Bayesian Data Analysis

Lyle Gurrin

Centre for MEGA Epidemiology,
School of Population Health, University of Melbourne
lgurrin@unimelb.edu.au

Bendix Carstensen

Steno Diabetes Center, Denmark
& Department of Biostatistics, University of Copenhagen
bxc@steno.dk www.biostat.ku.dk/~bxc

Søren Højsgaard

Bioinformatics, Genetics and Statistics Research Unit
Inst. Genetics & Biotechnology & Aarhus University
sorenh@agrsci.dk genetics.agrsci.dk/~sorenh

11 – 15 August, 2008, Copenhagen

Fundamentals & Bayesian Analysis of Single-Parameter Models (Part 1)

Monday 11th August, morning

Lyle Gurrin

Bayesian Data Analysis
11 – 15 August 2008, Copenhagen

Acknowledgments

People: Bendix Carstensen, Søren Højsgaard, John Carlin.

Institutions:

Denmark - Department of Biostatistics at Uni Copenhagen, Steno Diabetes Centre.

Australia - SPH at Uni Melbourne, Biostatistics Collaboration of Australia (BCA).

Authors: Gelman, Carlin, Stern & Rubin (2004); Spiegelhalter, Abrams, Myles (2004).

What is Bayesian statistical inference?

Reference: Gelman et al., 1.1-1.5

An approach to statistical data analysis whereby scientific conclusions are expressed in terms of probabilities.

Bayesian data analysis refers to practical methods for making inferences from data using probabilistic models for quantities we *observe* and for quantities about which we wish to *learn*.

Fit a model to the data and summarise the results using a *probability distribution* on the parameters of the model and unobserved quantities such as predictions for new observations.

Example

Sex ratio in births with *placenta previa*.

Study data: 980 births, 437 female.
Observed proportion female: $\frac{437}{980} = 44.6\%$.

How do we do a Bayesian statistical analysis of these data?

How does it compare with an analysis using a "known" proportion of female births $= 0.485$? (test of the null hypothesis, confidence interval)?

What are the relevant assumptions about where the data came from?

Bayesian analysis in outline

1. Set up a *probability model* – prob. dist'n for *all* numerical quantities considered, both observed and unobserved, based on substantive knowledge. This encompasses the "prior distribution" and the sampling distribution ("likelihood function").
2. Obtain the conditional probability distribution of the unobserved quantities given the observed (the "posterior dist'n").
3. Evaluate the fit of the model and implications of the results; possibly iterate steps 1 to 3.

Implications for applied statistics

1. Direct quantification of uncertainty and common-sense interpretation of statistical conclusions; e.g. Bayesian "confidence interval" can be interpreted as "probably containing the true value".
2. Essential work of Bayesian analysis is setting up models and computing the answers (posterior dist'ns), not deriving mathematical properties of procedures.
3. Can set up and fit models using complex multi-layered probability specification due to the conceptually simple method for multiparameter problems.

General notation

Context: scientific study. Wish to *infer* about unobservables based on a *sample* from the *population*, *eg* a clinical trial of 2 drugs: how does 5-year survival on one drug compare with other?

θ: unobservable parameter
(*eg* probability of survival at 5 years)

y: observed data
(*eg* individual data on survival times in 2 groups)

$\tilde{y}$: unknown but potentially observable quantity
(*eg* outcomes of other patients)

Units & exchangeability

Often data collected on n "units" (e.g. people): write data as $y = (y_1, \ldots, y_n)$; e.g. let $y_i = 1(0)$ if patient i alive (dead) after 5 years.

Usually assume the n values are *exchangeable*, i.e. model unaffected by changing the unit indexes (labels)... motivates the usual *"iid"* models:

$$p(y|\theta) = \prod_i p(y_i|\theta)$$

Exchangeability is a basic concept of modelling and is important when we come to building *hierarchical* models.

Mechanics of Bayesian analysis

Aim: Make probability statements about parameter θ or unobserved $\tilde{y}$ conditional on observed data y.

Start with joint probability dist'n written as product of *prior dist'n* $p(\theta)$ and *sampling dist'n* $p(y|\theta)$:

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

Once data y known, we can condition using *Bayes' rule* to generate the *posterior* distribution:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}, \qquad (1)$$

Here $p(y) = \sum_\theta p(\theta)p(y|\theta)$, and sum is over all possible values of θ.

Mechanics of Bayesian analysis

Equivalent form of (1) omits the factor $p(y)$ which doesn't depend on θ to generate the *unnormalized posterior dist*:

$$p(\theta|y) \propto p(\theta)p(y|\theta). \qquad (2)$$

This is Bayesian inference! We could stop here except for the practical tasks of developing the model $p(\theta, y)$ and performing computations to summarise $p(\theta|y)$ in useful ways.

Summarising posterior inference

The posterior distribution $p(\theta|y)$ contains all current information about θ: often useful to display graphically.

We can use the usual numerical summaries (may need computational help!):

- mean, median, mode
- standard deviation, inter-quartile range, quantiles, credible intervals

But we may also directly quote posterior probabilities associated with θ.

Genetic example

Reference: Gelman et al., 1.4

Background: human males have 1 X-chromosome and 1 Y-chromosome; females have 2 X-chromosomes (one from each parent).

Haemophilia exhibits "X-linked recessive inheritance": a male with a "bad" gene on his X-chromosome is affected, but a female with only one "bad" gene on one of her 2 X-chromosomes is not.

Prior distribution

Consider a woman with an affected brother: her mother must be carrier of the gene (one "good" and one "bad" hemophilia gene). Her father is *not* affected; thus the woman herself has a fifty-fifty chance of having the gene.

Unknown *"parameter"* = whether the woman carries the gene ($\theta = 1$) or not ($\theta = 0$).

Prior distribution: $\Pr(\theta = 1) = \Pr(\theta = 0) = \frac{1}{2}$.

Model and likelihood

Data used to update this prior information = affection status of the woman's sons.
Suppose there are two sons, neither affected.
Let $y_i = 1$ or 0 denote an affected or unaffected son, respectively. Here $y_1 = y_2 = 0$ and so the likelihood function is:

$$\begin{aligned} \Pr(y_1 = 0, y_2 = 0 \,|\, \theta = 1) &= (0.5)(0.5) = 0.25 \\ \Pr(y_1 = 0, y_2 = 0 \,|\, \theta = 0) &= (1)(1) = 1. \end{aligned}$$

Posterior distribution

Bayes' rule (using y to denote the joint data (y_1, y_2)):

$$\begin{aligned}
&\Pr(\theta = 1|y) \\
&= \frac{p(y|\theta = 1)\Pr(\theta = 1)}{p(y|\theta = 1)\Pr(\theta = 1) + p(y|\theta = 0)\Pr(\theta = 0)} \\
&= \frac{(0.25)(0.5)}{(0.25)(0.5) + (1.0)(0.5)} = \frac{0.125}{0.625} = 0.20.
\end{aligned}$$

Estimating a probability from binomial data

Reference: Gelman et al., 2.1-2.5

Problem: Estimate an unknown population proportion θ from the results of a sequence of "Bernoulli trials"; that is, data $y_1, \ldots, y_n$, each $=$ either 0 or 1.

The *binomial distribution* provides natural model for a sequence of n *exchangeable* trials each giving rise to one of two possible outcomes.

The Binomial Model

Summarise data by:
$y =$ total number of successes in the n Bernoulli trials.

Parametrise the model using:
$\theta =$ proportion of successes or prob. of success in a single trial.

Binomial sampling model:

$$p(y|\theta) = \text{Bin}(y|n, \theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}. \quad (3)$$

Example: Birth "ratio"

Consider estimating the sex ratio in a population of human births: Is Pr{female birth} $= 0.5$?

We define the parameter $\theta =$ proportion of female births
(We may work with transformation, e.g. the ratio of male to female birth rates, $\phi = (1-\theta)/\theta$).

Let $y =$ number of girls in n births.

When would binomial model be appropriate?

First analysis: discrete prior

Suppose only two values of θ are considered possible. e.g. in example: $\theta = 0.5$ (what we always thought) or $\theta = 0.485$ (someone told us but we're not sure whether to believe it).

Posterior distribution

$$p(\theta|y) \propto p(\theta)p(y|\theta) \tag{4}$$

This is best obtained by a table with one line per value of θ:

Suppose we specify prior dist'n uniformly across the 2 values:

If n=100, y=48:

θ	$p(\theta)$	$p(y\|\theta) = \theta^y(1-\theta)^{n-y}$	$p(\theta\|y)$
0.485	0.5	8.503×10^{-31}	0.52
0.500	0.5	7.889×10^{-31}	0.48
		16.39×10^{-31}	

These data don't shift our prior distribution much.

Now suppose that n=1000, y=480:

θ	$p(\theta)$	$\log p(y\|\theta)$	$p(\theta\|y)$
0.485	0.5	-692.397	0.68
0.500	0.5	-693.147	0.32

Data and prior now favour $\theta = .485$ by 2:1 (but still substantial probability on $\theta = 0.5$).

Is discrete prior distribution reasonable?

Second analysis: uniform continuous prior

Simplest example of a prior distribution is to assume $p(\theta) \propto 1$ (in fact $p(\theta) = 1$!) on the interval $[0, 1]$.

Bayes' rule gives

$$p(\theta|y) \propto \theta^y (1-\theta)^{n-y} \tag{5}$$

[Q: what happened to the factor $\binom{n}{y}$?]

This is a *beta* probability distribution:

$$\theta|y \sim \text{Beta}(y+1, n-y+1) \tag{6}$$

What is a beta distribution?

The beta distribution is a continuous dist'n on [0,1] with a wide variety of shapes, determined by 2 parameters:

$$p(\theta|\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{7}$$

where $\alpha > 0, \beta > 0$. $\alpha = \beta = 1$ is the uniform dist'n.

It is unimodal with mode $\in (0, 1)$ if $\alpha > 1, \beta > 1$, and approaches a normal curve for $\alpha, \beta \to \infty$.

The mean of the beta distribution is $\mathrm{E}(\theta|\alpha, \beta) = \frac{\alpha}{\alpha+\beta}$.

Results for the uniform prior distribution

Back to the sex ratio example: Under uniform prior dist'n,

$$\begin{aligned} \mathrm{E}(\theta|y) &= \frac{y+1}{n+2} \\ \mathrm{var}(\theta|y) &= \frac{(y+1)(n-y+1)}{(n+2)^2(n+3)} \end{aligned}$$

Posterior distribution as compromise

In binomial model with uniform prior distribution:
Prior mean $= \frac{1}{2} \rightarrow$ posterior mean $= \frac{y+1}{n+2}$
A *compromise* between the prior mean $\frac{1}{2}$ and the sample proportion, $\frac{y}{n}$.

This is a general feature of Bayesian inference: posterior distribution centered at a compromise between the prior information and the data, with "compromise" increasingly controlled by the data as sample size increases (can be investigated more formally using conditional expectation formulae).

In 2 hypothetical cases:

	n=100	n=1000
	y=48	y=480
$\mathrm{E}(\theta\|y)$	0.4804	0.4800
$\mathrm{SD}(\theta\|y)$	0.049	0.016

What would our conclusions about the sex ratio in each case?

Third analysis: conjugate prior

Based on the form of the likelihood (3) suppose prior density:

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \tag{8}$$

that is, $\theta \sim \text{Beta}(\alpha, \beta)$.
(equivalent to the binomial likelihood with $\alpha - 1$ prior successes and $\beta - 1$ prior failures)

Parameters of prior distribution called *hyperparameters*.
Two hyperparameters of beta prior distribution can be fixed by specifying two features of the distribution, e.g. its mean and variance.

Third analysis: conjugate prior

If α, β fixed at reasonable choices obtain

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ p(\theta|y) &\propto \theta^{y}(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \\ &= \text{Beta}(\theta|\alpha+y, \beta+n-y). \end{aligned}$$

The property that posterior distribution follows same parametric form as prior distribution is called *conjugacy*; the beta prior distribution is a *conjugate family* for the binomial likelihood.
(Maths is convenient, but not necessarily a good model!)

Third analysis: conjugate prior

Under the conjugate beta prior, posterior mean of θ (= posterior probability of success for a future draw from the population):

$$\text{E}(\theta|y) = \frac{\alpha+y}{\alpha+\beta+n}. \tag{9}$$

Posterior variance is

$$\begin{aligned} \text{var}(\theta|y) &= \frac{(\alpha+y)(\beta+n-y)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} \\ &= \frac{E(\theta|y)[1-E(\theta|y)]}{\alpha+\beta+n+1}. \end{aligned}$$

Estimating properties of the posterior distributions

Beta distribution: exact summaries (mean, SD etc) can be obtained but how do we determine quantiles and hence posterior probability intervals? We require either:

- Use numerical integration (incomplete beta integral).
- Approximate the beta integral (normal distribution?).
- Resort to simulation: obtain a *random sample* from the dist'n and use numerical summaries.

