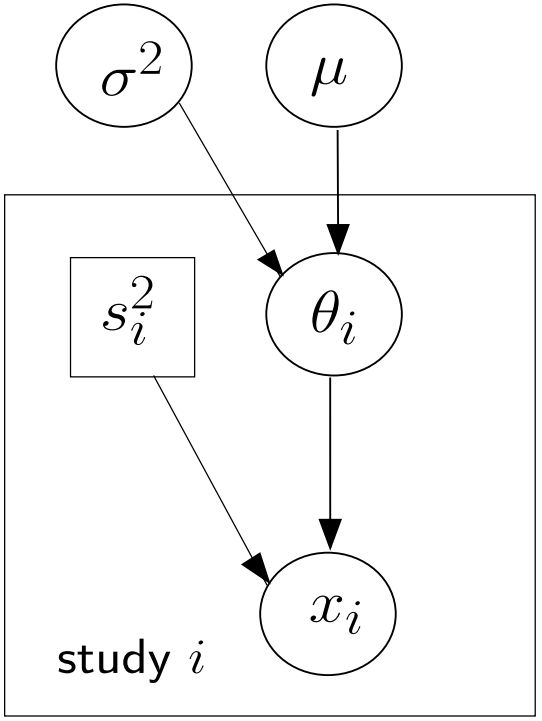
$$\begin{aligned} \mu &\sim \text{Unif}(-10, 10) \\ \sigma^{-2} &\sim \text{Gamma}(0.001, 0.001) \end{aligned}$$

Graphical models for the Magnesium meta-analysis example

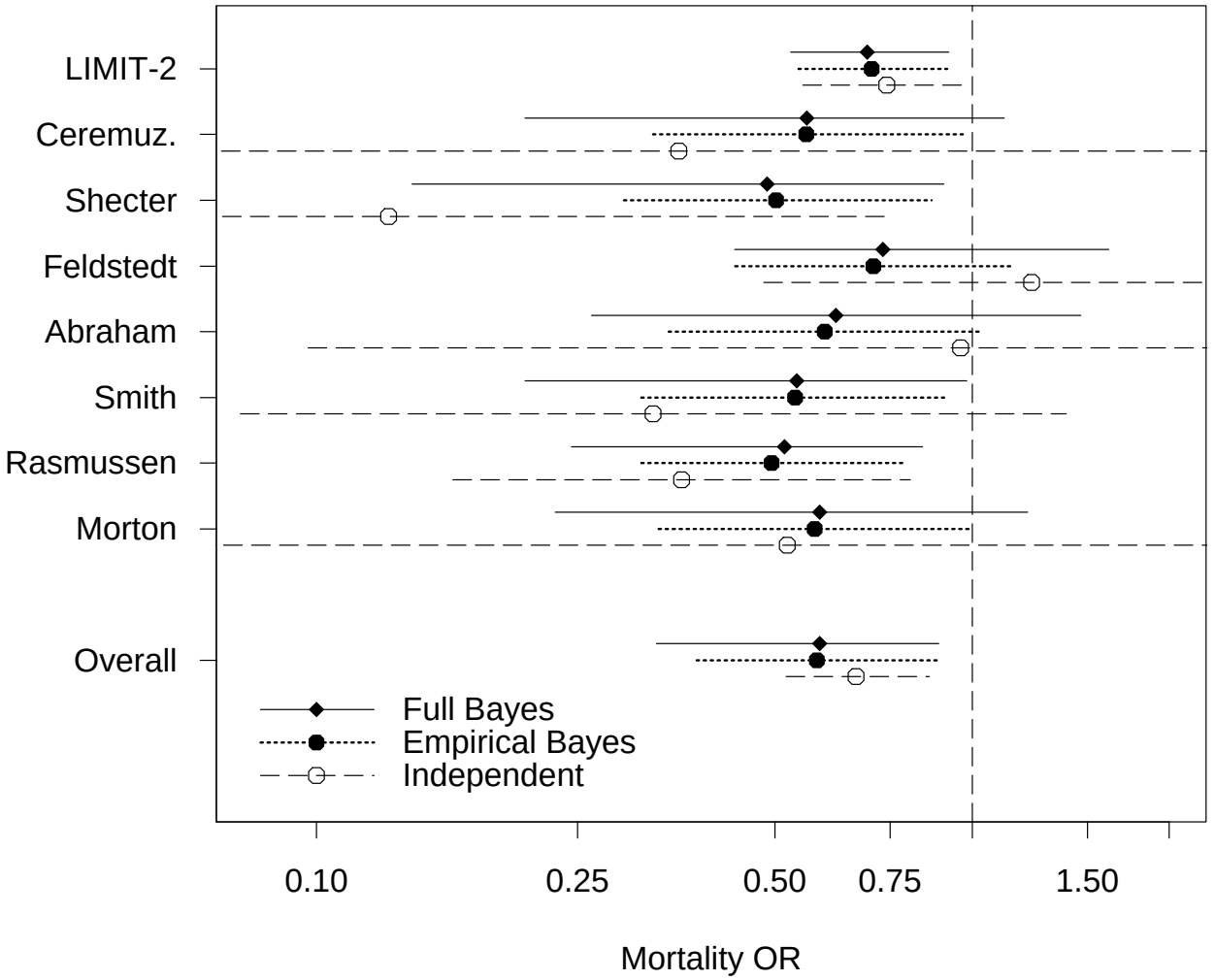
Empirical Bayes Model



Full Bayes Model



Trial	Magnesium group		Control group		Observed log OR	SE
	Deaths	Patients	Deaths	Patients	x_i	s_i
Morton	1	40	2	36	-0.65	1.06
Rasmussen	9	135	23	135	-1.02	0.41
Smith	2	200	7	200	-1.12	0.74
Abraham	1	48	1	46	-0.04	1.17
Feldstedt	10	150	8	148	0.21	0.48
Shechter	1	59	9	56	-2.05	0.90
Ceremuzynski	1	25	3	23	-1.03	1.02
LIMIT-2	90	1159	118	1157	-0.30	0.15


LIMIT-2
Ceremuz.
Shecter
Feldstedt
Abraham
Smith
Rasmussen
Morton
Overall
Full Bayes
Empirical Bayes
Independent
0.10
0.25
0.50
0.75
1.50
Mortality OR

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

Hierarchical regression models

Example: Hepatitis B Immunisation

Background

- Hepatitis B (HB) is endemic in Africa
- National program of childhood vaccination against HB introduced in Gambia
- Program effectiveness depends on duration of immunity afforded by vaccination

Data

- 106 children immunized against HB
- For each child: anti-HB titre measured at time of vaccination (baseline) and on 2 or 3 follow-up occasions

Study objective

- To obtain a model useful for predicting an individual child's protection against HB after vaccination

