Advanced Bayesian Modelling with BUGS
MRC Biostatistics Unit
University of Cambridge

Christopher Jackson, Robert Goudie and Anne Presanis

24–25 May 2017

Thanks also to Dave Lunn, Rebecca Turner, Ian White, Dan Jackson, Nicky Best, Alexina Mason, Jules Sanchez-Hernandez for comments and contributions, and to all the authors of BUGS, JAGS and related software.

Notes

Course timetable: Day 1

9.30–10.00 Introductions, BUGS principles (CJ)
10.00–11.00 Hierarchical models and meta-analysis (RG)
11.00–11.45 Practical: hierarchical models
11.45–12.45 Missing data (CJ)
12.45–13.30 Lunch
13.30–14.30 Practical: missing data
14.30–15.30 Censoring and truncation (CJ)
15.30–16.30 Practical: censoring and truncation
16.30–17.00 Measurement error (CJ)

Notes

Course timetable: Day 2

9.15–10.00 Practical: measurement error
10.00–11.00 Complex hierarchical models (RG)
11.00–12.00 Practical: complex hierarchical models
12.00–13.00 Evidence synthesis (1) (AP)
13.00–13.45 Lunch
13.45–14.45 Practical: evidence synthesis (1)
14.45–15.45 Evidence synthesis (2): model criticism (AP)
15.45–16.45 Practical: evidence synthesis (2)

Running themes: awareness of model assumptions, prior distributions, checking fitted models

Notes

What's assumed

First few chapters of the BUGS book (or our Intro course)

- Monte Carlo methods and basic Bayesian inference
- How to run models in WinBUGS, OpenBUGS or JAGS (or their R interfaces)
- Simple statistical models (eg linear regression)

We will briefly revise

- (now) How BUGS works...
- (later) Basics of specific advanced methods (hierarchical, missing data, model checking and comparison)

Course draws a lot on later chapters of BUGS Book

- examples in book all available to download for WinBUGS `http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-the-bugs-book/`

Notes

BUGS implementations

When we say "BUGS" we mean OpenBUGS, WinBUGS and JAGS

- All three programs based on the same modelling language and computational algorithms
- We will try to point out subtle differences between the three in specific situations
- Practical material in either
  - JAGS with `rjags` R interface
  - OpenBUGS Windows interface
  - OpenBUGS + `R2OpenBUGS` R interface

We don't cover the Stan software (`mc-stan.org`), though:

- Stan also can be used to fit all the models we describe
- the modelling language, and the algorithms used by the software, are different

Notes

BUGS code is equivalent to a model

BUGS code is equivalent to algebraic statements defining a statistical model

e.g. linear regression

$$\begin{aligned} y_i &\sim N(\mu_i, \sigma^2) \\ \mu_i &= \alpha + \beta x_i \\ i &= 1, \ldots, n \end{aligned}$$

plus priors on α, β, σ

- Unknown parameters are α, β, σ
- Known data are y_i, x_i
  (and parameters of priors for α, β, σ)

```
model {
 for (i in 1:n) {
   y[i]  ~ dnorm(mu[i], tau)
   mu[i] <- alpha + beta*x[i]
 }
 alpha   ~ dunif(-100,100)
 beta    ~ dunif(-100,100)
 sigma   ~ dunif(0, 100)
 tau     <- 1/(sigma*sigma)
}
```

~ stochastic relation: used for

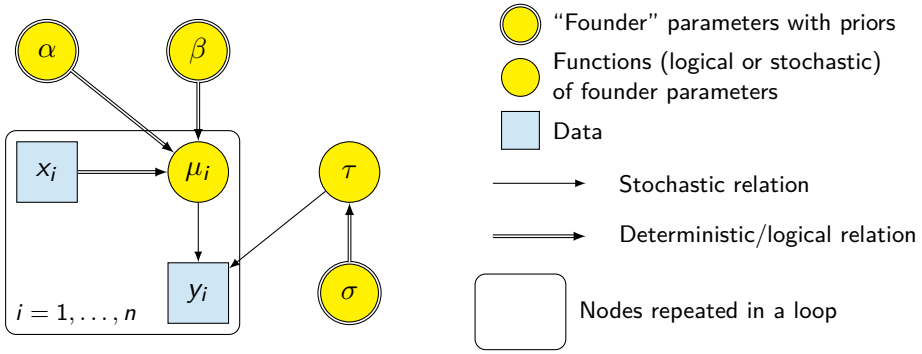
- models for data
- priors for parameters

<- logical/deterministic relation

Notes

Directed acyclic graph

BUGS can (only) fit models which can be represented as a directed acyclic graph (DAG).
All quantities are nodes in the graph.



"Founder" parameters with priors

Functions (logical or stochastic) of founder parameters

Data

Stochastic relation

Deterministic/logical relation

Nodes repeated in a loop

DAG notation used throughout course to represent models

Notes

BUGS does Bayesian inference

- θ all unknown parameters
- $\mathbf{y}$: all known data

BUGS estimates the posterior distribution

$$p(\theta|\mathbf{y}) \propto \underbrace{p(\theta)}_{\text{prior}} \underbrace{p(\mathbf{y}|\theta)}_{\text{likelihood}}$$

by Gibbs sampling (a kind of Markov Chain Monte Carlo, MCMC)

Notes

MCMC algorithms and convergence

Details of algorithms not covered in course

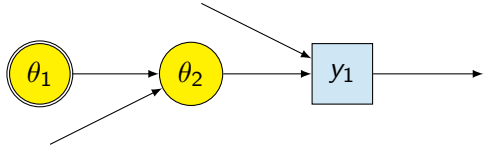
- BUGS philosophy is to hide these details, leaving users free to specify realistically complex model structures for their data
- There are tricks for improving MCMC convergence in some examples
  - reparameterisation: centering parameters, expanded parameterisations (see WinBUGS manual)
- Precise variety of sampler used by JAGS, OpenBUGS and WinBUGS may differ for the same example
  - Stan (`mc-stan.org`) uses Hamiltonian Monte Carlo, samples all parameters jointly, often more efficient.
  - NIMBLE `https://r-nimble.org/` software based on BUGS language, users can write own samplers.

Notes

DAG structure leads to the posterior distribution

Posterior $p(\theta|\mathbf{y}) \propto$ product $\prod p(v|pa[v])$ of distributions of all nodes $v \in \{\theta, \mathbf{y}\}$ in graph, each conditional on its parents $pa[v]$

Example:



$$p(\theta_1, \theta_2, ...|\mathbf{y}) = p(\theta_1) \underbrace{p(\theta_2|\theta_1, \ldots)}_{\text{prior for } \theta_2} \underbrace{p(y_1|\theta_2, \ldots)}_{\text{likelihood for } \theta_2} \ldots$$

Posterior distribution of a node (e.g. θ_2) depends on

- children: data/parameters generated by θ_2 ("likelihood", y_1)
- as well as parents: ("prior", θ_1).

Information flows "up, as well as down arrows" in a graph

Notes

Philosophy of Bayesian modelling with BUGS

Real data has many complicating features

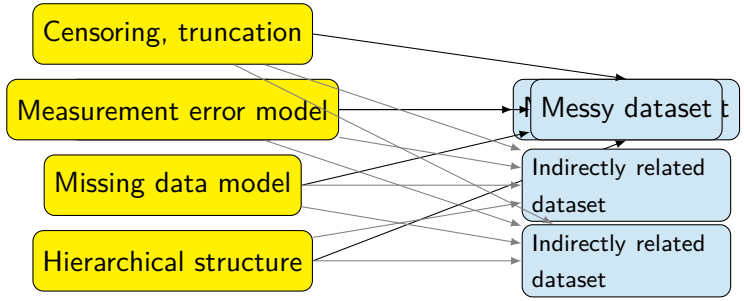
- missing data, measurement error
- censoring, truncation
- hierarchical structure
- multiple sources of data

Conventional statistical software and classical methods can be used for most of these things separately...

- In BUGS can have any or all of them at the same time...
- Complete freedom to write your own model to represent real world data

Notes

Piecing together model components with BUGS



- Multiple complicating features can be pieced together in the same BUGS code / DAG
- Multiple indirectly-related datasets can be modelled

Uncertainty is propagated

Uncertainty about inferences from one dataset/model component are accounted for in the rest of the model

Notes

Challenges of such flexibility

As models get bigger, assumptions accumulate:

- Prior distributions
  - some examples / principles for prior choice given later
- Model structures, assumptions in likelihoods
  - throughout course
- Inclusion of particular data
  - in general evidence synthesis

→ model checking and comparison, sensitivity analysis

Notes

Part I

Hierarchical models

Notes

Summary

- Basics of hierarchical models
- Hierarchical models for meta-analysis
  - Common effect meta-analysis
  - Random-effects meta-analysis
- Priors in hierarchical models
  - Residual variance priors
  - Location priors
  - Random-effect variance priors
    - Default priors
    - Informative priors

Notes

Assumptions for hierarchical data

Make inferences on parameters $\theta_1, ..., \theta_N$ measured on N 'units' (individuals, subsets, areas, time-points, trials, etc) which are related by the structure of the problem

Possible assumptions

1. **Identical parameters:** All the θs are identical $\rightarrow$ the data can be pooled and the individual units ignored.
2. **Independent parameters:** All the θs are entirely unrelated $\rightarrow$ the results from each unit can be analysed independently.
3. **Exchangeable parameters:** The θs 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

Notes

Bristol example

Data on mortality rates for 1991–1995 from $N = 12$ hospitals performing heart surgery on children (< 1 year old)

	Hospital	Operations n_i	Deaths y_i
1	Bristol	143	41
2	Leicester	187	25
3	Leeds	323	24
4	Oxford	122	23
5	Guys	164	25
6	Liverpool	405	42
7	Southampton	239	24
8	Great Ormond St	482	53
9	Newcastle	195	26
10	Harefield	177	25
11	Birmingham	581	58
12	Brompton	301	31

Notes

Bristol (continued)

Assume a binomial likelihood for the data from each hospital

$$y_i \sim \text{Bin}(n_i, \theta_i) \qquad i = 1, \ldots, 12$$

1. **Identical parameters** $\theta_1 = \theta_2 = \cdots = \theta_{12} = \theta$
   Assign a prior for θ: for example, $\theta \sim \text{Beta}(a, b)$
   where a and b are fixed constants
2. **Independent parameters**
   Assume independent priors for θ in each hospital, for example

   $$\theta_i \sim \text{Beta}(a, b) \qquad i = 1, \ldots, 12$$

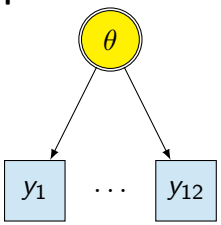
   where a and b are fixed constants
3. **Exchangeable parameters**
   Assume a hierarchical prior. For example, where μ and σ^2 are unknown parameters to be estimated:

   $$\text{logit}(\theta_i) = \phi_i$$
   $$\phi_i \sim \text{N}(\mu, \sigma^2) \qquad i = 1, \ldots, 12$$

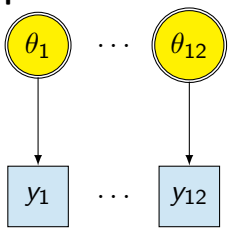
Notes

Graphical representation

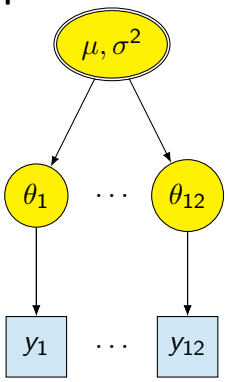
Identical parameters



Independent parameters



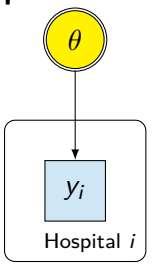
Exchangeable parameters



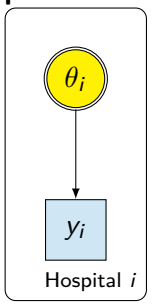
Notes

Graphical representation (plate notation)

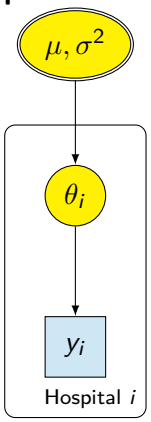
Identical parameters



Independent parameters



Exchangeable parameters



Notes

BUGS code

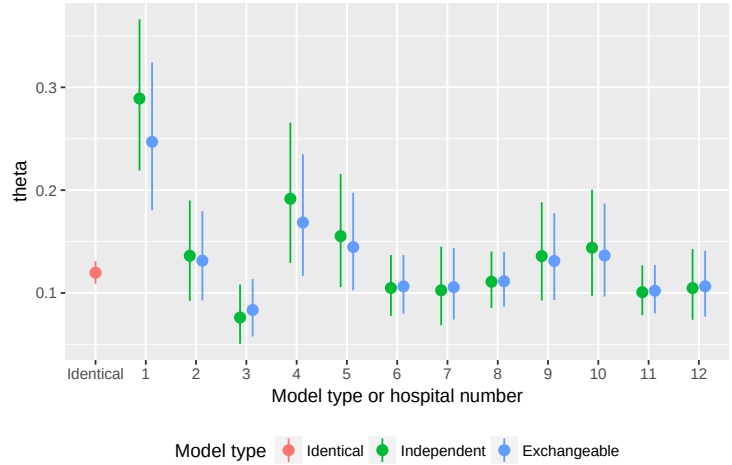
The exchangeable parameter model can be written as

```
model {
  for (i in 1:N){
    y[i] ~ dbin(theta[i], n[i])
    logit(theta[i]) <- phi[i]
    phi[i] ~ dnorm(mu, tau)
  }
  mu ~ dunif(-100, 100)
  tau <- 1/pow(sd, 2)
  sd ~ dunif(0, 100)
}
```

Notes

Results of Bristol analysis

Mortality rates θ in each hospital



Notes

Comments on shrinkage, borrowing of strength

- **Borrowing strength**
  - Smaller credible intervals in exchangeable model
  - Occurs because each posterior borrows strength from the others through their joint influence on the population parameters
- **Global smoothing of uncertainty**
  - Width of credible intervals is more equal
  - Occurs due to the borrowing of strength
- **Shrinkage to the mean**
  - Hospitals with a large number of observations contribute more to the overall mean than those with a small number of observations
  - Data for hospitals with a small number of observations are effectively supplemented with information from hospitals with a large number of observations
  - Outliers (which tend to be hospitals with few observations) are pulled towards the overall mean

Notes

Meta analysis

Meta analysis is a method for summarising results from a number of separate studies of treatments or interventions.

Forms part of the process of systematic review, which also includes the process of identifying and assessing the quality of relevant studies.

- Meta-analysis is very widely adopted in medical applications
- Each 'study' is often a single clinical trial
- Can be viewed as a special case of hierarchical modelling.

Notes

Meta analysis for odds ratios

Often we want to compare the number of 'successes' or 'events' between patients who were treated and those in a 'control' group

	Treatment		Control	
Study	Successes	Total	Successes	Total
Study 1	y_{1T}	n_{1T}	y_{1C}	n_{1C}
Study 2	y_{2T}	n_{2T}	y_{2C}	n_{2C}
⋮	⋮	⋮	⋮	⋮
Study s	y_{sT}	n_{sT}	y_{sC}	n_{sC}
⋮	⋮	⋮	⋮	⋮

Clearly the number of 'failures' in study s are

- Treatment group $n_{sT} - y_{sT}$
- Control group $n_{sC} - y_{sC}$

Compare the number of successes using the (log) odds ratio

Notes

Common effect meta-analysis for odds ratios

Assume a single underlying treatment effect δ common to all studies

For studies $s = 1, \ldots, N$

$$y_{sT} \sim \text{Bin}(n_{sT}, p_{sT}) \qquad y_{sC} \sim \text{Bin}(n_{sC}, p_{sC})$$
$$\text{logit}(p_{sT}) = \alpha_s + \delta \qquad \text{logit}(p_{sC}) = \alpha_s$$

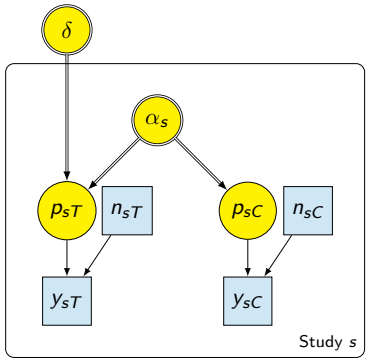
- α_s: log odds for a 'success' in control group in study s
- $\alpha_s + \delta$ is log odds for 'success' in treatment arm in study s
- δ is the log odds ratio

This is an identical parameter model.

Common effect meta-analysis is also called fixed-effects meta-analysis.

Notes

Common effect meta-analysis for odds ratios




```
for (s in 1:Ns){
  for (a in 1:2){
    y[s,a] ~ dbin(p[s,a], n[s,a])
    logit(p[s,a]) <- alpha[s] +
                     delta[a]
  }
  alpha[s] ~ dnorm(0, 0.01)
}
delta[1] <- 0
delta[2] ~ dnorm(0, 0.01)
```


Notes

Random effects meta-analysis for odds ratios

Procedures, patient characteristics etc often differ between studies, and often at least part of the variability is unexplained

For studies $s = 1, \ldots, N$

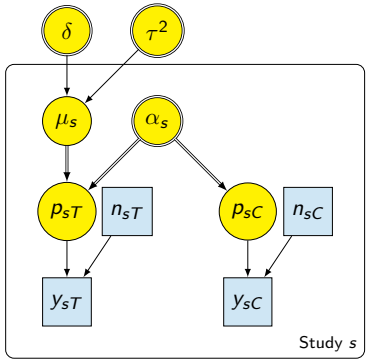
$$y_{sT} \sim \text{Bin}(n_{sT}, p_{sT}) \qquad y_{sC} \sim \text{Bin}(n_{sC}, p_{sC})$$
$$\text{logit}(p_{sT}) = \alpha_s + \mu_s \qquad \text{logit}(p_{sC}) = \alpha_s$$
$$\mu_s \sim N(\delta, \tau^2)$$

- α_s is the log odds for a 'success' in control arm in study s
- $\alpha_s + \mu_s$ is the log odds for a 'success' in treatment arm in study s
- δ is the overall log odds ratio

This is an exchangeable parameter model.

Notes

Random effects meta-analysis for odds ratios




```
for (s in 1:Ns){
  for (a in 1:2){
    y[s,a] ~ dbin(p[s,a], n[s,a])
    logit(p[s,a]) <- alpha[s] +
                     mu[s,a]
  }
  mu[s, 1] <- 0
  mu[s, 2] ~ dnorm(delta, prec.mu)

  alpha[s] ~ dnorm(0, 0.01)
}
delta ~ dnorm(0, 0.01)
```


Notes

Model comparison using the DIC

More complex models will fit data better, so need to trade off model fit and complexity.

- Deviance = -2 times the log-likelihood:

$$D(\boldsymbol{\theta}) = -2 \log\{p(\boldsymbol{y} \mid \boldsymbol{\theta})\}$$

- Measure of fit:
  - Posterior mean deviance: $\overline{D} = E_{\theta|\boldsymbol{y}}[D(\boldsymbol{\theta})]$
- Measure of complexity:
  - Effective number of parameters: $p_D = \overline{D} - D(\overline{\boldsymbol{\theta}})$ where "plug-in" deviance $D(\overline{\boldsymbol{\theta}})$ calculated at the posterior mean or median of the parameters: $D(\overline{\boldsymbol{\theta}}) = -2\log\{p(\boldsymbol{y} \mid \overline{\boldsymbol{\theta}})\}$
- The Deviance Information Criterion = fit + complexity:

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

- Smaller DIC better: very roughly, differences in DIC > 10 are substantial; DIC < 5 may be negligible

Notes

Deviance summaries in BUGS

In OpenBUGS, the deviance is automatically calculated, where the parameters θ are those that appear in the stated sampling distribution of y. The node `deviance` can be monitored.

DIC can also be calculated. After enough burn-in iterations have been run (the option will be greyed out before then), choose `Inference-DIC-Set` to start monitoring *DIC*. Once your posterior samples are obtained, choose `Inference-DIC-Stats` to see the built-in calculation of *DIC* and p_D.

Notes

DIC in JAGS

Two alternative methods:

1. $DIC = \bar{D} + p_D$
   - p_D calculated differently from WinBUGS/OpenBUGS
   - but same interpretation as "effective number of parameters"
2. "Penalized expected deviance": alternative to DIC
   - Penalizes complex models more severely:
   - usually higher than $\bar{D} + 2p_D$.

Both are options in `dic.samples` in `rjags` R package

See also Widely Applicable Information Criterion (WAIC) as recommended with the Stan software (see, e.g. Gelman *et al.* *Bayesian Data Analysis*, 3rd ed)

All approximations to different cross-validatory loss functions.

Notes

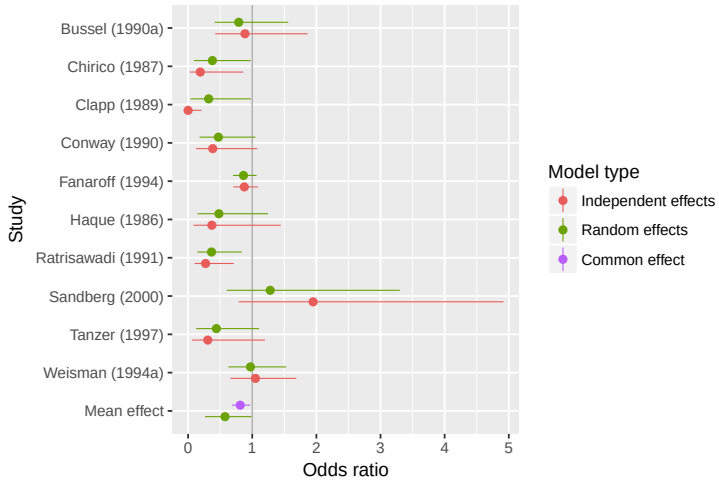
Example: Sepsis (Ohlsson and Lacy, 2013)

Outcome Infection (or not) in preterm/low birth weight infants
Arms Intravenous immunoglobulin (IVIG) vs placebo
Question Does administration of IVIG prevent infection in hospital, compared to placebo?

	Treatment		Control	
Study	Events	Total	Events	Total
Bussel (1990a)	20	61	23	65
Chirico (1987)	2	43	8	43
Clapp (1989)	0	56	5	59
Conway (1990)	8	34	14	32
Fanaroff (1994)	186	1204	209	1212
Haque (1986)	4	100	5	50
Ratrisawadi (1991)	10	68	13	34
Sandberg (2000)	19	40	13	41
Tanzer (1997)	3	40	8	40
Weisman (1994a)	40	372	39	381

Notes

Sepsis results



Notes

Sepsis results: DIC (from JAGS)

Model	$\overline{D}$	p_D	DIC
Common effect	115.7	11	126.7
Random effects	102.7	17.5	120.3

- The random effects model fits the data more closely: smaller $\overline{D}$ indicates better fit
- However, the common effect model is less complex: the effective number of parameters p_D is smaller

Overall, DIC prefers the random effects model, with a fairly substantial difference between the two DICs.

Notes

Extensions of Bayesian meta-analysis

- Allow baseline risks α_s and treatment effects μ_s to be correlated (more ill patients get more benefit?)
- Outcomes on different scales e.g. diagnostic tests[1]
- Combine individual patient data and aggregate trial results[2]
- Meta-regression relates the size of treatment effects to study covariates X_s. e.g.

$$\mu_s \sim N(\delta + \beta X_s, \tau^2)$$

- Modelling publication bias, non-exchangeable studies etc etc[3]

[1] Rutter and Gatsonis (Stat. Med., 2001)
[2] Riley *et al.* (Stat. Med., 2007, 2008)
[3] Sutton and Higgins (Stat. Med., 2008)

Notes

Meta-analysis summary

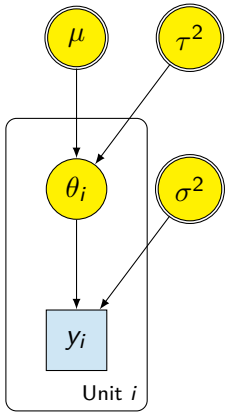
Hierarchical data asssumption	Meta analysis model
Identical parameters	Common efffect meta-analysis
Assume the parameters are identical, so just pool all the data/studies together	
Independent parameters	Study-specific models/raw data
Assume the parameters are unrelated, so analyse each study completely separately	
Exchangeable parameters	Random effects meta-analysis
Assume the parameters are 'similar' (or 'exchangeable'), use a hierarchical model	

Notes

Priors for hierarchical models

Normal-normal hierarchical model

$$y_i \sim \mathrm{N}(\theta_i, \sigma^2)$$
$$\theta_i \sim \mathrm{N}(\mu, \tau^2)$$



Three parameters that need prior distributions

1. σ^2 – variance of residual
2. μ – the population mean
3. τ^2 – variance of unit-specific parameters

Usually the choice of prior for (1) and (2) is not crucial (the data are usually informative enough)
But (3) requires considerable care

Notes

Classes of priors

- Default priors/vague/reference/'noninformative' priors
  Aim to be as neutral and generic as possible
- Informative priors
  Aim to 'capture' information/knowledge that is available outside the current dataset
  - Expert elicited priors
  - Data-based priors
- Sensitivity analysis priors
  Aim to reassure users of your results that the prior is not having undue effect
  - Enthusiastic priors/sceptical priors

Notes

1. Residual variance

A Jeffreys' prior is commonly used. In terms of the standard deviation σ

$$p(\sigma) \propto \frac{1}{\sigma}$$

In terms of the precision σ^{-2},

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

This is an improper distribution, but the posterior distribution is proper.

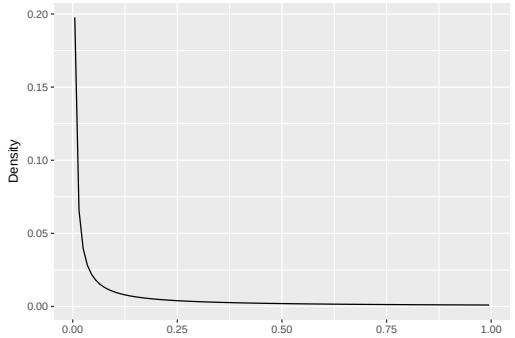
Jeffreys' prior: A recipe for finding a prior distribution that gives the same prior no matter what scale you work on (it is 'invariant to transformation')

Improper distribution: a density function that doesn't integrate to 1. Improper priors can be used as priors only if the *posterior* is proper

Notes

1. Residual variance - cont

- BUGS needs a proper distribution
- Approximate Jeffreys' prior by Gamma(ϵ, ϵ), $\epsilon = 0.001$, say.



Density between 0 and 1 shown here, but the density has support on $(0, \infty)$

An alternative choice is $\sigma \sim \text{Unif}(0, A)$ for large A (more interpretable scale? useful for informative priors!)

Notes

2. Location parameters

- **Informative priors**
  - Specify a range e.g. $\mu \sim \text{Unif}(a, b)$ where a and b are chosen. In BUGS: `mu ~ dunif(a, b)`
  - Specify quantiles, then identify a distribution with those quantiles
  - Implicit data method with conjugate priors (see BUGS book p90-91)
- **Vague priors**
  - Normal prior with large variance, e.g. $\mu \sim \text{N}(0, 100^2)$ In BUGS set a small precision: `mu ~ dnorm(0, 0.0001)`
  - Uniform prior with large range, e.g.

$$\mu \sim \text{Unif}(-A, A), \text{ where } A = 100$$

Important: definition of 'large' depends on the scale of the quantity you are modelling!