Using simulation to estimate posterior distributions

Last strategy is the most general (and requires least analytical effort: computers replace algebra!).

We can simulate from the beta distribution using either `R` or `BUGS`.

Further advantage of simulation: distribution of functions of θ can be obtained with little further effort, e.g. sex ratio $= (1-\theta)/\theta$.

Example: Female births and placenta praevia

Sex ratio in placenta praevia (see lecture 1).

Uniform prior $\rightarrow p(\theta|y) =$ Beta(438,544).
Mean = 0.446, sd = 0.016, median = 0.446,
central 95% posterior interval = (0.415,0.477).

The following figure is a histogram of 1000 simulated values. These give sample mean, median and sd almost identical to exact values and interval (0.415,0.476).

Female births and placenta praevia

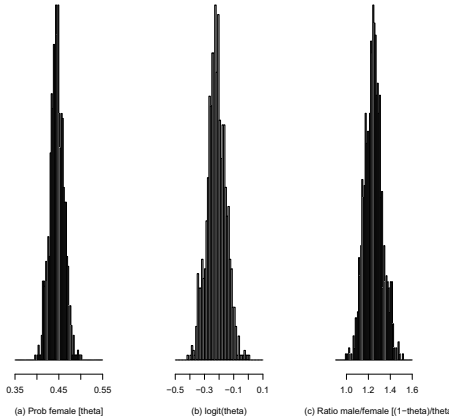


Figure: Draws from the posterior distribution of (a) the probability of female birth, θ; (b) the logit transform, logit(θ);

Female births and placenta praevia

Figure (c) shows histogram of simulated sex ratio. 95% posterior interval for this = (1.10,1.41); median = 1.24.

Note: intervals are well removed from $\theta = 0.485$ (ratio = 1.06), implying that probability of female birth in placenta praevia is lower than in general pop'n.

See Table 2.1 in Gelman *et al.* for sensitivity analysis to varying prior distribution;
Figure 2.4 in Gelman *et al.* for non-conjugate analysis.

APPENDIX I

Mathematical Background

Reference: Gelman et al. 1.7
In calculations relating to a joint density $p(u, v)$ often refer to

- *conditional* distribution or density function: $p(u|v)$
- *marginal* density: $p(u) = \int p(u, v)dv$.

(Range of integration = entire range of variable being integrated out.)

Often *factor* joint density as product of marginal and conditional densities:
$p(u, v, w) = p(u|v, w)p(v|w)p(w)$.

Mean of a conditional distribution

In joint distribution of (u, v) mean of u can be obtained by averaging the conditional mean over the marginal distribution of v:

$$\mathrm{E}(u) = \mathrm{E}(\mathrm{E}(u|v)), \tag{10}$$

[where inner expectation averages over u, conditional on v, and outer expectation averages over v].

Derivation:

$$\begin{aligned}\mathrm{E}(u) = \int\int up(u, v)dudv &= \int\int up(u|v)dup(v)dv \\ &= \int \mathrm{E}(u|v)p(v)dv.\end{aligned}$$

Variance of a conditional distribution

Corresponding result for variance has 2 terms: mean of conditional variance *plus* variance of conditional mean:

$$\mathrm{var}(u) = \mathrm{E}(\mathrm{var}(u|v)) + \mathrm{var}(\mathrm{E}(u|v)). \tag{11}$$

$$\begin{aligned}&\mathrm{E}[\mathrm{var}(u|v)] + \mathrm{var}[\mathrm{E}(u|v)] \\ =\ &\mathrm{E}[\mathrm{E}(u^2|v) - (\mathrm{E}(u|v))^2] \\ &+\mathrm{E}[(\mathrm{E}(u|v))^2] - (\mathrm{E}[\mathrm{E}(u|v)])^2 \\ =\ &\mathrm{E}(u^2) - \mathrm{E}[(\mathrm{E}(u|v))^2] + \mathrm{E}[(\mathrm{E}(u|v))^2] - (\mathrm{E}(u))^2 \\ =\ &\mathrm{E}(u^2) - (\mathrm{E}(u))^2 = \mathrm{var}(u).\end{aligned}$$

APPENDIX II

Prior Prediction

Before any data is observed the distribution of unknown but observable y is

$$p(y) = \int p(y, \theta) d\theta = \int p(\theta) p(y|\theta) d\theta. \qquad (12)$$

This is the *marginal* or *prior predictive distribution* of y.

Posterior Prediction

After y is observed, we can derive the dist'n of unknown but potentially observable $\tilde{y}$ using the same process:

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}, \theta|y) d\theta \\ &= \int p(\tilde{y}|\theta, y) p(\theta|y) d\theta \\ &= \int p(\tilde{y}|\theta) p(\theta|y) d\theta. \end{aligned}$$

This is the *posterior predictive distribution* of $\tilde{y}$.

Posterior Prediction

The natural application here is for $\tilde{y}$ to be the result of one new trial, exchangeable with first n.

$$\begin{aligned} \Pr(\tilde{y}=1|y) &= \int_0^1 \Pr(\tilde{y}=1|\theta,y)p(\theta|y)d\theta \\ &= \int_0^1 \theta p(\theta|y)d\theta = \mathrm{E}(\theta|y) = \frac{y+1}{n+2}. \end{aligned}$$

This is Laplace's notorious "Law of Succession".

Choice of prior distribution

Bayes is believed to have justified choosing the uniform prior dist'n in the binomial model because the *prior predictive dist'n* is

$$\begin{aligned} p(y) &= \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} d\theta \\ &= \frac{1}{n+1} \quad \text{for } y = 0, \ldots, n. \end{aligned}$$

Thus all possible values of y are equally likely *a priori*.

Not necessarily a compelling argument – "true" Bayesian analysis must use a subjectively assessed prior dist'n.

Introduction to Computation

Monday 11th August 2008, afternoon

Lyle Gurrin

Bayesian Data Analysis
11 – 15 August 2008, Copenhagen

The need for simulation

The posterior distribution $p(\theta|y)$ describes our beliefs about the possible values of θ.

Calculations for inference using this distribution may not be straightforward, especially when θ is a vector representing multiple parameters.

Simulation is central to Bayesian analysis - it's easy to draw samples from "difficult" probability distributions.

Simulation in practice

Simulation uses the duality between the probability density function (PDF) and a histogram of random draws from that distribution, which provides information about the distribution to any level of accuracy.

eg To estimate the 95^{th} percentile of the distribution for θ, draw a random sample of size L and use the $0.95L^{\text{th}}$ order statistic.

L = 1,000 is usually sufficient, with the 95^{th} centile represented by the 50^{th} largest value.

Example: Coin tossing (Spiegelhalter et al. (2004))

Suppose we want to know the probability of getting 8 or more heads when we toss a fair coin 10 times. An *algebraic* approach would use the formula for the binomial distribution:

$$\begin{aligned} P(\text{8+ heads}) &= \binom{10}{8}\left(\frac{1}{2}\right)^8\left(\frac{1}{2}\right)^2 + \binom{10}{9}\left(\frac{1}{2}\right)^9\left(\frac{1}{2}\right)^1 \\ &\quad + \binom{10}{10}\left(\frac{1}{2}\right)^{10}\left(\frac{1}{2}\right)^0 \\ &= \frac{1}{2^{10}}(45+10+1) = 0.0547 \end{aligned}$$

Physical approach

An alternative, physical approach: Repeatedly throw a set of 10 coins and count the proportion of throws where there were 8 or more heads. After sufficient throws, this proportion will tend to the correct result of 0.0547.

We can imitate this by a *simulation* approach in which a computer program generates the throws according to a reliable random mechanism, say by generating a random number U between 0 and 1, and declaring "head" if $U \geq 0.5$.

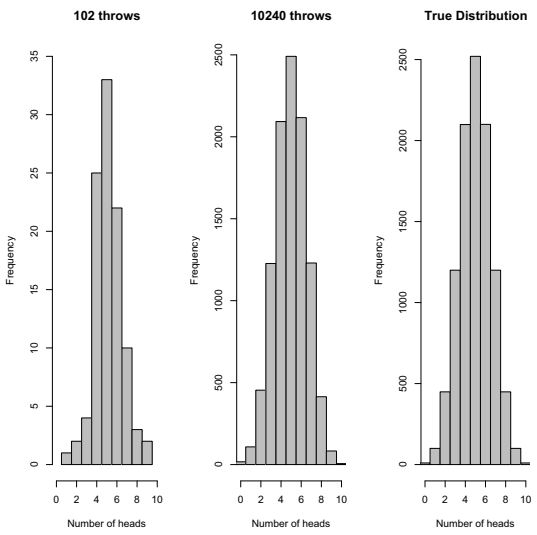
Simulation results

In fig (a):
3, 2 and 0 instances
of 8, 9 and 10 heads.

$\hat{p}$ is $5/102 = 0.0490$
(true prob 0.0547).

In fig (b):
$\hat{p} = 0.0510$

In fig (c):
p $= 0.0547$



Limitations of simple Monte Carlo

We're most interested in simulating draws from the posterior distribution of parameters θ and predictive distribution of unknown observables $\tilde{y}$.

Monte Carlo or direct simulation methods are useful provided we know the distributional form explicitly.

In conjugate Bayesian analysis we can derive the posterior distribution algebraically and hence use Monte Carlo methods to find tail areas (or directly using R)

Markov chain Monte Carlo

For non-conjugate distributions or nuisance parameters in more complex Bayesian analysis it will not be possible to derive the posterior distribution in an algebraic form.

Markov chain Monte Carlo (MCMC) provide a means of sampling from the posterior distribution of interest even when that posterior has no known algebraic form.

Essential components of MCMC

1. Determine a way to sample from the posterior distribution using a *Markov chain*, in which the distribution for the next simulated value $\theta^{(j+1)}$ depends only on the current $\theta^{(j)}$.
2. Generate initials values for parameters and if necessary unobserved data values.
3. Run the simulation and monitor convergence.
4. Summarise results of draws from desired posterior distributions.

BUGS

BUGS is a piece of software designed to make MCMC analyses straightforward.

It carries out Bayesian inference on statistical problems by selecting from a number of simulation techniques.

The first version of the software used *Gibbs sampling* exclusively, hence the name **B**ayesian inference **U**sing **G**ibbs **S**ampling.

For which problems is BUGS best suited?

BUGS is intended for problems where
(i) There is no analytic solution.
(ii) Standard approximation techniques have difficulties.

Especially suited for use with complex models with many unknown quantities but substantial structural relationships, which are expressed as conditional independence assumptions.

BUGS works with a *directed acyclic graph* (or "DAG") representing the random quantities, where missing edges between *nodes* of the graph represent conditional independence.

BUGS in brief

BUGS requires a Bayesian or full probability model, generally defined by combining a sampling model (for data) with prior distributions (for parameters).

It conditions on the observed data in order to obtain a posterior distribution over the parameters and unobserved data.

BUGS cycles around the quantities of interest, simulating from the simpler conditional distributions, which ultimately generates samples from the unconditional or *marginal* distributions.

The BUGS engine

We will use WinBUGS version 1.4 (which can be accessed from R using the R2WinBUGS package) to analyse statistical models expressed using the very flexible BUGS *language*.

A *compiler* processes the model and available data into an internal data structure suitable for efficient computation.

A *sampler* operates on this structure to access the relevant distributions and generate a stream of simulated values for each quantity of interest (automatically working out appropriate sampling methods).

Extra features

BUGS has a suite of functions provided for analysis and plotting of the output files via built-in graphics and convergence diagnostics, and a large range of examples and web presence that covers many different subject areas.

Much of the post-processing of output from BUGS can be done directly in R

COINS example in BUGS

Recall the simulated repeated tossing of 10 "balanced coins" discussed earlier. These simulations can be carried out in WinBUGS using the program:

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


Explaining the COINS model code

y is binomial with $p = \frac{1}{2}$ and $n = 10$.

P8 is a step function which will take on the value 1 if $y - 7.5$ is non-negative, that is, if y is 8 or more, 0 if 7 or less.

'$\sim$' indicates a distribution.

'$<-$' indicates a logical identity.