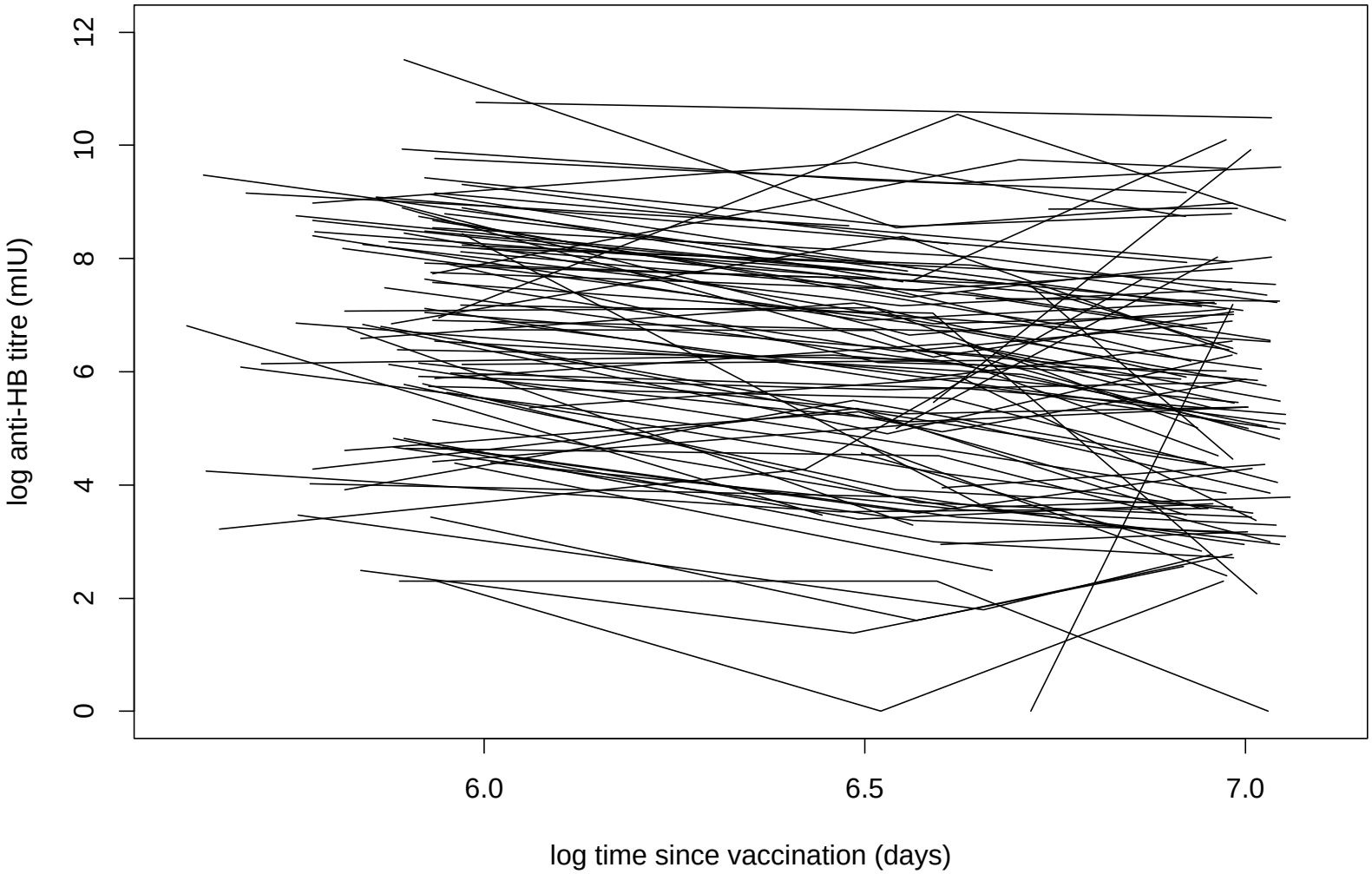
Related studies

- Similar study in Senegal found:

$$\text{anti-HB titre} \propto \frac{1}{T}$$

where T = time since HB vaccination

Raw data for a subset of 106 individuals



Non hierarchical linear model (LM) for the HB data

1. Probability dist^n (likelihood) for responses:

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

where y_{ij} = *log* of the jth anti-HB titre measurement for child i

2. Linear predictor:

$$\mu_{ij} = \alpha + \beta(t_{ij} - \bar{t}) + \gamma(y_{0i} - \overline{y_0})$$

where
t_{ij} = *log* of time (days since vaccination) of jth measurement for child i
y_{0i} = *log* of baseline anti-HB titre for child i

Problems

- Assumes a common regression line for all children
- Takes no account of the repeated measurements within children

$\Rightarrow$ modify LM to allow separate intercept and slope for each child:

$$\begin{aligned} y_{ij} &\sim \text{Normal}(\mu_{ij}, \sigma^2) \\ \mu_{ij} &= \alpha_i + \beta_i(t_{ij} - \bar{t}) + \gamma(y_{0i} - \overline{y_0}) \end{aligned}$$

Assumes that *conditionally* on α_i and β_i, $\{y_{ij}, j = 1, 2, ..\}$ are independent

- Assume α_i's are exchangeable and β_i's are exchangeable, e.g.

$$\begin{aligned} \alpha_i &\sim \text{Normal}(\mu_\alpha, \sigma^2_\alpha) \quad i = 1, ..., 106 \\ \beta_i &\sim \text{Normal}(\mu_\beta, \sigma^2_\beta) \quad i = 1, ..., 106 \end{aligned}$$

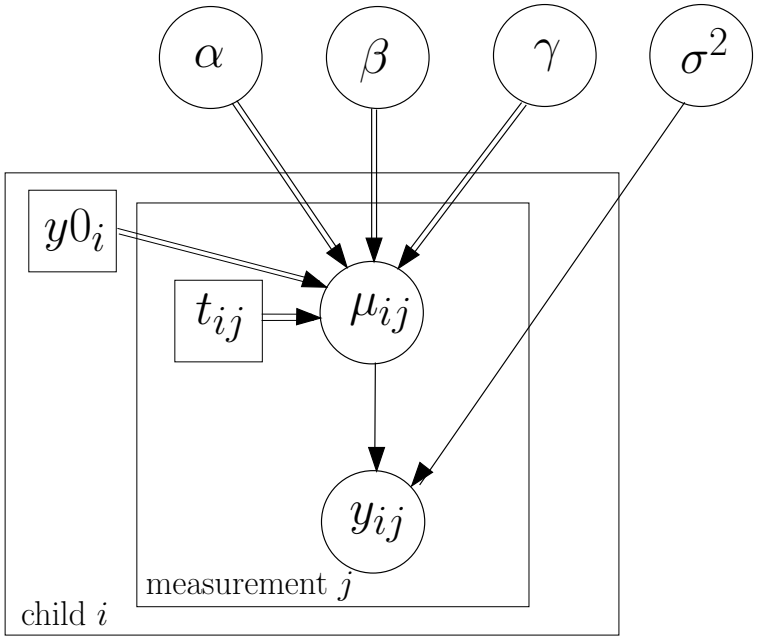
Note — alternatively, could allow slopes and intercepts to be *correlated* and assume that pairs (α_i, β_i) are exchangeable
$\rightarrow$ bivariate normal prior (see Practical Exercises)