Notes

3. Random effect variances - vague prior

There is no vague prior that will always be best in all settings

- A widely recommended choice is a uniform prior on the standard deviation scale, e.g.

$$\tau \sim \text{Unif}(0, 100)$$

  This is a proper prior that is an approximation to the improper prior $\tau \propto 1$
- Alternatively, Gelman (Bayesian Analysis, 2006) recommends a half-t or half-normal distribution on the standard deviation scale ($\rightarrow$ see Censoring and Truncation lecture)

Notes

3. Random effect variances - Why not Jeffreys' prior?

With a Jeffreys' prior for τ: $p(\tau) \propto \frac{1}{\tau}$

- The marginal likelihood $p(y \mid \tau)$ approaches a non-zero finite value as $\tau \rightarrow 0$, because the data can never rule out $\tau = 0$
- Jeffreys' prior has infinite mass at $\tau = 0$
- $p \times \infty$ with $p \neq 0$ causes problems: leads to an improper distribution

With a 'just proper' approximation $\tau^{-2} \sim \text{Gamma}(\epsilon, \epsilon)$, with ϵ small fixed value, e.g. $\epsilon = 0.001$:

- Posterior is sensitive to choice of ϵ

Note

Jeffreys' prior (and the 'just proper' Gamma approximation) is OK for residual variances because it is usually implausible that $\sigma = 0$ (i.e. the residual variance is zero) because there is almost always noise of some kind.

Notes

Example: Schools data (Gelman, Bayesian Analysis, 2006)

Outcome SAT achievement measured by a scale 200–800, mean about 500, during tests of educational programs

Units 8 schools

Data Observed effects y_i (and SE $\hat{\sigma}$) for 8 schools: -3 (15), -1 (10), 1 (16), 7 (11), 8 (9), 12 (11), 18 (10), 28 (18)

$$
\begin{aligned}
y_i &\sim \mathrm{N}(\theta_i, \hat{\sigma}^2) \\
\theta_i^2 &\sim \mathrm{N}(\mu, \tau^2) \\
\mu &\sim N(0, 1000^2)
\end{aligned}
$$

Notes

Prior comparisons

Compare

- $\tau \sim \mathrm{Unif}(0, 100)$
- $\tau \sim \mathrm{Unif}(0, 10)$
- $\tau^{-2} \sim \mathrm{Gamma}(1, 1)$
- $\tau^{-2} \sim \mathrm{Gamma}(0.1, 0.1)$

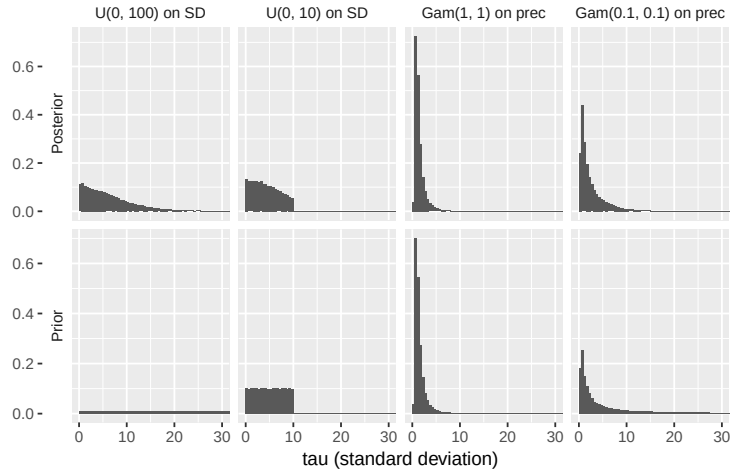
Tip:

- You might find it easier to fit all options for the priors in a single BUGS model file
- So long as each model has completely separate parameters, this will give the correct answers
- (but for a complicated model it might make MCMC updates slow).

Notes

Schools data

Priors and posteriors under the four different priors:



Note similarity of the prior and posterior for Gamma priors

Notes

Data-driven informative priors

Meta-analysis is often one of the trickiest cases – there are often very few trials – can we use previous meta-analyses to construct a reasonable prior for the random-effect variance?

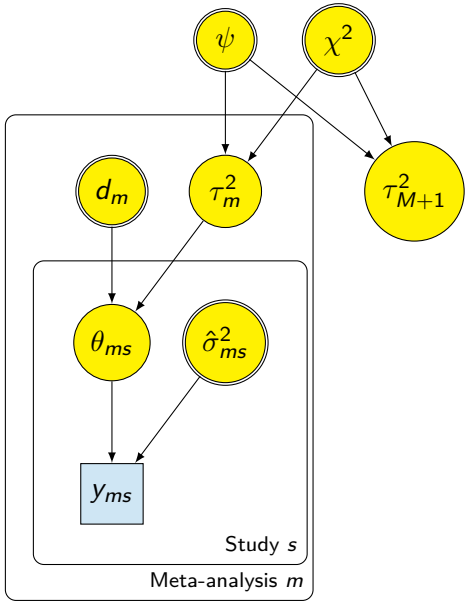
Turner *et al.* (Int. J. Epid., 2012) used data from over 14,000 meta-analyses from Cochrane Database of Systematic Reviews

1. Start with the standard random effects meta-analysis model for each meta-analysis
2. Add an additional level to the model, describing variation in τ^2
3. Obtain predictive distribution for τ^2 for a new meta-analyis

This predictive distribution can be used as an informative prior

Notes

Model for τ^2



For study $s = 1, \ldots, N_m$ within meta-analysis $m = 1, \ldots, M$

$$y_{ms} \sim \text{N}(\theta_{ms}, \hat{\sigma}^2_{ms})$$
$$\theta_{ms} \sim \text{N}(d_m, \tau^2_m)$$
$$\log(\tau^2_m) \sim \text{N}(\psi, \chi^2)$$

For a new meta-analysis ($m = M + 1$), we can predict τ^2_{M+1} in the usual way:

$$\log(\tau^2_{M+1}) \sim \text{N}(\psi, \chi^2)$$

Notes

Example data-driven priors

However, the some types of outcomes and comparisions may be more heterogenous than others.
We can add a regression model component, and provide suggested informative priors for τ^2 in each case.

	Pharmacological vs Placebo/Control	Pharmacological vs Pharmacological
All-cause mortality	Log-normal ($-4.06, 1.45^2$): median = 0.017; 95% range = (0.001–0.30)	Log-normal ($-4.27, 1.48^2$): median = 0.014; 95% range = (0.0008–0.25)
Semi-objective[b]	Log-normal ($-3.02, 1.85^2$): median = 0.049; 95% range = (0.001–1.83)	Log-normal ($-3.23, 1.88^2$): median = 0.040; 95% range = (0.001–1.58)

Notes

Priors for hierarchical models: summary

1. σ^2 – variance of residual
   - Vague: $\sigma^{-2} \sim \text{Gamma}(\epsilon, \epsilon)$, with $\epsilon = 0.001$, say. Or $\sigma \sim \text{Unif}(0, A)$.
   - Informative: $\sigma \sim \text{Unif}(0, A)$
2. μ – the population mean
   - Vague: Normal or uniform with large range
   - Informative: uniform ranges, specify quantiles, or use implicit data conjugate method
3. τ^2 – variance of unit-specific parameters
   - Requires extra care!
   - Vague: $\tau \sim \text{Unif}(0, A)$, or half-normal
   - Informative: can you construct a model using similar data?

Notes

Practical: hierarchical models

- Random and common effect meta-analysis
  - Sepsis example: exploring different models and scales
- Prior sensitivity in hierarchical models
  - Formulating an informative prior, comparing the posteriors under differing priors

For all practicals

- Follow the question sheet at the back of the course notes.
- Computer material (BUGS code/data and solutions) in choice of three formats:
  - OpenBUGS with standard Windows interface
  - JAGS with `rjags` R interface
  - OpenBUGS with `R2OpenBUGS` R interface

Notes

Other resources

- Welton, NJ, Sutton, AJ, Cooper, NJ, Abrams, KR & Ades, AE (2012) *Evidence Synthesis for Decision Making in Healthcare*. Wiley.
- NICE Technical Support Documentation for Evidence Synthesis `http://www.nicedsu.org.uk`
- Congdon, P (2006) *Bayesian Statistical Modelling*. Wiley. Chapter 5.
- Borenstein, M, Hedges LV, Higgins JPT, Rothstein, HR (2009) *Introduction to Meta-analysis*. Wiley. (mostly non-Bayesian)

Notes

Part II

Missing Data in BUGS

Notes

Missing data: overview

General regression models
Outcomes or covariates could be missing for some observations

1. Outcomes missing at random $\equiv$ simple prediction (56–57)
2. Outcomes missing not at random (58–62)
   - Aside on choosing priors, applied to logistic regression (63–65)
3. Covariates missing (67–77)

BUGS view of missing data

Missing observations treated as any other unknown parameter

- A model/prior is assumed for them
- Posterior estimated jointly with all other parameters

(BUGS Book chapter 9.1)

Notes

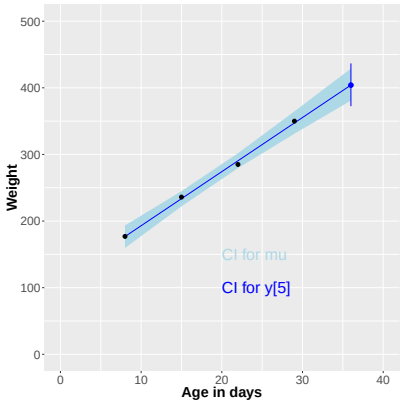
Missing outcomes: BUGS simply predicts them

- Given a model for an outcome $\mathbf{y}$ with parameters θ
- Some of the $\mathbf{y}$ data are missing so $\mathbf{y} = (\mathbf{y}^{\mathbf{obs}}, \mathbf{y}^{\mathbf{mis}})$
- $\mathbf{y}^{\mathbf{mis}}$ assumed to come from same model as $\mathbf{y}^{\mathbf{obs}}$
- Set these $\mathbf{y}^{\mathbf{mis}}$ to `NA` in the data input to BUGS
- BUGS will simultaneously
  - Estimate the posterior of $\theta|\mathbf{y}^{\mathbf{obs}}$
  - Predict from the posterior predictive distribution of $\mathbf{y}^{\mathbf{mis}}|\theta$
- Inferences for θ will be the same as if we had removed the $\mathbf{y}^{\mathbf{mis}}$ from the data

Notes

Missing outcome example: growth of a rat

```
for (i in 1:5) {
  y[i]  ~ dnorm(mu[i], tau)
  mu[i] <- alpha + beta*x[i]
}
.....

list(x = c(  8, 15, 22, 29, 36),

     y = c(177,236,285,350, NA) )
```



BUGS simply predicts the missing y_5 from posterior predictive distribution, using posterior of $\theta = (\alpha, \beta, \tau)$

$$p(y_5|y_1,\ldots,y_4) = \int p(y_5|\theta)p(\theta|y_1,\ldots,y_4)d\theta$$

Notes

Outcomes missing not at random

We assumed that the chance of an observation being missing does not depend on the true weight.

- What if heavier (or lighter?) rats tend to drop out before their weight can be measured?

Many real examples of data missing not at random

Medical Miss clinic appointments when disease less severe? Or when more severe, e.g. dementia?

Social Sensitive questions in surveys: income, political attitudes, sexual behaviour

Notes

Types of missing data

Informally [4] the chance of an observation being missing:

Missing Completely At Random : doesn't depend on observed or unobserved data

Missing At Random : doesn't depend on the unobserved value, but may depend on other observed data

Missing Not At Random : may depend on the unobserved value

These categories are a continuum: strength of any dependence may vary

[4] see e.g. Rubin 1976; Little & Rubin book; Seaman, Stat. Sci, 2013

Notes

Ignorable missing data in a Bayesian model

Define missing data indicators $\mathbf{m}$: $m_i = 1$ if y_i is missing, $m_i = 0$ otherwise. Consider a $\mathbf{m}$ as an additional outcome. If:

- data are MAR or MCAR, and
- parameters θ_y, θ_m of models for $\mathbf{y}, \mathbf{m}$ are distinct,

then the joint likelihood factorises as:

$$f(\mathbf{y}^{\mathbf{obs}}, \mathbf{m}|\theta) = f(\mathbf{y}^{\mathbf{obs}}|\theta_y) f(\mathbf{m}|\mathbf{y}^{\mathbf{obs}}, \theta_m)$$

- If, also, priors $p(\theta_y)$, $p(\theta_m)$ independent,

posterior factorises similarly
$\rightarrow$ inference for $\theta_y|\mathbf{y}^{\mathbf{obs}}$ same as under joint model for $(\mathbf{y}^{\mathbf{obs}}, \mathbf{m})$

Then missing data are ignorable.
Otherwise, need to model the missingness mechanism to avoid bias

Notes

Selection models for non-random missing outcomes

Rats: assume odds(missing) increase by 2% for extra gram of weight?

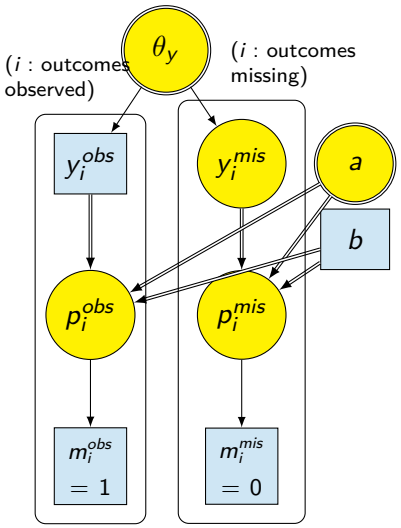
Model of interest

$$\begin{aligned} y_i &\sim N(\alpha + \beta x_i, \sigma^2 = 1/\tau) \\ \theta_y &= (\alpha, \beta, \sigma) \end{aligned}$$

Missingness model

$$\begin{aligned} m_i &\sim Bern(p_i) \\ \text{logit}(p_i) &= a + b y_i \\ \theta_m &= (a, b), \quad b = \log(1.02) \end{aligned}$$

- Joint posterior of $\left(\boxed{\theta_y, a, \mathbf{y^{mis}}} \mid \boxed{\mathbf{y^{obs}}, \mathbf{m}} \right)$ estimated
- Missing outcomes $\mathbf{y}_i^{mis}$ "imputed" given knowledge of missingness mechanism θ_m



Notes

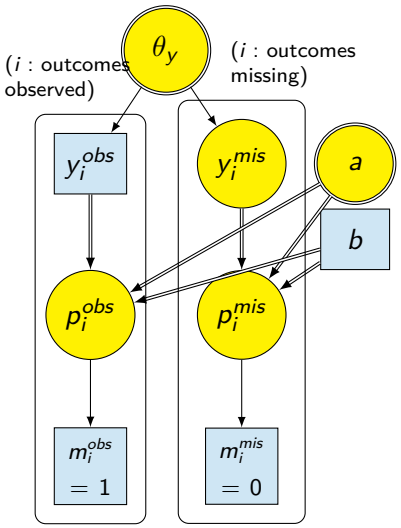
Selection models for non-random missing outcomes

Non-random missing data can't generally be estimated without knowing the missing data mechanism

Parameters of missingness model θ_m

- should usually be fixed, or
- given strong priors based on expert belief → sensitivity analysis

Posterior learning about θ_m possible in theory, but highly dependent on model for $\mathbf{y}$ (Daniels & Hogan 2008)



Notes

Non-random missing outcomes: BUGS implementation

```
for (i in 1:5) {
  y[i]   ~ dnorm(mu[i], tau)
  mu[i] <- alpha + beta*x[i]
  miss[i]      ~ dbern(p[i])
  logit(p[i]) <- a + b*y[i]
}
b             <- log(1.02)
a              ~ dlogis(0, 1)
.....
list(x = c(  8, 15, 22, 29, 36),
     y = c(177,236,285,350, NA),
  miss = c(  0,  0,  0,  0,  1) )
```

Rat growth: results robust to non-random missing data

- posterior median of missing data point goes from about 404 to 405
- posterior median of $\hat{\beta}$ goes from about 8.11 to 8.13
- regression fits visually indistinguishable

Notes

Aside: priors in logistic regression

$$\text{logit}(p_i) = \alpha + \beta(x_i - \bar{x})$$

Priors for α and β?

Want priors which express our beliefs on an interpretable scale

Typical to use normal priors with huge variances
But what would that imply about:

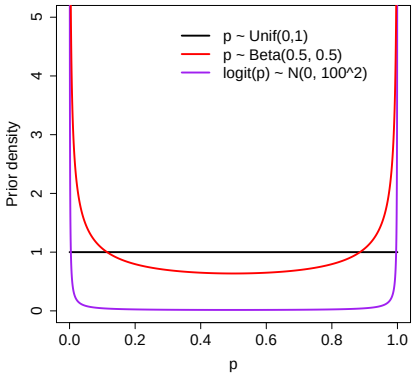
- $p = expit(\alpha) = \exp(\alpha)/(1 + \exp(\alpha))$
  - = outcome probability for average covariate value $x_i = \bar{x}$?
- odds ratio $\exp(\beta)$ for 1 unit of the covariate?

Notes

Prior for intercepts α in logistic regression

$$\text{logit}(p_i) = \alpha + \beta(x_i - \bar{x})$$

- $\alpha \sim N(0, \omega^2)$ for large ω^2?
  - $p = expit(\alpha)$ skewed to 0 and 1
- Uniform distribution for p equivalent to logistic distribution for $\alpha = logit(p)$.
  - `dlogis(0,1)` in BUGS
  - similar shape to $N(0, 1.6^2)$
- "Jeffreys' prior" $p \sim Beta(0.5, 0.5)$ another alternative



Notes

Priors for odds ratios in logistic regression

Prior for log odds ratio β? Alternative views: could use

1. Substantive information about similar applications.
   - e.g. in epidemiology, rarely find effects as big as smoking on lung cancer (e.g. OR=40 `http://www.ncbi.nlm.nih.gov/pubmed/11700268`)
   - $\beta \sim N(0, 2.3^2)$ represents 95% probability that $OR = \exp(\beta)$ is between about $1/100$ and 100.

Larger prior variance $\rightarrow$ large probability of implausibly big odds ratios!

2. Default, weakly informative prior
   - e.g. Gelman, Jakulin et al. (Ann. App. Stat 2008): $\beta \sim t_7(0, 2.5)$ encodes "implicit data" equivalent to half an additional success and half a failure for 1 unit of the covariate

Notes

Prior choice: general

Prior choice makes a difference in smaller datasets
Tighter priors regularise inference / give better-behaved samplers

Further resources:

- BUGS Book (chapter 5)
- `https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations` (Andrew Gelman and Stan development team)

Notes

Missing covariates: overview

- Outcome (vector) **y** and covariates (matrix) **x**
- Some of the covariates may be missing intermittently
  - again set to `NA` in BUGS or R data input

Outcome	Covariates		
y_1	x_{11}	x_{12}	...
y_2	x_{12}	NA	...
y_3	NA	x_{23}	...
...			
y_n	x_{1n}	x_{2n}	...

- Again, treat missing data as random variables
- Build an additional covariate model with the partially observed covariate(s) as the outcome

Notes

Harms of ignoring missing covariates

Loss of efficiency/power if exclude all data from cases with partially-observed data (e.g. some but not all covariates missing)

Bias: complete-case analysis with outcome **y** and missing covariates **x** is

- unbiased if missingness in **x** is MCAR, or MAR/MNAR dependent on **x**
- biased if missingness in **x** is MAR dependent on **y**, or MNAR dependent on **x** and **y**

(White & Carlin, Stat. Med. 2010)

→ Can be alleviated by imputing missing covariates using a model

Notes

Example: low birth weight and trihalomethanes (THMs) in water

- Outcome: Low birth weight (binary) y_i
- Covariates: THM exposure $> 60\mu$g/L, baby's sex, non-white ethnicity, smoking during pregnancy, deprived local area (all binary).
- Question: Relative odds of low birth weight associated with excess THM exposure?

80% of data on smoking and ethnicity missing: important confounders

(see BUGS Book chapter 9.1.4: simulated data following Molitor et al. 2009)

Notes

Imputing a single missing covariate

Analysis model (ignoring ethnicity for the moment)

$$\text{logit}(P(\text{LBW}_i)) = \alpha + \beta_T \text{THM}_i + \beta_D \text{dep}_i + \beta_S \text{smoke}_i$$

Covariate model: $\text{logit}(P(\text{smoke}_i)) = \gamma + \delta_T \text{THM}_i + \delta_D \text{dep}_i$

- smoking status depends on THM and area deprivation

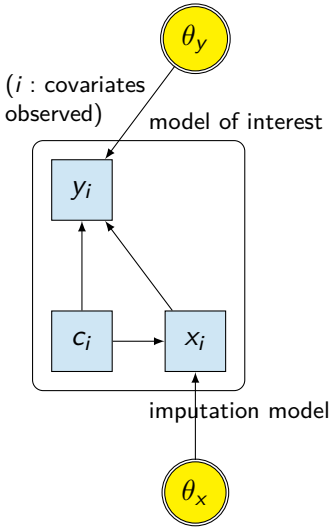
Generally, should include in the covariate model

- predictors of the missing data
- predictors of the chance of being missing (deprivation here)

Notes

Bayesian imputation of missing covariates

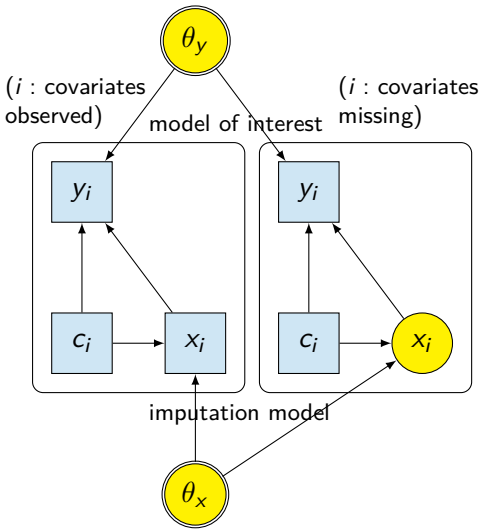
- Covariates with missing data: x_i (smoking here)
- Covariates used for imputation: c_i (THM, deprivation)
- Parameters θ_x of covariate model learnt from completely-observed data



Notes

Bayesian imputation of missing covariates

- Estimated θ_x provides a "strong prior" for missing elements x_i.
- Posterior of missing x_i combines prior with "likelihood" from y_i
  - feedback from y_i.
- The missing x_i are essentially a "parameter" of the regression model for y_i
  - in the same way as the θ_y
  - $\rightarrow$ "estimated" from y_i



Notes

Bayesian imputation of missing covariates

- Don't include outcome y as a predictor c in a Bayesian covariate imputation model.
- Influence of outcome on (missing) covariate is always implicitly included through the model for the outcome.

Notes

BUGS implementation

```
for (i in 1:n) {
  ## model of interest for low birth weight
  lbw[i]       ~ dbern(p[i])
  logit(p[i]) <- beta[1] +
                 beta[2]*THM[i] +
                 beta[4]*dep[i] +
                 beta[5]*smk[i]

  ## model for "imputing" missing smoking covariate
  smk[i]       ~ dbern(q[i])
  logit(q[i]) <- delta[1] + delta[2]*THM[i] + delta[3]*dep[i]
}
... [ priors ...]
```


Notes

Imputing multiple missing covariates

- Accounting for correlation between multiple covariates improves imputation
- Example: smoking and ethnic group both intermittently missing in the low birth weight example

Simplest approach for few covariates: conditional regressions

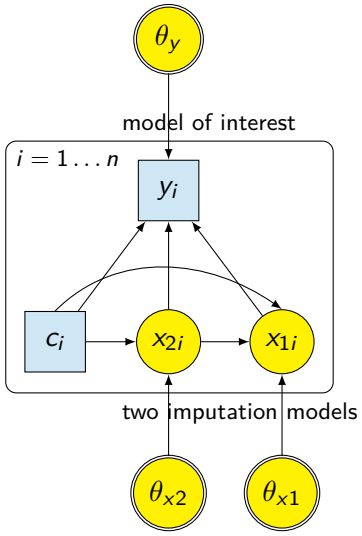
- model each incomplete variable in turn, conditionally on the remaining unmodelled incomplete variables, e.g.:

$$\begin{aligned} \text{logit}(P(\text{smoke}_i)) &= \gamma + \delta_{1T}\text{THM}_i + \delta_{1D}\text{dep}_i + \delta_{1E}\text{eth}_i \\ \text{logit}(P(\text{eth}_i)) &= \gamma + \delta_{2T}\text{THM}_i + \delta_{2D}\text{dep}_i \end{aligned}$$

- or for three incomplete variables x_1, x_2, x_3 and x_o complete, model $P(x_1|x_2, x_3, x_o)$, then $P(x_2|x_3, x_o)$, then $P(x_3|x_o)$.

Notes

Feedback when imputing multiple missing covariates



x_{1i}, x_{2i} observed for some i and missing for others (rough DAG notation)

Note again, just as the response y_i feeds back through graph to x_{1i}:

- information about x_{1i} feeds back to improve estimation of x_{2i}
- don't include x_{1i} in the imputation model for x_{2i}

Notes

Alternative: multivariate model for missing covariates

Cleaner approach when many covariates have missing data
Vector of covariates $\mathbf{x}_i$ for observation i:

- Multivariate normal model if covariates continuous / normal

$$\mathbf{x}_i \sim MVN(\boldsymbol{\mu_i}, \Sigma)$$

- Latent variable model if mixture of binary / continuous, e.g.

$$\mathbf{U}_i \sim MVN(\boldsymbol{\mu_i}, \Sigma)$$

$$r\text{th covariate } x_{ir} = \begin{cases} u_{ir} \\ I(u_{ir} > 0) \end{cases} \quad \text{if } x_{ir} \begin{cases} \text{continuous} \\ \text{binary} \end{cases}$$

- Model μ_i in terms of predictors

See BUGS Book 9.1.4 for an example with two missing binary variables

Notes

Other issues with missing covariates

- Bayesian approach similar to classical multiple imputation (MI)
  - MI: create multiple datasets with different values for missing variables based on the imputation model → analyse each dataset separately → pool results
  - Bayesian approach: achieve same aim by sampling from full joint posterior of parameters of interest and missing variables
  - MI: include outcome explicitly in imputation model
    Bayesian: outcome included through feedback
- Covariates missing not at random
  - same principles as for missing outcomes
  - model missingness status in terms of predictors of missingness
- Missing data in hierarchical models
  - imputation model should respect hierarchical structure.