COINS example output

```
node  mean  sd       MC error
P8    0.05  0.2179   0.002859
```

```
node  2.5%  median 97.5% start sample
P8    0.0   0.0    1.0   5001  5000
```


LINES: A simple example in BUGS

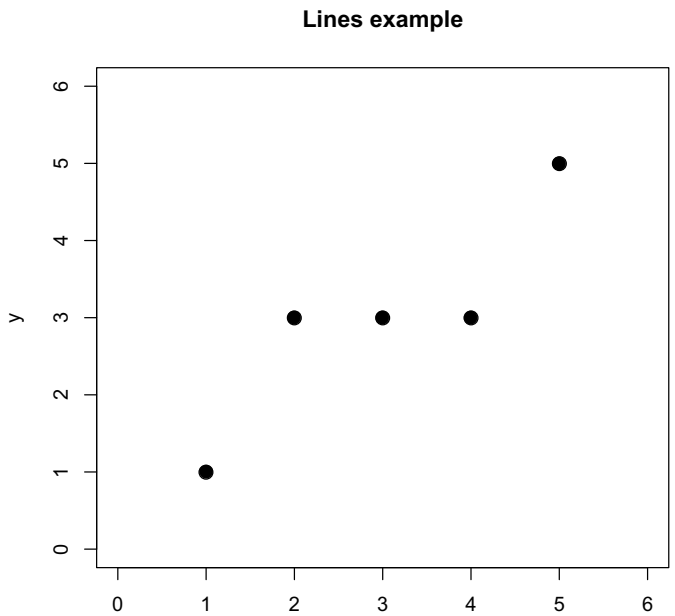
We introduce a trivial problem for which exact solutions are possible in order to illustrate the nature of the Gibbs sampling approach in BUGS.

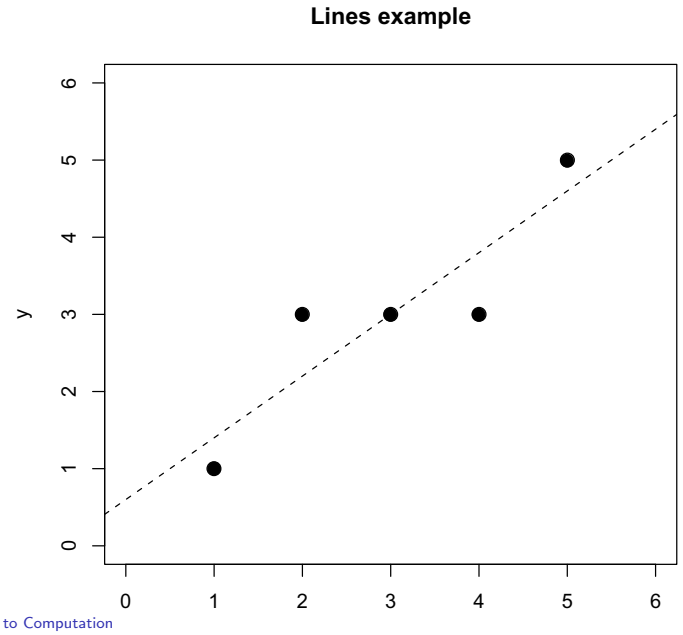
Consider a set of 5 observed (x,y) pairs (1,1), (2,3), (3,3), (4,3), (5,5).

We shall fit a simple linear regression of y on x , using the notation

$$Y_i \sim \text{Normal}(\mu_i, \tau) \quad (13)$$

$$\mu_i = \alpha + \beta(x_i - \overline{x}) \quad (14)$$





Classical analysis

Classical unbiased estimates are

$$\begin{aligned}
\hat{\alpha} &= \overline{y} \\
&= 3.00 \\
\hat{\beta} &= \sum_i y_i(x_i - \overline{x}) / \sum_i (x_i - \overline{x})^2 \\
&= 0.80 \\
\hat{\sigma}^2 &= \sum (y_i - \hat{y}_i)^2/(n-2) \\
&= 0.533
\end{aligned}$$

$\widehat{\text{var}}(\hat{\alpha}) = \hat{\sigma}^2/n = 0.107,$
$\widehat{\text{var}}(\hat{\beta}) = \hat{\sigma}^2/\sum_i (x_i - \overline{x})^2 = 0.053.$

Inference for parameters

Both frequentist and Bayesian 'noninformative' priors lead to inference being based on the pivotal quantities:

$(\hat{\alpha} - \alpha)/\sqrt{\widehat{\text{var}(\hat{\alpha})}}$ and $(\hat{\beta} - \beta)/\sqrt{\widehat{\text{var}(\hat{\beta})}}$, which both having t_3 distributions with mean 0 and variance 3, and

$\hat{\sigma}^2(n-2)/\sigma^2$, which has a χ^2_3 distribution, leading to the following 95% confidence/credible intervals:

Confidence intervals for α, β and τ

95% interval for α : $\hat{\alpha} \pm 3.18\sqrt{\widehat{\text{var}}(\hat{\alpha})} = (1.96, 4.04)$

95% interval for β : $\hat{\beta} \pm 3.18\sqrt{\widehat{\text{var}}(\hat{\beta})} = (0.07, 1.53)$

95% interval for τ : $(0.22, 9.35)/(3\hat{\sigma}^2) = (0.14, 5.85)$

LINES code

The BUGS language allows a concise expression for the model with the core relations 13 and 14 described as follows:

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


LINES output

```
node    mean    sd      MC error
alpha   2.972   0.9433  0.01862
beta    0.806   0.6482  0.008952
sigma   1.461   1.613   0.05916

node    2.5%    median  97.5%
alpha   1.444   2.988   4.622
beta   -0.3617 0.8012  1.983
sigma   0.4619  1.019   4.953
```

Bayesian Analysis of Single-Parameter Models (Part 2)

Tuesday 12th August 2008, morning

Lyle Gurrin

Bayesian Data Analysis
11 – 15 August 2008, Copenhagen

The standard single-parameter models

Recall general problem that the posterior density, $p(\theta|y)$, often has no closed-form expression, and it can be especially difficult to evaluate the normalising constant $p(y)$.

Much formal Bayesian analysis concentrates on situations where closed (conjugate) forms are available as useful starting point for constructing more realistic models.

We will look at another one parameter example of a conjugate Bayesian analysis before turning to the normal model, multiparameter and hierarchical models.

The exponential family (1)

The familiar "exponential family" distributions—binomial, normal, Poisson, and exponential—have natural derivations from simple probability models:

- Binomial distribution from counting exchangeable outcomes.
- Normal dist'n applies to a random variable that is sum of a large number of exchangeable or independent terms.
- Normal dist'n also arise naturally when the logarithm of all-positive data are modelled as *product* of many independent multiplicative factors.

The exponential family (2)

The familiar "exponential family" distributions—binomial, normal, Poisson, and exponential—have natural derivations from simple probability models:

- Poisson and exponential distributions arise as the number of counts and the waiting times, respectively, for events modelled as occurring *exchangeably in all time intervals*; i.e. independently in time, with constant rate of occurrence.

Conjugate priors for standard models

Each of these standard models has an associated family of conjugate prior distributions (see below for Poisson).

Realistic models for more complicated outcomes may be constructed using combinations of these basic distributions.

The Poisson distribution

The Poisson distribution rises naturally in study of data taking the form of counts. In epidemiology, we might postulate that the number of cases of disease in a population follow a Poisson model.

If $y \sim \text{Poi}(\theta)$ (a single observation y follows a Poisson distribution with mean θ) then:

$$p(y|\theta) = \frac{\theta^y e^{-\theta}}{y!}, \quad \text{for} \qquad y = 0, 1, 2, \ldots$$

Poisson likelihood for multiple observations

For a vector $y = (y_1, \ldots, y_n)$ of i.i.d. observations, the likelihood based on the Poisson distribution is:

$$\begin{aligned} p(y|\theta) &= \prod_{i=1}^{n} \frac{1}{y_i!} \theta^{y_i} e^{-\theta} \\ &\propto \theta^{t(y)} e^{-n\theta}, \end{aligned}$$

where $t(y) = \sum_{i=1}^{n} y_i = n\overline{y}$ is a sufficient statistic.

Conjugate prior from the likelihood

A more conventional parametrisation of this conjugate prior distribution is:

$$p(\theta) \propto e^{-\beta\theta} \theta^{\alpha-1}$$

which is a *gamma* density, Gamma(α, β). Comparing $p(y|\theta)$ and $p(\theta)$ reveals that the prior density is equivalent to a total count of $\alpha - 1$ in β prior observations.

The corresponding posterior distribution is then:

$$\theta|y \sim \text{Gamma}(\alpha + n\overline{y}, \beta + n).$$

Poisson model parameterized in terms of rate and exposure

It is often convenient to extend Poisson model for data points $y_1, \ldots, y_n$ to

$$y_i \sim \text{Poi}(x_i\theta),$$

where x_i are known positive values of an explanatory variable, x.

In epidemiology, θ is the *rate*, and x_i is *exposure* of the i^{th} unit. Note that this model is *not* exchangeable in the y_i's but it is exchangeable in the (x_i, y_i) pairs.

Likelihood for the Poisson rate and exposure model

The likelihood for θ is now:

$$p(y|\theta) \propto \theta^{(\sum_{i=1}^{n} y_i)} e^{-(\sum_{i=1}^{n} x_i)\theta}$$

ignoring factors that do not depend on θ. Again the gamma distribution is conjugate. With prior $\theta \sim \text{Gamma}(\alpha, \beta)$, the posterior distribution is

$$\theta|y \sim \text{Gamma}\left(\alpha + \sum_{i=1}^{n} y_i,\ \beta + \sum_{i=1}^{n} x_i\right). \quad (15)$$

Example: Estimating a rate from count data

An idealized example: Suppose causes of death were reviewed for a city for a single year and 3 people out of population of $200{,}000$, died of asthma.

Crude estimated asthma mortality rate $= 1.5$ cases per $100{,}000$ persons per year.

Model for $y =$ number of deaths in city of $200{,}000$ in one year, is Poi$(2.0 \times \theta)$, where $\theta =$ underlying asthma mortality rate (cases per $100{,}000$ persons per yr).

Setting up a prior distribution

Reviews of asthma mortality rates suggest that rates greater than 1.5 per $100{,}000$ people are rare (in Western countries), with typical asthma mortality rates around 0.6 per $100{,}000$.

Trial-and-error exploration suggests $p(\theta) = \text{Gamma}(3.0, 5.0)$ (mean $= 0.6$, mode $= 0.4$ and 97.5% of the mass of the density lies below 1.44) is a reasonable prior distribution.

In general, specifying a prior mean sets the ratio of the two gamma parameters; then shape parameter can be altered by trial and error to match prior knowledge about the tail of the distribution.

Posterior distribution

The posterior distribution for θ is Gamma$(6.0, 7.0)$

Mean $= 0.86$, posterior probability that $\theta > 1.0$ per $100{,}000$ per year $= 0.30$ (substantial shrinkage towards prior distribution).

Additional data: Suppose ten years of data obtained and mortality rate of 1.5 per $100{,}000$ is maintained; i.e. $y = 30$ deaths over 10 years. Under same model, posterior distribution of θ now Gamma$(33.0, 25.0)$

Mean $= 1.32$, posterior probability that $\theta > 1.0 = 0.93$.

APPENDIX I

Marginal distribution for the Poisson model

With conjugate families, we can use known form of prior and posterior densities to find the marginal distribution, $p(y)$, of the count outcome y using:

$$p(y) = \frac{p(y|\theta)p(\theta)}{p(\theta|y)}.$$

The Negative Binomial distribution

For single observation from the Poisson distribution where the rate is governed by a gamma prior, the *prior* predictive distribution $p(y)$ is

$$\begin{aligned} p(y) &= \frac{\text{Poi}(y|\theta)\text{Gamma}(\theta|\alpha,\beta)}{\text{Gamma}(\theta|\alpha+y,1+\beta)} \\ &= \frac{\Gamma(\alpha+y)\beta^{\alpha}}{\Gamma(\alpha)y!(1+\beta)^{\alpha+y}}, \end{aligned}$$

which reduces to

$$p(y) = \binom{\alpha+y-1}{y}\left(\frac{\beta}{\beta+1}\right)^{\alpha}\left(\frac{1}{\beta+1}\right)^{y},$$

This is the probability mass function of the *negative binomial* distribution:

$$y \sim \text{Neg-Bin}(\alpha,\beta).$$

Neg-bin as a Gamma mixture of Poissons

The above derivation shows that the negative binomial distribution is a *mixture* of Poisson distributions with rates, θ, that follow the gamma distribution:

$$\text{Neg-Bin}(y|\alpha,\beta) = \int \text{Poi}(y|\theta)\text{Gamma}(\theta|\alpha,\beta)d\theta.$$

This provides a *robust* alternative to the Poisson distribution as a sampling model that can be used to capture *overdispersion*.