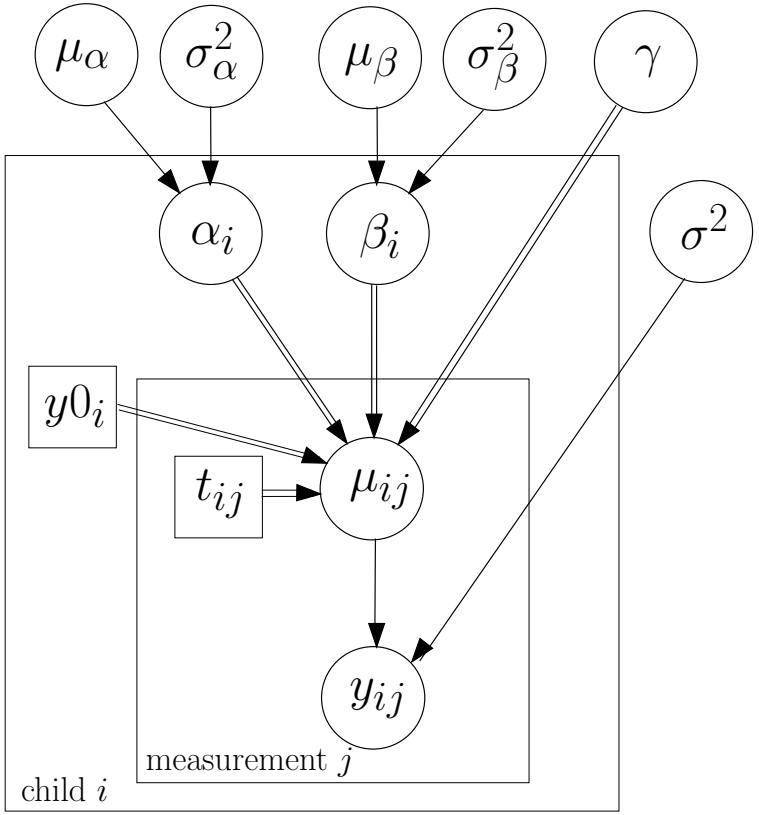
- We may then assume vague priors for the *hyperparameters* of the population distribution, e.g.

$$\begin{aligned} \mu_\beta, \mu_\alpha &\sim \text{Normal}(0, 10000) \\ \tau_\alpha = \sigma_\alpha^{-2}, \tau_\beta = \sigma_\beta^{-2} &\sim \text{Gamma}(0.001, 0.001) \end{aligned}$$

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

Graph of a LM and LMM for the HB data





Implementation in WinBUGS

Data contain 2 or 3 observations per child, so ragged array again

Child	Log Titre	Log Time
1	4.99, 8.02	6.54, 6.96
2	6.83, 4.91, 6.29	5.84, 6.52, 6.98
3	3.95, 4.35	6.60, 7.02
4	....	.....

Model code using nested index formulation

```
for(k in 1:TotalObs) {
  y[k] ~ dnorm(mu[k],tau)
  mu[k] <- alpha[child[k]]+ beta[child[k]]*(t[k]-tbar) + gamma*(y0[child[k]]-y0bar)
}
for(i in 1:N) {
  alpha[i] ~ dnorm(mu.alpha, tau.alpha)
  beta[i] ~ dnorm(mu.beta, tau.beta)
}
.... etc.
```

Data

```
list(N=106, TotalObs=288)
```

```
y[]     t[]     child[]    y0[]
4.99    6.54    1          8.61
8.02    6.96    1          8.61
6.83    5.84    2          7.10
4.91    6.52    2          7.10
.....
.....
END
```

Results for the LM and LMM models fitted to the HB data

Parameters	LM	Parameters	LMM
α	6.03 (0.10)	μ_α	6.04 (0.15)
β	-1.05 (0.22)	μ_β	-1.08 (0.13)
γ	0.67 (0.06)	γ	0.67 (0.08)
σ^2	3.00 (0.26)	σ^2	1.01 (0.11)
		σ^2_α	2.02 (0.35)
		σ^2_β	0.06 (0.09)
DIC	1136	DIC	913
p_D	4.0	p_D	95.1

One can see considerable improvement from fitting the LMM model, the variability being substantial only for the intercept

Note how the residual variance σ^2 has been reduced.

Multiple random effects and cross classified data

- Straightforward to extend basic 2-level hierarchical model to include multiple random effects at different levels:
  - nested hierarchies, e.g. THM measurements within zones within regions; pupils within classes within schools
  - cross-classified hierarchies, 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 BUGS using nested index notation