Notes

Other resources

- Daniels MJ and Hogan JW (2008) *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis* Chapman & Hall / CRC.
- Former Imperial College course on missing data with BUGS (Nicky Best & Alexina Mason), material available at `http://www.bias-project.org.uk/Missing2012/MissingIndex.htm`
- Mason, Best, Richardson & Plewis, "Strategy for modelling non-random missing data mechanisms in observational studies using Bayesian methods" J. Official Stat (2012)
- `http://onlinelibrary.wiley.com/doi/10.1002/sim.6944/abstract` comparison between Bayesian and classical multiple imputation

Notes

Practical: missing data

Missing data practical

1. Outcomes missing not at random
   - choosing priors, coding/interpreting a model for missingness.
2. Missing covariate imputation
   - coding/interpreting a model for imputing missing covariates

Notes

Part III

Censoring and truncation

Notes

Censoring and truncation: overview

- Censored data
  - Left/right/interval censoring
  - Rounded / grouped data
- Truncated data or parameters
  - Difference from censoring
- Specifying new distributions

How to implement in WinBUGS, OpenBUGS and JAGS

Illustrated in BUGS Book, chapter 9.

Notes

Censoring: definition

A data point x is (interval-)censored when we

don't know its exact value...
but we know it is within some interval: $x \in (a, b)$

- Right-censoring: we just know $x > a$ (thus $b = \infty$, assuming x comes from unbounded distribution)
- Left-censoring: we just know $x < b$ (thus $a = -\infty \ldots$)

Censored data are partially observed:

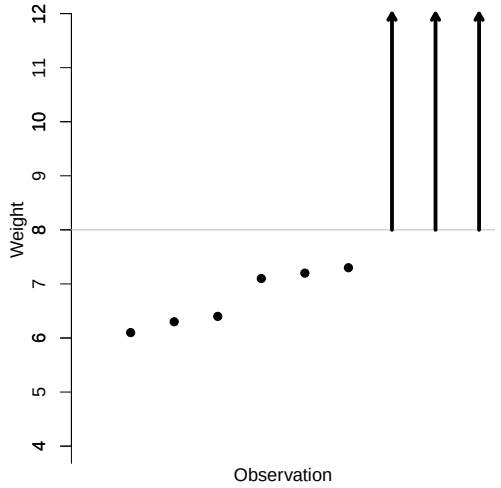
- treated as unknowns in BUGS, given a distribution and estimated, but
- they always contribute information, unlike missing data.

Notes

Censoring: example

Weigh 9 chickens, but the scale only goes up to 8 units

Observed data: 6.1, 6.3, 6.4, 7.1, 7.2, 7.3, 8+, 8+, 8+



- Don't believe all weights from 8 to ∞ equally likely
- Learn about both observed and censored data through modelling
- Could assume censored data from same distribution as observed data
  - assuming non-informative censoring, as missing data

Notes

Joint model for observed and censored data

Let $\{y_i^* : i = 1, \ldots 9\}$ be the true underlying data.

- Assume generated as $y_i^* \sim f(y|\mu)$

What we observe instead is y_i, where

- $y_i = y_i^*$ for $i = 1, \ldots 6$
- $y_i = c_i = 8$ for $i = 7, 8, 9$.

Likelihood contribution: $\prod_{i=1}^{6} f_i(y_i|\mu) \prod_{i=7}^{9} (1 - F_i(y_i|\mu))$

- where $f()$ is the PDF and $F()$ is the CDF

Thus we learn about μ through both observed and censored data

- $\rightarrow$ we can predict the true values of the censored data.

Notes

Censoring: WinBUGS and OpenBUGS syntax (1)

```
model {
 for (i in 1:6) { y[i] ~ dnorm(mu, 1) }
 for (i in 7:9) { y[i] ~ dnorm(mu, 1)I(8,) }
 mu ~ dunif(0, 20)
}
# data
list(y = c(6.1, 6.3, 6.4, 7.1, 7.2, 7.3, NA, NA, NA))
```

- Censored data points included as `NA` in data.
- `I(a,b)`: All values sampled outside `(a,b)` are rejected
- `C()` preferred as an equivalent to `I()` in OpenBUGS
  - to emphasise it is for censoring, not truncation — more later...

Notes

Censoring: WinBUGS and OpenBUGS syntax (2)

More generally, supply censoring points with data

- allows different censoring points for different observations

```
model {
 for (i in 1:6) { y[i] ~ dnorm(mu, 1) }
 for (i in 7:9) { y[i] ~ dnorm(mu, 1)I(c[i],) }
 mu ~ dunif(0, 20)
}
# data
list(y = c(6.1, 6.3, 6.4, 7.1, 7.2, 7.3, NA, NA, NA),
     c = c(0,   0,   0,   0,   0,   0,    8,  8,  8))
```

e.g. survival analysis: some people observed to die, some still alive at end of study

- observed deaths supplied in the data vector `y[]`
- censoring times supplied in the data vector `c[]`

Notes

JAGS syntax: right-censoring at common time

```
model {
 for (i in 1:9) {
  y[i]           ~ dnorm(mu, 1)
  is.censored[i] ~ dinterval(y[i], 8)
 }
 mu ~ dunif(0, 20)
}

data <- list(
    y           = c(6, 6, 6, 7, 7, 7, NA, NA, NA),
    is.censored = c(0, 0, 0, 0, 0, 0, 1,  1,  1) )
```

Here `break` is a single number

`dinterval(y, break)` is $\begin{cases} 0 \text{ if } \texttt{y[i]} <= \texttt{break} \text{ (uncensored)} \\ 1 \text{ if } \texttt{y[i]} > \texttt{break} \text{ (censored)} \end{cases}$

Notes

Censoring in JAGS: Why the strange syntax? (advanced)

`is.censored[i] ~ dinterval(y[i], break)` ?

- Syntax emphasises that censoring status is a kind of observation.

Note $\sim$ is used instead of -> to define `is.censored`

- `is.censored` is a "piece of data" which we use to learn about the unknown "parameter" of `dinterval` (the `y[i]` and hence μ)
- `dinterval()` is an observable deterministic function
  - in contrast with e.g. `z <- sqrt(x)`: can't supply data on `z` to learn about `x`

Notes

JAGS syntax: censoring in general

Could also have different censoring times for each i

- e.g. survival analysis: observed deaths supplied in `y[i]`, censoring times in `c[i]`
- for `i` where deaths observed, define `c[i] > y[i]` arbitrarily.

```
model {
 for (i in 1:9) {
  y[i] ~ dnorm(mu, 1)
  is.censored[i] ~ dinterval(y[i], c[i])
 ...
```

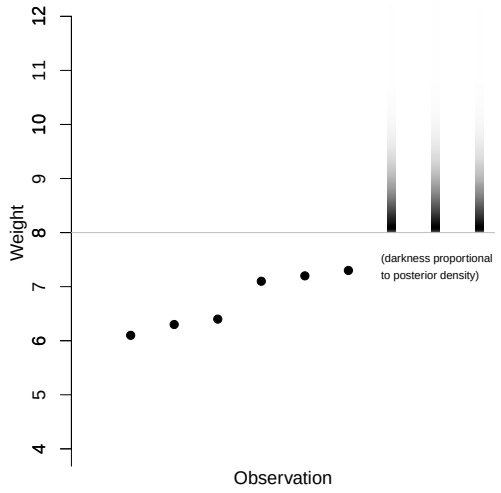
Or interval-censoring: if `break` is a vector:

`is.censored(t, break)` is

- 0 if `t[i] <= break[1]` (uncensored)
- 1 if `break[1] <= t[i] < break[2]` (censored on interval 1)
- 2 if `break[2] <= t[i] < break[3]` (censored on interval 2)
- ...

Notes

Chickens: results



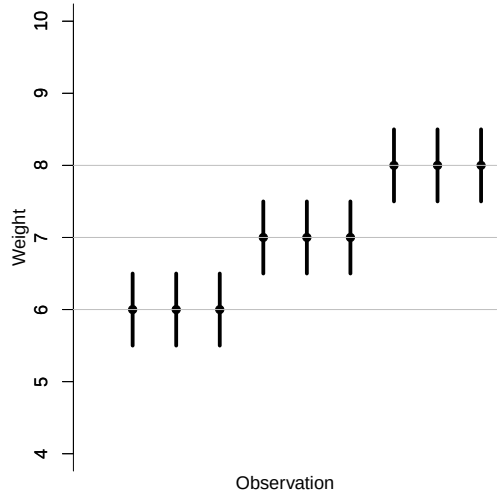
- $y_i \sim N(\mu, 1)$, posterior median $\hat{\mu} = 7.4$ (6.7, 8.0)
- Censored observations: posterior median 8.5 (8.0, 9.9)

Notes

Rounded data

Nine chickens are weighed as before, on a new set of scales, which

- can go over 8,
- but only report integer values.



- We observe 6, 6, 6, 7, 7, 7, 8, 8, 8
- Treat as interval-censored on $[5.5, 6.5]$, $[6.5, 7.5]$ or $[7.5, 8.5]$
- Same BUGS implementation as for interval-censoring...

Notes

Rounded data in BUGS

- Could implement rounding via interval-censoring, as before
- JAGS has a convenient shorthand:
  - `yround[i] ~ dround(y[i], digits)`

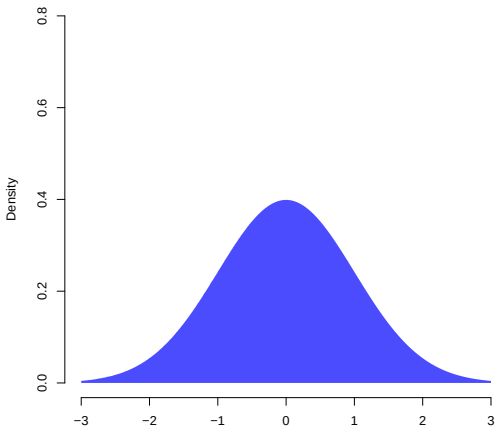
  means that `yround[i]` is `y[i]` rounded to `digits` digits.

```
model {
 for (i in 1:9) {
  y[i]      ~ dnorm(mu, 1)
  yround[i] ~ dround(y[i], 0)

 ...
## Data
list(yround = c(6, 6, 6, 7, 7, 7, 8, 8, 8))
## Initial values
list(mu=7, y=c(6, 6, 6, 7, 7, 7, 8, 8, 8))
```

- Sensible to specify initial values for `y`: auto-generated ones may be inconsistent with `yround`.

Notes

Truncated distributions

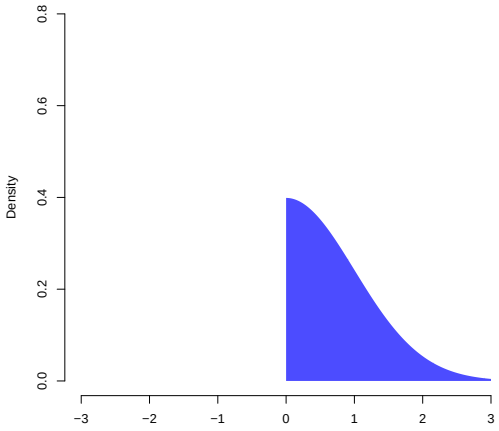


Standard normal distribution

$$f(y) = \frac{1}{\sqrt{2\pi}} exp(-y^2)$$

Notes

Truncated distributions



Truncated normal distribution?

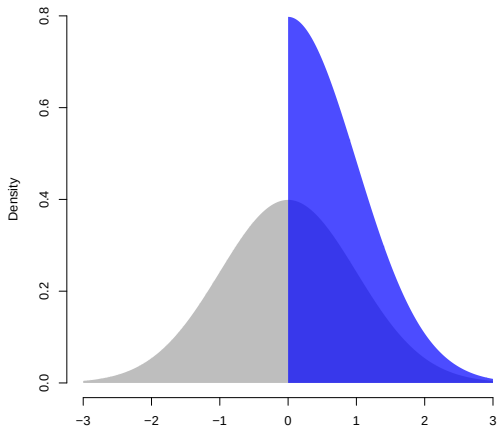
$$f(y) = \frac{1}{\sqrt{2\pi}} exp(-y^2)$$

for $y \in [U, L]$, 0 otherwise??

No — need normalising constant

Notes

Truncated distributions



Truncated standard normal distribution

$$f(y|U,L) = \frac{\frac{1}{\sqrt{2\pi}} exp(-y^2)}{\Phi(U) - \Phi(L)}$$

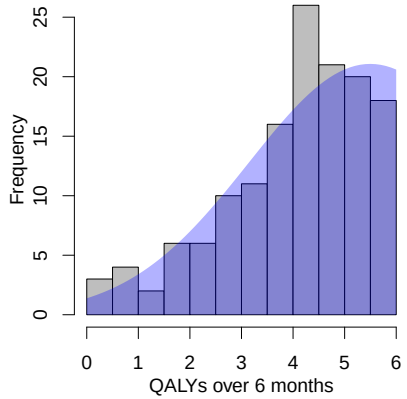
for $y \in [U, L]$, 0 otherwise, where $\Phi()$ is the Normal cumulative distribution.

- Ensures $\int_{-\infty}^{\infty} f(y|U,L) = \int_{U}^{L} f(y|U,L) = 1$.

Notes

Use of truncated distributions

- As a prior distribution
  - e.g. half-normal or half-t prior for standard deviation of random effects, as recommended by Gelman (Bayesian Analysis, 2006)
- To model data, e.g.



- Quality-adjusted life (over 6 months) from lung cancer patients [a]
- Survival over 6 months adjusted by patient-reported health-related quality of life
- Bounded from 0–6 months by definition

[a] Sharples et al., (2012) Health Technology Assessment, 16(18)

Notes

Truncated distributions in JAGS and OpenBUGS

JAGS: Any distribution can be truncated by using `T()`

e.g. for normal distribution truncated to [0, 6]

```
y ~ dnorm(mu, tau)T(0, 6)
```

OpenBUGS: A few distributions can be truncated by using `T()`

in the same way, though this is not documented.

Restricted to `dbeta, dexp, dnorm, dpois, dt, dweib`

Notes

Truncated distributions in WinBUGS

WinBUGS doesn't support the `T()` construct for truncation

Can we use the `I()` construct instead?

Sometimes...

Notes

When can we use I() for truncation in WinBUGS

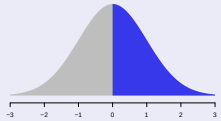
(note I() simply rejects all samples outside the range)

YES: When distribution used as a prior

Parameters are fixed, e.g.

```
sigma ~ dnorm(0, 1)I(0,)
```

Resulting sample after rejecting out of range values will have the correct truncated distribution



Notes

When can we use I() for truncation in WinBUGS

(note I() simply rejects all samples outside the range)

NO: When distribution used as a likelihood

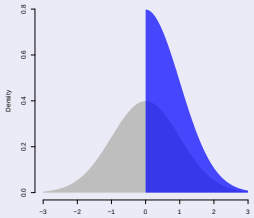
e.g. want to learn the parameters mu and tau from data y[]

```
y[i] ~ dnorm(mu, tau)I(l,u)
```

Wrong distribution sampled for mu, tau.

Likelihood $l(\mu, \tau|\mathbf{y})$ will ignore the normalising constant that depends on μ, τ:

$$f(y|u, l) = \frac{\sqrt{\frac{\tau}{2\pi}} exp(-\tau(y-\mu)^2)}{\Phi(u|\mu, \tau) - \Phi(l|\mu, \tau)}$$



Notes

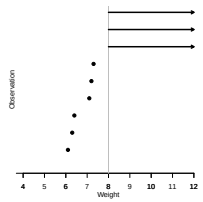
How do we give a truncated model to data in WinBUGS?

The "zeros trick" or "ones trick" ...

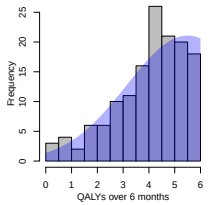
Notes

Summary: Difference between censoring and truncation

Censoring: data is partially observed: known to lie in a range



Truncation: impossible to observe data outside the range:



"Would I make the same assumption if predicting a new observation?"

No Censoring
Yes Truncation

Notes

Specifying a new sampling distribution in BUGS

Model data $y_i \sim f(y|\theta)$, a distribution that is not built in to BUGS, with parameters θ

The zeros trick (WinBUGS, OpenBUGS)

```
for (i in 1:n) {
  z[i] <- 0
  z[i] ~ dpois(phi[i])
  phi[i] <- -logL[i] + C
  logL[i] <- # code for log f(y[i], theta)
}
C <- 10000 ## (or any large number to ensure phi[i]>0)
```

How does it work? If $z \sim Poisson(\phi)$, and observe $z = 0$, then

$$\begin{aligned} p(z) &= \phi^z exp(-\phi)/z! = exp(-\phi) \\ \phi &= -log\{p(z)\} \end{aligned}$$

So if we define ϕ using the log-likelihood from the data $y|\theta$, then $z = 0$ will contribute the correct likelihood to the posterior $p(\theta|z)$.

Notes

The ones trick, alternative to the zeros trick

The ones trick (WinBUGS, OpenBUGS)

```
for (i in 1:N) {
   z[i]   <- 1
   z[i]   ~ dbern(phi[i])
   phi[i] <- L[i] / C
   L[i]   <- # code for f(y[i], theta)
}
C <- 10000 ## (or any large number to ensure p[i]<1)
```

Similarly, for $z = 1$, $p(z) = \phi$. So if define ϕ using the likelihood from the real data y,

- $z = 1$ contributes the correct likelihood to the posterior $p(\theta|y)$.

Notes

OpenBUGS and JAGS variations

OpenBUGS also has a shortcut `dloglik`

```
z[i] <- 0
z[i] ~ dloglik(logL[i])
logL[i] <-  # code for log f(y[i], theta)
```

In JAGS, can't supply both data and a model for a node like `z[i]` in the model block

- Supply zero values for `z` in the data list
- Or use a `data` block:

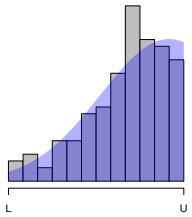
```
data {
 for (i in 1:n){
   z[i] <- 0
 }
}
model {
 ...
```


Notes

Truncated normal model using zeros trick

If $y \sim TN(\mu, \sigma^2, L, U)$, truncated Normal on (L, U)

$$f(y|\mu, \tau, U, L) = \frac{\sqrt{\frac{\tau}{2\pi}} exp(-\tau(y-\mu)^2)}{\Phi(U|\mu, \tau) - \Phi(L|\mu, \tau)}, \quad \tau = 1/\sigma^2$$




```
logL[i] <- 0.5*log(tau) -
           0.5*tau*(y[i] - mu[i])*(y[i] - mu[i]) -
           log(phi((U - mu[i])/sig) -
               phi((L - mu[i])/sig))
tau <- 1/(sig*sig)
```


Notes

Specifying a new prior distribution

Define $\theta \sim f()$, a prior distribution that is not built-in.
Zeros trick applies again, with extra line:

WinBUGS and OpenBUGS

```
for (i in 1:n) {
  z[i] <- 0
  z[i] ~ dpois(phi[i])
  phi[i] <- -logL[i] + C
  logL[i] <- # code for log f(theta)
}
theta ~ dflat()
```

Dummy $z = 0$ and an improper uniform prior $\theta \sim 1$ give the correct "likelihood" for `theta` (i.e. the desired prior)

JAGS has no `dflat`, but can substitute `dunif(-U, U)` where U is big

Notes

Defining new distributions (and functions) using low-level programming (advanced)

WinBUGS WBDev. Written in Component Pascal
`http://winbugs-development.mrc-bsu.cam.ac.uk/wbdev.html`

OpenBUGS UDev. Essentially same as WBDev
`UDev/Docu` subdirectory in OpenBUGS installation

JAGS : possible but no official manual. Written in C++
`http://www.cidlab.com/prints/wabersich2014extending.pdf` gives tutorial for defining distributions

Advantages over zeros trick:

- more compact code, quicker execution
- . . . especially with complicated functions involving many nodes

Notes

Practical: censoring and related topics

Three questions with slight modifications of the chickens example — do in any order, depending what you want to practise.

- Censoring
- Truncation
- Zeros trick

All ask you to write BUGS code from scratch, but feel free to work through the solutions document instead.

Fourth example: Weibull survival models

(a) Choice of priors for Weibull parameters

(b) Time-dependent covariates demonstration (JAGS only)

Notes

Part IV

Measurement error

Notes

Measurement error: overview (30 mins)

> "For my money, the #1 neglected topic in statistics is measurement ... Bad things happen when we don't think seriously about measurement"
> *(Prof. Andrew Gelman, Columbia* : `http://andrewgelman.com`*)*

- Data we want is not exactly the data we can measure
- We can model the discrepancy in BUGS
  - given data / beliefs about the measurement process
- Can cause bias if don't adjust for measurement error
  - particularly covariates in regression models
- Examples (BUGS Book chapter 9.3)
  - binary data / diagnostic tests
  - continuous variables: "classical" vs "Berkson" error

Notes

Examples of measurement error: continuous

- Measuring instrument is noisy: ("classical" error)

  Interested in: "usual" daytime blood pressure (e.g. SBP$>$ 140 $\rightarrow$ hypertension)
  Measure: fluctuates throughout the day, instrument tricky to use

- Measuring instrument coarsens the data: ("Berkson" error)

  Interested in: "usual" pollution level at particular address
  Measure: pollution levels averaged over area, using smoothed data from nearby monitors

Notes

Examples of measurement error: binary

Diagnostic test to detect something (like a disease): yes/no outcome.

- Can fail to detect something that's there
  - false negatives, imperfect sensitivity
- Can "detect" something that's not there
  - false positives, imperfect specificity

Binary measured outcome and binary quantity of interest

Notes

Covariate measurement error biases regression models

- Causes attenuation of regression slope towards null
- Ideally model the error, using either
  - "gold standard" data with both observed and true
  - prior on error process

  or simply do sensitivity analysis

Notes

Measurement error in outcomes?

- Handled naturally through choice of distribution in standard models
- e.g. $y_i = \alpha + \beta x_i + \epsilon_i$, error term $\epsilon_i \sim N(\mu, \sigma^2)$.
- Though be careful the distribution and other model choices (e.g. covariate selections, interactions) are appropriate.

Notes

Binary covariate misclassification: cervix example

Case-control study, logistic regression with:

Outcome Cervical cancer

Covariate Exposure to herpes simplex virus (HSV).

- HSV measured by inaccurate "western blot" test alone for 1929 women
- Have "gold standard" test (as well as inaccurate test) for 115 women

Notes

Cervix: data

Accurate HSV z	Inaccurate HSV x	Disease d	Count
Gold standard data			
0	0	1	13
0	1	1	3
1	0	1	5
1	1	1	18
0	0	0	33
0	1	0	11
1	0	0	16
1	1	0	16
Incomplete data			
	0	1	318
	1	1	375
	0	0	701
	1	0	535

- Use complete data to estimate P(true HSV) given
  - inaccurate HSV test result
  - and disease status ("differential error")
- Predict HSV status for each person with incomplete data

Notes

Model parameters

- HSV test errors, allowing differential error between people with/without cancer d

P(diagnose HSV | no HSV) 1-specificity, or false positive probability

$$\begin{aligned} \phi_{11} &= P(x = 1|z = 0, d = 0) \\ \phi_{12} &= P(x = 1|z = 0, d = 1) \end{aligned}$$

P(diagnose HSV | true HSV) sensitivity, or 1 - false negative probability

$$\begin{aligned} \phi_{21} &= P(x = 1|z = 1, d = 0) \\ \phi_{22} &= P(x = 1|z = 1, d = 1) \end{aligned}$$

- β_0: Baseline log odds of cervical cancer
- β: Log odds ratio of cervical cancer between HSV / non-HSV
- ψ: Prior probability of true HSV presence

Notes

JAGS code for cervix model

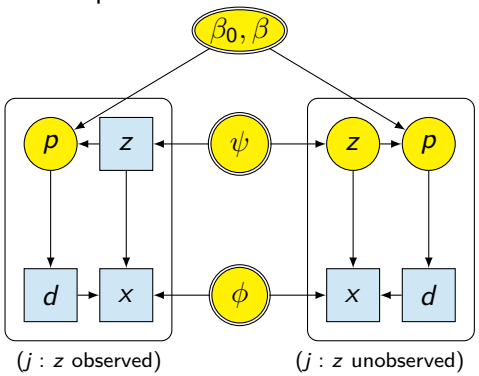
```
model{
  for (i in 1:n) { # individuals
    d[i]          ~ dbern(p[i])
    logit(p[i]) <- beta0 + beta*z[i]
    z[i]          ~ dbern(psi)
    # index j in phi[j,..] has to start
    # at 1, not 0
    x[i]    ~ dbern(phi[z[i]+1, d[i]+1])
  }
  for (j in 1:2) {
    for (k in 1:2) {
      phi[j, k]  ~ dunif(0, 1)
    }
  }
  psi            ~ dunif(0, 1)
  beta0          ~ dlogis(0, 1)
  # Cauchy with scale 2.5 (Gelman 2008)
  beta           ~ dt(0, 0.16, 1)
}
```

- Error rates `phi` estimated from observed true HSV `z[i]` and misclassified HSV `x[i]`
- Unknown `z[i]` estimated from likelihood $p(x_i|z_i)$ given observed x_i and prior HSV probability ψ
- Thus we can estimate the log odds ratio β from
  - Observed disease status `d[i]`
  - Observed and estimated HSV exposure `z[i]`

Notes

JAGS code for cervix model

DAG representation



- Error rates `phi` estimated from observed true HSV `z[i]` and misclassified HSV `x[i]`
- Unknown `z[i]` estimated from likelihood $p(x_i|z_i)$ given observed x_i and prior HSV probability ψ
- Thus we can estimate the log odds ratio β from
  - Observed disease status `d[i]`
  - Observed and estimated HSV exposure `z[i]`

Notes

Modification of cervix model for WinBUGS/OpenBUGS

Replace

```
x[i]    ~ dbern(phi[z[i]+1, d[i]+1])
```

by

```
x[i]    ~ dbern(phi[z1[i], d1[i]])
z1[i]       <- z[i] + 1
d1[i]       <- d[i] + 1
```

since WinBUGS / OpenBUGS don't allow functions of unknowns inside indices

Notes

Cervix: results

When measurement error ignored, estimate of odds ratio $\exp(\beta)$ is

- attenuated towards null
- with uncertainty underestimated

	Posterior median (95% CI)
Error-adjusted model	1.78 (0.94–3.47)
Naive regression on misclassified x	1.57 (1.31–1.89)

Estimates of ϕ from error-adjusted model

- substantial misclassification
- differential misclassification between controls/cases

	Controls	Cases
False positive probability	0.32 (0.21–0.42)	0.21 (0.08–0.40)
HSV test sensitivity	0.57 (0.44–0.69)	0.76 (0.63–0.89)

Notes

Classical error in continuous variable: dugongs

Growth of dugongs (sea cows)

- Length y_j (in m) modelled as nonlinear function of age

$$y_j \sim N(\mu_j, \sigma^2), \quad \mu_j = \alpha - \beta\gamma^{z_j}$$

- True age z_j, observed age x_j, with known measurement SD $\omega = 1$:

$$x_j \sim N(z_j, \omega^2)$$

- Vague priors $z_j \sim U(0, 100)$ for $j = 1, \ldots, n$.

```
for(j in 1:n) {
  y[j]    ~ dnorm(mu[j], tau)
  mu[j] <- alpha -
           beta*pow(gamma, z[j])
  x[j]      ~ dnorm(z[j], 1)
  z[j]      ~ dunif(0, 100)
}
alpha       ~ dunif(0, 100)
beta        ~ dunif(0, 100)
gamma       ~ dunif(0, 1)
tau         <- 1/sigma2
log(sigma2) <- 2*log.sigma
log.sigma   ~ dunif(-10, 10)
```


Notes

Classical error in continuous variable: dugongs

Growth of dugongs (sea cows)

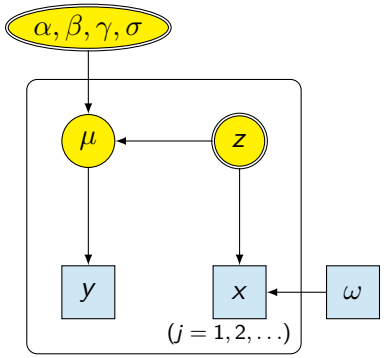
- Length y_j (in m) modelled as nonlinear function of age

$$y_j \sim N(\mu_j, \sigma^2), \quad \mu_j = \alpha - \beta\gamma^{z_j}$$

- True age z_j, observed age x_j, with known measurement SD $\omega = 1$:

$$x_j \sim N(z_j, \omega^2)$$

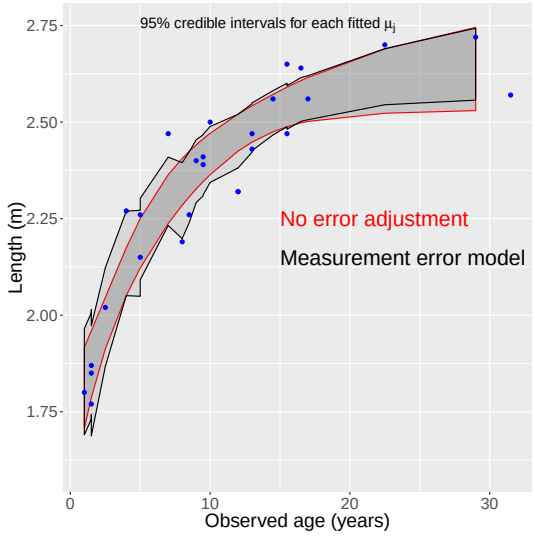
- Vague priors $z_j \sim U(0, 100)$ for $j = 1, \ldots, n$.