Normal and multiparameter models

Tuesday 12th August 2008, afternoon

Lyle Gurrin

Bayesian Data Analysis
11 – 15 August 2008, Copenhagen

Analysis for the normal distribution: (i) Unknown mean, known variance

Reading: Gelman et al., 2.6

Normal model underlies much statistical modelling (why?)

We start with the simplest case, assuming the variance is known:

1. Just one data point.
2. General case of a "sample" of data with many data points.

Likelihood of one data point

Consider a single observation y from normal distribution with mean θ and variance σ^2, with σ^2 known. The sampling distribution is:

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\theta)^2}. \qquad (16)$$

Conjugate prior and posterior distributions

This likelihood is the exponential of a quadratic form in θ, so conjugate prior distribution must have same form; parameterize this family of conjugate densities as

$$p(\theta) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right); \qquad (17)$$

i.e. $\theta \sim \mathrm{N}(\mu_0, \tau_0^2)$, with hyperparameters μ_0 and τ_0^2.

For now we assume μ_0 and τ_0 to be known.

Posterior distribution

From the conjugate form of prior distribution, the posterior distribution for θ is also normal:

$$p(\theta|y) \propto \exp\left(-\frac{1}{2}\left[\frac{(y-\theta)^2}{\sigma^2}+\frac{(\theta-\mu_0)^2}{\tau_0^2}\right]\right). \qquad (18)$$

Some algebra is required, however, to reveal its form (recall that in the posterior distribution everything except θ is regarded as constant).

Parameters of the posterior distribution

Algebraic rearrangement gives

$$p(\theta|y) \propto \exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right), \qquad (19)$$

that is, the posterior distribution $\theta|y$ is $\mathrm{N}(\mu_1, \tau_1^2)$ where

$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2}+\frac{1}{\sigma^2}} \quad \text{and} \quad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2}+\frac{1}{\sigma^2}. \qquad (20)$$

Precisions of the prior and posterior distributions

In manipulating normal distributions, the inverse of the variance or *precision* plays a prominent role. For normal data and normal prior distribution, each with known precision, we have

$$\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}.$$

posterior precision $=$ prior precision $+$ data precision.

Interpreting the posterior mean, μ_1

There are several ways of interpreting the form of the posterior mean μ_1. In equation (20):

$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}.$$

posterior mean $=$ weighted average of prior mean and observed value, y, with weights proportional to the precisions.

Interpreting the posterior mean, μ_1

Alternatively, $\mu_1 =$ prior mean "adjusted" toward observed y:

$$\mu_1 = \mu_0 + (y - \mu_0)\frac{\tau_0^2}{\sigma^2 + \tau_0^2}, \qquad (21)$$

or $\mu_1 =$ data "shrunk" toward the prior mean:

$$\mu_1 = y - (y - \mu_0)\frac{\sigma^2}{\sigma^2 + \tau_0^2}. \qquad (22)$$

In both cases, the posterior mean μ_1 is a compromise between the prior mean and the observed value.

Interpretation of μ_1 for extreme cases

In the extreme cases, the posterior mean μ_1 equals the prior mean or the observed value y.

$$\mu_1 = \mu_0 \quad \text{if} \quad y = \mu_0 \ \text{ or } \ \tau_0^2 = 0;$$

$$\mu_1 = y \quad \text{if} \quad y = \mu_0 \ \text{ or } \ \sigma^2 = 0.$$

What is the correct interpretation for each scenario?

Posterior predictive distribution 1

The posterior predictive distribution of a future observation, $\tilde{y}$, can be calculated directly be integration:

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}|\theta)p(\theta|y)d\theta \\ &\propto \int \exp\left(-\frac{1}{2\sigma^2}(\tilde{y}-\theta)^2\right) \times \\ &\quad \exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right) d\theta. \end{aligned}$$

Avoid algebra in simplifying this by using properties of the bivariate normal distribution.

Posterior predictive distribution 2

$\tilde{y}$ and θ have a *joint* normal posterior distribution (why?), which implies that the *marginal* posterior predictive distribution of $\tilde{y}$ is normal.

We can now determine the mean and variance using the fact that $\mathrm{E}(\tilde{y}|\theta) = \theta$ and $\mathrm{var}(\tilde{y}|\theta) = \sigma^2$, along with the iterative expectation and variance identities given in an earlier lecture:

$$\mathrm{E}(\tilde{y}|y) = \mathrm{E}(\mathrm{E}(\tilde{y}|\theta, y)|y) = \mathrm{E}(\theta|y) = \mu_1, \qquad (23)$$

and

$$\begin{aligned}\mathrm{var}(\tilde{y}|y) &= \mathrm{E}(\mathrm{var}(\tilde{y}|\theta,y)|y) + \mathrm{var}(\mathrm{E}(\tilde{y}|\theta,y)|y) \\ &= \mathrm{E}(\sigma^2|y) + \mathrm{var}(\theta|y) \\ &= \sigma^2 + \tau_1^2.\end{aligned}$$

This shows that the posterior predictive distribution for unobserved $\tilde{y}$ has mean equal to posterior mean of θ and two components of variance:

- predictive variance σ^2 from the sampling model
- variance τ_1^2 due to posterior uncertainty in θ

Normal model with multiple observations

The normal model with a single observation can easily be extended to the more realistic situation where we have a sample of independent and identically distributed observations $y = (y_1, \ldots, y_n)$. We can proceed formally, from

$$p(\theta|y) \propto p(\theta)p(y|\theta) = p(\theta)\prod_{i=1}^{n} p(y_i|\theta)$$

where $p(y_i|\theta) = \mathsf{N}(y_i|\theta, \sigma^2)$ with algebra similar to that above. The posterior distribution depends on y only through the sample mean, $\overline{y} = \frac{1}{n}\sum_i y_i$, which is a *sufficient statistic* in this model.

Normal model via the sample mean

In fact, since $\overline{y}|\theta, \sigma^2 \sim \mathsf{N}(\theta, \sigma^2/n)$, we can apply results for the single normal observation $p(\theta|y_1 \ldots, y_n) = p(\theta|\overline{y}) = \mathsf{N}(\theta|\mu_n, \tau_n^2)$, where

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\overline{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

and

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}.$$

Limits for large n and large τ^2

The prior precision, $\frac{1}{\tau_0^2}$, and data precision, $\frac{n}{\sigma^2}$, play equivalent roles; if n large, the posterior distribution is largely determined by σ^2 and the sample value $\overline{y}$.

As $\tau_0 \to \infty$ with n fixed, *or* as $n \to \infty$ with τ_0^2 fixed, have:

$$p(\theta|y) \approx \mathsf{N}(\theta|\overline{y}, \sigma^2/n) \tag{24}$$

Limits for large n and large τ^2

A prior distribution with large τ^2 and thus low precision captures prior beliefs are diffuse over the range of θ where the likelihood is substantial.
Compare the well-known result of classical statistics:

$$\overline{y}|\theta, \sigma^2 \sim \mathsf{N}(\theta, \sigma^2/n) \tag{25}$$

leads to use of

$$\overline{y} \pm 1.96\frac{\sigma}{\sqrt{n}} \tag{26}$$

as a 95% confidence interval for θ.

Bayesian approach gives the same result for noninformative prior.

Analysis for the normal distribution: (ii) Known mean, unknown variance

This is not directly useful for applications but is an important building block especially for the normal distribution with unknown mean and variance.

It also introduces estimation of a *scale* parameter, a role played by σ^2 for the normal distribution.

Normal likelihood

For $y \sim \mathrm{N}(\theta, \sigma^2)$ with known θ and unknown σ^2, the likelihood for vector y of n i.i.d. observations:

$$\begin{aligned} p(y|\sigma^2) &\propto \sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right) \\ &= (\sigma^2)^{-n/2}\exp\left(-\frac{n}{2\sigma^2}v\right) \end{aligned}$$

where the sufficient statistic is

$$v = \frac{1}{n}\sum_{i=1}^{n}(y_i-\theta)^2. \tag{27}$$

Conjugate prior for σ^2

Conjugate prior distribution is the *inverse-gamma*:

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)}e^{-\beta/\sigma^2}, \tag{28}$$

which has hyperparameters (α, β).

A convenient parametrisation is the scaled inverse-χ^2 distribution with scale σ_0^2 and ν_0 degrees of freedom.

Then prior distribution of σ^2 is the same as the distribution of $\sigma_0^2\nu_0/X$, where $X \sim \chi^2_{\nu_0}$. We use the convenient (but nonstandard) notation, $\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$.

Posterior distribution for σ^2

Resulting posterior distribution:

$$\begin{aligned} p(\sigma^2|y) &\propto p(\sigma^2)p(y|\sigma^2) \\ &\propto \left(\frac{\sigma_0^2}{\sigma^2}\right)^{\nu_0/2+1}\exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right) \cdot \\ &\quad (\sigma^2)^{-n/2}\exp\left(-\frac{n}{2}\frac{v}{\sigma^2}\right) \\ &\propto (\sigma^2)^{-((n+\nu_0)/2+1)}\exp\left(-\frac{1}{2\sigma^2}(\nu_0\sigma_0^2+nv)\right). \end{aligned}$$

Thus...

Posterior distribution for σ^2

$$\sigma^2|y \sim \text{Inv-}\chi^2\left(\nu_0 + n, \frac{\nu_0\sigma_0^2 + nv}{\nu_0 + n}\right) \qquad (29)$$

—scaled inverse-χ^2 distribution:

Posterior scale = precision-weighted average of prior and data scales.

posterior degrees of freedom = sum of prior and data degrees of freedom.

Prior distribution $\approx$ information equivalent to ν_0 observations with average squared deviation σ_0^2.

Multiparameter models: Introduction

The reality of applied statistics: there are always several (maybe many) unknown parameters!

BUT the interest usually lies in only a few of these (parameters of interest) while others are regarded as *nuisance parameters* for which we have no interest in making inferences but which are required in order to construct a realistic model.

At this point the simple conceptual framework of the Bayesian approach reveals its principal advantage over other forms of inference.

The Bayesian approach

The Bayesian approach is clear: Obtain the *joint* posterior distribution of all unknowns, then *integrate* over the nuisance parameters to leave the *marginal* posterior distribution for the parameters of interest.

Alternatively using simulation, draw samples from the entire joint posterior distribution (even this may be computationally difficult), look at the parameters of interest and ignore the rest.

Averaging over nuisance parameters

To begin exploring the ideas of joint and marginal distributions, suppose that θ has two parts, so $\theta = (\theta_1, \theta_2)$. We are interested only in θ_1, with θ_2 considered a "nuisance" parameter.

For example:

$$y|\mu, \sigma^2 \sim \mathsf{N}(\mu, \sigma^2),$$

with both μ $(=\theta_1)$ and σ^2 $(=\theta_2)$ unknown. Interest usually focusses on μ.

Averaging over nuisance parameters

AIM: To obtain the conditional distribution $p(\theta_1|y)$ of the parameters of interest θ_1.

This can be derived from *joint posterior density*,

$$p(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2)p(\theta_1, \theta_2),$$

by averaging or *integrating* over θ_2:

$$p(\theta_1|y) = \int p(\theta_1, \theta_2|y)d\theta_2.$$

Factoring the joint posterior

Alternatively, the joint posterior density can be factored to yield:

$$p(\theta_1|y) = \int p(\theta_1|\theta_2, y)p(\theta_2|y)d\theta_2, \qquad (30)$$

showing posterior distribution, $p(\theta_1|y)$, as a *mixture* of the conditional posterior distributions given the nuisance parameter, θ_2, where $p(\theta_2|y)$ is a weighting function for the different possible values of θ_2.

Mixtures of conditionals

- The weights depend on the posterior density of θ_2—so on a combination of evidence from data and prior model.
- What if θ_2 known to have a particular value?

The averaging over nuisance parameters can be interpreted very generally: θ_2 can be categorical (discrete) and may take only a few possible values representing, for example, different sub-models.

A strategy for computation

We rarely evaluate integral (30) explicitly, but it suggests an important strategy for constructing and computing with multiparameter models, using simulation:

1. Draw θ_2 from its marginal posterior distribution.
2. Draw θ_1 from conditional posterior distribution, given the drawn value of θ_2.

Conditional simulation

In this way the integration in (30) is performed indirectly.

In fact we can alter step 1 to draw θ_2 from its *conditional* posterior distribution given θ_1. Iterating the procedure will ultimately generate samples from the *marginal* posterior distribution of *both* θ_1 and θ_2.