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

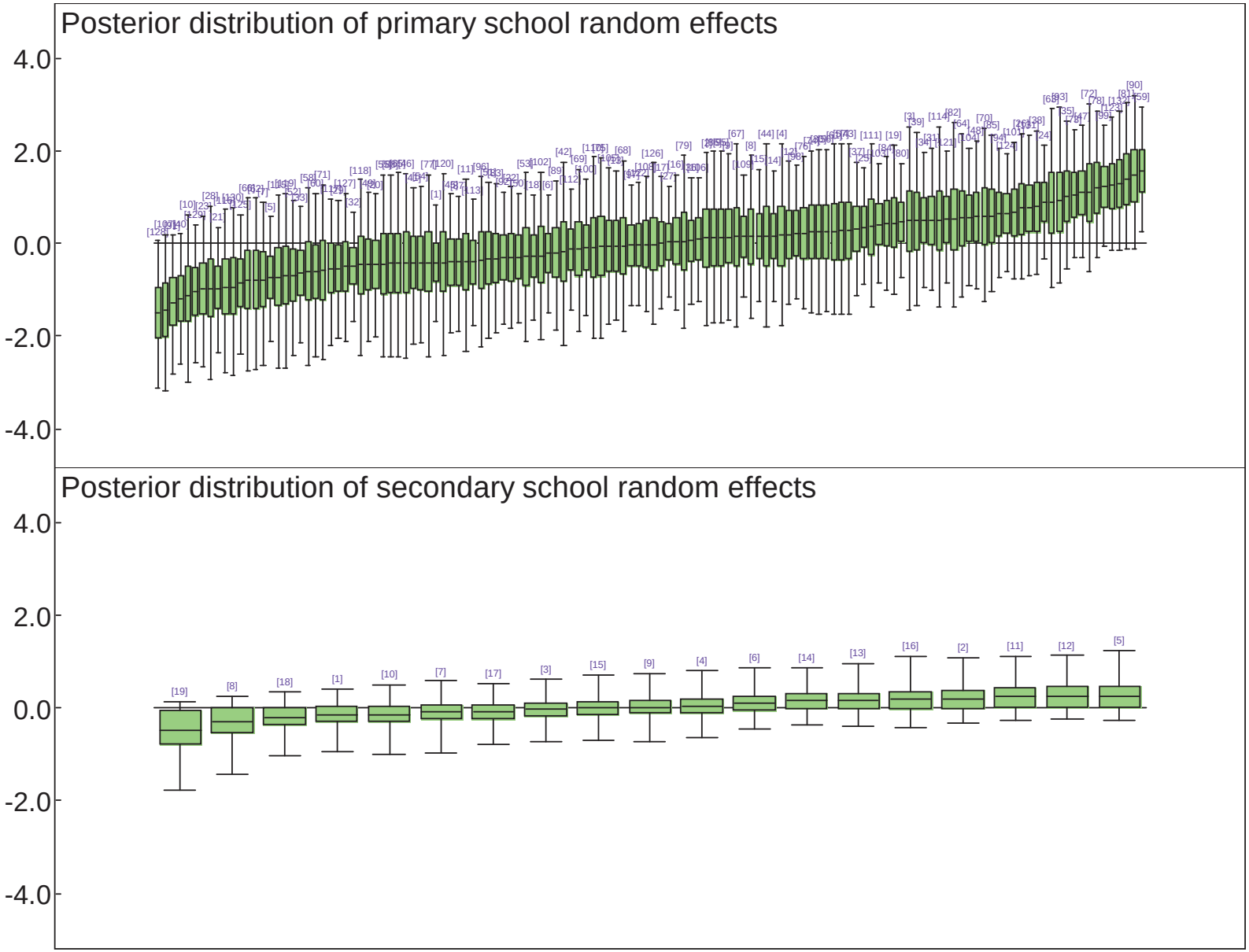
- Verbal reasoning score and gender are included as 'fixed' covariate effects (but note that in Bayesian framework, 'fixed' effect coefficients are still assigned prior distributions)

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 (note: non centered)
for (j in 1:Nprim) { theta.ps[j] ~ dnorm(0, tau.ps) } # primary school effects
for (k in 1:Nsec) { theta.ss[k] ~ dnorm(0, tau.ss) } # secondary school effects

# 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 variance
tau.ss ~ dgamma(0.001, 0.001)
sigma2.ss <- 1/tau.ss          # between secondary school variance
alpha ~ dnorm(0, 0.000001)     # intercept
for(q in 1:2) { beta[q] ~ dnorm(0, 0.000001)   }    # regression coefficients