Measurement error 120 (108–126)

Notes

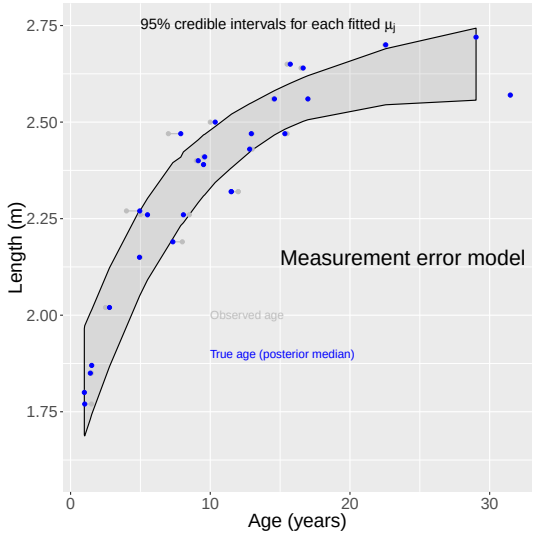
Dugongs: results



- With measurement error, posterior distribution of mean regression line is wider, fitting the noisy data better
- Measurement error model enables "true" ages to be estimated
  - less noisy

Notes

Dugongs: results



- With measurement error, posterior distribution of mean regression line is wider, fitting the noisy data better
- Measurement error model enables "true" ages to be estimated
  - less noisy

Notes

Berkson error: air pollution example

- Observed covariates less variable than the true values
  - coarsened / aggregated data

Example: effect of exposure to NO_2 on respiratory illness.

- Data from 103 children

	Bedroom NO_2 level in ppb (x)		
Respiratory illness (y)	< 20	20–40	40+
Yes	21	20	15
No	27	14	6
Total (n)	48	34	21

- Measurement error known through a calibration study: true exposure $z \sim N(\alpha + \beta x, \sigma^2)$,
  - where x is midpoint of aggregate range (or 50 for 40+).
  - crude model, interval-censoring methods also possible
- Assume (for illustration) point estimates $\alpha = 4.48, \beta = 0.76, \sigma = 9$
  - would ideally push through uncertainty about these.

Notes

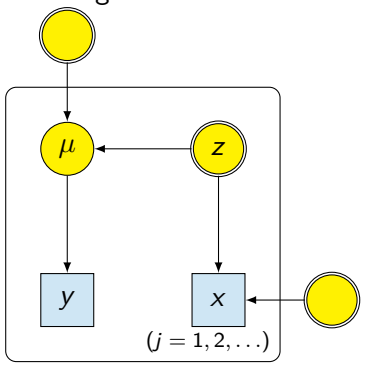
BUGS code

```
model{
 for (j in 1:3) {
   y[j]          ~ dbin(p[j], n[j])
   logit(p[j]) <- theta[1] + theta[2]*z[j]
   z[j]          ~ dnorm(mu[j], 0.01232)
   mu[j]       <- 4.48 + 0.76*x[j]
 }
 theta[1]        ~ dlogis(0, 1)
 theta[2]        ~ dnorm(0, 0.2)
 or10           <- exp(10*theta[2])
}
# data
list(y = c(21, 20, 15), n = c(48, 34, 21),
     x = c(10, 30, 50))
```

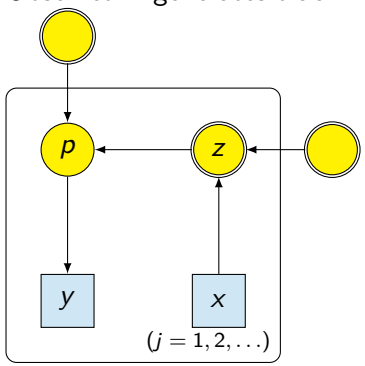

Notes

DAG — contrast classical and Berkson

Classical error (dugongs model)
True z generates observed x



Berkson error (pollution model)
Observed x generates true z



Notes

Results

Odds ratio for 10 units of pollution exposure, $\exp(10 \times \theta_2)$:

	Posterior median	95% CI
Error-adjusted model	1.40	1.00–2.58
Naive regression on midpoints	1.34	1.04–1.75

After accounting for measurement error:

- less effect dilution
- but more uncertainty

Notes

Practical: measurement error

Three questions – can do in any order

1. Binary misclassification (cervix)
2. Classical measurement error (dugongs)
3. Berkson measurement error (pollution)

All based on the examples in the lectures

- Getting used to coding and interpreting measurement error models
- Influence of model assumptions

Notes

Notes

Part V

Complex hierarchical models

Complex hierarchical models: overview

1. Hierarchical models for longitudinal data
2. Joint models for longitudinal and survival data
3. Model checking for hierarchical models

Notes

Complex hierarchical models

Many elaborations and applications of hierarchical models, e.g.

- Further levels: e.g. pupils within schools within local authorities/group of schools etc (like the data-driven prior in previous lecture)
- Spatial structure (see BUGS book p262-272)
- Non-normal random effects distributions (e.g. outlier-robust t-distributions)
- Hierarchical models for variances (see BUGS book p237-240)
- Cross-classified data e.g. pupils within primary and secondary schools

We consider longitudinal observations on units in this lecture

- measurements of the same unit (or individual) over time
- (positive) correlation between the observations

Notes

AIDS longitudinal data (Abrams *et al.*, NEJM, 1994)

Population 467 patients with AIDS during antiretroviral treatment (who had failed or were intolerant to zidovudine therapy)

Aim To compare efficacy and safety of two drugs:

- Didanosine (ddI)
- Zalcitabine (ddC)

Treatment Random assignment of ddI or ddC

Data CD4 cell counts recorded at study entry, and after 2, 6, 12 and 18 months

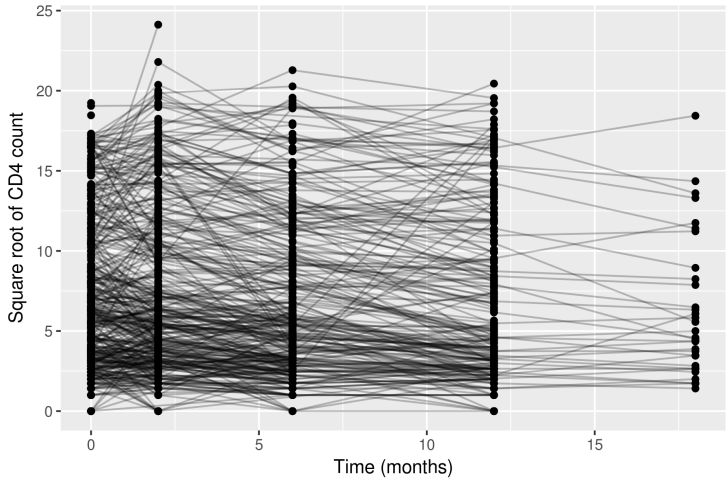
The data are quite complex

- Missing data – 188 had died by the end of the study, 59.7% censoring
- The data are on $[0, \infty)$ - i.e. positive and zeros.

For simplicity, we ignore all of these issues, and assume the missingness is ignorable, and the observations are Normal.

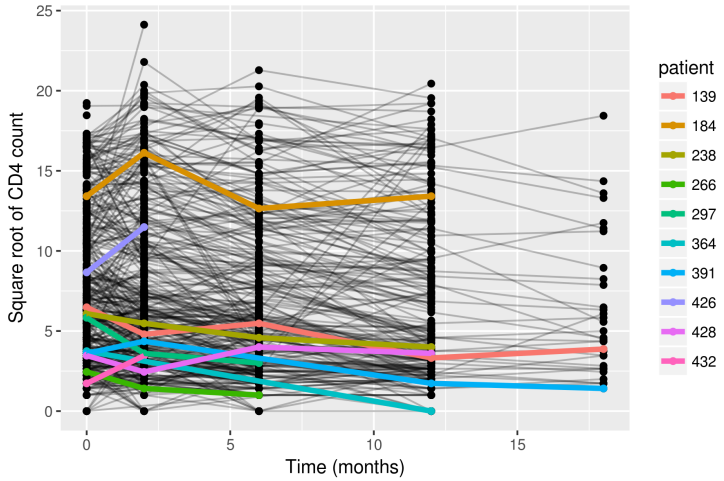
Notes

AIDS longitudinal data



Notes

AIDS longitudinal data (cont)



Notes

AIDS longitudinal data - non-hierarchical model

A non-hierarchical linear model — no account is made for the structure of the data.

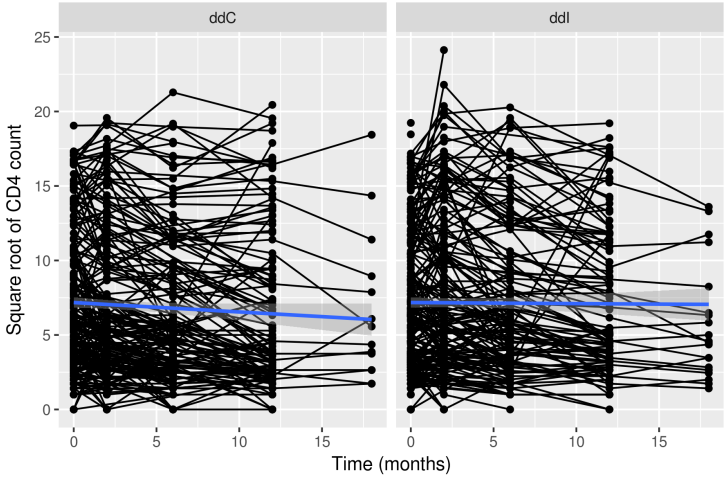
$$y_{ij} \sim \mathsf{N}(\mu_{ij}, \sigma^2) \qquad \beta_1, \beta_2, \beta_3 \sim \mathsf{N}(0, 100^2)$$
$$\mu_{ij} = \beta_1 + \beta_2 t_{ij} + \beta_3 \mathsf{drug}_i t_{ij} \qquad \sigma \sim \mathsf{Unif}(0, 100)$$

```
for (i in 1:I){
  for (j in 1:J){
    CD4[i, j] ~ dnorm(CD4.mu[i, j], CD4.sigma.prec)
    CD4.mu[i, j] <- beta1 +
                    beta2 * t.obs[j] +
                    beta3 * t.obs[j] * drug.ddI[i]
  }
}
beta1 ~ dnorm(0, 0.0001)
beta2 ~ dnorm(0, 0.0001)
beta3 ~ dnorm(0, 0.0001)
CD4.sigma.prec <- 1/pow(CD4.sigma.sd, 2)
CD4.sigma.sd ~ dunif(0, 100)
```


Notes

AIDS non-hierarchical model: mean

Posterior median and 95% CI for the linear predictor μ_{ij} (`CD4.mu[i, j]`)



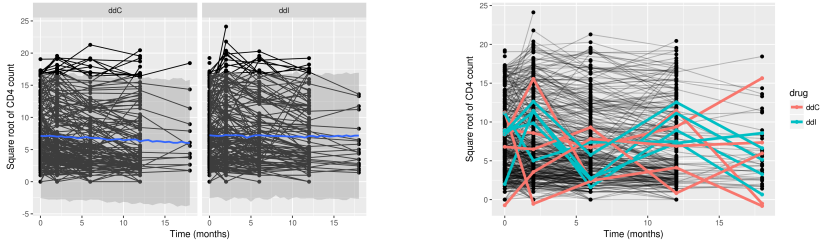
Notes

AIDS non-hierarchical model: making predictions

Predict $y^\star_{kl} \sim \mathrm{N}(\mu^\star_{kl}, \sigma^2)$, where $\mu^\star_{kl} = \beta_1 + \beta_2 t_{kl} + \beta_3 \mathrm{drug}^\star_k t_{kl}$

for new patients $k = 1, \ldots, K$, with drug $\mathrm{drug}^\star_k$ at time points t_{kl}

```
CD4.pred[k, l] ~ dnorm(CD4.pred.mu[k, l], CD4.sigma.prec)

CD4.pred.mu[k, l] <- beta1 +
                     beta2 * t.pred[l] +
                     beta3 * t.pred[l] * drug.ddI.pred[k]
```



Notes

AIDS random intercept model

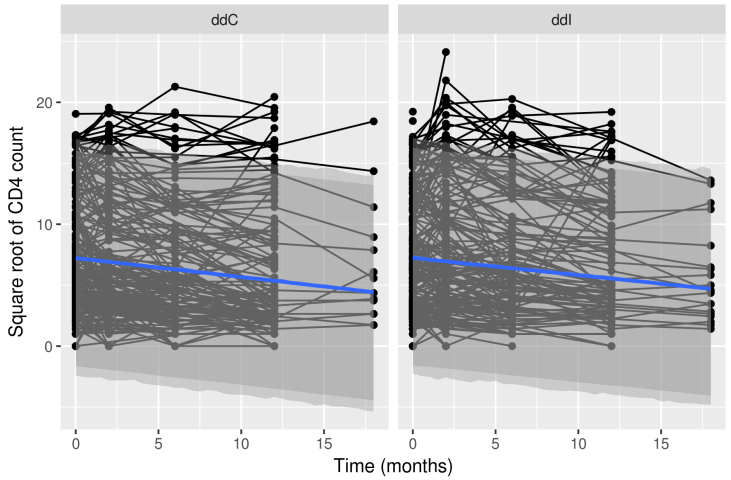
Now consider a hierarchical linear model

$$y_{ij} \sim \mathsf{N}(\mu_{ij}, \sigma^2) \qquad \beta_{1i} \sim \mathsf{N}(\beta_1, \sigma^2_{\beta_1})$$
$$\mu_{ij} = \beta_{1i} + \beta_2 t_{ij} + \beta_3 \text{drug}_i t_{ij} \qquad \beta_1, \beta_2, \beta_3 \sim \mathsf{N}(0, 100^2)$$
$$\sigma, \sigma_{\beta_1} \sim \mathsf{Unif}(0, 100)$$

```
for (i in 1:I){
  for (j in 1:J){
    CD4[i, j] ~ dnorm(CD4.mu[i, j], CD4.sigma.prec)
    CD4.mu[i, j] <- beta1[i] +
                    beta2 * t.obs[j] +
                    beta3 * t.obs[j] * drug.ddI[i]
  }
  beta1[i] ~ dnorm(beta1.mu, beta1.sigma.prec)
}
beta1.mu ~ dnorm(0, 0.001)
beta1.sigma.prec <- 1/pow(beta1.sigma.sd, 2)
beta1.sigma.sd ~ dunif(0, 100)
```

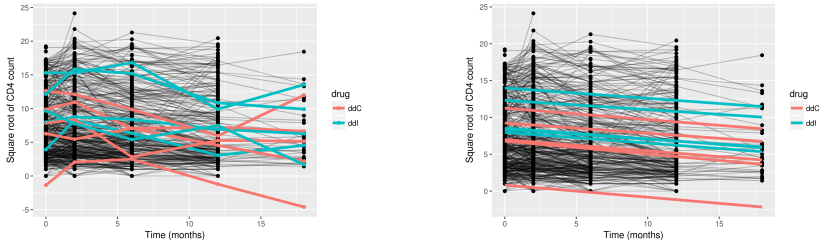

Notes

AIDS random intercept - uncertainty



Notes

AIDS random intercept - predictions



Notes

AIDS random intercept and slope model

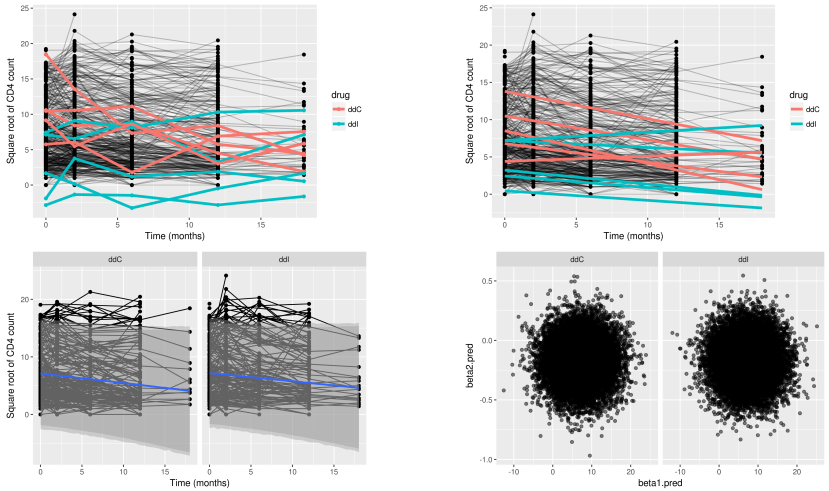
Now allow the slope to vary as well

$$y_{ij} \sim \mathsf{N}(\mu_{ij}, \sigma^2) \qquad \beta_{1i} \sim \mathsf{N}(\beta_1, \sigma^2_{\beta_1})$$
$$\mu_{ij} = \beta_{1i} + \beta_2 t_{ij} + \beta_3 \mathsf{drug}_i t_{ij} \qquad \beta_1, \beta_2, \beta_3 \sim \mathsf{N}(0, 100^2)$$
$$\sigma, \sigma_{\beta_1} \sim \mathsf{Unif}(0, 100)$$

```
for (i in 1:I){
  for (j in 1:J){
    CD4[i, j] ~ dnorm(CD4.mu[i, j], CD4.sigma.prec)
    CD4.mu[i, j] <- beta1[i] +
                    beta2[i] * t.obs[j] +
                    beta3 * t.obs[j] * drug.ddI[i]
  }
  beta1[i] ~ dnorm(beta1.mu, beta1.sigma.prec)
  beta2[i] ~ dnorm(beta2.mu, beta2.sigma.prec)
}
```


Notes

AIDS random intercept and slope - uncertainty



Notes

Priors for covariance matrices

Wishart$_p(k, R)$ - multivariate extension of the Gamma distribution

- Distribution on symmetric, positive definite $p \times p$ matrices
- Two parameters
  - R – a symmetric positive-definite matrix
  - $k \geq p$ - becomes more informative as k increases
- Expectation $E(\Sigma) = kR^{-1}$
- Is the conjugate prior for the precision of a multivariate normal
- `sigma.inv[1:p, 1:p] ~ dwish[R[,], k]` in BUGS. k and R must be constants.
- More flexible alternatives: scaled inverse Wishart for covariances (e.g. Gelman and Hill, 2007)

Notes

Multivariate priors code

Model

$\beta \sim \text{MVN}(\mu, \Sigma_\beta)$

$\Sigma \sim \text{Wishart}(R, k)$

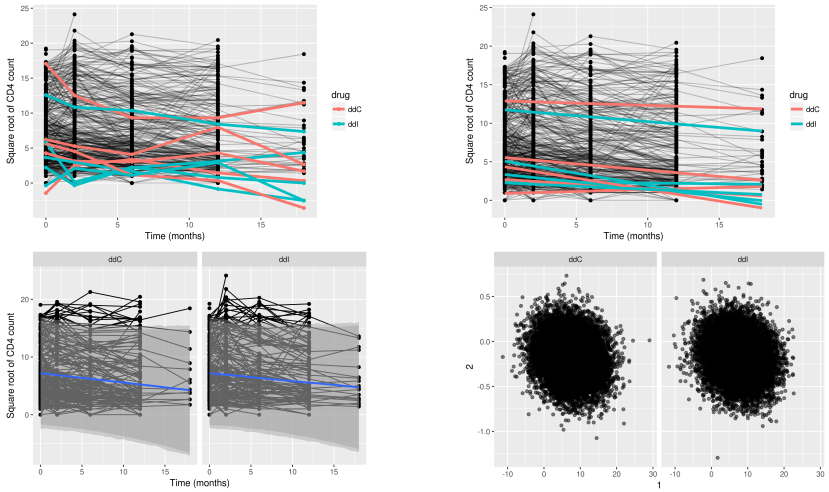
$R = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

$k = 2$

```
beta[i, 1:2] ~ dmnorm(mu[], sigma.inv[,])
sigma.inv[1:2,1:2] ~ dwish(R[,], k)

d <- 2
R[1, 1] <- 1
R[2, 1] <- 0
R[1, 2] <- 0
R[2, 2] <- 1
# (Or supply R and k as data)
```


Notes

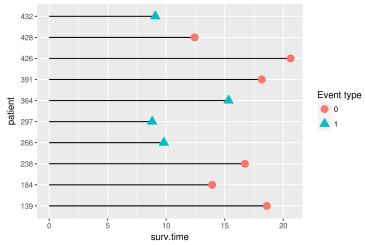
AIDS random intercept and slope - Wishart prior



Notes

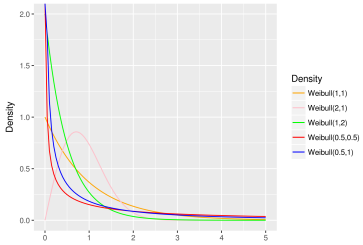
Survival models: Weibull model

Data $t_1, \ldots, t_n$ measuring time-to-event
Often censoring: event times are only known within a certain range.



0 denotes censoring, 1 denotes death.

Often modelled by Weibull(a, b) $(a, b > 0)$



- In BUGS:
  `t ~ dweib(a, b)`
- $p(t \mid a, b) = a\, b\, t^{a-1}\, e^{-b\, t^a}$
- $E(t) = b^{-1/a}\Gamma\left(1 + \frac{1}{a}\right)$
- In R `(dweibull)`: shape $= a$; scale $= b^{-1/a}$

Notes

Survival models: frailty models

$t_i \sim$ Weibull(r, μ_i) with $\log(\mu_i) = \beta_{1i} + \beta_{2i} x_i$

```
## JAGS code
is.cen[i] ~ dinterval(surv.time[i], break.time[i])
surv.time[i] ~ dweib(1, surv.mu[i])

log(surv.mu[i]) <- beta[i, 1] +
                   beta[i, 2] * drug.ddI[i]
```

In OpenBUGS replace lines 1–2 with:

```
surv.time[i] ~ dweib(1, surv.mu[i])I(surv.c[i], )
```

A frailty model is a survival model in which the parameters are random effects

```
beta[i, 1:2] ~ dmnorm(beta.mu[], beta.sigma.prec[,])
# priors for beta.mu[] and beta.sigma.prec[,] as before
```


Notes

Joint longitudinal and survival models

- Survival in the AIDS study is related to CD4 count.
- Joint modelling aims to describe both processes in a single model.
- Dependence (and missingness) between the two processes is captured by shared random effects

$$y_{ij} \sim \mathsf{N}(\mu_{ij}, \sigma^2) \qquad \beta_{1i} \sim \mathsf{N}(\beta_1, \sigma^2_{\beta_1})$$
$$\mu_{ij} = \beta_{1i} + \beta_2 t_{ij} + \beta_3 \mathsf{drug}_i t_{ij} \qquad \beta_1, \beta_2, \beta_3 \sim \mathsf{N}(0, 100^2)$$
$$\sigma, \sigma_{\beta_1} \sim \mathsf{Unif}(0, 100)$$

$t_i \sim \mathsf{Weibull}(r, \mu_i)$ with $\log(\mu_i) = \beta_4 + \beta_5 x_i + r_1 * \beta_1 + r_2 * \beta_2$

Notes

Joint longitudinal and survival models

```
for (i in 1:I){
  for (j in 1:J){
    CD4[i, j] ~ dnorm(CD4.mu[i, j], CD4.sigma.prec)
    CD4.mu[i, j] <- beta[i, 1] +
                    beta[i, 2] * t.obs[j] +
                    beta3 * t.obs[j] * drug.ddI[i]
  }

  is.cen[i] ~ dinterval(surv.time[i], break.time[i])
  surv.time[i] ~ dweib(1, surv.mu[i])

  log(surv.mu[i]) <- beta4 +
                     beta5 * drug.ddI[i] +
                     r1 * beta[i, 1] +
                     r2 * beta[i, 2]

  beta[i, 1:2] ~ dmnorm(beta.mu[], beta.sigma.prec[,])
}
```


Notes

Joint longitudinal and survival models: results

	2.5%	25%	50%	75%	97.5%
r1	-0.29	-0.25	-0.23	-0.21	-0.18
r2	-3.85	-2.92	-2.44	-1.92	-0.98
beta.mu[1]	6.77	7.05	7.20	7.34	7.62
beta.mu[2]	-0.24	-0.21	-0.19	-0.18	-0.15
beta.sigma.cov[1,1]	18.55	20.11	21.08	22.14	24.36
beta.sigma.cov[2,1]	-0.22	-0.11	-0.06	-0.01	0.09
beta.sigma.cov[1,2]	-0.22	-0.11	-0.06	-0.01	0.09
beta.sigma.cov[2,2]	0.03	0.04	0.04	0.05	0.06
beta4	-3.17	-2.84	-2.69	-2.54	-2.26
beta5	-0.06	0.15	0.26	0.36	0.57

- Negative r1 suggests that a higher initial CD4 count leads to lower surv.mu, i.e. longer expected survival.
- Negative r2 suggests that a steeper upwards slope in CD4 similarly leads to longer expected survival.