This is the much vaunted **Gibbs sampler**.

Multiparameter models: Normal mean & variance

Consider a vector y of n i.i.d. observations (univariate) distributed as $\mathsf{N}(\mu, \sigma^2)$.

We begin by analysing the model under the convenient assumption of a noninformative prior distribution (which is easily extended to informative priors).

We assume prior independence of location and scale parameters and take $p(\mu, \sigma^2)$ to be uniform on $(\mu, \log\sigma)$:

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-1}.$$

The joint posterior distribution, $p(\mu, \sigma^2|y)$

Under the improper prior distribution the joint posterior distribution is proportional to the likelihood $\times$ the factor $1/\sigma^2$:

$$\begin{aligned} p(\mu, \sigma^2|y) &\propto \sigma^{-n-2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right) \\ &= \sigma^{-n-2}\exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\overline{y})^2 + n(\overline{y}-\mu)^2\right]\right) \\ &= \sigma^{-n-2}\exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\overline{y}-\mu)^2]\right), \quad (31) \end{aligned}$$

where $s^2 = 1/(n-1)\sum_{i=1}^{n}(y_i-\overline{y})^2$ is the sample variance of y_i's.

The sufficient statistics are _____ and _____?

The conditional posterior dist'n, $p(\mu|\sigma^2, y)$

We can factor the joint posterior density by considering first the conditional distribution $p(\mu|\sigma^2, y)$, and then the marginal $p(\sigma^2|y)$.

We can use a previous result for the mean μ of a normal distribution with *known* variance.

$$\mu|\sigma^2, y \sim \mathsf{N}(\overline{y}, \sigma^2/n). \quad (32)$$

The marginal posterior distribution, $p(\sigma^2|y)$

This requires averaging the joint distribution (31) over μ, that is, evaluating the simple normal integral

$$\int \exp\left(-\frac{1}{2\sigma^2}n(\overline{y}-\mu)^2\right)d\mu = \sqrt{2\pi\sigma^2/n};$$

thus,

$$p(\sigma^2|y) \propto (\sigma^2)^{-(n+1)/2}\exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right), \quad (33)$$

which is a scaled inverse-χ^2 density:

$$\sigma^2|y \sim \text{Inv-}\chi^2(n-1, s^2). \quad (34)$$

Product of conditional and marginal

We have therefore factored (31) as the product of conditional and marginal densities

$$p(\mu, \sigma^2|y) = p(\mu|\sigma^2, y)p(\sigma^2|y).$$

Parallel between Bayes & frequentist results

As with one-parameter normal results, there is a remarkable parallel with sampling theory:

Bayes:

$$\frac{(n-1)s^2}{\sigma^2}|y \sim \chi^2_{n-1}$$

Frequentist:

$$\frac{(n-1)s^2}{\sigma^2}|\mu, \sigma^2 \sim \chi^2_{n-1}$$

Conditional on the values of μ and σ^2 the sampling distribution of the appropriately scaled sufficient statistic $(n-1)s^2/\sigma^2$ is chi-squared with $n-1$ d.f.

Analytic form of marginal posterior distribution of μ

μ is typically the estimand of interest, so ultimate objective of the Bayesian analysis is the marginal posterior distribution of μ. This can be obtained by integrating σ^2 out of the joint posterior distribution. Easily done by *simulation*: first draw σ^2 from (34), then draw μ from (32).

The posterior distribution of μ can be thought of as a mixture of normal distributions mixed over the scaled inverse chi-squared distribution for the variance - a rare case where analytic results are available.

Performing the integration

We start by integrating the joint posterior density over σ^2

$$p(\mu|y) = \int_0^\infty p(\mu, \sigma^2|y)d\sigma^2$$

This can be evaluated using the substitution

$$z = \frac{A}{2\sigma^2}, \quad \text{where } A = (n-1)s^2 + n(\mu - \overline{y})^2.$$

Marginal posterior distribution of μ

We recognise (!) that the result is an unnormalized gamma integral:

$$\begin{aligned} p(\mu|y) &\propto A^{-n/2} \int_0^\infty z^{(n-2)/2}\exp(-z)dz \\ &\propto [(n-1)s^2 + n(\mu - \overline{y})^2]^{-n/2} \\ &\propto \left[1 + \frac{n(\mu - \overline{y})^2}{(n-1)s^2}\right]^{-n/2}. \end{aligned}$$

This is $t_{n-1}(\overline{y}, s^2/n)$ density.

Marginal posterior distribution of μ

Equivalently, under the noninformative uniform prior distribution on $(\mu, \log\sigma)$, the posterior distribution of μ is

$$\left.\frac{\mu - \overline{y}}{s/\sqrt{n}}\right| y \sim t_{n-1},$$

where t_{n-1} is the standard Student-t density (location 0, scale 1) with $n-1$ degrees of freedom.

Comparing the sampling theory

Again it is useful to compare the sampling theory: Under the sampling distribution, $p(y|\mu, \sigma^2)$,

$$\left.\frac{\overline{y} - \mu}{s/\sqrt{n}}\right| \mu, \sigma^2 \sim t_{n-1}.$$

The ratio $(\overline{y} - \mu)/(s/\sqrt{n})$ called a *pivotal quantity*: Its sampling distribution does not depend on the nuisance parameter σ^2, and posterior distribution does not depend on data (helps in sampling theory inference by eliminating difficulties associated with the nuisance parameter σ^2).

APPENDIX

Noninformative prior distributions

Prior distributions may be hard to construct if there is no population basis.

Statisticians have long sought prior distributions guaranteed to play a minimal role in determining the posterior distribution.

The rationale is that we should "let the data speak for themselves". Such as "objective" Bayesian analysis would use a reference or *noninformative* prior with a density described as vague, flat or diffuse.

Proper and improper prior distributions

Recall in estimating mean θ of a normal model with known variance σ^2, if prior precision, $1/\tau_0^2$, is small relative to the data precision, n/σ^2, then posterior distribution is approximately as if $\tau_0^2 = \infty$:

$$p(\theta|y) \approx \mathsf{N}(\theta|\overline{y}, \sigma^2/n).$$

That is, the posterior distribution is approximately what would result from assuming that $p(\theta) \propto$ constant for $\theta \in (-\infty, \infty)$.

Proper and improper prior distributions

The integral of the "flat" distribution $p(\theta) = 1$ for $\theta \in (-\infty, \infty)$ is not finite. In this case the distribution is referred to as *improper*.

A prior density $p(\theta)$ is *proper* if it does not depend on data and integrates to 1 (provided the integral is finite the density can always be normalised to integrate to 1).

Despite impropriety of prior distribution in the example with the normal sampling model, the posterior distribution is proper, given at least one data point.

Special cases: Location & Scale

For binomial and other single-parameter models, different principles give (slightly) different noninformative prior distributions. But for two cases—location parameters and scale parameters—all principles seem to agree.

1. If θ is a location parameter (the density of y is such that $p(y-\theta|\theta)$ is free of θ and y) then $p(\theta) \propto$ constant over the range $(-\infty, \infty)$ can be shown to be only reasonable choice of noninformative prior.

Special cases: Location & Scale

2. If $\theta =$ scale parameter (the density of y is such that $p(y/\theta|\theta)$ is free of θ and y) then $p(\theta) \propto 1/\theta$ (equivalently, $p(\log\theta) \propto 1$ or $p(\theta^2) \propto 1/\theta^2$) is the "natural" choice of noninformative prior.

General principles?

But beware that *even these principles can be misleading in some problems*, in the critical sense of suggesting prior distributions that can lead to improper and thus uninterpretable posterior distributions.

The basic point is that *all* noninformative prior specifications are arbitrary and if the results are sensitive to the particular choice, then more effort in specifying genuine prior information is required to justify any particular inference.

Multiparameter models example: Bioassay experiment

Tuesday 12th August 2008, afternoon

Lyle Gurrin

Bayesian Data Analysis
11 – 15 August 2008, Copenhagen

Multiparameter models

Reference: Gelman et al. 3.7

Few multiparameter sampling models allow explicit calculation of the posterior distribution.

Data analysis for such models is usually achieved with simulation (especially MCMC methods).

We will illustrate with a nonconjugate model for data from a bioassay experiment using a two-parameter generalised linear model.

Scientific problem

In drug development, acute toxicity tests are performed in animals.

Various dose levels of the compound are administered to batches of animals.

Animals responses typically characterised by a binary outcome: alive or dead, tumour or no tumour, response or no response etc.

Data Structure

Such an experiment gives rise to data of the form

$$(x_i, n_i, y_i); \qquad i = 1, \ldots, k \tag{35}$$

where

x_i is the i^{th} dose level $(i = 1, \ldots, k)$.

n_i animals given i^{th} dose level.

y_i animals with positive outcome (tumour, death, response).

Example Data

For the example data, twenty animals were tested, five at each of four dose levels.

Dose,x_i (log g/ml)	Number of animals, n_i	Number of deaths, y_i
-0.863	5	0
-0.296	5	1
-0.053	5	3
0.727	5	5

Racine A, Grieve AP, Fluhler H, Smith AFM. (1986). Bayesian methods in practice: experiences in the pharmaceutical industry (with discussion). *Applied Statistics* **35**, 93-150.

Sampling model at each dose level

Within dosage level i:

The animals are assumed to be exchangeable (there is no information to distinguish among them).

We model the outcomes as independent given same probability of death θ_i, which leads to the familiar binomial sampling model:

$$y_i|\theta_i \sim \text{Bin}(n_i, \theta_i) \tag{36}$$

Setting up a model across dose levels

Modelling the response at several dosage levels requires a relationship between the θ_i's and x_i's.

We start by assuming that each θ_i is an independent parameter. We relax this assumption tomorrow when we develop hierarchical models.

There are many possibilities for relating the θ_i's to the x_i's, but a popular and reasonable choice is a *logistic regression* model:

$$\text{logit}(\theta_i) = \log(\theta_i/(1-\theta_i)) = \alpha + \beta x_i \qquad (37)$$

Setting up a model across dose levels

We present an analysis based on a prior distribution for (α, β) that is independent and locally uniform in the two parameters, that is, $p(\alpha, \beta) \propto 1$, so an improper "noninformative" distribution.

We need to check that the posterior distribution is proper (details not shown).

Describing the posterior distribution

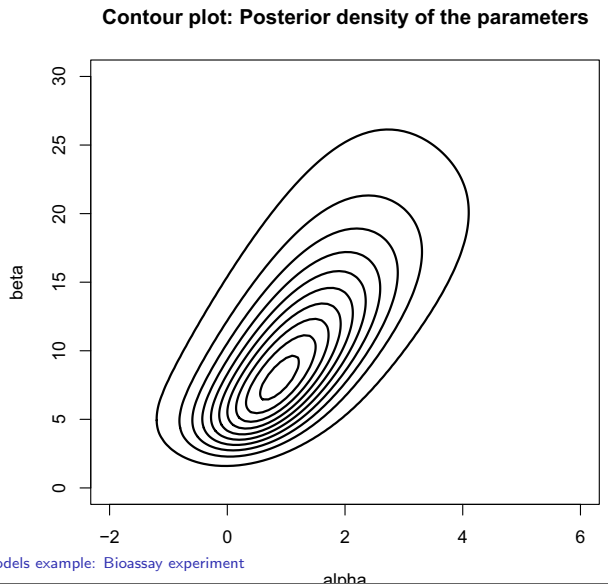
The form of posterior distribution:

$$\begin{aligned} p(\alpha, \beta|y) &\propto p(\alpha, \beta)p(y|\alpha, \beta) \\ &\propto \prod_{i=1}^{k} \left(\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)^{y_i} \left(\frac{1}{1+e^{\alpha+\beta x_i}}\right)^{n_i-y_i} \end{aligned}$$

One approach would be to use a normal approximation centered at posterior mode $(\tilde{\alpha} = 0.87, \tilde{\beta} = 7.91)$

This is similar to the classical approach of obtaining maximum likelihood estimates (*eg* by running `glm` in `R`) Asymptotic standard errors can be obtained via ML theory.

Bioassay graph 2



Discrete approx. to the post. density (1)

We illustrate computing the joint posterior distribution for (α, β) at a grid of points in 2-dimensions:

1. We begin with a rough estimate of the parameters.
   - Since $\text{logit}(\text{E}(y_i/n_i)) = \alpha + \beta x_i$ we obtain rough estimates of α and β using a linear regression of $\text{logit}(y_i/n_i)$ on x_i
   - Set $y_1 = 0.5, y_4 = 4.5$ to enable calculation.
   - $\hat{\alpha} = 0.1, \hat{\beta} = 2.9$ (standard errors 0.3 and 0.5).