# percentage of total variance explained ...
VPC.ps <- sigma2.ps / (sigma2.e + sigma2.ps + sigma2.ss) #..by primary school effects
VPC.ss <- sigma2.ss / (sigma2.e + sigma2.ps + sigma2.ss) #..by secondary school effects
```

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 example

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 coefficient of 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

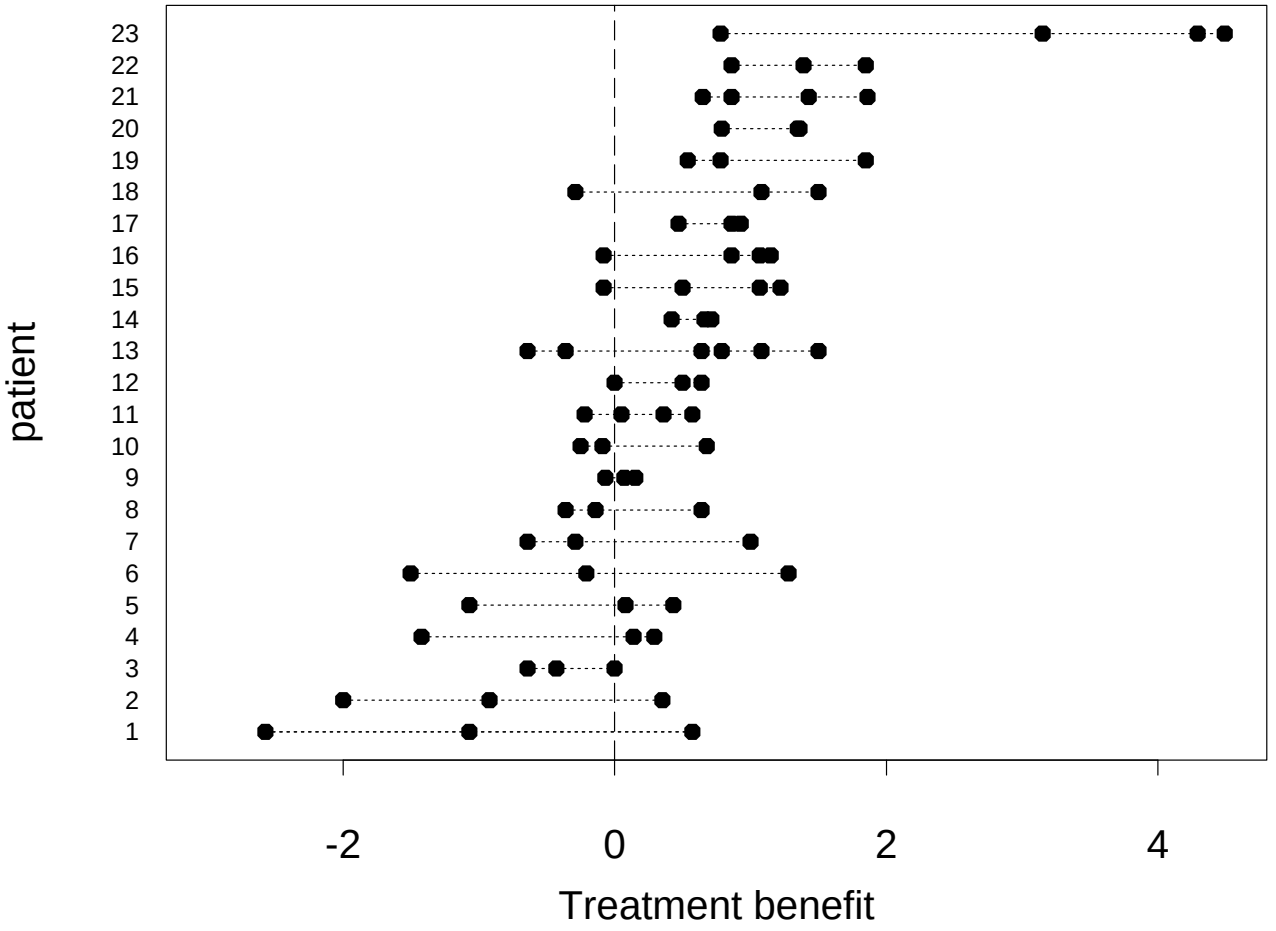
Hierarchical models for variances

Example: N-of-1 trials

Spiegelhalter et al (2004) Example 6.10

- 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

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

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.

Note: alternative assumptions are either that θ_k and σ_k^2 are same for all patients (pooled model) or that they are independent (fixed effects) for each patient

We make the specific distributional assumption that

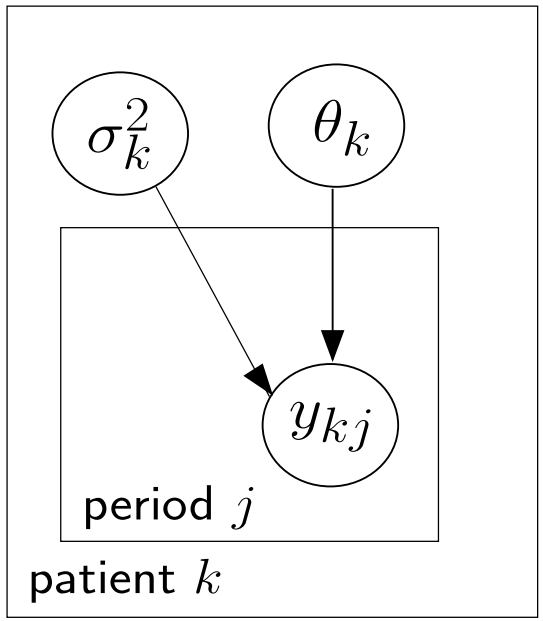
$$\begin{aligned} \theta_k &\sim \mathsf{N}(\mu_\theta, \phi_\theta^2) \\ \log(\sigma_k^2) &\sim \mathsf{N}(\mu_\sigma, \phi_\sigma^2) \end{aligned}$$

A normal distribution for the log-variances is equivalent to a log-normal distribution for the variances

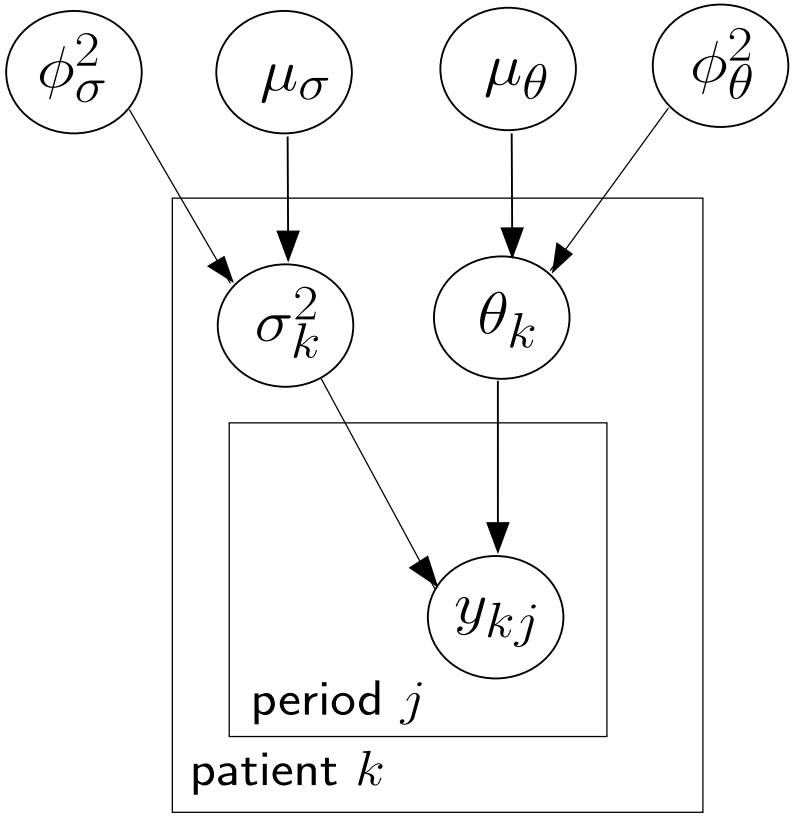
Uniform priors adopted for $\mu_\theta, \phi_\theta, \mu_\sigma$ and ϕ_σ.

Graphical model

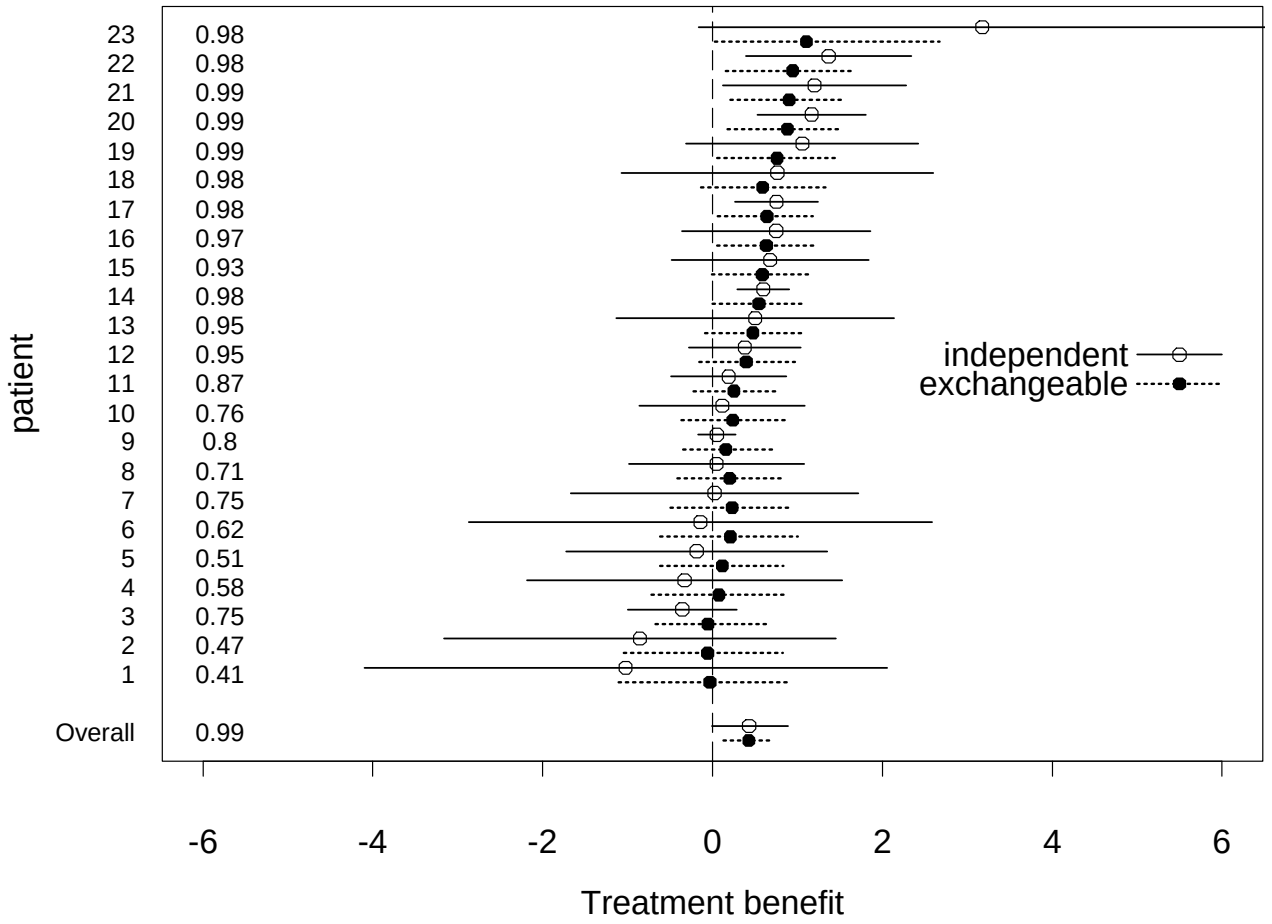
Independent effect



Exchangeable means and variances



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.

Further reading

WinBUGS examples volumes I and II (lots of examples of Bayesian hierarchical models)

Congdon (2001) (lots of examples of Bayesian hierarchical models)

Gelman et al (2004) Chapters 5, 13, 14

References

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman and Hall, New York.

Bartholomew, DJ, Steele, F, Moustaki, I and Galbraith, JI. (2002). *The analysis and interpretation of multivariate data for social scientists*, Chapman & Hall, London.

Berry, DA (1996). *Statistics: A Bayesian Perspective*, Duxbury, London.

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

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

Brooks, SP (1998). Markov chain Monte Carlo method and its application. *The Statistician*, **47**, 69-100.

Brooks, SP and Gelman, A (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7**, 434-455.

Casella, G and George, EI (1992). Explaining the Gibbs sampler. *The American Statistician*, **46**, 167–174.

Celeux, G, Forbes F, Robert CP and Titterington DM. (2003). *Deviance information criteria for missing data models.* Institut National de la Statistique et des Etudes Economiques. Serie des Documents de Travail du CREST, No. 2003-30.

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

Cowles, MK and Carlin, BP (1996) Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, **91**, 883–904.

Dempster, A (1998). Bayesian methods. In *Encyclopedia of Biostatistics*, (eds. P Armitage and T Colton). Wiley, Chichester, pp. 263–271.

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

Diggle, P, Moyeed, R, Rowlingson, B and Thomson, M (2002). Childhood malaria in the Gambia: a case-study in model-based geostatistics. *Applied Statistics*, **51**, 493–506.

DerSimonian, R and Laird, N (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, **7**, 177-188.

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.

Foster, RA, Stine, RP and Waterman, DP (1998). *Business Analysis Using Regression*. Springer-Verlag

Gelfand, AE and Smith, AFM (1990). Sampling-based approaches to calculating marginal densities. *J Amer Statistic Assoc*, **85**, 398–409.

Gelman, A (2005). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, to appear.