Notes

Model checking

Ideally we should check the model with new data

- y_{fit} – used to fit the data
- y_{crit} – used for model criticism

Out-of-sample prediction

1. Obtain the posterior distribution $p(\theta \mid y_{\text{fit}})$
2. Get posterior predictive for 'criticism' data

$$p(y_{\text{crit}}^{\text{pred}}) = \int p(y_{\text{crit}}^{\text{pred}} \mid y_{\text{fit}}, \theta)\, p(\theta \mid y_{\text{fit}}) \mathrm{d}\theta$$
$$= \int p(y_{\text{crit}}^{\text{pred}} \mid \theta)\, p(\theta \mid y_{\text{fit}}) \mathrm{d}\theta$$

3. Compare $y_{\text{crit}}^{\text{pred}}$ and y_{crit}

Called 'model checking', but is a check of both the model and the prior

Notes

Splitting the data

Often no new data are available, so we must use the same data to build and check the model

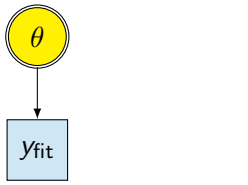
Data splitting options

Leave-one-out cross-validation Set $y_{\text{crit}} = y_i$ and $y_{\text{fit}} = y_{\setminus i}$
Here y_i a single observation and $y_{\setminus i}$ is the rest
Repeat for each observation, leaving one observation out at a time

k-fold cross-validation Set y_{fit} to be k% of the data
Here y_{crit} is the remaining $(1-k)$% of the data
Repeat for each 'fold', leaving k% out at a time

Notes

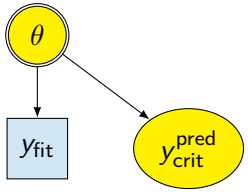
1+2: Posterior predictive distribution




```
yfit  ~ dpois(theta)
theta ~ dgamma(0.5, 1)
```

Posterior predictive

$$p(y_{\text{crit}}^{\text{pred}}) = \int p(y_{\text{crit}}^{\text{pred}} \mid \theta)\, p(\theta \mid y_{\text{fit}}) \mathrm{d}\theta$$



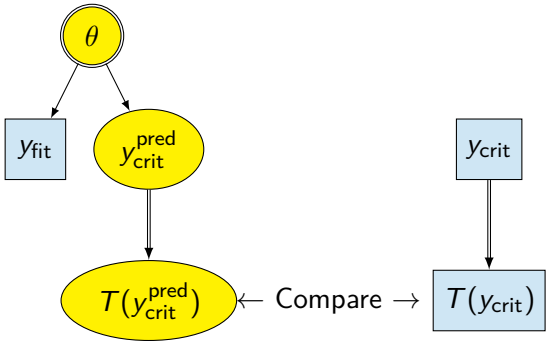

```
yfit  ~ dpois(theta)
theta ~ dgamma(0.5, 1)
ypred ~ dpois(theta)
```


Notes

3. Comparing predictions with data

Need to compare $y_{\text{crit}}^{\text{pred}}$ and y_{crit} – use a checking function T. e.g.

- $T(y_{\text{crit}}) = y_{\text{crit}}$ to check for individual outliers




```
yfit  ~ dpois(theta)
theta ~ dgamma(0.5, 1)
ypred ~ dpois(theta)
Typred <- ypred
Tycrit <- ycrit
```

(for $T(y_{\text{crit}}) = y_{\text{crit}}$)

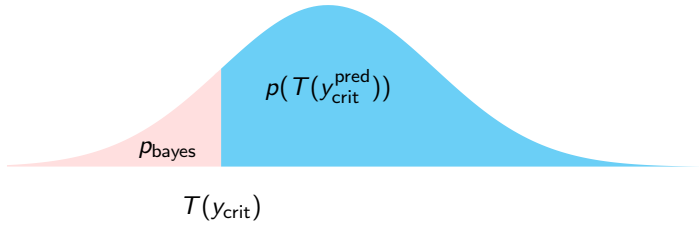
Then assess the compatibility of $T(y_{\text{crit}}^{\text{pred}})$ and $T(y_{\text{crit}})$

Notes

Comparing $T(y_{\text{crit}}^{\text{pred}})$ and $T(y_{\text{crit}})$

Graphically e.g. Density plot of $p(y_{\text{crit}}^{\text{pred}})$ with y_{crit} marked

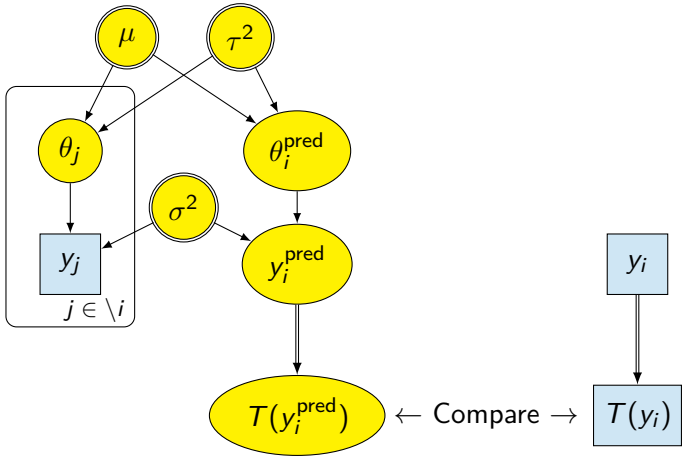
Bayesian p-value $p_{\text{bayes}} = P\left(T(y_{\text{crit}}^{\text{pred}}) \leq T(y_{\text{crit}}) \mid y_{\text{fit}}\right)$



Notes

Cross-validation in hierarchical models

Typically use cross-validation with the i^{th} unit left out

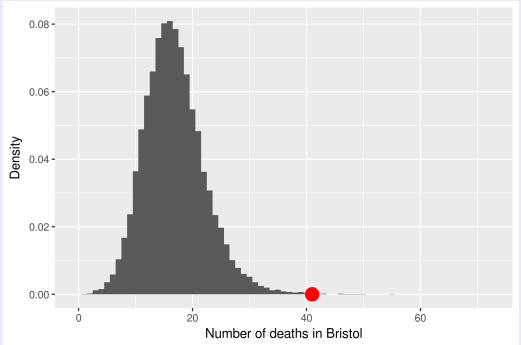


Notes

Mixed cross-validation of Bristol model

Single fold of a leave-one-out cross-validation, leaving Bristol out
Number of deaths in Bristol: observed is 41.

Predictive distribution



Median 16
95% CI 8–29

One-sided p-value 0.0015

Notes

Approximate mixed cross-validation

Full cross-validation can be time-consuming

- Need to repeat the analysis lots of times
- In principle, cross-validation is embarrassingly parallel
- But still need to check convergence of MCMC each time etc etc

Approximate mixed cross-validation

1. Analyse *all* data, but create a 'ghost' θ_i^{rep} for each unit
2. Generate observations for each ghosted unit
3. Compare the ghost observations with the real observations

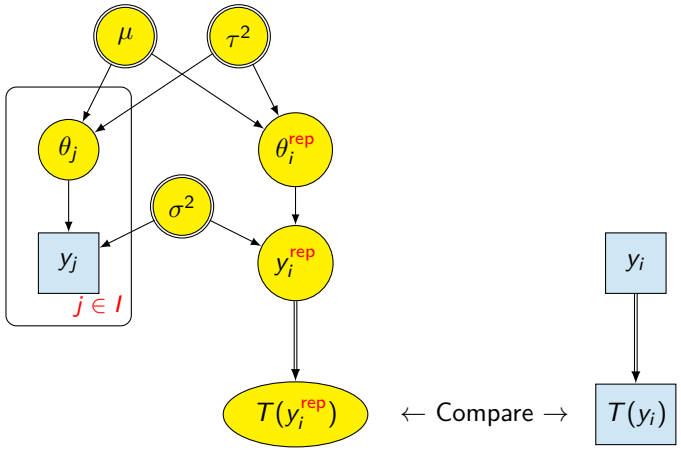
This is likely to be conservative, but this may be only moderate since the influence of y_i on θ_i^{rep} is only through μ and τ^2

Notes

Approximate mixed cross-validation

Compare $T(y_i)$ with the mixed predictive for i^{th} unit

$$p(T(y_i^{\text{pred}})) = \iint p(T(y_i^{\text{rep}}) \mid \theta_i^{\text{rep}}, \sigma^2)\, p(\theta_i^{\text{rep}} \mid \mu, \tau^2) p(\mu, \tau^2, \sigma^2 \mid y) \mathrm{d}\theta \mathrm{d}\Omega$$



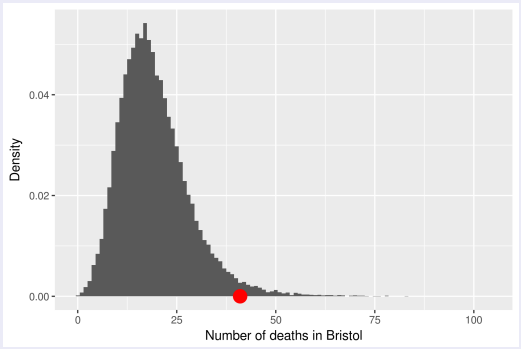
← Compare →

Unit i's data are now included

Notes

Approximate mixed cross-validation of Bristol model

Predictive distribution



Median 18
95% CI 6–41

One-sided p-value 0.024

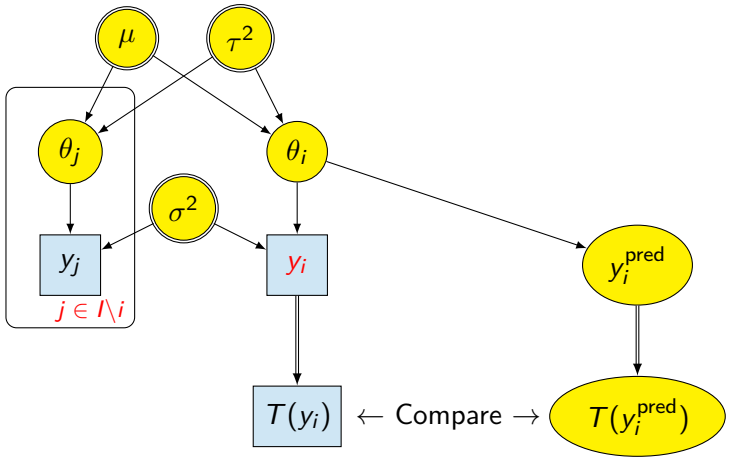
So clearly some conservatism introduced, but still suggesting that Bristol is an outlier

Notes

Other approaches: posterior predictive

Compare $T(y_i)$ with the posterior predictive for i^{th} unit

$$p(T(y_i^{\text{pred}})) = \int p(T(y_i^{\text{pred}}) \mid \theta_i, \sigma^2)\, p(\theta_i, \sigma^2 \mid y) \mathrm{d}\Omega$$

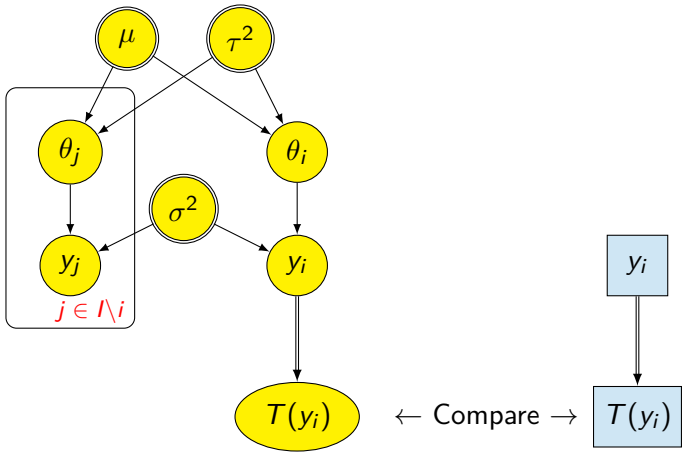


Notes

Other approaches: prior predictive

Compare $T(y_i)$ with the prior predictive for i^{th} unit

$$p(T(y_i^{\text{pred}})) = \int p(T(y_i^{\text{pred}}) \mid \Omega)\, p(\Omega) \mathrm{d}\Omega$$



Notes

Summary of predictive model assessment

- Cross validation usually viewed as the gold standard
- Prior predictive compares observations with predictions from the prior
  - Will only be useful for checking informative priors
- Posterior predictive compares observations with predictions from the posterior
  - Will tend to be conservative
- Mixed predictive is a compromise

Notes

Practicals: complex hierarchical models

Three exercises to try:

- Longitudinal data
  - Fitting and predicting ratings of teachers with random intercept/slope models
- Model checking
  - Which of the methods identify an (obvious) outlier?
- Joint modelling
  - Explore the AIDS example (on 20 patients), understand the model code

Notes

Further reading

- Gelman and Hill (2007) *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.
- Rizopoulos (2012) *Joint models for longitudinal and time-to-event data: with applications in R*. Chapman & Hall / CRC.
- Model checking: The BUGS Book chapter 8
- Hierarchical models: further references in the BUGS Book p252

Notes

Part VI

Evidence Synthesis

Notes

Overview

- Introduction
- Network Meta-Analysis (NMA)
- Generalised evidence synthesis
- Practical/further resources

Notes

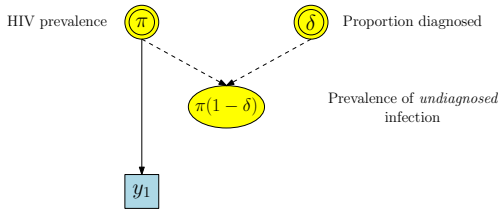
What is evidence synthesis?

- generalisation of (hierarchical) models to inference based on *multiple* data sources
- usually *multi-parameter* models

Spiegelhalter et al 2004, Ades & Sutton 2006

Notes

Simple example: HIV prevalence



- δ only *weakly identifiable*, i.e. requires informative prior for identification
- y_1 independent of δ

$$p(y_1, \pi, \delta) = p(\pi)p(\delta)p(y_1 \mid \pi, \delta) = p(\pi)p(\delta)p(y_1 \mid \pi)$$

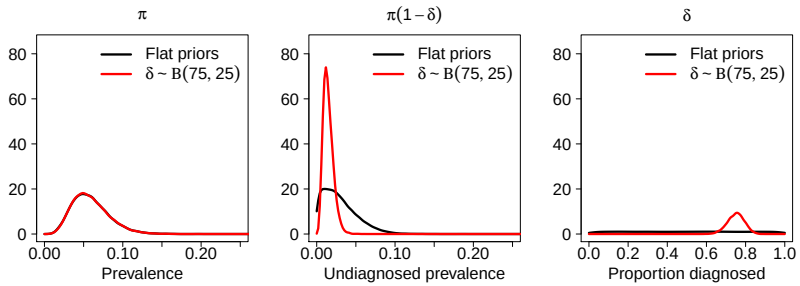
Notes

Model code

```
## Flat priors
pi    ~ dbeta(1,1)      # or informative: dbeta(15,85)
delta ~ dbeta(1,1)      # or informative: dbeta(75,25)

## Likelihood: prevalence data
for(i in 1:1)
{
  y[i] ~ dbin(p[i], n[i])      # (y1,n1) = (5,100)
}

## Proportions in terms of basic and
## functional parameters
p[1] <- pi
p[2] <- pi * (1 - delta)
p[3] <- delta
```


Notes

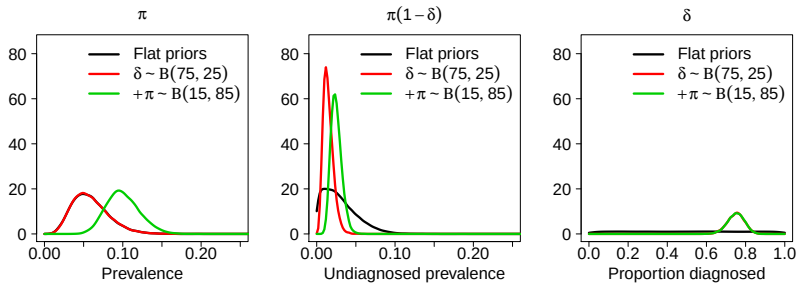
Simple example: HIV prevalence



- $\pi(1-\delta)$ *weakly identified* with flat priors, since $\delta \in [0,1]$
- δ identified when use informative prior, additional information *increases precision* of $\pi(1-\delta)$

Notes

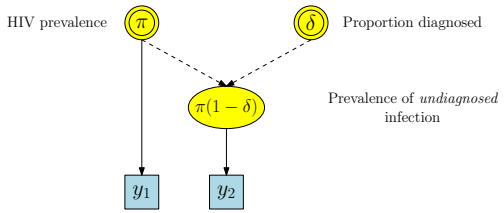
Simple example: HIV prevalence



- informative prior for π inconsistent with likelihood, so posterior is a *compromise*
- conflict *reduces precision* of $\pi(1-\delta)$ again

Notes

Simple example: HIV prevalence



- y_1, y_2 *conditionally* independent, given π, δ
- π, δ both now identifiable *without* informative priors, potential for *conflict* if priors are informative

$$\begin{aligned} p(y_1, y_2, \pi, \delta) &= p(\pi)p(\delta)p(y_1, y_2 \mid \pi, \delta) \\ &= p(\pi)p(\delta)p(y_1 \mid \pi)p(y_2 \mid \pi, \delta) \end{aligned}$$

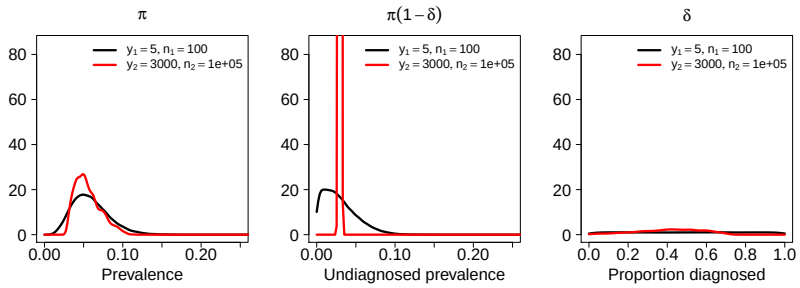
Notes

Model code

```
...

## Likelihood: prevalence data
## Likelihood: prevalence of undiagnosed infection
for(i in 1:2)
{
  y[i] ~ dbin(p[i], n[i])     # (y1,n1) = (5,100),
                              # (y2,n2) = (3000,100000)
}

## Proportions in terms of basic and
## functional parameters
p[1] <- pi
p[2] <- pi * (1 - delta)
p[3] <- delta
```


Notes

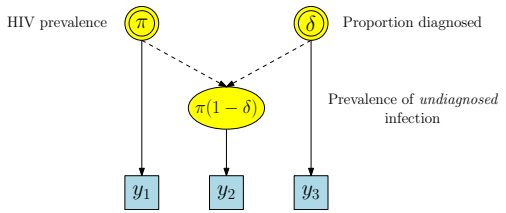
Simple example: HIV prevalence



- both basic and functional parameters now *identified*, even with flat priors
- although δ still relatively uncertain, despite large sample size informing $\pi(1-\delta)$, since information *indirect* and small sample size informing π

Notes

Simple example: HIV prevalence



- in theory, more data $\Rightarrow$ more *precise* estimates
- even with uninformative priors, potential for *conflict*: 3 data items informing 2 parameters

$$\begin{aligned} p(y_1, y_2, y_3, \pi, \delta) &= p(\pi)p(\delta)p(y_1, y_2, y_3 \mid \pi, \delta) \\ &= p(\pi)p(\delta)p(y_1 \mid \pi)p(y_2 \mid \pi, \delta)p(y_3 \mid \delta) \end{aligned}$$

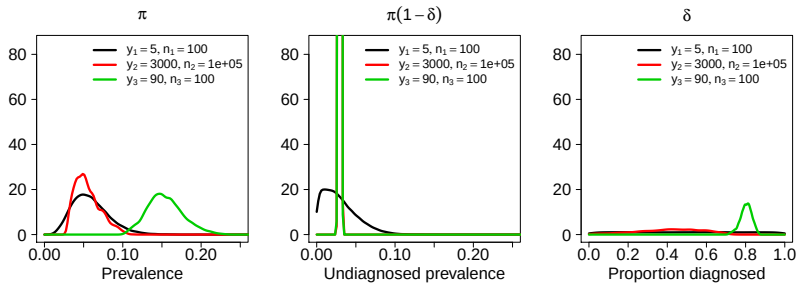
Notes

Model code

```
...

## Likelihood: prevalence data
## Likelihood: prevalence of undiagnosed infection
## Likelihood: proportion diagnosed
for(i in 1:3)
{
  y[i] ~ dbin(p[i], n[i])    # (y1,n1) = (5,100),
                             # (y2,n2) = (3000,100000),
}                            # (y2,n2) = (90,100)

## Proportions in terms of basic and
## functional parameters
p[1] <- pi
p[2] <- pi * (1 - delta)
p[3] <- delta
```


Notes

Simple example: HIV prevalence



- δ is now better identified (*more peaked* posterior)
- but *conflict* between the three data points leads to a larger, more uncertain estimate of π

Notes

Features of evidence synthesis

Evidence synthesis leads to complex probabilistic models

- Combination of all available relevant data sources *ideally* should lead to more *precise* estimates
- Multiple sources informing a single parameter ⇒ potential for *conflicting evidence*
- Sparsity of data ⇒ parameters *unidentifiable* without further model constraints, e.g. *informative priors, exchangeability*

Notes

Statistical formulation

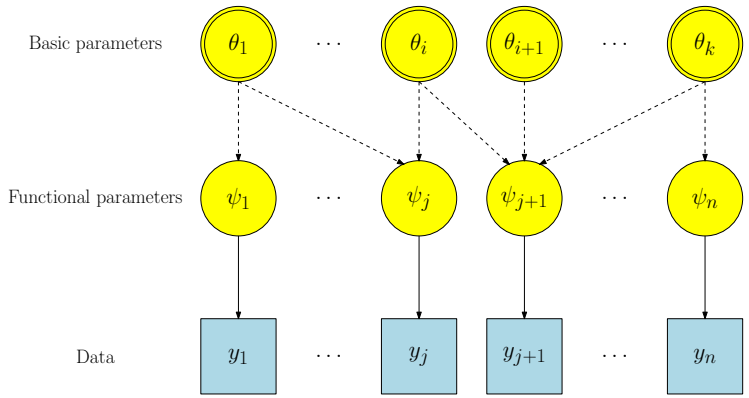
- Interest: estimation of $\boldsymbol{\theta} = (\theta_1, \theta_2 \ldots, \theta_k)$ on the basis of a collection of data $\boldsymbol{y} = (y_1, y_2 \ldots, y_n)$
- Each y_i provides information on
  - a *single* component of $\boldsymbol{\theta}$ (*"direct"* data), or
  - a *function* of one or more components, *i.e.* on a quantity $\psi_i = f(\boldsymbol{\theta})$ (*"indirect"* data)

Thus inference is conducted on the basis of both **direct** and **indirect** information.

- Maximum likelihood: $L(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} L_i(y_i \mid \boldsymbol{\theta})$
- Bayesian: $p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto p(\boldsymbol{\theta}) \times L(\boldsymbol{y} \mid \boldsymbol{\theta})$

Notes

Graphical representation (DAG)



Notes

- Introduction
- Network Meta-Analysis (NMA)
- Generalised evidence synthesis
- Practical/further resources

Notes

Network meta-analysis

- Generalisation of meta-analysis to synthesise trials that might include different *"designs"*
- e.g. for treatments A, B, C and D, we might have a network of trials that make the pairwise comparisons AB, AC and CD, but not any other pairwise comparisons.
- So the comparisons AD and BC are never directly made in a single trial
- *But*, we can still estimate these *indirectly*, by synthesising the network of trials.

Notes

Smoking cessation data

Lu & Ades JASA 2006, Higgins et al RSM 2012, Jackson et al SiM 2014

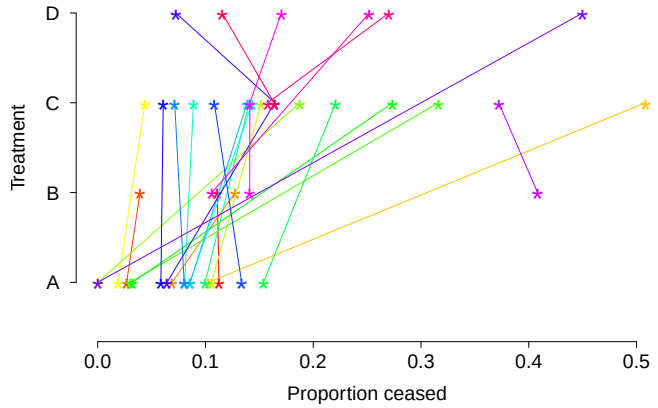
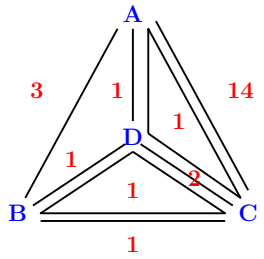


Figure : 24 trials (2 3-arm studies: ACD and BCD), 50 pairs compared, A no contact; B self-help; C individual counselling; D group counselling

Notes

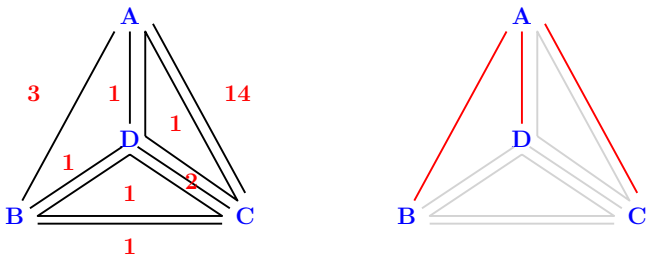
NMA graphs



- method to visualise network and number of studies of each *design*

Notes

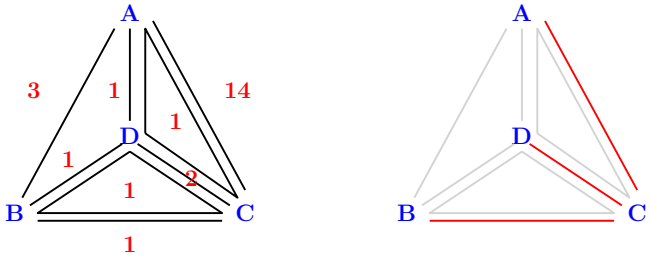
NMA graphs



- choose *overall baseline* treatment, e.g. A
- *spanning tree* formed by effects for each treatment relative to baseline are the *basic* parameters, e.g. (AB, AC, AD)

Notes

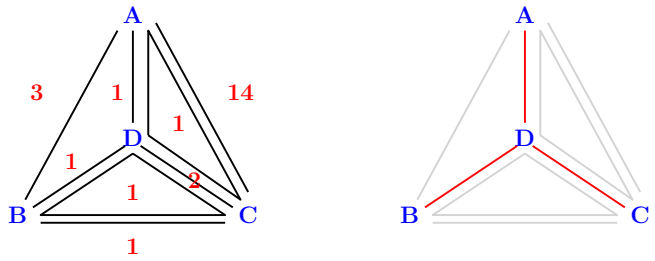
NMA graphs



- or baseline C, spanning tree (CA, CB, CD)

Notes

NMA graphs



- or baseline D, spanning tree (DA, DB, DC)

Notes

Model under consistency assumption

Aim

Estimate treatment effects $\delta^{(JK)}$ for each pair JK in the network.