Discrete approx. to the post. density (2)

2. Evaluate the posterior on a 200×200 grid; use range $[-5, 10] \times [-10, 40]$.
3. Use `R` to produce a contour plot (lines of equal posterior density).
4. Renormalize on grid so $\sum_\alpha \sum_\beta \; p(\alpha, \beta|y) = 1$ (i.e., create discrete approx to posterior)
5. Sample from *marginal* dist'n of one parameter $p(\alpha|y) = \sum_\beta \; p(\alpha, \beta|y)$.

Discrete approx. to the post. density (3)

6. Sample from *conditional* dist'n of second parameter $p(\beta|\alpha, y)$
7. We can improve sampling slightly by drawing from linear interpolation between grid points.

Alternative: exact posterior using advanced computation
(methods covered later)

Posterior inference

Quantities of interest:

- parameters (α, β).
- LD50 = dose at which $\Pr(\text{death})$ is 0.5
  $= -\alpha/\beta$
  - This is meaningless if $\beta \leq 0$ (substance not harmful).
  - We perform inference in two steps:
    (i) $\Pr(\beta > 0|y)$
    (ii) posterior dist'n of LD50 *conditional* on $\beta > 0$

Results

We take 1000 simulation draws of (α, β) from the grid (different posterior sample than results in book)

Note that $\beta > 0$ for all 1000 draws.

Summary of posterior distribution

	posterior quantiles				
	2.5%	25%	50%	75%	97.5%
α	−0.6	0.6	1.3	2.0	4.1
β	3.5	7.5	11.0	15.2	26.0
$LD50$	−0.28	−0.16	−0.11	−0.06	0.12

Lessons from simple examples

Reference: Gelman et al., 3.8.

The lack of multiparameter models with explicit posterior distributions not necessarily a barrier to analysis.

We can use simulation, maybe after replacing sophisticated models with hierarchical or conditional models (possibly invoking a normal approximation in some cases).

The four steps of Bayesian inference

1. Write the likelihood $p(y|\theta)$.
2. Generate the posterior as $p(\theta|y) = p(\theta)p(y|\theta)$ by including well formulated information in $p(\theta)$ or else use $p(\theta) =$ constant.
3. Get crude estimates for θ as a starting point or for comparison.
4. Draw simulations $\theta^1, \theta^2, \ldots, \theta^L$ (summaries for inference) and predictions $\tilde{y}^1, \tilde{y}^2, \ldots . \tilde{y}^K$ for each θ^l.

Assessing Convergence

Wednesday 13th August 2008, morning

Lyle Gurrin

Bayesian Data Analysis
11 – 15 August 2008, Copenhagen

Inference from iterative simulation

Reference: Gelman et al., 11.6

Basic method of inference from iterative simulation:

Use the collection of all simulated draws from the posterior distribution $p(\theta|y)$ to summarise the posterior density and to compute quantiles, moments etc.

Posterior predictive simulations of unobserved outcomes $\tilde{y}$ can be obtained by simulation conditional on the drawn values of θ.

Inference using iterative simulation draws does, however, require care...

Difficulties with iterative simulation

1. Too few iterations generate simulations that are not representative of the target distribution. Even at convergence the early iterations are still influenced by the starting values.
2. Within-sequence correlation: Inference based on simulations from correlated draws will be less precise than those using the same number of independent draws.
3. What initial values of the parameters should be use to start the simulation?

Within-sequence correlation

Serial correlation in the simulations is not necessarily a problem since:

- At convergence the draws are all identically distributed as $p(\theta|y)$.
- We ignore the order of simulation draws for summary and inference.

But correlation causes inefficiencies, reducing the effective number of simulation draws.

Within-sequence correlation

Should we *thin* the sequences by keeping every k^{th} simulation draw and discarding the rest?

Useful to skip iterations in problems with a large number of parameters (to save computer storage) or built-in serial correlation due to restricted jumping/proposal distributions.

Thinned sequences treated in the same way for summary and inference.

Challenges of iterative simulation

The Markov chain must be started somewhere, and initial values must be selected for the unknown parameters.

In theory the choice of initial values will have no influence on the eventual samples from the Markov chain.

In practice convergence will be improved and numerical problems avoided if reasonable initial values can be chosen.

Diagnosing convergence

It is generally accepted that the only way to diagnose convergence is to

1. Run multiple chains from a diverse set of initial parameter values.
2. Use formal diagnostics to check whether the chains, up to expected chance variability,come from the same equilibrium distribution which is assumed to be the posterior of interest.

Diagnosing convergence

Checking whether the values sampled from a Markov chain (possibly with many dimensions) has converged to its equilibrium distribution is not straightforward.

Lack of convergence might be diagnosed simply by observing erratic behaviour of the sampled values...

...**but** a steady trajectory does not necessarily mean that it is sampling from the correct posterior distribution - is it stuck in a particular area of the parameter space? Is this a result of the choice of initial values?

Handling iterative simulations

A strategy for inference from iterative simulations:

1. Simulate multiple sequences with starting points dispersed throughout the sample space.
2. Monitor the convergence of all quantities of interest by comparing variation between and within simulated sequences until these are almost equal.
3. If no convergence then alter the algorithm.
4. Discard a *burn-in* (and/or thin) the simulation sequences prior to inference.

Discarding early iterations as a burn-in

Discarding early iterations, known as *burn-in*, can reduce the effect of the starting distribution.

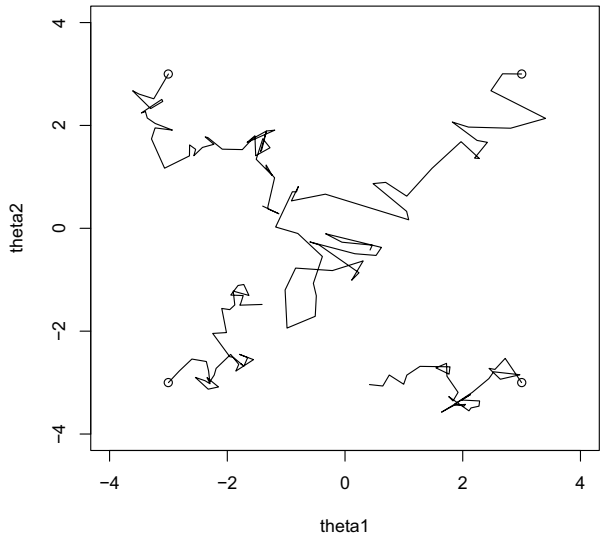
Simulated values of θ^t, for large enough t, should be close to the target distribution $p(\theta|y)$.

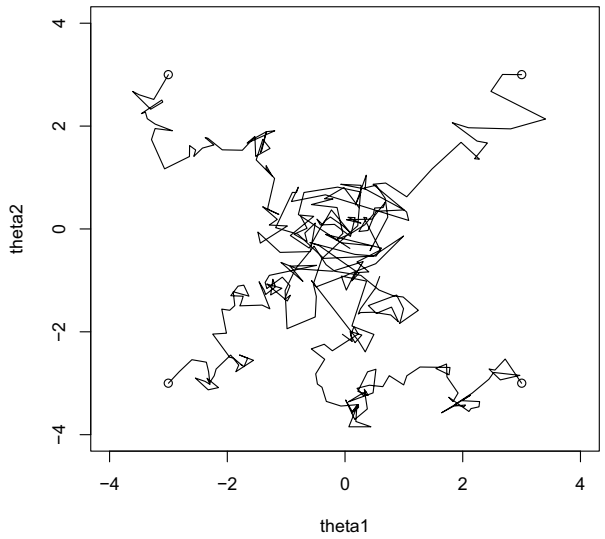
Depending on the context, different burn-in fractions can be appropriate. For any reasonable number of simulations discarding the first half is a conservative choice.

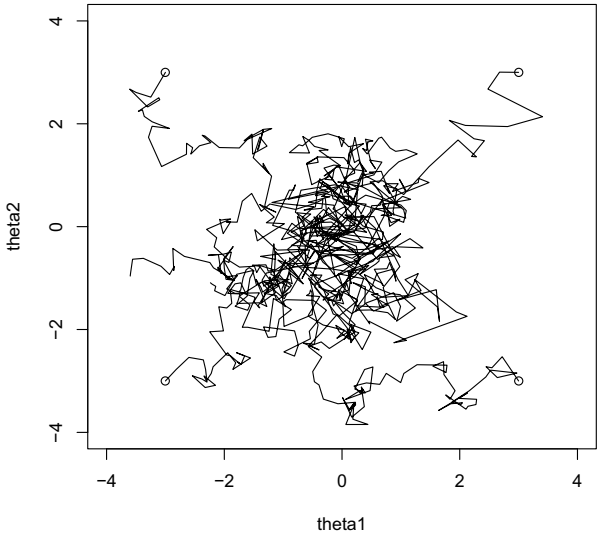
Formally assessing convergence

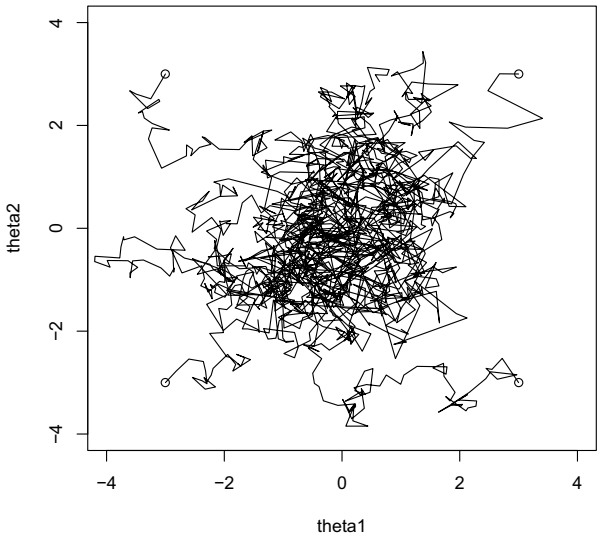
For overdispersed starting points, the within-sequence variation will be much less than the between sequence variation.

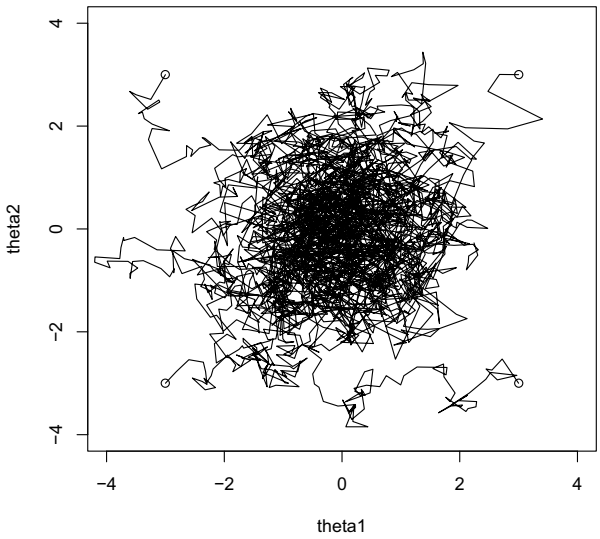
Once the sequences have mixed, the two variance components will be almost equal.






theta2
theta1
−4
−2
0
2
4


theta2
theta1
−4
−2
0
2
4


theta2
theta1
−4
−2
0
2
4

Monitoring convergence using multiple chains

- Run several sequences in parallel
- Calculate two estimates of standard deviation (SD) of *each component* of $(\theta|y)$:
  - An underestimate from SD within each sequence
  - An overestimate from SD of mixture of sequences
- Calculate the potential scale reduction factor:

$$\widehat{R} = \frac{\text{mixture-of-sequences estimate of SD}(\theta|y)}{\text{within-sequence estimate of SD}(\theta|y)}$$

Monitoring convergence (continued)

- Initially $\widehat{R}$ is large (use overdispersed starting points)
- At convergence, $\widehat{R} = 1$ (each sequence has made a complete tour)
- Monitor $\widehat{R}$ for all parameters and quantities of interest; stop simulations when they are all near 1
  (*eg* below 1.2)
- At approximate convergence, simulation noise ("MCMC error") is minor compared to posterior uncertainty about θ

Monitoring scalar estimands

Monitor each scalar estimand and other scalar quantities of interest separately.

We may also monitor the log of the posterior density (which is computed as part of the Metropolis algorithm).

Since assessing convergence is based on means and variances it is sensible to transform scalar estimands to be approximately normally distributed.