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

Greenland, S (1997). Probability logic and probabilistic induction. *Epidemiology*, **9**, 322–332.

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

Katz, JA and King, G. (1999). A Statistical Model for Multiparty Electoral Data. *The American Political Science Review*, **93**, 15-32.

Lee, PM (2004). *Bayesian Statistics: An Introduction*, 3rd edition, Arnold, London.

Lilford, RJ and Braunholtz, D (1996). The statistical basis of public policy: a paradigm shift is overdue. *British Medical Journal*, **313**, 603–607.

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

O'Hagan, A (1988). *Probability: Methods and Measurement*, Chapman and Hall, London.

Senn, S (1997). Statistical basis of public policy — present remembrance of priors past is not the same as a true prior. *British Medical Journal*, **314**, 73.

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, Thomas, A, and Best, NG (1995). Computation on Bayesian graphical models. In *Bayesian Statistics 5* (eds. JM Bernardo, JO Berger, AP Dawid and AFM Smith). Oxford University Press, Oxford), pp. 407-425.

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.

Tomz, M, Tucker, JA and Wittenberg, J. (2002). An easy and accurate regression model for multiparty electoral data. *Political Analysis*, **10**, 66–83.

Western, B and Jackman, S. (1994). Bayesian Inference for Comparative Research Source. *The American Political Science Review*, **88**, 412

Zellner, A. (1962). An efficient method for estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.*, **57**, 348–368.

Appendix.
Supplementary material

Extra slides for Lecture 2

Bayes theorem and its link with Bayesian inference

Axioms of probability

Let A, B be events, and $\{A_i, i = 1, 2, 3, ...\}$ be a set of events. The *probability* of A, $p(A)$, is a number which satisfies:

Axiom 1: $0 \leq p(A) \leq 1$ and $p(A) = 1$ if A is certain.

Axiom 2: If the events A_i are mutually exclusive, $p(\bigcup_i A_i) = \sum_i p(A_i)$

Axiom 3: $p(A \bigcap B) = p(B|A)p(A)$

Bayesian inference with binary data

Example: Inference on proportions using discrete prior

Assume treatment may have response rate θ of .2, .4, .6 or .8., each of equal prior probability. If we observe a single positive response ($x = 1$), how is our belief revised?

Likelihood, $p(x \mid \theta) = \theta^x(1-\theta)^{(1-x)}$

θ	Prior $p(\theta)$	Likelihood $p(x = 1 \mid \theta) = \theta$	Likelihood $\times$ prior $p(x = 1\|\theta)p(\theta)$	Posterior $p(\theta\|x = 1) = \frac{p(x=1\|\theta)p(\theta)}{\sum_j p(x=1\|\theta_j)p(\theta_j)}$
.2	.25	.2	.05	.10
.4	.25	.4	.10	.20
.6	.25	.6	.15	.30
.8	.25	.8	.20	.40
$\sum_j$	1.0		.50	1.0

Note: a single positive response makes it four times as likely that the true response rate is 80% rather than 20%.

Prediction

With a Bayesian approach, prediction is straightforward. Suppose we wish to predict the outcome of a new observation $\tilde{x}$ (say), given what we have already observed.

For discrete θ we have

$$p(\tilde{x}|x) = \sum_{\theta_j} p(\tilde{x}, \theta_j|x)$$

which, assuming $\tilde{x}$ and x are conditionally independent given θ, is equal to

$$p(\tilde{x}|x) = \sum_{\theta_j} p(\tilde{x}|\theta_j)p(\theta_j|x)$$

where the $p(\theta_j|x)$ can be thought of as 'posterior weights'.

In example, predictive probability of treatment outcome for a new patient is:

$$\begin{aligned} p(\tilde{x} = 0|x = 1) &= \sum_{\theta_j} (1 - \theta_j) p(\theta_j|x = 1) \\ &= (0.8) \times 0.1 + (0.6) \times 0.2 + (0.4) \times 0.3 + (0.2) \times 0.4 = 0.4 \\ p(\tilde{x} = 1|x = 1) &= \sum_{\theta_j} \theta_j p(\theta_j|x = 1) \\ &= 0.2 \times 0.1 + 0.4 \times 0.2 + 0.6 \times 0.3 + 0.8 \times 0.4 = 0.6 \end{aligned}$$