- Treatment effects might be (log) odds ratios or relative risks
- Not all pairs JK directly observed, so choose the *spanning tree* of effects relative to an overall baseline treatment B to be our basic parameters:

$$\boldsymbol{\delta_b} = (\delta^{BJ})_{J \neq B}$$

- Then remaining effects are defined *functionally* in terms of the basic parameters (the *consistency* assumption) and B:

$$\boldsymbol{\delta_f} = f(\boldsymbol{\delta_b}) = (\delta^{JK} := \delta^{BK} - \delta^{BJ})_{JK \notin ST}$$

Notes

Smoking example

Likelihood:

$$y_{sa} \sim Bin(n_{sa}, p_{sa}) \qquad \text{study } s\text{, arm } a$$

Functional parameters (b_s is a study-specific baseline):

$$logit(p_{sa}) = \alpha_s + \mu_s^{b_s a}$$

NB intercept is log-odds for study-specific baseline $\alpha_s = logit(p_{sb_s})$

Common effects:

$$\mu_s^{b_s a} = \delta^{Ba} - \delta^{Bb_s}$$

Random effects:

$$\mu_s^{b_s a} \sim N(\delta^{Ba} - \delta^{Bb_s}, \sigma^2_{b_s a})$$

Priors:

$$\alpha_s \sim N(0, 10^2)$$
$$\delta^{Ba} \sim N(0, 10^2), \quad Ba \in ST$$
$$\sigma_{b_s a} = \sigma \sim U(0, 5) \qquad \text{common heterogeneity variance}$$

Notes

Random effects for multi-arm trials

Within a study s, the random effects assumption for each treatment pair $b_s a$ for each non-baseline arm a is written multi-variately as

$$\boldsymbol{\mu_s} = (\mu_s^{b_s a})_a \sim MVN(\boldsymbol{\delta_f}, \boldsymbol{\Sigma})$$

It is common (Salanti et al SMMR 2008, Jackson et al SiM 2014) to assume that the diagonal entries of $\boldsymbol{\Sigma}$ are σ^2 and the non-diagonal entries are $\sigma^2/2$, which implies that the heterogeneity variance is the same for all treatment comparisons for every study.

Notes

Model code - RE assumption

Dias et al SiM 2010, Jackson et al SiM 2014

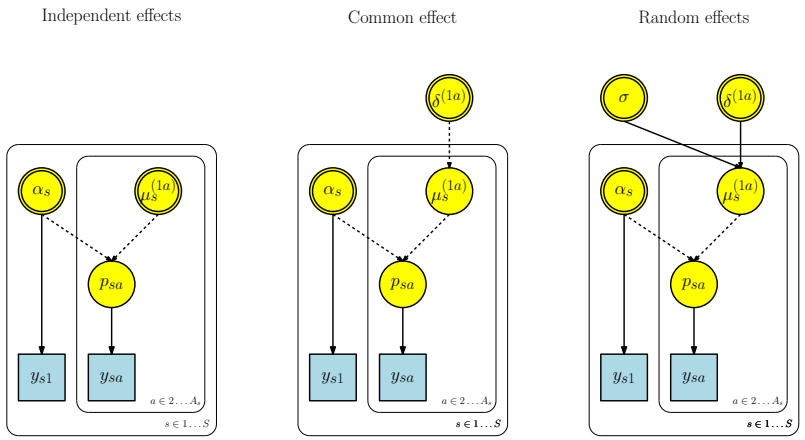
```
## Two-arm studies
mu[s,t[s,a]] ~ dnorm(mean.d[s,t[s,a]], prec.d[s,t[s,a]])
mean.d[s,t[s,a]] <- delta[t[s,a]] - delta[t[s,1]]
prec.d[s,t[s,a]] <- 1 / pow(sd,2)
```

```
## Multi-arm studies
mu[s,t[s,2:Na[s]]] ~ dmnorm(mean.d[s,t[s,2:Na[s]]], Prec[s,,])
for(a in 2:Na[s])
{
  mean.d[s,t[s,a]] <- delta[t[s,a]] - delta[t[s,1]]
  for(i in 2:Na[s])
  {
    Sigma[s,a,i] <- (equals(i,a)*pow(sd,2))   ## variances
                  + ((1 - equals(i,a))*pow(sd,2)/2)   ## covariances
  }
}
Prec[s,,] <- inverse(Sigma[s,,])
```


Notes

Recall: meta-analysis DAGs

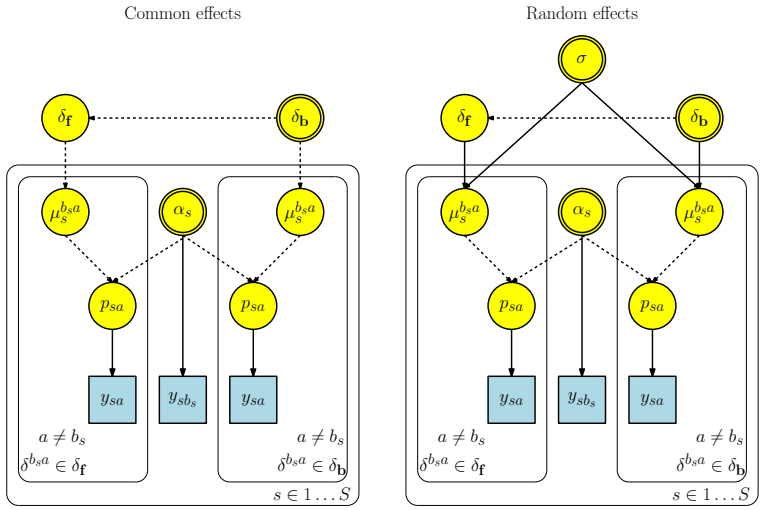
Same baseline treatment (arm 1) in all studies, *same* other treatment arms, all *observed*:



Notes

NMA DAGs

By contrast, *study-specific baselines* as well as overall baseline, *not all contrasts* observed: some basic $\boldsymbol{\delta_b}$, some functional $\boldsymbol{\delta_f}$



Notes

Smoking example results - model comparison

Smaller *DIC* implies random effect model preferred

Model:	Fixed effects Posterior:		Random effects Posterior:	
δ^{JK}:	Mean	SD	Mean	SD
AB	0.224	(0.124)	0.496	(0.405)
AC	0.765	(0.059)	0.843	(0.236)
AD	0.840	(0.174)	1.103	(0.439)
BC	0.541	(0.132)	0.347	(0.419)
BD	0.616	(0.192)	0.607	(0.492)
CD	0.075	(0.171)	0.260	(0.418)
$\mathbb{E}_{\theta\|y}(D)$	267		54	
$D(\mathbb{E}_{\theta\|y}\theta)$	240		10	
p_D	27		44	
DIC	294		98	

Notes

- Introduction
- Network Meta-Analysis (NMA)
- Generalised evidence synthesis
- Practical/further resources

Notes

GES: estimating influenza severity

AIM: estimate *case-severity risks*
- case-*fatality* risk
- case-*ICU-admission* risk
- case-*hospitalisation* risk

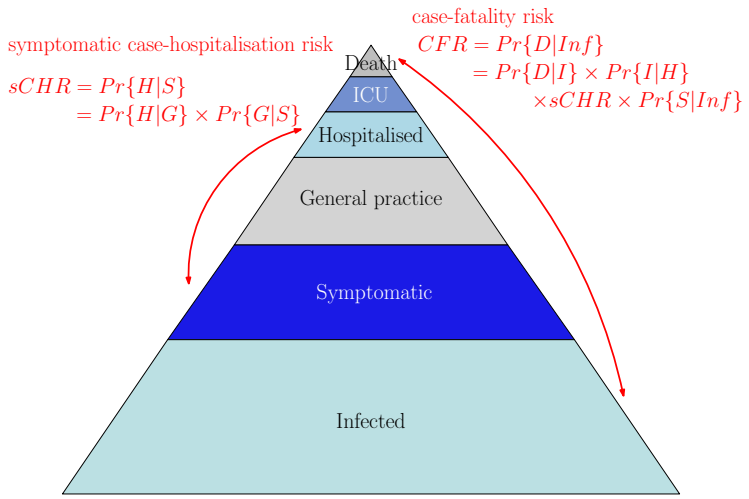
i.e. probabilities of a *severe* event given an influenza infection
- or given a *symptomatic* infection

from: imperfect observations of numbers of infections at different levels of severity from various *surveillance* data sources

Presanis et al PloS Med 2009, AoAS 2014, Shubin et al Epidem & Inf 2013, Wong et al Epidem 2013, McDonald et al IORV 2014

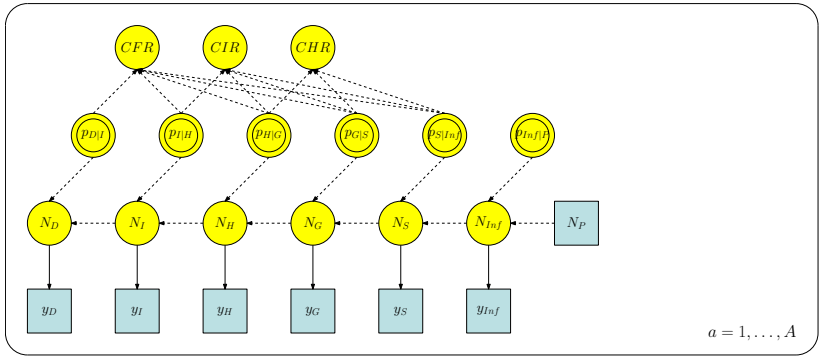
Notes

GES: estimating influenza severity



Notes

Ideally...



- observations at *each* level of severity
- *directly* informing number of infections at each severity level

Notes

Direct observations at each level...

Poisson/log-linear regression

Observations $y_{a,\ell}$ in age group a at severity level ℓ

$$y_{a,\ell} \sim Pois(N_{a,\ell})$$
$$\log(N_{a,\ell}) = \log(N_{a,P}) + \alpha_\ell + \beta_a + \gamma_{a,\ell}$$

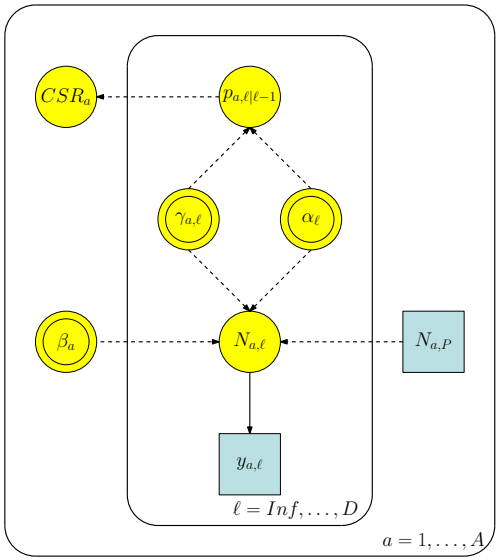
Functional parameters

$$p_{a,\ell|\ell-1} = \exp\{\alpha_{a,\ell} - \alpha_{a,\ell-1} + \gamma_{a,\ell} - \gamma_{a,\ell-1}\}$$
$$CFR_a = p_{a,D|I} \times p_{a,I|H} \times p_{a,H|G} \times p_{a,G|S} \times p_{a,S|Inf}$$

NB this parameterisation implies $N_{a,\ell} = p_{a,\ell|\ell-1} N_{a,\ell-1}$, i.e. *expected* number at each level.

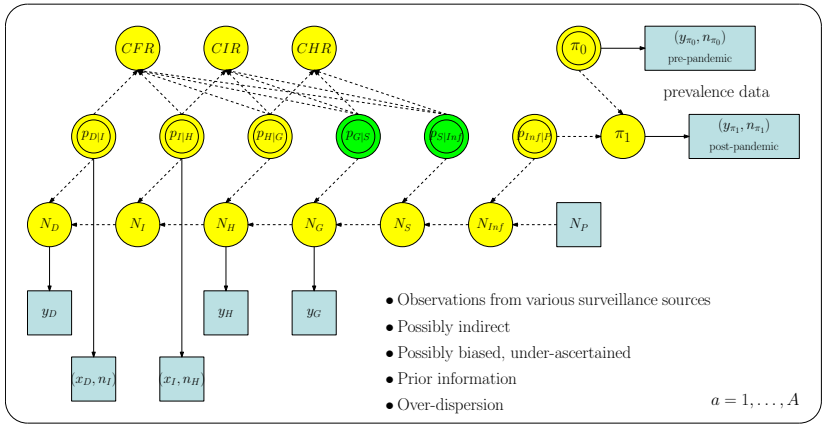
Notes

DAG of this parameterisation



Notes

Reality...



Retain *expected* parameterisation: $N_{a,\ell} = p_{a,\ell|\ell-1} N_{a,\ell-1}$

Simplified/idealised version of Presanis et al, AoAS 2014

Notes

Model

Priors:

$$p_{a,\ell|\ell-1} \sim Beta(\lambda_\ell, \kappa_\ell)$$
$$\pi_{a,0} \sim Beta(1,1)$$
$$1/\log(r_\ell) \sim Unif(0.05, 0.5)$$

Likelihood:

$$y_{a,\ell} \sim NegBin(\psi_{a,\ell}, r_\ell), \qquad \ell = G, H, D$$
$$x_{a,\ell} \sim Bin(n_{a,\ell-1}, p_{a,\ell|\ell-1}), \qquad \ell = I, D$$
$$y_{a,\pi_t} \sim Bin(n_{a,\pi_t}, \pi_t), \qquad t = 0, 1$$

Functional parameters:

$$N_{a,\ell} = p_{a,\ell|\ell-1} \times N_{a,\ell-1}$$
$$CSR_a = \prod_\ell p_{a,\ell|\ell-1}$$
$$\pi_{a,1} = \pi_{a,0} + p_{a,Inf|P}$$
$$\psi_{a,\ell} = r_\ell/(N_{a,\ell} + r_\ell)$$

Interpretation:

$$\mathbb{E}(Y_{a,\ell}) = r_\ell(1 - \psi_{a,\ell})/\psi_{a,\ell} = N_{a,\ell}$$
$$Var(Y_{a,\ell}) = \mathbb{E}(Y_{a,\ell})/\psi_{a,\ell} = N_{a,\ell}/\psi_{a,\ell}$$

Smaller r_ℓ implies more over-dispersion. Informative prior on inverse of log r_ℓ, smaller value implies less over-dispersion.

Notes

Aside: negative binomial parameterisation

Note that the negative binomial may be *parameterised* in different ways:

- In BUGS/JAGS: `y ~ dnegbin(p,r)`

$$p(Y = y) = \binom{y + r - 1}{y} p^r (1 - p)^y$$

r is interpreted as a number of failures need to observe y successes, p is probability of failure

- If p is interpreted instead as a probability of *success*, then

$$p(Y = y) = \binom{y + r - 1}{y} (1 - p)^r p^y$$

Notes

Aside: negative binomial parameterisation

Note that the negative binomial may be *parameterised* in different ways:

- Can be parameterised in terms of *mean* $\mu = r(1 - p)/p$ and *variance* $\sigma^2 = r(1 - p)/p^2$:

$$p(Y = y) = \binom{y + \frac{\mu^2}{\sigma^2 - \mu} - 1}{y} \left(\frac{\mu}{\sigma^2}\right)^{\frac{\mu^2}{\sigma^2 - \mu}} \left(\frac{\sigma^2 - \mu}{\sigma^2}\right)^y$$

- Can be expressed as a *Gamma-Poisson* mixture, i.e. compound of Poisson distributions with mixing distribution of the Poisson rate being Gamma:

$$Y \sim Pois(\lambda)$$
$$\lambda \sim \Gamma(r, \theta)$$

with shape r and scale $\theta = (1 - p)/p$.

Notes

Aside: negative binomial parameterisation

Note that the negative binomial may be *parameterised* in different ways:

- In the version we've used, an alternative is to express r as a function of the mean and p:

$$r = \frac{p}{1-p}\mathbb{E}(Y)$$

- and then to place a prior on p, e.g. Uniform on some reasonable subset of $[0, 1]$
- This means $Var(Y) = \mathbb{E}(Y)/p$, i.e. $1/p$ is a measure of the over-dispersion.
  - $p = 1 \Rightarrow$ *Poisson*
  - $p = 0.5 \Rightarrow Var(Y) = 2\mathbb{E}(Y)$
  - $p = 0.1 \Rightarrow Var(Y) = 10\mathbb{E}(Y)$
  - $p \Rightarrow 0$ implies huge over-dispersion
- *But* in this 'flu example, trouble with mixing when specifying Uniform prior on p, hence use of prior for $1/\log(r)$

Notes

Model code: functional parameters

```
## For each age group
for(a in 1:A)
{
    ## Population size is a constant
    for(l in INF:INF)
    {
        Ncont[a,l] <- p[a,l] * Npop[a]
        N[a,l] <- round(Ncont[a,l])
    }
    ## Number of infections at each other severity level l
    for(l in SYM:DEA)
    {
        Ncont[a,l] <- Ncont[a,l-1] * p[a,l-1]
        N[a,l] <- round(Ncont[a,l])
    }
    ## post-pandemic prevalence
    pi[a,2] <- pi[a,1] + p[a,INF]
}
```


Notes

Model code: priors

```
## For each age group
for(a in 1:A)
{
    ## ~40% (30-50%) for S|I
    ## ~10% (0-33%) for G|S, flat for rest
    for(l in INF:DEA)
    {
        p[a,l] ~ dbeta(pA[l], pB[l]) ## set in data file
    }
    ## Flat prior for pre-pandemic prevalence,
    pi[a,1] ~ dbeta(1,1)
}
## over-dispersion parameter
for(l in 1:3)
{
    r[l] <- round(exp(lnR[l]))
    lnR[l] <- 1 / invLnR[l]
    invLnR[l] ~ dunif(odL,odU) ## prior on interpretable scale
}
```

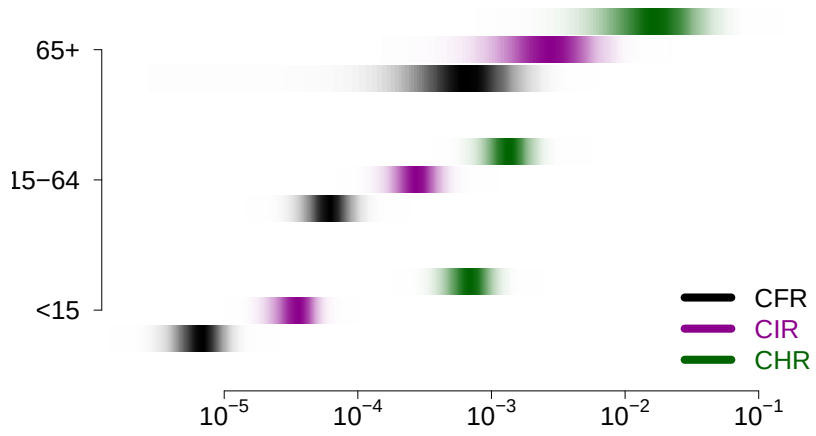

Notes

Model code: likelihood

```
## Observed prevalences pre- and post-pandemic
for(t in 1:2)
{
    ypi[a,t] ~ dbin(pi[a,t], npi[a,t])
}
## Observed number at 3 levels, G, H, D
## Negative binomial, allowing for over-dispersion
for(l in 1:3)
{
    y[a,l] ~ dnegbin(psi[a,l],r[l])
    psi[a,l] <- r[l] / (N[a,counts[l]] + r[l])
    ## smaller r, more over-dispersion
}
## Observations informing 2 conditional probabilities
## ICU|HOS, DEA|ICU
for(l in 1:2)
{
    yp[a,l] ~ dbin(p[a,probs[l]], np[a,l])
}
```


Notes

Severity estimates



Notes

Exploratory model criticism - recall

- *Deviance* = -2 times the log-likelihood:
$$D(\boldsymbol{\theta}) = -2 \log\{p(\boldsymbol{y} \mid \boldsymbol{\theta})\}$$
- *Standardised* deviance = -2 times the log-likelihood, standardised by a term that is a function of the data alone:
$$D(\boldsymbol{\theta}) = -2 \left[\log\{p(\boldsymbol{y} \mid \boldsymbol{\theta})\} - \log\{f(\boldsymbol{y})\}\right]$$
- *Saturated* deviance, for exponential family with $\mathbb{E}(\boldsymbol{Y}) = \mu(\boldsymbol{\theta})$, where we set $f(\boldsymbol{y}) = p(\boldsymbol{y} \mid \mu(\boldsymbol{\theta}) = \boldsymbol{y})$, e.g. observation y for Normal or MLE y/n for Binomial:
$$\begin{aligned} D(\boldsymbol{\theta}) &= -2[\log\{p(\boldsymbol{y} \mid \boldsymbol{\theta})\} - \log\{p(\boldsymbol{y} \mid \mu(\boldsymbol{\theta}) = \boldsymbol{y})\}] \\ &= \sum_{i=1}^{n} D_i(\boldsymbol{\theta}) \end{aligned}$$
- *Posterior mean* deviance: $\overline{D} = E_{\theta|\boldsymbol{y}}[D(\boldsymbol{\theta})]$

Notes

Exploratory model criticism

For exponential family, when model *true* and under standard regularity conditions: $E_Y[\overline{D}] = E_Y E_{\theta|y}[D(\boldsymbol{\theta})] \approx n$

> $\Rightarrow$ *Informally assess model fit by comparison of* $\overline{D}$ *or* $\overline{D_i}$ *to number of data points* n *or* 1 *respectively*

Note however that since distribution of posterior deviance not well understood, we can't say how *far* away from n or 1 $\overline{D}$ or $\overline{D_i}$ need to be to count as lack of fit $\Rightarrow$ *exploratory* only

Saturated deviance examples:

$$y \sim Poi(\theta) : D(\theta) = 2\left\{y \log\left[\frac{y}{\theta}\right] - (y - \theta)\right\}$$

$$y \sim Bin(n, \theta) : D(\theta) = 2\left\{y \log\left[\frac{y/n}{\theta}\right] + (n - y) \log\left[\frac{1 - (y/n)}{1 - \theta}\right]\right\}$$

$$y \sim N(\theta, \sigma^2) : D(\theta) = \left[\frac{y - \theta}{\sigma}\right]^2$$

Notes

BUGS/JAGS model code: deviances

```
for(t in 1:2)
{
   ## Binomial likelihood
   ypi[a,t] ~ dbin(pi[a,t], npi[a,t])

   ## for model criticism - on original scale
   ## mean of sampling distribution and deviance
   yhatpi[a,t] <- pi[a,t] * npi[a,t]
   devpi[a,t] <- 2 * (ypi[a,t] * (log(ypi[a,t]) - log(yhatpi[a,t]))
               + (npi[a,t] - ypi[a,t]) * (log(npi[a,t] - ypi[a,t])
                                  - log(npi[a,t] - yhatpi[a,t])))

   ## on logit scale
   lgtpi[a,t] <- logit(pi[a,t])
   devlgtpi[a,t] <- -2 * ((ypi[a,t] * (lgtpi[a,t] - log(ypi[a,t]
                  / npi[a,t]) + log((npi[a,t] - ypi[a,t]) /
                  npi[a,t]))) - (npi[a,t] * (log(1 + exp(lgtpi[a,t]))
                  + log((npi[a,t] - ypi[a,t]) / npi[a,t]))))
}
```


Notes

Deviance summaries

	y	n	y/n	$\overline{y}$	$\overline{\theta}$	$\overline{D(\theta)}$	$D(\overline{\theta})$	p_D	DIC
yhatc[1,1]	41	883	0.05	44	0.05	1.13	0.20	0.92	2.05
yhatc[1,2]	41	210	0.20	41	0.20	0.99	0.01	0.98	1.96
yhatc[2,1]	350	1722	0.20	352	0.20	0.98	0.01	0.97	1.95
yhatc[2,2]	374	1676	0.22	374	0.22	1.01	0.00	1.01	2.01
yhatc[3,1]	22	168	0.13	27	0.16	2.08	1.23	0.85	2.94
yhatc[3,2]	13	52	0.25	13	0.25	0.97	0.01	0.96	1.93
yhatN[1,1]	101039			798200	0.00	18.80	19.33	-0.53	18.26
yhatN[1,2]	1247			2098	0.00	2.48	1.50	0.99	3.47
yhatN[1,3]	94			84	0.17	1.08	0.58	0.50	1.58
yhatN[2,1]	331773			673550	0.00	3.34	3.36	-0.02	3.32
yhatN[2,2]	2256			3358	0.00	1.70	0.86	0.83	2.53
yhatN[2,3]	576			547	0.03	0.90	0.31	0.59	1.48
yhatN[3,1]	6982			11660	0.00	2.83	1.86	0.96	3.79
yhatN[3,2]	210			776	0.01	9.80	8.48	1.32	11.13
yhatN[3,3]	144			111	0.14	2.37	2.47	-0.10	2.27
yhatpi[1,1]	10	359	0.03	12	0.03	1.27	0.25	1.02	2.28
yhatpi[1,2]	236	423	0.56	228	0.54	1.63	0.65	0.98	2.60
yhatpi[2,1]	66	495	0.13	71	0.14	1.49	0.48	1.01	2.50
yhatpi[2,2]	379	978	0.39	369	0.38	1.36	0.41	0.95	2.31
yhatpi[3,1]	128	549	0.23	147	0.27	4.00	3.36	0.64	4.64
yhatpi[3,2]	168	445	0.38	147	0.33	4.78	4.30	0.49	5.27
TOTAL						64.97	49.68	15.30	80.27

Notes

Deviance summaries

	y	$\overline{y}$	$\overline{\theta}$	$\overline{D(\theta)}$	$D(\overline{\theta})$	p_D	DIC
yhatN[1,1]	101039	798200	0.00	18.80	19.33	-0.53	18.26
yhatN[1,2]	1247	2098	0.00	2.48	1.50	0.99	3.47
yhatN[1,3]	94	84	0.17	1.08	0.58	0.50	1.58
yhatN[2,1]	331773	673550	0.00	3.34	3.36	-0.02	3.32
yhatN[2,2]	2256	3358	0.00	1.70	0.86	0.83	2.53
yhatN[2,3]	576	547	0.03	0.90	0.31	0.59	1.48
yhatN[3,1]	6982	11660	0.00	2.83	1.86	0.96	3.79
yhatN[3,2]	210	776	0.01	9.80	8.48	1.32	11.13
yhatN[3,3]	144	111	0.14	2.37	2.47	-0.10	2.27

Note: plug-in deviance $D(\overline{\theta})$ uses posterior *median* $\overline{\theta}$ as plug-in estimate, but still obtain some *negative* pD contributions: may occur when posterior is *skewed* or if sampling distribution is *not log-concave*, or if there is *strong prior-data conflict*

Notes

Alternatively - log/logit scale

	y	$\overline{y}$	$\overline{\theta}$	$\overline{D(\theta)}$	$D(\overline{\theta})$	p_D	DIC
yhatN[1,1]	101039	805080	0.00	19.29	19.22	0.07	19.36
yhatN[1,2]	1247	2236	0.00	2.54	2.13	0.41	2.95
yhatN[1,3]	94	83	0.32	1.09	0.60	0.48	1.57
yhatN[2,1]	331773	688573	0.00	3.52	3.33	0.19	3.71
yhatN[2,2]	2256	3563	0.00	1.74	1.35	0.39	2.14
yhatN[2,3]	576	538	0.17	0.90	0.29	0.61	1.51
yhatN[3,1]	6982	13331	0.00	2.99	2.07	0.91	3.90
yhatN[3,2]	210	810	0.01	9.93	9.75	0.18	10.10
yhatN[3,3]	144	108	0.29	2.38	3.56	-1.19	1.19

Note: deviances calculated using $log(r_\ell)$ and $logit(p_{a,\ell})$, due to posterior skewness, but using posterior mean as plug-in. $p_{Di} < 1$ since each N is a function of N at lower severity levels.