Monitoring convergence of each scalar estimand

Suppose we've simulated m parallel sequences or *chains*, each of length n (after discarding the burn-in).

For each scalar estimand ψ we label the simulation draws as $\psi_{ij}(i = 1, 2, \ldots, n; j = 1, 2, \ldots, m)$, and we compute B and W, the between- and within-sequence variances:

Between- and within-sequence variation

Between-sequence variation:

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\overline{\psi}_{.j} - \overline{\psi}_{..})^2,$$

where $\overline{\psi}_{.j} = \frac{1}{n} \sum_{i=1}^{n} \psi_{ij}$ and $\overline{\psi}_{..} = \frac{1}{m} \sum_{j=1}^{m} \overline{\psi}_{.j}$

Within-sequence variation:

$$W = \frac{1}{m} \sum_{j=1}^{m} s_j^2, \text{ where } s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\psi_{ij} - \overline{\psi}_{.j})^2$$

Estimating the marginal posterior variance

We can estimate $\mathrm{var}(\psi|y)$, the marginal posterior variance of the estimand using a weighted average of B and W:

$$\widehat{\mathrm{var}}^{+}(\psi|y) = \frac{n-1}{n} W + \frac{1}{n} B$$

This *overestimates* the posterior variance assuming an overdispersed starting distribution, but is unbiased under *stationarity* (start with the target distribution) or in the limit as $n \to \infty$.

Estimating the marginal posterior variance

For finite n, the within-sequence variance will be an *underestimate* of $\mathrm{var}(\psi|y)$ because individual sequences will not have ranged over the target distribution and will be less variable.

In the limit the expected value of W approaches $\mathrm{var}(\psi|y)$.

The scale reduction factor

We monitor convergence of iterative simulation by estimating the factor $\hat{R}$ by which the scale of the current distribution for ψ might be reduced in the limit as the number of iterations $n \to \infty$:

$$\widehat{R} = \sqrt{\frac{\widehat{\mathrm{var}}^{+}(\psi|y)}{W}},$$

which declines to 1 as $n \to \infty$.

If $\widehat{R}$ and hence the potential scale reduction is high, further simulations may improve inference about the target distribution of the estimand ψ.

The scale reduction factor

It is straightforward to calculate $\widehat{R}$ for all scalar estimands of interest (and comes automatically with `R2WinBUGS`).

The condition of $\widehat{R}$ being "near" 1 depends on the problems at hand; values below 1.1 are usually acceptable.

Avoids the need to examine time-series graphs etc.

Simulations may still be far from convergence if some areas of the target distribution was not well captured by the starting values and are "hard to reach".

The effective sample size

If n chains were truly independent then the between-sequence variance B is an unbiased estimate of $\mathrm{var}(\psi|y)$; we'd have mn simulations from n chains.

If simulations of ψ within each sequence are autocorrelated, B will be larger (in expectation) than $\mathrm{var}(\psi|y)$.

Define the effective number of independent draws as:

$$n_{\text{eff}} = mn\frac{\widehat{\mathrm{var}}^{+}(\psi|y)}{B}.$$

Hierarchical models

Wednesday 13th August 2008, morning

Lyle Gurrin

Bayesian Data Analysis
11 – 15 August 2008, Copenhagen

Hierarchical models - introduction

Reference: Gelman et al., 5.1-5.3

Often we would like the parameters of a prior or population distribution to be estimated from the data.

Many problems have multiple parameters that are related; we'd like a joint probability model that reflects this dependence.

It is useful to think **hierarchically**: The distribution of observed outcomes are conditional on parameters which themselves have a probability specification.

Bioassay example continued

Let's generalize our simple bioassay example:

A single (α, β) may be inadequate to fit a combined data set (several experiments).

Imagine repeated bioassays with same compound, where (α_j, β_j) parameters from different bioassays.

Separate unrelated (α_j, β_j) are likely to "overfit" data (only 4 points in each data set).

Information about the parameters of one bioassay can be obtained from others' data.

The hierarchical approach

A natural prior distribution arises by assuming the (α_j, β_j)'s are a sample from a common population distribution.

We'd be better off estimating the parameters governing the population distribution of (α_j, β_j) rather than each (α_j, β_j) separately.

This introduces new parameters that govern this population distribution, called **hyperparameters**.

Hierarchical models uses many parameters but imposing a population distribution induces enough structure to avoid overfitting.

Hierarchical models and empirical Bayes

Hierarchical models retain the advantages of using the data to estimate the population parameters, but eliminate the disadvantages (of dealing with many parameters) by putting a joint probability model on the entire set of parameters *and* the data.

We can then do a Bayesian analysis of the joint distribution of all model parameters.

Using data to estimate the prior (hyper)parameters beforehand is called *empirical Bayes* and is an approximation to the complete hierarchical Bayesian analysis.

Setting up hierarchical models: exchangeability (review)

Consider a set of experiments $j = 1, 2, \ldots, J$ in which experiment j has data (vector) y_j, parameter (vector) θ_j and likelihood $p(y_j|\theta_j)$.

If no information (other than the y_j's) is available to distinguish the θ_j's from each other, and no ordering or grouping of the parameters can be made, then we can assume symmetry among the parameters in the prior distribution.

This symmetry is represented probabilistically by *exchangeability*:

Defining exchangeability

A set of random variables $(\theta_1, \ldots, \theta_k)$ is said to be *exchangeable* if the joint distribution is invariant to permutations of the indexes $(1, \ldots, k)$, that is,the indexes contain no information about the data values.

- The simplest form: i.i.d. given some unknown parameter.
- Seemingly non-exchangeable random variables may become exchangeable if we condition on all available information (*e.g.* covariates regression analysis)
- Hierarchical models often use exchangeable models for prior distribution of model parameters.

Basic exchangeable model

The basic form of an exchangeable model has the parameter θ_j as an independent sample from a prior distribution governed by some unknown parameter ϕ.

$\theta = (\theta_1, \ldots, \theta_k)$ are independent conditional on additional parameters ϕ (the hyperparameters):

$$p(\theta|\phi) = \prod_{j=1}^{k} p(\theta_j|\phi) \tag{38}$$

In general ϕ is unknown so our distribution for θ must average over uncertainty in ϕ:

Basic exchangeable model

$$p(\theta) = \int \left[\prod_{j=1}^{k} p(\theta_j|\phi) \right] p(\phi) d\phi \qquad (39)$$

This *mixture* of i.i.d.'s is usually all we need to capture exchangeability in practice.

Bruno de Finetti proved a theorem that as $J \to \infty$ any suitably well-behaved exchangeable distribution on $\theta_1, \theta_2, \ldots, \theta_J$ can be written as an i.i.d. mixture.

Setting up hierarchical models: typical structure

The model is specified in nested stages

- $p(y|\theta)$ = the sampling distribution of the data.
- $p(\theta|\phi)$ = the prior (population) distribution for θ given ϕ.
- $p(\phi)$ = the prior (hyperprior) distribution for ϕ
- More levels are possible!
- The hyperprior at highest level is often diffuse.

A normal-normal hierarchical model

Inference based on posterior distribution of unknowns:

$$\begin{aligned} p(\theta, \phi|y) &\propto p(\theta, \phi) p(y|\theta, \phi) \\ &\propto p(\theta, \phi) p(y|\theta) \qquad y \text{ ind. of } \phi \text{ given } \theta \\ &\propto p(\phi) p(\theta|\phi) p(y|\theta), \end{aligned}$$

Inference (and computation) is often carried out in two steps:

1. Inference for θ as if we knew ϕ using the posterior conditional distribution $p(\theta|y, \phi)$.
2. Inference for ϕ based on posterior marginal dist'n $p(\phi|y)$.

Example 1: Meta-analysis of clinical trials

Reference: Gelman et al., 5.6 & 19.4 Spiegelhalter et al., 3.17

Meta-analysis aims to summarise and integrate findings from research studies in a particular area.

It involves combining information from several parallel data sources, so is closely connected to hierarchical modelling (but there are well known frequentist methods as well).

We'll re-inforce some of the concepts of hierarchical modelling in a meta-analysis of clinical trials data.

The aims of meta-analysis

Use a combined analysis of the studies to measure the strength of evidence for (and magnitude of) any beneficial effect of the treatment under study.

Any formal analysis must be preceded by the application of inclusion rigorous criteria (see Gelman *et al.* chapter 7).

Two clinical trial examples

Both examples deal with reducing mortality after myocardial infarction:

1. 22 clinical trials, each with two groups of heart attack patients receiving (or not) beta-blockers (rates 3% to 21%, samples sizes from < 100 to almost 2000) (BUGS online, Gelman *et al.* 5.6)
2. 8 clinical trials, each with each with two groups of heart attack patients receiving (or not) IV magnesium sulfate (rates 1% to 17%, samples sizes from < 50 to more than 2300)

Parameters for each clinical trial

Meta-analysis often involves data in the form of several 2×2 tables.

In trial j there are n_{0j} control subjects and n_{1j} treatment subjects, with y_{0j} and y_{1j} deaths respectively.

Sampling model: y_{0j} and y_{1j} have independent binomial sampling distributions with probabilities of death p_{0j} and p_{1j} respectively.

Trial	Magnesium group		Control group	
	deaths	patients	deaths	patients
Morton	1	40	2	36
Rasmussen	9	135	23	135
Smith	2	200	7	200
Abraham	1	48	1	46
Feldstedt	10	150	8	148
Schechter	1	59	9	56
Ceremuzynski	1	25	3	23
LIMIT-2	90	1159	118	1157

Odds ratios as a measure of effectiveness

We'll use the natural logarithm θ_j of the *odds ratio* $\rho_j = (p_{1j}/(1-p_{1j}))/(p_{0j}/(1-p_{0j}))$ as a measure of effect size comparing treatment to control groups:

- Interpretability in a range of study designs (cohorts, case-control and clinical trials).
- Posterior distribution of $\theta_j = \ln(\rho_j)$ close to normality even for small sample sizes.
- Canonical (natural) parameter for logistic regression.

Normal approximation to the likelihood

Summarise the results of each trial with an approximate normal likelihood for θ_j.

Let y_j represent the **empirical logit**, a point estimate of the effect θ_j in the j^{th} study where $j = 1, \ldots, J$:

$$y_j = \log\left(\frac{y_{1j}}{n_{1j} - y_{1j}}\right) - \log\left(\frac{y_{0j}}{n_{0j} - y_{0j}}\right),$$

with approximate sampling variance:

$$\sigma_j^2 = \frac{1}{y_{1j}} + \frac{1}{n_{1j} - y_{1j}} + \frac{1}{y_{0j}} + \frac{1}{n_{0j} - y_{0j}}.$$

Normal approximation to the likelihood

Here we use the results of one analytic approach to produce a point estimate and standard error that can be regarded as approximately a normal mean and standard deviation.

We use the notation y_j and σ_j^2 to be consistent with the earlier lecture.

We do not use the continuity correction of adding a fraction such as $1/2$ to the four counts of the contingency table to improve the asymptotic normality of the sampling distributions.

More data for the Magnesium example

Trial	Magnesium group		Control group		Estimated log(OR) y_k	Estimated SD s_k
	deaths	patients	deaths	patients		
Morton	1	40	2	36	−0.83	1.25
Rasmussen	9	135	23	135	−1.06	0.41
Smith	2	200	7	200	−1.28	0.81
Abraham	1	48	1	46	−0.04	1.43
Feldstedt	10	150	8	148	0.22	0.49
Schechter	1	59	9	56	−2.41	1.07
Ceremuzynski	1	25	3	23	−1.28	1.19
LIMIT-2	90	1159	118	1157	−0.30	0.15

A hierarchical model: Stage 1

The first stage of the hierarchical model assumes that:

$$y_j|\theta_j, \sigma_j^2 \sim \mathsf{N}(\theta_j, \sigma_j^2), \tag{40}$$

The simplification of known variances is reasonable with large sample sizes (but see the online examples that use the "true" binomial sampling distribution).

Possible assumptions about the $\theta_j's$

1. Studies are identical replications, so $\theta_j = \mu$ for all j (no heterogeneity)...**or**
2. No comparability between studies so that each study provides no information about the other (complete heterogeneity)...**or**
3. Studies are exchangeable but not identical or completely unrelated (compromise between 1 and 2).