More complex priors for proportions

Suppose we want to express more complex prior opinion that cannot be adequately summarised by a beta distribution

For example, might suspect that either drug will produce similar effect to other related compounds, or if it doesn't behave like these compounds we are unsure about its likely effect

Prior is then a *mixture*

$$p(\theta) = qp_1(\theta) + (1-q)p_2(\theta)$$

where $p_i = \text{Beta}(a_i, b_i)$

If we now observe r successes out of n cases, it turns out that the posterior is

$$p(\theta|r,n) = q'p_1(\theta|r,n) + (1-q')p_2(\theta|r,n)$$

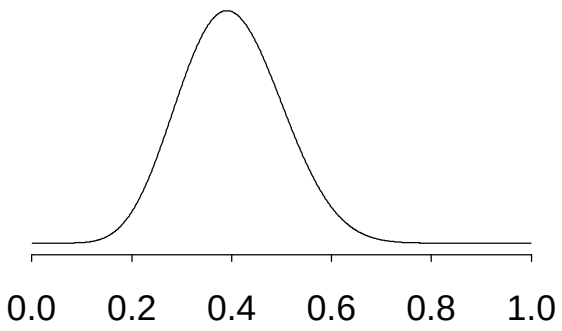
where

$$\begin{aligned} p_i(\theta|r,n) &\propto p(r|\theta,n)p_i(\theta) \\ q' &= \frac{qp_1(r|n)}{qp_1(r|n) + (1-q)p_2(r|n)} \end{aligned}$$

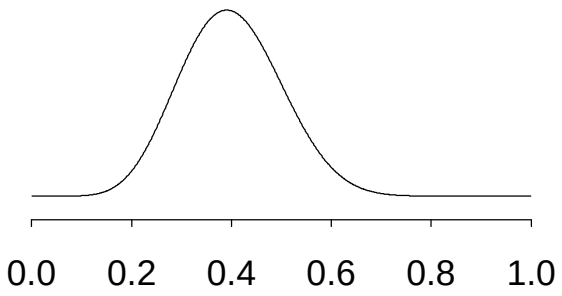
Note: $p_i(r|n)$ is the beta-binomial predictive probability of r successes in n cases assuming θ has distribution $p_i(\theta)$

$\Rightarrow$ posterior is mixture of respective beta posteriors, with mixture weights adapted to support prior that provides best prediction for the observed data

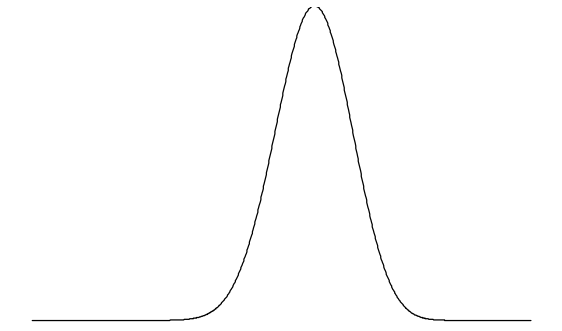
(a) Beta(9.2, 13.8) prior



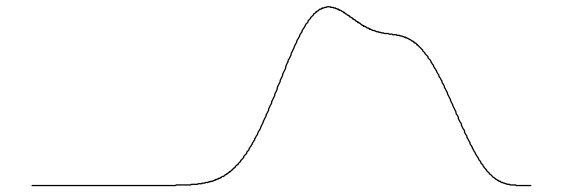
(b) Mixture prior with 20% uniform



(c) Posterior



(d) Posterior



Mixture posterior has 44% on beta(24.2, 18.8) and 56% on beta(16, 6)

Derivation of posterior for Normal data with unknown mean, known variance

Suppose we have a sample of Normal data $x_i \sim \mathrm{N}(\theta, \sigma^2) \quad (i = 1, ..., n)$. For now assume σ^2 is known and θ has a Normal prior $\theta \sim \mathrm{N}(\mu, \sigma^2/n_0)$

Then the posterior distribution is

$$\begin{aligned} p(\theta|\boldsymbol{x}) &\propto \prod_i p(x_i \,|\, \theta)\; p(\theta) \\ &\propto \exp\left[-\frac{\sum_i (x_i - \theta)^2}{2\sigma^2}\right] \times \exp\left[-\frac{(\theta - \mu)^2 n_0}{2\sigma^2}\right] \end{aligned}$$

By matching terms in θ and writing $\sum x_i = n\overline{x}$ it can be shown that

$$\sum_i (x_i - \theta)^2 + (\theta - \mu)^2 n_0 = \left(\theta - \frac{n_0\mu + n\overline{x}}{n_0 + n}\right)^2 (n_0 + n) + \text{constant}$$

The term involving θ is exactly that arising from a Normal distribution, so

$$p(\theta|\boldsymbol{x}) = \mathrm{N}\left(\frac{n_0\mu + n\overline{x}}{n_0 + n}, \frac{\sigma^2}{n_0 + n}\right)$$

Extra slides for Lecture 4

Transformations and Jacobians

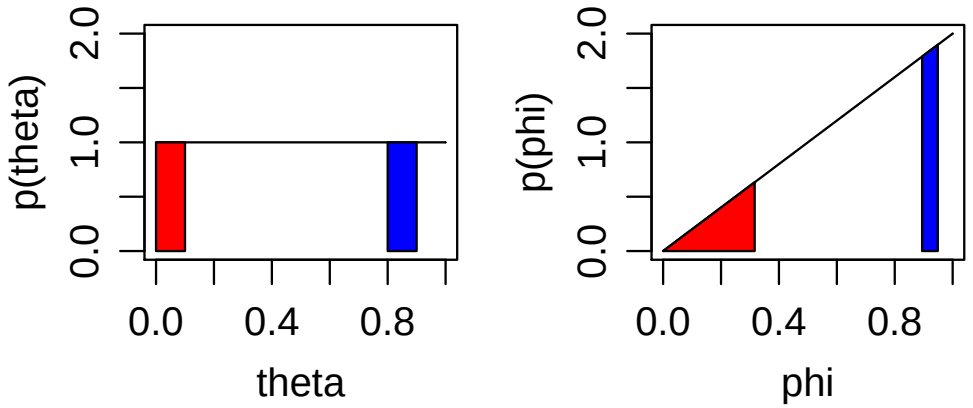
Suppose $p(\theta) = \text{Uniform}(0,1)$, i.e. $p(\theta) = 1$ for $\theta \in (0,1)$

Now consider $\phi = \sqrt{\theta}$. What is the distribution of ϕ?

The probability that $\theta \in (a,b)$ must equal probability that $\phi \in (\sqrt{a}, \sqrt{b})$
$\Rightarrow$ area under curve of $p(\theta)$ in the interval (a,b) = area under curve of $p(\phi)$ in the interval $(\sqrt{a}, \sqrt{b})$

Turns out that the amount we need to stretch or compress the height of the density $p(\theta)$ to give the corresponding height of the density $p(\phi)$ is given by the Jacobian of the transformation, $\left|\frac{d\theta}{d\phi}\right|$

Here, $\phi = \sqrt{\theta} \Rightarrow \theta = \phi^2$, so $\left|\frac{d\theta}{d\phi}\right| = 2\phi$ and hence $p(\phi) = 2\phi \times 1 = 2\phi$ for $\phi \in (0,1)$



Further examples of Jeffreys' priors

- Poisson case with parameter θ

$$\log p(x|\theta) = -\theta + x \log \theta + C \qquad \Rightarrow I(\theta) = 1/\theta$$

So Jeffreys' prior for θ is $\propto \theta^{-1/2}$
This improper distribution is approximated by a Gamma distribution with $\alpha = 1/2$ and $\beta \to 0$

Some inconsistencies associated with Jeffreys' priors