Notes

Practical: evidence synthesis

- *HIV example:* Explore evidence synthesis
- *'Flu example:*
  - Exploring *deviance* summaries
  - Calculating *functional parameters* representing the case-severity risks
  - *Informative prior on over-dispersion*: what happens when we allow for more over-dispersion (r_ℓ smaller) or for less over-dispersion (r_ℓ large enough that negative binomial likelihood approaches Poisson)?

Notes

Further resources

BUGS book chapter 11.4 on evidence synthesis

NMA references Lu & Ades JASA 2006, Dias et al SiM 2010, Higgins et al RSM 2012, White et al RSM 2012, Jackson et al SiM 2014

More complex evidence synthesis e.g. synthesising individual- and aggregate-level data, see ecological bias model in BUGS book (11.4.1)

Other fields Evidence synthesis references in epidemiology, medical decision-making, genomics, ecology etc: Spiegelhalter et al 2004, Ades & Sutton 2006, Jackson et al 2008, Clark et al 2010, Henderson et al 2010, Tom et al 2011, Welton et al 2012, De Angelis et al 2014

Deviance/DIC Spiegelhalter et al JRSS(B) 2002, Celeux BA 2006, Plummer Biostatistics 2008, Watanabe JMRL 2013

Thanks to Dan Jackson

Notes

Notes

Part VII

Model criticism

Overview

- Influence & Conflict
- Bias modelling - accommodation
- Practical/further resources

Notes

What are the components of inference?

Model criticism requires understanding the role of three components: the model *structure* (part of prior? Box 1980, Efron 2010); the *prior*; and the *likelihood.*

Influence & Conflict

- *Detection/quantification: heterogeneity vs conflict?*
  - Cross-validatory / posterior / mixed *predictive* methods
    Rubin 1984, Gelman 1996, Bayarri & Berger 1999, Marshall & Spiegelhalter 2007
  - *Conflict diagnostics*, including prior-predictive
    Box 1980, O'Hagan 2003, Evans & Moshonov 2006, Gåsemyr & Natvig 2009, Presanis et al 2013, Gåsemyr 2015
- *Resolution:*
  - *Drop* suspect/biased data
  - *Expand* the model to account for biased data
  - May require *evidence expansion* (external evidence) for *identifiability*

Notes

Defining/measuring conflict

Existing methods can be generalised to the question:

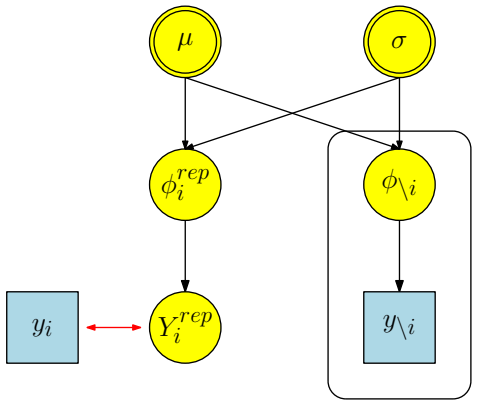
- *How do we compare two sets of evidence?*
  - *prior-data* [prior-predictive methods]
  - *posterior-data* [posterior-predictive methods]
  - *data-data* [outlier detection, conflict diagnostics]

Bayesian predictive diagnostics all involve comparison of a test statistic $T(\boldsymbol{y})$ with a reference distribution $p_T\{T(\boldsymbol{Y}^{rep}) \mid M\}$ of the test statistic for hypothetical (replicated) data under the assumed (null) model M.

Notes

Recall: CV mixed-predictive

Is unit i *consistent* with other units, i.e. comes from RE dist'n?
(e.g. hospital mortality in *Bristol*)



Notes

Model code: Bristol example

```
## Cross-validation: repeat data set leaving one out each time
for(k in 1:N)
{
   ## 12 hospitals
   for(i in 1:N)
   {
      ## copy data 12 times, once for each cross-validation
      y[k,i] <- ydata[i]
      n[k,i] <- ndata[i]

      ## p is the proportion of surgeries ending in death
      y[k,i] ~ dbin(theta[k,i], n[k,i])

      ## on logit scale except k'th left out
      logit(theta[k,i]) <- ((1 - equals(k,i)) * phi.re[k,i])

      ## RE distribution
      phi.re[k,i] ~ dnorm(mu[k], tau[k])
```

Note in JAGS, copying data in code won't work, need instead to either supply copied data in data list or in data{} transformation block.

Notes

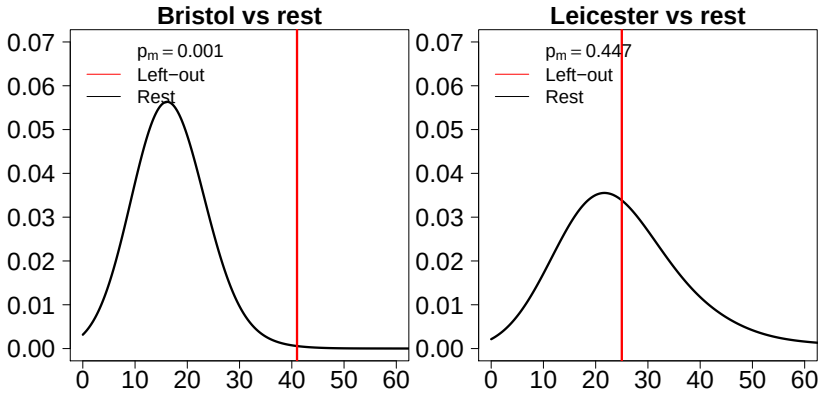
Model code

```
   }

   ## Priors for parameters of RE distribution
   mu[k] ~ dunif(-100,100)
   tau[k] <- 1 / pow(sd[k],2)
   sd[k] ~ dunif(0,100)

   ## replicated data for kth cross-validation
   yrep[k] ~ dbin(theta.rep[k], ndata[k])
   logit(theta.rep[k]) <- phi.re[k,k]

   ## cross-validatory mixed-predictive one-sided mid-p-value
   pval.cvmix[k] <- step(yrep[k] - ydata[k] - 0.5)
                              + (0.5*equals(yrep[k], ydata[k]))
}
```


Notes

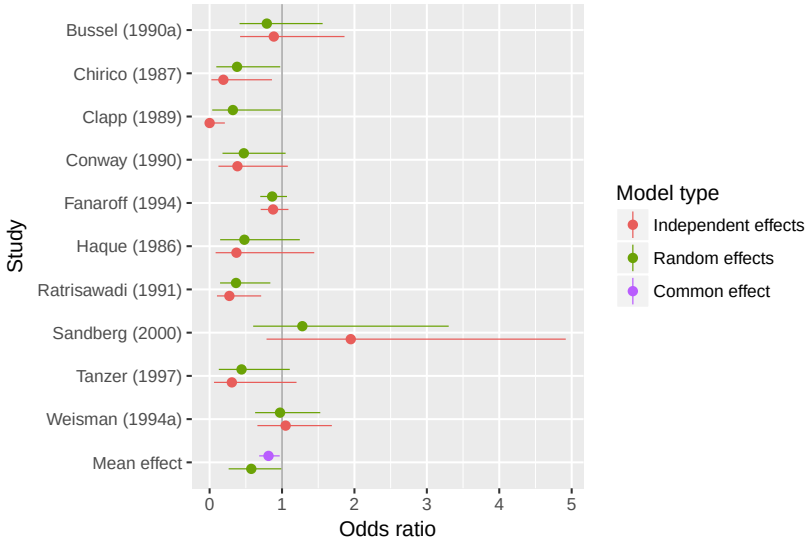
Bristol example

Cross-validatory mixed-predictive p-value



Notes

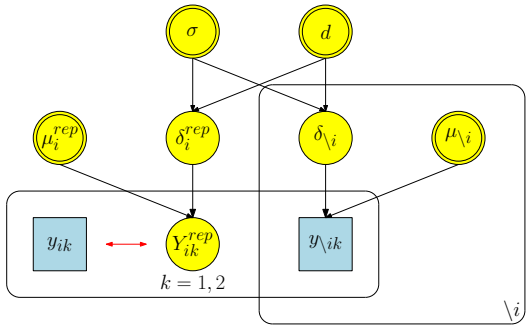
Sepsis example - recall...



Notes

CV mixed-predictive: sepsis example

Is unit i *consistent* with other units, i.e. comes from RE dist'n?

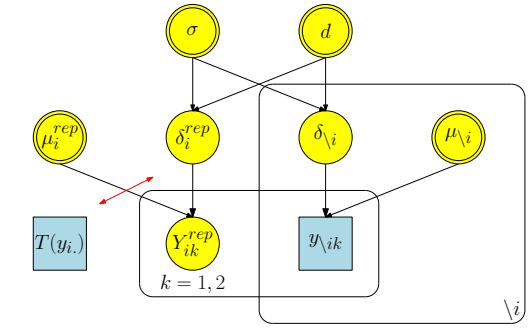


$\Rightarrow$ choice of *appropriate* test statistic

Notes

CV mixed-predictive: sepsis example

Is unit i *consistent* with other units, i.e. comes from RE dist'n?



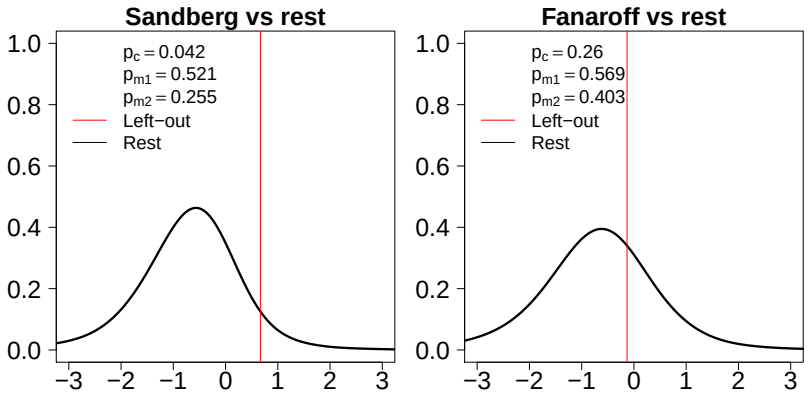
$\Rightarrow$ choice of *appropriate* test statistic

$T(y_{i\cdot}) = \texttt{logit}(y_{i,2}/n_{i,2}) - \texttt{logit}(y_{i,1}/n_{i,1})$

Notes

Sepsis example: choice of test statistic

Comparing y_{ik} to Y_{ik}^{rep} (p_m), vs comparing
$T(y_{i\cdot}) = \texttt{logit}(y_{i,2}/n_{i,2}) - \texttt{logit}(y_{i,1}/n_{i,1})$ to δ_i^{rep} (p_c)



Notes

- Influence & Conflict
- Bias modelling - accommodation
- Practical/further resources

Notes

Resolution of conflict

Different options

- *Exclude* suspect/biased data (*subjective* judgement)
- *Robustify* e.g. random effects distributions by allowing heavier tails to *accommodate* conflict/outliers
- Introduce extra parameters to *model suspected biases*
  - Important to use any possible *external* evidence to derive informative priors for any bias parameters (even if just on *direction* of bias)
  - Don't just "mop up" any lack of fit/conflict without *understanding* what the biases may be

Notes

Systematic bias adjustment

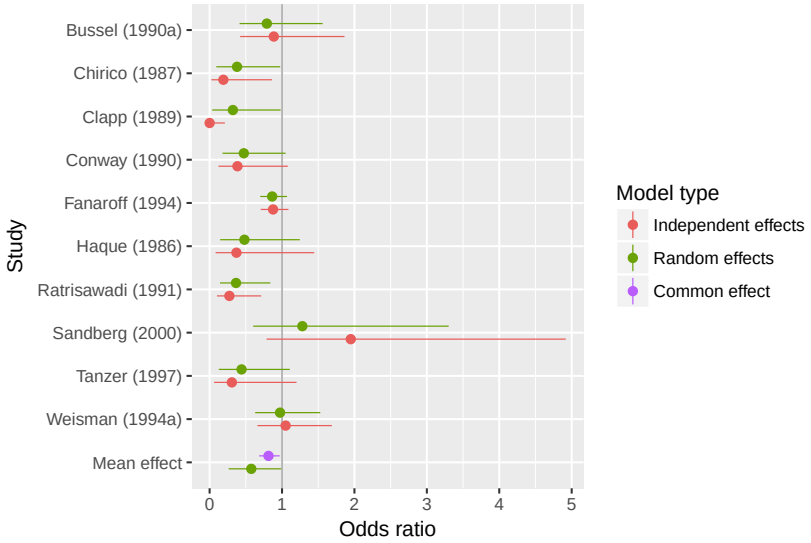
e.g. in meta-analysis of clinical trials, there are a number of recognised issues in poorer quality trials that might bias results:

- unclear/inadequate *sequence generation*
- unclear/inadequate *allocation concealment*
- unclear/inadequate *blinding*
- *incomplete/missing* data

⇒ Cochrane Collaboration's *"Risk of bias"* tool Turner et al JRSS(A) 2009, Higgins et al BMJ 2011, Savovic et al Ann Int Med 2012 allows systematic assessment/judgement of bias in clinical trials.

Notes

Sepsis example - recall...



Notes

Aside: Leverage and deviance residuals

Note that the total deviance can be interpreted as the sum of the squared *deviance residuals*:

$$D(\boldsymbol{\theta}) = \sum_{i=1}^{n} D_i(\boldsymbol{\theta}) = \sum_{i=1}^{n} (sign_i \times \sqrt{D_i})^2$$

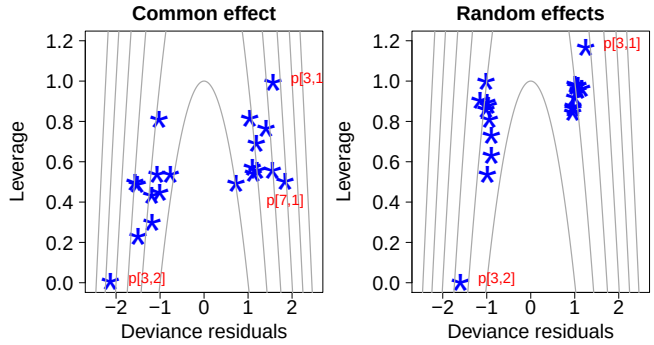
where $d_i = sign_i \times \sqrt{D_i}$ is a deviance residual.

Similarly, the effective number of parameters, $p_D = \sum_{i=1}^{n} p_{Di}$, has an interpretation where the contribution of each data point, p_{Di}, is a measure of the *leverage* of that data point. Spiegelhalter et al JRSS(B) 2002

Plotting d_i against p_{Di} gives a visual look at the goodness of fit and influence of different data points. Can also visualise contribution to *DIC*.

Notes

Aside: Leverage vs deviance residuals



- larger (absolute) deviance residuals ⇒ poorer (*absolute*) fit
  - but fit less important if data don't *influence* the posterior
  - hence *penalise* fit by leverage = p_D contribution
- contours show constant contributions to DIC = penalised deviance
  - points on farthest contours have largest/poorest DIC

Notes

Prior judgements on risk of bias - Cochrane

Systematic *elicitation* of expert opinion on potential biases, each of which is modelled. (Higgins et al BMJ 2011, Ohlssen & Lacy 2013).

Study	SeqGen	AllCon	BlindUnspec	IncompData
Bussel (1990a)	Unclear	Low risk	Low risk	High risk
Chirico (1987)	Unclear	Low risk	High risk	Low risk
Clapp (1989)	Unclear	Low risk	Unclear	Low risk
Conway (1990)	Unclear	Low risk	High risk	Low risk
Fanaroff (1994)	Low risk	Low risk	Low risk	Low risk
Haque (1986)	Unclear	Low risk	High risk	Low risk
Ratrisawadi (1991)	Unclear	Unclear	High risk	Unclear
Sandberg (2000)	Low risk	Low risk	Low risk	Low risk
Tanzer (1997)	High risk	High risk	High risk	Low risk
Weisman (1994a)	Unclear	Low risk	Low risk	Low risk

Notes

Bias-adjustment model

Expert judgements on risk of (*internal*) biases summarised by $\beta_{ij} \sim f(\nu_{ij}, \tau^2_{ij})$, for each bias j in study i

Turner et al JRSS(A) 2009, Savovic et al Ann Int Med 2012

After elicitation process, assume internal biases *independent*, so that total internal bias per study i is:

$$\beta_i \sim f(\nu_i = \sum_j \nu_{ij}, \tau^2_i = \sum_j \tau^2_{ij})$$

Then an *additive bias* model is assumed for the treatment effect in arm k versus baseline treatment 1:

$$logit(p_{i1}) = \mu_i$$
$$logit(p_{ik}) = \mu_i + \delta^{bias}_{ik}, \quad k > 1$$
$$\delta^{bias}_{ik} = \delta_{ik} + \beta_i$$

$$\delta_{ik} \sim N(d_k, \sigma^2)$$
$$\beta_i \sim N(\nu_i, \tau^2_i)$$

Notes

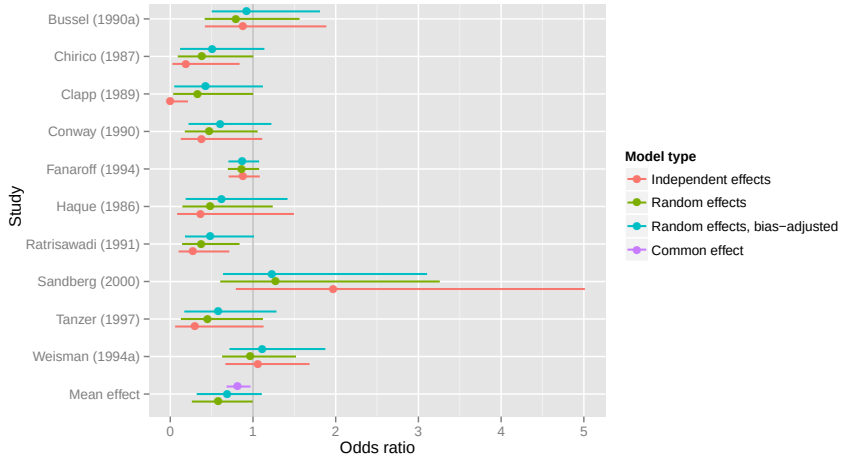
Model code

```
## bias/random effect model
logit(p[i,k]) <- mu[i] + deltaB[i,k]
deltaB[i,k] <- delta[i,k] + bias[i]
delta[i,k] ~ dnorm(d[k], prec.d)

## priors - risky studies are those with at least one high
## risk judgement
for(i in 1:Nsafe) { bias[safe[i]] <- 0 }
for(i in 1:Nrisky)
{  bias[risky[i]] ~ dnorm(bias.mu, bias.prec)  }

## same informative prior for all studies implies average
## multiplicative effect of 0.82 on odds ratio for treatment
## effect, approx prior 95% range (0.67,1) from Savovic (2012),
## implying a prior belief that treatment effects are
## exaggerated by flawed study conduct.
bias.mu <- -0.2
bias.prec <- 1 / (0.1^2)
```


Notes

Bias-adjusted results

Mean bias-adjustment of 0.82 on treatment effect *moderates* estimate of treatment effect towards 1 for *"risky"* studies (Bussel, Chirico, Conway, Haque, Ratrisawadi, Tanzer):



Notes

Practical: conflict and bias

Two exercises:

- *Sepsis example*: calculate *cross-validatory mixed-predictive* p-values for each study, to detect outliers.
- *Sepsis example*: extending from a *single* elicited prior for all four bias types (averaged over biases) to *different* elicited priors for each bias type, summed to a single prior rather than averaged.

Thanks to Rebecca Turner

Notes

Model criticism - further resources

- *Deviance summaries/information criteria*: Spiegelhalter et al JRSS(B) 2002, Celeux et al BA 2006, Plummer Biostat 2008
- Bayesian *predictive* diagnostics: Box 1980, Rubin 1984, Gelman 1996, Bayarri & Berger 1999, O'Hagan 2003, Evans & Moshonov 2006, Marshall & Spiegelhalter 2007
- including *conflict p-values*: Gåsemyr & Natvig 2009, Presanis et al 2013, Gåsemyr 2015, Presanis et al 2017
- Conflict/inconsistency in *network meta-analysis*: Lu & Ades 2006, Dias et al 2010, White et al 2012, Higgins et al 2012, Jackson et al 2014
  - `www.bristol.ac.uk/social-community-medicine/projects/mpes/courses/treatment-comparisons/`
- *Bias* modelling: Turner et al 2008, Presanis et al 2008
- *Identifiability*: Gustafson Stat Sci 2005, IJB 2010, Greenland Stat Sci 2009
- *Sensitivity analysis*: Saltelli Sensitivity Analysis 2000, Oakley & O'Hagan JRSS(B) 2004
- How to *weight* evidence?: e.g. power priors Neuenschwander et al SiM 2009

Notes

Notes

Advanced Bayesian Modelling with BUGS
Practical Exercises

- The practical exercises involve editing and running BUGS models, and interpreting the results.
- **Just follow this question sheet.**
- For each exercise, computer material is provided in the folder `practical_material`, giving template BUGS models and data to help you answer the questions. These are given for three alternative varieties of the BUGS software. Use whichever you are most comfortable with.

  OpenBUGS Windows interface :
  filenames ending `-OpenBUGS.odc`. OpenBUGS compound documents containing code, data and annotations.

  JAGS run through the `rjags` R package :
  filenames ending `-jags.r`. R scripts containing code, data, annotations, and commands to run the models.

  OpenBUGS run through the `R2OpenBUGS` R package :
  filenames ending `-R2OpenBUGS.r`. R scripts containing code, data, annotations, and commands to run the models.
- The solutions and discussions of the answers are provided in the same folders, with similar file names containing `-solutions`. Feel free to look at these solutions as you are going along.
- There won't be enough time to do all the questions, so just start with the topics that you want to learn better.

1 Hierarchical models: meta-analysis and priors

Model templates	Solutions
`hier-jags.r`	`hier-jags-solutions.r`
`hier-R2OpenBUGS.r`	`hier-R2OpenBUGS-solutions.r`
`hier-OpenBUGS.odc`	`hier-OpenBUGS-solutions.odc`

1.1 Meta-analysis

The Sepsis data (discussed in the lecture) aim to answer the question: 'does administration of intravenous immunoglobulin (IVIG) prevent infection in hospital, compared to placebo?' The raw data from the 10 studies are reproduced in the table below.

		Treatment (IVIG)		Control	
Study	Country	Events	Total	Events	Total
Bussel (1990a)	US	20	61	23	65
Chirico (1987)	Italy	2	43	8	43
Clapp (1989)	US	0	56	5	59
Conway (1990)	UK	8	34	14	32
Fanaroff (1994)	US	186	1204	209	1212
Haque (1986)	Saudi Arabia	4	100	5	50
Ratrisawadi (1991)	Thailand	10	68	13	34
Sandberg (2000)	Sweden/Austria	19	40	13	41
Tanzer (1997)	Turkey	3	40	8	40
Weisman (1994a)	US	40	372	39	381

Table 1: Sepsis data

1. We wish to fit a random-effects meta-analysis model for odds ratio.

$$y_{sT} \sim \text{Bin}(n_{sT}, p_{sT}) \qquad y_{sC} \sim \text{Bin}(n_{sC}, p_{sC})$$
$$\text{logit}(p_{sT}) = \alpha_s + \mu_s \qquad \text{logit}(p_{sC}) = \alpha_s$$
$$\mu_s \sim N(\delta, \tau^2)$$

   Here y_{sT} is the number of events in study s in the IVIG group, and n_{sT} is the total number in the IVIG group; y_{sC} and n_{sC} are the corresponding numbers in the control group.

   Fit this meta-analysis model using the model and data provided.

   (a) The parameter δ is the overall log odds ratio treatment effect. What does the posterior distribution of this parameter suggest about the effectiveness of the treatment?

(b) The figure in the lecture shows the odds ratio treatment effect rather than the log odds ratio. Obtain a posterior sample for this quantity.

2. Modify the code to fit a common effect meta-analysis for the odds ratios. How does the estimated treatment effect change? Can you explain why this happens?

3. The table above gives the country in which each study was conducted. Is there evidence that the treatment effect is higher in US hospitals?

   [Hint: A meta-regression model, with an indicator variable for whether the study was conducted in the US, would be one way to assess this. Alternatively, consider the model separately for US and non-US studies - how does would this approach differ to the meta-regression approach?]

1.2 Prior sensitivity in hierarchical models

Transfusions of granulocytes (a type of white blood cell) have been used clinically for many years to support and treat severe infection in high risk groups of patients with neutropenia or neutrophil dysfunction. However, there is still uncertainty about whether such transfusions are beneficial. In 2005, the Cochran Collaboration published a systematic review of the available evidence. Six studies were included in their study of all-cause mortality. We will explore the dependence of the results on the priors are chosen.

1. Consider the meta-analysis model used in the Sepsis example above. Imagine you want to be sceptical about the effect δ of transfusions of granulocytes. Specifically, you think *apriori* that

   - there is only a 2.5% chance that transfusions will reduce the odds of mortality by 25% or more
   - and that it is most likely there is no effect i.e. the mode of the prior is 1 on the odd ratio scale, and 0 on the log odds ratio scale.

   Formulate a prior distribution matching these beliefs.

   [Hint: the first condition is equivalent to a 2.5% chance that the log odd ratio is less than -0.29]

2. Fit the meta-analysis model with the 'vague' prior $\delta \sim \mathrm{N}(0, 10^2)$ using the supplied model and data files. Compare the results to the results with your sceptical prior, and using an alternative vague prior $\delta \sim \mathrm{Unif}(-100, 100)$.

3. Compare the following priors for the random effect variance sd.mu^2.

   (a) Uniform$(0, 10)$ for the standard deviation
   (b) Uniform$(0, 100)$ for the standard deviation
   (c) Gamma$(1, 1)$ for the precision
   (d) Gamma$(0.001, 0.001)$ for the precision

Either run each of these in turn, or try to run them all in the same model file, taking care to ensure the models under each prior are completely separate. The data set with the data repeated in the supplied file (and initial values) might be useful. Ask (or see the solutions file) if you are unsure!

4. Turner et al (2012) recommend for meta-analysis of all-cause mortality with non-pharmacological interventions a log-normal$(-3.93, 1.51^2)$ informative prior for the distribution of the variance τ^2, where 3.93 is the mean on the log scale, and 1.51^2 is the variance on the log scale. How does that change the results?

   [Hint: log normal distributions are specified as `dlnorm(a, b)` *in BUGS, where* `a` *is the mean on the log scale and* `b` *is the precision on the log scale]*

2 Missing Data

Model templates	Solutions
`missing-jags.r`	`missing-jags-solutions.r`
`missing-R2OpenBUGS.r`	`missing-R2OpenBUGS-solutions.r`
`missing-OpenBUGS.odc`	`missing-OpenBUGS-solutions.odc`

2.1 Outcomes missing not at random

The Mini Mental State Examination (MMSE) is a measure of cognitive function used in the diagnosis of dementia. It can take values from 30 (representing no cognitive impairment) to 0 (the most severe), and we assume it is continuous. A person in a longitudinal study had four MMSE observations of 28, 26, 27 and 25 respectively at baseline, and years 5, 10 and 15 after baseline. A planned fifth observation at 20 years was missing.

We wish to fit a standard linear regression model of the form $y_i \sim N(\mu_i, \sigma_i^2)$, $\mu_i = \alpha + \beta x_i$, where y_i is the MMSE value, and x_i is the year after the baseline measurement.

1. Devise suitable priors for the intercept α, slope β and error standard deviation $\sigma = 1/\sqrt{\tau}$, given

   - MMSE is bounded by definition
   - from previous knowledge, we are 97.5% certain that MMSE cannot increase over time, and 97.5% certain that after 1 year it cannot drop more than 20 points *[note that a 95% credible interval is about* ± 2 *standard deviations in the normal distribution]*

   [Note there is no single correct answer: the point is to get used to the idea of translating prior beliefs to distributions. Indeed the normal error model is not strictly realistic given that MMSE is bounded.]

2. Fit the regression model with the given priors, and the template model and data provided, and compare the posteriors to the priors.

   Note that the code also includes a binary variable `p20` to monitor, whose posterior mean will be the posterior probability that the missing MMSE value is less than 20, defining a moderate cognitive impairment.

3. Suppose now that the odds of missing a MMSE measurement depend on the measurement itself, such the odds are doubled if MMSE is 5 units lower. Adapt the model, data and initial values to include the non-random missingness mechanism (see slide 62Doc-Start), run and compare the results.

4. Suppose now that instead of a normal error distribution, we assume the error distribution is a t distribution with 4 degrees of freedom. Implement this in the model. What difference does this make to the chance of the

missing value being below 20, and the regression slope? *[Hint: the t distribution in BUGS is* `dt(mu, tau, df)`*. As* `df` *increases, the t distribution converges to the normal distribution* `dnorm(mu, tau)`*, and for low* `df` *the fatness of the tails compared to the centre of the distribution increases, so outlying values are more likely.]*

2.2 Missing covariates

This example uses a dataset on childhood malaria in the Gambia [1]. Data are available for 805 children including a binary indicator Y for whether the child has malaria, and four predictors:

- AGE: child's age in years,
- BEDNET: binary indicator for use of bed nets
- GREEN: measure of greenness of the child's environment (related to mosquito prevalence)
- PHC: binary indicator for whether child's village is in the primary health care system

We wish to fit a logistic regression model, however about 40% of the values of the covariate BEDNET are missing. We assume it is missing at random, but not completely at random.

The data are provided as `"malaria_data.txt"` for R, or embedded in the `.odc` file if running from the OpenBUGS Windows interface.

1. The R script provided contains the BUGS code to fit a logistic regression to the full data. Add to this code an imputation model for the missing variable BEDNET. This should be a logistic regression model with AGE, GREEN and PHC as predictors.

$$\begin{aligned} \text{BEDNET}_i &\sim Bern(q_i) \\ logit(q_i) &= \gamma_1 + \gamma_2 \text{AGE}_i + \gamma_3 \text{GREEN}_i + \gamma_4 \text{PHC}_i \end{aligned}$$

   and use any vague priors for the coefficients. The initial values will also need to be modified.

   Monitor and produce summary statistics for the odds ratios, coefficients of the imputation model, and the imputed values of `BEDNET[1:10]` and `BEDNET[531:540]`. *[Assume that 10000 iterations with a burn-in of 1000 are sufficient to understand the results.]*

[1] many thanks to Nicky Best and Alexina Mason for this example, taken from their course `http://www.bias-project.org.uk/Missing2012/MissingIndex.htm`. The original example is from 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, 493506.

Compare with an analysis of only the data where all covariates were completely observed, which gave the following estimates for the odds ratios:

	Posterior median	95% credible interval
1 year of age	1.42	1.17–1.72
BEDNET	1.10	0.70–1.77
Green	0.96	0.90–1.02
PHC	0.55	0.36–0.86

2. Rows 1–10 and 531–540 of the original data are presented below. Can you explain the posterior means of the imputed values for BEDNET under your fitted model? *[hint, consider the estimated coefficients of the imputation model, and how BEDNET influences the outcome Y]*

	Y	AGE	BEDNET	GREEN	PHC
1	0	2	NA	40.85	1
2	1	2	NA	40.85	1
3	1	2	NA	40.85	1
4	1	3	NA	40.85	1
5	1	4	NA	40.85	1
6	1	4	NA	40.85	1
7	0	2	NA	40.85	1
8	1	2	NA	40.85	1
9	1	3	NA	40.85	1
10	1	3	NA	40.85	1
531	0	1	NA	40.85	0
532	0	2	NA	40.85	0
533	1	2	NA	40.85	0
534	0	3	NA	40.85	0
535	0	3	NA	40.85	0
536	0	3	NA	40.85	0
537	1	3	NA	40.85	0
538	0	1	NA	40.85	0
539	0	1	NA	40.85	0
540	0	2	NA	40.85	0

3 Censoring and related topics

The following exercises ask for BUGS code to be written from scratch. Feel free to work through the solutions document instead if you don't feel like doing this!

Model templates	Solutions
`censtrunc-jags.r`	`censtrunc-jags-solutions.r`
`censtrunc-R2OpenBUGS.r`	`censtrunc-R2OpenBUGS-solutions.r`
`censtrunc-OpenBUGS.odc`	`censtrunc-OpenBUGS-solutions.odc`

3.1 Censoring

1. (a) For the rounded chickens data presented in the lecture, slides 91Doc-Start–92Doc-Start, fit a model to estimate the mean weight and predict the true weight of each chicken, using the techniques described for interval-censoring. Assume again that the weights are normally distributed with variance 1, and initial values will be required for the nodes representing the underlying weights. A uniform prior on a restricted range for the mean, as in the lectures, is reasonable.

   Discuss any interesting features of the posterior distributions — how might the results change if the posterior variance were different?

   (b) If you have time, adapt the code to estimate the posterior precision (using a uniform prior on the SD scale).

2. Suppose now that the scales can not show values over 8, so that weights over 8 are rounded down to 8. Weights below 8 are rounded to the nearest integer. We observe nine data points: 6, 6, 6, 7, 7, 7, 8, 8, 8. Adapt the model fitted in part 1(a), fit it again and compare the results.

3.2 Truncation

1. In the chicken weights example, suppose we know that a particular kind of chicken can never weigh more than 8 units. Thus we can only observe chickens weighing 6 or 7 units. Suppose we observe three chickens weighing 6 and three weighing 7, and assume these are exact with no censoring. Implement a truncated normal model for these data, with known SD of 1.

   Take care with the choice of prior distributions and initial values. In OpenBUGS, μ will need to be restricted to have a prior upper bound of 10 or not much more than 10 to avoid sampling errors.

3.3 Zeros trick

1. Implement the truncated normal model for the chickens data in the previous section using the zeros or ones trick, and check the results match those from using `T()`.

2. *(more advanced)* The zeros trick may also be used as an alternative to `I()`, `C()` or `dinterval` to represent censored data.

   The principle (see slide 84Doc-Start of lectures) is that an observation y censored on `[a,b]` contributes a term $P(a < Y < b) = F(b) - F(a)$ to the likelihood. Thus if we know the probability distribution function $F()$ we can use this to define the term `logL` (see also slide 104Doc-Start)

   Use this technique to implement the censored chickens model from the slides, and try to reproduce the results in slide 90Doc-Start. *[Hint: the Normal(0,1) cumulative distribution function in BUGS is* `phi()`. To predict the value of an observation censored above 8, you can simply generate another observation from the assumed normal distribution, using T() to restrict the generated values above 8.

3.4 Survival analysis

[no code needs to be run or written in this section, though the full code, data and discussion are given in the solution files for you to study]

3.4.1 Standard Weibull survival model

We have data giving a set of survival times *in months* (and censoring indicators) for patients with HIV, since entry into a clinical trial [2]. This dataset will also be used in the second lecture on hierarchical models.

Here is BUGS code for a Weibull survival model for these data (see the lecture slide 89Doc-Start for more explanation).

```
model {
 for (i in 1:n) {
  ### JAGS version
  y[i] ~ dweib(nu, lambda)
  is.censored[i] ~ dinterval(y[i], c[i])

  ### OpenBUGS (or WinBUGS) version
#  y[i] ~ dweib(nu, lambda)I(cens[i],)

 }
 lambda <- exp(loglambda)
 nu <- exp(lognu)
 loglambda ~ dnorm(0, 0.0001)
 lognu ~ dnorm(0, 0.0001)
}
```

This question explores the influence of the prior distributions for the parameters of the Weibull. Given the following points:

[2] See the R package `JM`, `help(aids)`, for references

- In BUGS, the survival probability at time t under the Weibull distribution is parameterised as $S(t) = \exp(-\lambda t^{\nu})$.
- The mean survival time is $\lambda^{-1/\nu}\Gamma(1 + 1/\nu)$, or simply $1/\lambda$ for $\nu = 1$ (the exponential distribution).
- The parameter ν controls the rate of increase or decrease of the hazard $h(t) = \lambda\nu t^{\nu-1}$ with time t. Doubling the time since $t = 0$ multiplies the hazard by $hr = 2^{\nu-1}$. Thus by considering plausible ranges for this hazard ratio, we can obtain prior distributions for $\nu = \log(hr)/\log(2) + 1$.

answer these questions:

1. Are the priors for λ and ν given in the BUGS code above plausible for human survival?
2. Suppose we were certain that the mean survival time was between 0 and 60 *years* after study entry, and certain that the hazard ratio for a doubled time was between 1/100 and 100. Express these beliefs approximately as priors in BUGS code, using uniform distributions on the mean survival (under an exponential distribution) and $\log(hr)$, and ignoring any prior correlation between the two parameters. *[no need to actually run the code: it is slow to converge in this example, and with 1000 observations, there's probably enough data to overcome any influence of the prior]*

3.4.2 Weibull survival model with time-dependent covariates (JAGS only)

The second part of this "question" is simply a demonstration of a survival model with a time-dependent covariate for you to study. R users could also check that the results from this model agree roughly with those from the `flexsurv` package that allows parametric survival modelling with truncated data (the code is given).

For this, we use an extended version of the AIDS data with multiple rows for each patient. Each row represents an occasion where the CD4 cell count was measured. The survival data are now given in *counting process* format: instead of simply an event time and censoring indicator, we have three variables `start`, `stop` and `event` giving the start and end time of the interval, and whether the interval ended with death `event=1` or censoring `event=0`. For example, the data for patient 1, censored at 16.97 months, becomes three rows, corresponding to CD4 measurements at 0, 6 and 12 months.

```
   patient       CD4  start  stop event
1        1 10.677078      0  6.00     0
2        1  8.426150      6 12.00     0
3        1  9.433981     12 16.97     0
```

The following JAGS code implements a Weibull survival model with one time-dependent covariate: the hazard in the ith interval is assumed to be proportional to the CD4 x_i measured at the start of this interval. As before, `c[i]` is set to be `stop[i]` for intervals that end in censoring, and an arbitrary number higher than all observed death times for intervals that end in death.

```
model {
 for (i in 1:n) {
  stop[i] ~ dweib(nu, lambda[i])T(start[i], )
  is.censored[i] ~ dinterval(stop[i], c[i])

  log(lambda[i]) <- a[1] + a[2]*CD4[i]
 }
## [ priors ... ]
}
```

Each row i encodes the following contribution to the likelihood:

$$\left.\begin{array}{l} P(T_i > \text{stop}_i | T_i > \text{start}_i, x_i) \\ P(T_i = \text{stop}_i | T_i > \text{start}_i, x_i) \end{array}\right\} \text{if the survival time is} \left\{\begin{array}{l} \text{censored} \\ \text{observed} \end{array}\right. \text{at stop}_i.$$

The `T()` construct for left-truncation is used to represent the condition that at each i the survival time is defined to be greater than start_i.

In OpenBUGS, ideally we would be able to implement this model as

```
  stop[i] ~ dweib(nu, lambda[i])T(start[i], )C(cens[i], )
```

but nodes with both truncation and censoring are not supported. The zeroes trick might be used to implement the truncation and/or censoring: alternatively just use JAGS!

4 Measurement Error

Model templates	Solutions
`measerr-jags.r`	`measerr-jags-solutions.r`
`measerr-R2OpenBUGS.r`	`measerr-R2OpenBUGS-solutions.r`
`measerr-OpenBUGS.odc`	`measerr-OpenBUGS-solutions.odc`

4.1 Binary misclassification

In the cervix example in the slides, consider only the subset of 115 women who had both "gold standard" accurate and inaccurate western blot HSV tests. Run the same model on this subset data only, and compare the results with the results presented in the slides. What are the advantages and potential disadvantages of using the full data while correcting for measurement error, compared to using the gold-standard subset alone?

4.2 Classical measurement error

In the dugongs example from the slides, increase the standard deviation for measurement error around ages from 1 to 5 (i.e. precision 0.04). Monitor and summarise the fitted true ages, and plot the posterior median true age against dugong length. (Example code is provided for R users, and instructions are given in the `odc` file for Windows OpenBUGS users.) What difference does this make and why?

4.3 Berkson measurement error

The calibration data for the pollution example in the slides are provided in the R script, giving 23 observations of true exposure z_j and corresponding observed exposures x_j. Extend the BUGS model code provided to include the normal calibration model which relates observed to true exposures, using vague priors for α, β, σ. (This will not necessarily make much difference to the results – the point of this exercise is to practice implementing this kind of evidence synthesis.)

5 Complex hierarchical models

Model templates	Solutions
`hier2-jags.r`	`hier2-jags-solutions.r`
`hier2-R2OpenBUGS.r`	`hier2-R2OpenBUGS-solutions.r`
`hier2-OpenBUGS.odc`	`hier2-OpenBUGS-solutions.odc`

5.1 Longitudinal data: ratings of teachers by pupils

The `teachers` dataset is a study conducted by Brekelmans and Créton (1993) of interpersonal skills of teachers, as judged by their pupils. We will examine data about the 'proximity' dimension in the data, which aims to represent the degree of cooperation or closeness between a teacher and the pupil in their classes. Each of the 51 teachers were rated when they were newly-qualified, and then after 1, 2 and 3 years of experience. There is some missingness in the data, but this can reasonably be regarded as missing completely at random. Snijders and Bosker (1999) previously analysed these data in a classical framework (see chapter 12).

1. We will first consider a standard linear regression model

$$\begin{aligned} y_{ij} &\sim \mathrm{N}(\mu_{ij}, \sigma^2) & \beta_1, \beta_2 &\sim \mathrm{N}(0, 100^2) \\ \mu_{ij} &= \beta_1 + \beta_2 \mathrm{experience}_{ij} & \sigma &\sim \mathrm{Unif}(0, 100) \end{aligned}$$

   Use the provided model file and data to fit this model. Monitor the regression parameters, the mean `mu`, and the predictions `score.pred`. Obtain summary statistics for the regression parameters.

   Also create 'model fit' plots for both `mu` and `score.pred` for some of the individuals. Code is provided to create model fit plots in R in the supplied file, or in point-and-click OpenBUGS, choose 'Compare' from the 'Inference' menu and then enter, for example, `mu[38:41]` in 'node', `score[38:41]` in 'other' and `experience[38:41]` in 'axis' to obtain a plot for for the mean `mu` for individual 13. Replace `38:41` with appropriate indicies to view other individuals, e.g. `5:6` for individual 3.

2. Add in a random intercept, so that the following model is fitted.

$$\begin{aligned} y_{ij} &\sim \mathrm{N}(\mu_{ij}, \sigma^2) \\ \mu_{ij} &= \beta_{1i} + \beta_2 \mathrm{experience}_{ij} \\ \beta_{1i} &\sim \mathrm{N}(\beta_1, \sigma^2_{\beta_1}) \\ \beta_1, \beta_2 &\sim \mathrm{N}(0, 100^2) \\ \sigma, \sigma_{\beta_1} &\sim \mathrm{Unif}(0, 100) \end{aligned}$$

   Obtain summary statistics and model fits as before. How do these compare to the standard linear regression model? What kind of predictive check is `score.pred` like: prior predictive, posterior predictive or approximate mixed?

3. Repeat the analysis with a random slope for `experience` as well as the random intercept, allowing for correlation between the regression parameters for the intercept and slope. What does this suggest about the relationship between a teacher's initial rating and the 'slope' of their ratings?

4. Obtain predictions under this model for the intercept and slope for a 'new' teacher, and credible intervals for their rating after 0, 1, 2 and 3 years of experience.

5.2 Model checking

This example uses the results of the May 2015 general election in the UK aggregated to region level ('NUTS1' level). There are 12 NUTS1 regions in total: 9 regions in England, plus one for each of Wales, Scotland and Northern Ireland. For each region, the total number of valid votes cast and the number of votes for the Conservative Party are recorded.

A statistician has been asked to understand variability in the proportion of votes that were cast for the Conservative Party. The statistician has left Northern Ireland in the analysis, despite the Conservative Party not fielding any candidates in this region. In this example, we will investigate whether this obvious outlier is revealed using various model checking approaches. Note that all vote counts in this questions are thousands.

1. Use the supplied model code to run an analysis leaving out Northern Ireland. What is the p-value for Northern Ireland?

2. Using all of the regions (including Northern Ireland), edit the supplied model so that 'mixed' approximate cross-validation p-values are calculated. Does the mixed p value suggest any outliers? How do the results compare to the non-approximated check? Can you explain what is happening?

3. What would happen if you used the prior predictive? Or the posterior predictive?

4. Before the May 2015, opinion polls published in UK newspapers indicated that a 'hung parliament' was the most likely outcome, and there was discussion of the possibility of the Conservative party forming a coalition government including the Democratic Unionist Party, which is the largest unionist political party in Northern Ireland. The DUP received 184 thousand votes in Northern Ireland. If this figure were used as `y` (in thousands) for Northern Ireland, how would your results change?

5.3 Joint modelling

1. Use the model and data files supplied to fit the joint model for the first 20 patients in AIDS dataset described in the lectures. Obtain summary

statistics for the parameters `r1` and `r2` that capture the dependence between the processes, and interpret your results.

2. Simplify the joint model to a model that is just for the longitudinal data, and fit this model.

6 Evidence Synthesis

Model templates	Solutions
`evsyn-jags.r`	`evsyn-jags-solutions.r`
`evsyn-R2OpenBUGS.r`	`evsyn-R2OpenBUGS-solutions.r`
`evsyn-OpenBUGS.odc`	`evsyn-OpenBUGS-solutions.odc`

6.1 HIV model

This example, inspired by e.g. Rosinska et al, Epidemiology & Infection (2015) but using made-up data and simplified, is designed for you to understand the basics of evidence synthesis, and the challenges/intricacies that can arise in synthesising disparate, possibly inconsistent, data sources. The table below summarises the available data:

Parameter	y	n	y/n
π	5	100	0.05
$\pi(1-\delta)$	3,000	100,000	0.03
δ	90	100	0.90

Table 2: HIV evidence

1. Implement the code for the model on slide 149, assuming we have only a single data point 5/100 informing prevalence π. You can either implement the code from scratch, or use the provided templates, or look directly at the solution files, for whichever version of `BUGS/JAGS` you are using.

    - Have a look at the traces/history of each chain, at the summary statistics for both the basic and functional parameters, and the densities of each. If you are working in `R`, you can use the provided functions in the `fns.R` file to look at traces and summaries.
    - Change the prior for δ to be informative, with a prior mean of 0.75 and prior standard deviation of approximately 0.04 (the mean of a $Beta(a, b)$ distribution is $a/(a+b)$ and the variance is $ab/((a+b)^2(a+b+1))$). Re-run with this prior, and compare the posterior summaries/densities to the model with vague priors.
    - Now change the prior for π to be informative, with a prior mean of 0.15 and a prior standard deviation of approximately 0.036. Compare again the posterior summaries/densities.

2. Returning to the model with vague priors, now introduce the second data point 3000/100000 informing the functional parameter $\pi(1-\delta)$.

    - Run the model and have a look now at the traces/history. What do you think is happening?

- Look at the joint posterior distribution of π and δ, via a scatter plot of the posterior samples for π against the posterior samples for δ. In point-and-click `OpenBUGS`, you can use the `Inference-Correlations` tool to produce the plot.

3. Finally, introduce the third data point, 90/100, directly informing δ.

    - Have a look again at the history of the chains, the correlation plot and the posterior summaries. Has the mixing of the chains improved?

6.2 'Flu Severity

The purpose of this example is to explore different aspects of a relatively complex evidence synthesis: exploratory model criticism; functional parameters; and likelihood assumptions. In 2009, a new strain of A/H1N1 influenza emerged, that in the UK caused three initial waves of infection, in the summer of 2009, autumn/winter 2009/10 and in the winter season of 2010/11. Various surveillance and survey data sources, as well as prior information from previous 'flu outbreaks, were available for use in an evidence synthesis to estimate severity. Each suffered from under-ascertainment or other biases (Presanis et al, AoAS, 2014). This example uses data similar to, but amended from the original data, to illustrate a simplified version of the original model.

1. Run the code provided for the influenza severity model introduced in the lecture (slides 177-183).

    - Explore the posterior summaries using either the `Inference-Samples` menu in `OpenBUGS` or the summary command in the `R2OpenBUGS` or `rjags` packages.
    - Look also at the total (unstandardised) deviance and DIC estimated directly by `OpenBUGS` or `JAGS`. If you are working in `R`:
        - compare these to calculating the total saturated deviance and DIC based on the saturated deviances explicitly included in the model code - you can make use of the provided functions in the `fns.R` file to do so, see the `R` template/solution files for how to do so.
        - Explore what happens when you use either the posterior mean or the posterior median of the parameters on their original scale as the plug-in estimate to calculate the plug-in deviance $D(\hat{\theta})$. Explore also calculating the deviance using the parameters on a logit or log scale, so that the posterior distributions are less likely to be skewed.

2. Notice that in the model code provided, we have not explicitly calculated the case-severity risks (case-fatality risk CFR, case-ICU admission risk CIR, case-hospitalisation risk CHR). As per the lecture notes, calculate these functional parameters as products of the relevant conditional

probabilities $p[a, l]$. You can either do these in R without re-running the model code, from the posterior samples you already have, or (particularly if you are working directly in `OpenBUGS`), you can amend the model code to calculate the functional parameters.

3. The negative binomial likelihood for the count data allows for over-dispersion in the counts.

   - Have a look at the posterior distribution (by whichever summary/plotting method you prefer) for the over-dispersion parameter r_ℓ.
   - Look at the posterior for $1/log(r_\ell)$ and compare it to its Uniform(0.05,0.5) prior distribution.
   - Explore what happens when
     (a) you allow for more over-dispersion, by increasing the upper bound of the Uniform prior;
     (b) you allow for less over-dispersion, by decreasing the upper bound of the Uniform prior. Note that the smaller $1/log(r_\ell)$ (or equivalently the larger r_ℓ), the closer the likelihood approaches a Poisson likelihood. Check what happens to convergence for different parameters in this case.

7 Conflict & bias modelling in evidence synthesis

Model templates	Solutions
`conflict-jags.r`	`conflict-jags-solutions.r`
`conflict-R2OpenBUGS.r`	`conflict-R2OpenBUGS-solutions.r`
`conflict-OpenBUGS.odc`	`conflict-OpenBUGS-solutions.odc`

7.1 Sepsis meta-analysis example - outlier detection

For the sepsis meta-analysis example you considered in the first hierarchical modelling practical, we are going to carry out leave-one-out cross-validatory mixed-prediction, to test systematically for outliers among the included studies. We will look particularly at the choice of test statistic.

1. In the provided model code, replace the question marks with expressions for the two different choices of test statistic:
   - comparing $y_{i,k}$ for study i in each treatment arm k to the posterior distribution of its replication $Y_{i,k}^{rep}$ (*hint: adapt the expression for the one-sided cross-validatory mixed-predictive mid-p-value given in the Bristol example in the slides*);
   - comparing $\texttt{logit}(y_{i,2}/n_{i,2}) - \texttt{logit}(y_{i,1}/n_{i,1})$ to the posterior distribution of δ_i^{rep} (*hint: since the comparison here is of a continuous rather than integer quantity, a mid-p-value is not needed, instead a one-sided p-value is fine.*).
2. Run the model and obtain summary statistics for each quantity of interest. What do the summaries of your p-values suggest? Which of the two alternative test statistics gives a more meaningful test for outliers?
3. If you are working in R, try plotting the test statistic $\texttt{logit}(y_{i,2}/n_{i,2}) - \texttt{logit}(y_{i,1}/n_{i,1})$ overlaid on the posterior distribution of δ_i^{rep}, as in the lecture slides, to see how far in the tails of the distribution the statistic lies, for the Sandberg and Fanaroff studies (studies 8 and 5 respectively).

7.2 Sepsis meta-analysis example - systematic bias modelling

Turner et al, JRSS(A) (2008) introduced the idea of systematic bias adjustment for meta-analysis where some of the included studies are considered less rigorous or relevant than others. By incorporating prior information on the effect possible biases in such studies might have on the estimated treatment effect, the evidence from less rigorous or relevant studies may still be included, but down-weighted. Since then, much work has been done on systematic bias adjustment, see the Cochrane Collaboration's "Risk of bias" tool (`handbook.cochrane.org/index.htm#chapter_8/8_assessing_risk_of_bias_in_included_studies.htm`). The

"BRANDO" study (Savovic et al, Ann Intern Med, 2012) estimated the effect that different flaws in study design have on the estimated treatment effect: these estimates can be used as prior information to incorporate in a systematic bias adjustment model as suggested by Turner et al (2008).

1. Run the provided code to obtain estimates of the treatment effects when we systematically adjust for bias in studies with at least one "high risk" judgement. The code assumes a single bias parameter on the log-scale for each study at risk of bias, each with the same informative $N(-0.2, 0.1^2)$ prior. This implies a multiplicative effect of 0.82 on the odds ratio representing the treatment effect, with an approximate prior 95% range $(0.67, 1)$, and was obtained as an average of the relative effects of different types of bias on treatment effects estimated from the "BRANDO" study by Savovic et al (2012).

    - Obtain estimates of the bias-adjusted treatment effect and deviance summaries for this random effects model. Compare the bias-adjusted estimate to your unadjusted estimate from the hierarchical model practical.
    - The "BRANDO" study in fact estimated relative effects of bias due to inadequate/unclear sequence generation, allocation concealment and blinding as $0.88(0.76 - 1.00), 0.82(0.70 - 0.94)$ and $0.77(0.61 - 0.93)$ respectively. Instead of using a prior based on an average of these three effects, try using a prior based on an additive bias model, as in slide 244 of the lecture notes. (Hint: translate the estimated effects which are multiplicative on the odds ratio scale to additive on the log odds ratio scale). See how much effect the different prior has on the estimate of the treatment effect.