For example, applying this rule to the normal case with both mean and variance parameters unknown does not lead to the same prior as applying separately the rule for the mean and the variance and assuming a priori independence between these parameters.

Various 'non-informative' priors for the binomial parameter

Consider r successes from n trials: $r \sim \text{Binom}(n, \theta)$, then

$$\log p(r|\theta) = r \log \theta + (n - r) \log(1 - \theta) + C$$

and $I(\theta) = \frac{n}{\theta(1-\theta)}$.

Thus Jeffreys' prior is

$$p(\theta) \propto (\theta(1 - \theta))^{-\frac{1}{2}},$$

which is a Beta(1/2, 1/2) distribution .

The Bayes-Laplace Uniform density is a Beta(1,1) distribution.

A prior density that is uniform for logitθ is $\propto$ to $(\theta(1-\theta))^{-1}$, which is the improper Beta(0,0) distribution.

In practise, there will not be much difference between these alternatives, but the improper Beta(0,0) prior distribution leads to an improper posterior if $r = 0$ (or $n = 0$)!

Extra slides for Lecture 6

Bayesian measures of model dimensionality (Spiegelhalter et al, 2002)

$$\begin{aligned} p_D &= E_{\theta|y}[d_\Theta(y,\theta,\tilde{\theta}(y))] \\ &= E_{\theta|y}[-2\log p(y|\theta)] + 2\log p(y|\tilde{\theta}(y)). \end{aligned}$$

If we take $\tilde{\theta} = E[\theta|y]$, then

$$p_D = \text{“posterior mean deviance - deviance of posterior means”}.$$

In normal linear hierarchical models:

$$p_D = tr(H)$$

where $Hy = \widehat{y}$. Hence H is the ‘hat’ matrix which projects data onto fitted values.

Thus $p_D = \sum h_{ii} = \sum$leverages.

Approximately Normal likelihoods

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

A Taylor expansion of $D(\theta)$ around $D(\bar{\theta})$, followed by posterior expectation, gives

$$\begin{aligned} E_{\theta|y}[D(\theta)] &\approx D(\bar{\theta}) - \mathsf{E}\left[\mathrm{tr}\left((\theta-\bar{\theta})^T L''_{\bar{\theta}}(\theta-\bar{\theta})\right)\right] \\ &= D(\bar{\theta}) + \mathrm{tr}\left(-L''_{\bar{\theta}}\, V\right) \end{aligned}$$

where $V = \mathsf{E}\left[(\theta-\bar{\theta})(\theta-\bar{\theta})^T\right]$ is the posterior covariance matrix of θ, $-L''_{\bar{\theta}}$ is the observed Fisher's information evaluated at the posterior mean of θ.

Thus

$$p_D = \mathrm{tr}\left(-L''_{\bar{\theta}} V\right),$$

can be thought of as the ratio of the information in the likelihood about the parameters as a fraction of the total information in the posterior (likelihood + prior). parameters - this result in likelihood parameters (degrees of freedom)

$\rightarrow$ can also think of p_D as the dimensionality of the parameter space that is identifiable by the data

Which plug-in estimate to use in p_D?

- p_D is not invariant to reparameterisation, *i.e.* which estimate is used in $D(\tilde{\theta})$
- WinBUGS currently uses posterior mean of stochastic parents of θ, *i.e.* if there are stochastic nodes ψ such that $\theta = f(\psi)$, then $D(\tilde{\theta}) = D(f(\overline{\psi}))$
- 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

- If $\theta \sim U[0,1]$, then $\psi = \theta^a$ is beta$(a^{-1}, 1)$.
- Suppose we observe $r = 1$ successes out of $n = 2$ Bernoulli trials, so that $r \sim \text{Bin}[\theta, n]$
- Consider putting prior on $\psi = \theta$, θ^5 and θ^{20}, each equivalent to uniform prior on θ

```
r <- 1;  n<- 2 a[1]<-1 ; a[2] <- 5; a[3] <- 20
for (i in 1:3){
     a.inv[i]<- 1/a[i]
     theta[i] <- pow(psi[i], a.inv[i])
     psi[i]  ~ dbeta(a.inv[i] , 1)
     }
r1<- r; r2<-r ; r3 <- r
r1 ~ dbin(theta[1],n)
r2 ~ dbin(theta[2],n)
r3 ~ dbin(theta[3],n)
```

After 21000 iterations

```
    Dbar     Dhat      pD
r1  1.947   1.386    0.561 r2  1.912   1.547    0.365 r3  1.921
2.239   -0.318
```

Mean deviances (Dbar) and posteriors for all θ's are the same, but using $\overline{\psi}$ as a plug-in is clearly a bad idea.