Classical model for a fixed treatment effect

If all of the θ_j are identical and equal to a common treatment effect μ, and therefore equation (40) becomes

$$y_j|\mu, \sigma_j^2 \sim \mathsf{N}(\mu, \sigma_j^2). \tag{41}$$

The classical pooled estimate $\hat{\mu}$ of μ weights each trial estimate inversely by its variance:

$$\hat{\mu} = \frac{\sum_{j=1}^{J} y_j/\sigma_j^2}{\sum_{j=1}^{J} 1/\sigma_j^2}.$$

Assumptions imply $\hat{\mu}$ normal with variance $[\sum_{j=1}^{J} 1/\sigma_j^2]^{-1}$.

Classical test of heterogeneity

A classical test for heterogeneity, that is, whether it is reasonable to assume all the trials are measuring the same quantity, is provided by

$$Q = \sum_{j=1}^{J} \frac{(y_j - \hat{\mu})^2}{\sigma_j^2} \tag{42}$$

which has a χ^2_{J-1} distribution under the null hypothesis of homogeneity. It is well known that this is not a very powerful test (Whitehead (2002)).

A hierarchical model: Stage 2

The second stage of the hierarchical model assumes that the trial means θ_j are exchangeable with a normal distribution

$$\theta_j \sim \mathsf{N}(\mu, \tau^2). \tag{43}$$

A hierarchical model: Stage 2

If μ and τ^2 are fixed and known, then the conditional posterior distribution of the θ_j's are independent, and

$$\theta_j|\mu, \tau, y \sim \mathsf{N}(\hat{\theta}_j, V_j),$$

where

$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2} y_j + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \quad \text{and} \quad V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}.$$

Note that the posterior mean is a precision-weighted average of the prior population mean and the observed y_j representing the treatment effect in the j^{th} group.

Posterior distn's for the θ_j's given y, μ, τ

The expression for the posterior distribution of θ_j can be rearranged as

$$\theta_j|y_j \sim \mathsf{N}(B_j\mu + (1-B_j)y_j, (1-B_j)\sigma_j^2) \quad (44)$$

where $B_j = \sigma_j^2/(\sigma_j^2+\tau^2)$ is the weight given to the prior mean.

Ignoring data from the other trials is equivalent to setting $\tau^2 = \infty$, that is, $B_j = 0$.

The classical pooled result results from $\tau^2 \to 0$, that is, $B_j = 1$.

Conditional prior distribution for μ

A uniform conditional prior distribution $p(\mu|\tau) = 1$ for μ leads to the following posterior distribution:

$$\mu|\tau, y \sim \mathsf{N}(\hat{\mu}, V_\mu)$$

where $\hat{\mu}$ is the precision weighted average of the y_j values, and V_μ^{-1} is the total precision:

$$\hat{\mu} = \frac{\sum_{j=1}^J \frac{1}{\sigma_j^2+\tau^2} y_j}{\sum_{j=1}^J \frac{1}{\sigma_j^2+\tau^2}} \quad \text{and} \quad V_\mu^{-1} = \sum_{j=1}^J \frac{1}{\sigma_j^2+\tau^2}.$$

$\tau^2 \to \infty$ gives the classical result where $B_j = 1$.

The exchangeable model and shrinkage

The exchangeable model therefore leads to *narrower* posterior intervals for the θ_j's than the "independence" model, but they are *shrunk* towards the prior mean response.

The degree of shrinkage depends on the variability between studies, measured by τ^2, and the precision of the estimate of the treatment effect from the individual trial, measured by σ_j^2.

The full hierarchical model

The hierarchical model is completed by specifying a prior distribution for τ - we'll use the noninformative prior $p(\tau) = 1$.

Nevertheless, $p(\tau|y)$ is a complicated function of τ:

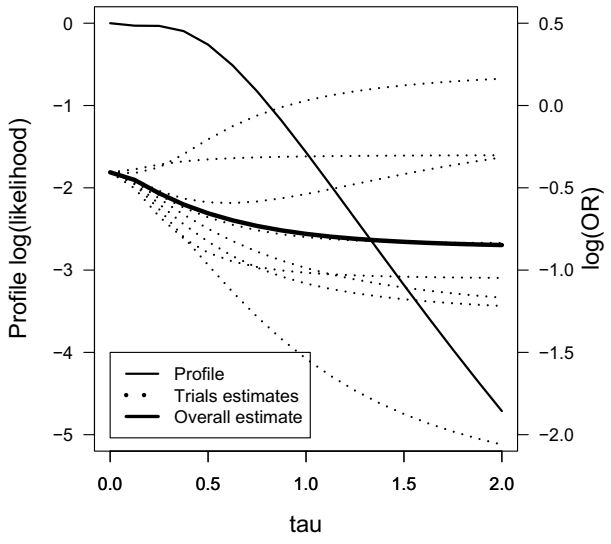
$$\begin{aligned} p(\tau|y) &\propto \frac{\prod_{j=1}^{J} \mathsf{N}(y_j|\hat{\mu}, \sigma_j^2 + \tau^2)}{\mathsf{N}(\hat{\mu}|\hat{\mu}, V_\mu)} \\ &\propto V_\mu^{1/2} \prod_{j=1}^{J} (\sigma_j^2 + \tau^2)^{-1/2} \exp\left(-\frac{(y_j - \hat{\mu})^2}{2(\sigma_j^2 + \tau^2)}\right) \end{aligned}$$

The profile likelihood for τ

A tractable alternative to the marginal posterior distribution is the *profile* likelihood for τ, derived by replacing μ in the joint likelihood for μ and τ by its conditional maximum likelihood estimate $\hat{\mu}(\tau)$ *given* the value of τ.

This summarises the support for different values of τ and is easily evaluated as

$$\prod_{j=1}^{J} \mathsf{N}(y_j|\hat{\mu}(\tau), \sigma_j^2 + \tau^2).$$



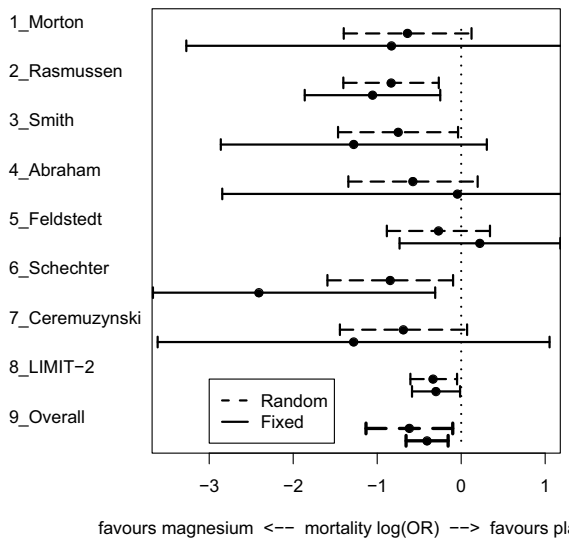
Estimates of τ

The maximum likelihood estimate is $\hat{\tau} = 0$ although values of τ with a profile log(likelihood) above $-1.96^2/2 \approx -2$ might be considered as being reasonably supported by the data.

$\hat{\tau} = 0$ would not appear to be a robust choice as an estimate since non-zero values of τ, which are well-supported by the data, can have a strong influence on the conclusions. We shall assume, for illustration, the method-of-moments estimator $\hat{\tau} = 0.35$.

Results of the meta-analysis

Trial	Magnesium group		Control group		Estimated log(OR) y_k	Estimated SD s_k	Shrinkage B_k
	deaths	patients	deaths	patients			
Morton	1	40	2	36	−0.83	1.25	0.90
Rasmussen	9	135	23	135	−1.06	0.41	0.50
Smith	2	200	7	200	−1.28	0.81	0.80
Abraham	1	48	1	46	−0.04	1.43	0.92
Feldstedt	10	150	8	148	0.22	0.49	0.59
Schechter	1	59	9	56	−2.41	1.07	0.87
Ceremuzynski	1	25	3	23	−1.28	1.19	0.89
LIMIT-2	90	1159	118	1157	−0.30	0.15	0.11



Example 2: SAT coaching

Reference: Gelman et al., 5.5

The Study:

- Separate randomized experiments were conducted in 8 high schools.
- The outcome measure is the improvement in SAT-Verbal score.
- The intervention effect is estimated using analysis of covariance to adjust for PSAT (preliminary SAT) involving a separate regression for each school.

SAT coaching example: The model

- The quantities of interest are the θ_j: Average "true" effects of coaching programs.
- Data y_j: separate estimated treatment effects for each school.
- The standard errors σ_j are assumed known (large samples).
- This is a randomized experiment with large samples, no outliers, so we appeal to the central limit theorem:

$$y_j|\theta_j \sim \mathsf{N}(\theta_j, \sigma_j^2)$$

SAT coaching example: The data

School	Estimated treatment effect, y_j	Standard error of effect estimate, σ_j	True treatment effect, θ_j
A	28	15	?
B	8	10	?
C	−3	16	?
D	7	11	?
E	−1	9	?
F	1	11	?
G	18	10	?
H	12	18	?

Nonhierarchical approach 1

Consider the 8 programs separately:

- Two programs appear to work (18-28 points)
- Four programs appear to have a small effect
- Two programs appear to have negative effects
- Large standard errors imply overlapping CIs

Nonhierarchical approach 2

Use a pooled estimate of the coaching effect:

- "Classical" test of homogeneity fails to reject that all θ_j's are equal.
- Pooled estimate $\hat{\mu} = \sum_j (y_j/\sigma_j^2) / \sum_j (1/\sigma_j^2) = 7.9$ (pooled standard error is s.e.$(\hat{\mu}) = 4.2$).
- Pooled estimate applies to each school.

Separate and pooled estimates are both unreasonable! A hierarchical model provides a compromise.

Model specification

- Observed data are normally distributed with a different mean in each group:
  $p(y_j|\theta_j) \sim N(\theta_j, \sigma_j^2) \quad j = 1, \ldots, J.$
  $y_j = 1/n_j \sum_{i=1}^{n_j} y_{ij}.$
  $\sigma_j^2 = \sigma^2/n_j$ (assumed known).
- Prior model for θ_j's is based on a normal population distribution (conjugate)

$$p(\theta_1, \ldots, \theta_J|\mu, \tau) = \prod_{j=1}^{J} N(\theta_j|\mu, \tau)$$

Model specification

- Hyperprior distribution can be factored as $p(\mu, \tau) = p(\tau)p(\mu|\tau)$
- $p(\mu|\tau) \propto 1$ (noninformative, this won't matter much because the data supply a great deal of information about μ)
- $p(\tau) \propto 1$ (must be sure the posterior is proper)

Computation

The joint posterior distribution:

$$
\begin{aligned}
& p(\theta, \mu, \tau|y) \\
& \propto p(\mu, \tau)p(\theta|\mu, \tau)p(y|\theta) \\
& \propto \prod_{j=1}^{J} N(\theta_j|\mu, \tau^2) \prod_{j=1}^{J} N(y_j|\theta_j, \sigma_j^2) \\
& \propto \tau^{-J} \exp\left[-\frac{1}{2}\sum_j \frac{1}{\tau^2}(\theta_j - \mu)^2\right] \exp\left[-\frac{1}{2}\sum_j \frac{1}{\sigma_j^2}(y_j - \theta_j)^2\right]
\end{aligned}
$$

Factors depending only on y and $\{\sigma_j\}$ treated as constant.

Conditional posterior dist'n of θ given μ, τ, y

- Treat (μ, τ) as fixed in the previous expressions.
- Given (μ, τ), the J separate parameters θ_j are independent in their posterior distribution since they appear in different factors in the likelihood (which factors into J components).
- $\theta_j|y, \mu, \tau \sim N(\hat{\theta}_j, V_j)$ with

$$
\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\, y_j + \frac{1}{\tau^2}\, \mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \quad \text{and} \quad V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}
$$

Marginal posterior dist'n of μ, τ given y

Derive this analytically by integrating $p(\theta, \mu, \tau|y)$ over θ:

Data distribution:
$p(y|\mu, \tau) = \prod_{j=1}^{J} N(y_j|\mu, \sigma_j^2 + \tau^2)$

$$\begin{aligned} p(\mu, \tau|y) &\propto \prod_{j=1}^{J} N(y_j|\mu, \sigma_j^2 + \tau^2) \\ &\propto \prod_{j=1}^{J} (\sigma_j^2 + \tau^2)^{-1/2} \exp\left(-\frac{(y_j - \mu)^2}{2(\sigma_j^2 + \tau^2)}\right) \end{aligned}$$

Posterior distribution of μ given τ, y

Factor the distribution: $p(\mu, \tau|y) = p(\tau|y)p(\mu|\tau, y)$.

$p(\mu|\tau, y)$ is obtained by looking at $p(\mu, \tau|y)$ and thinking of τ as known. With a uniform prior for $\mu|\tau$, the log posterior is quadratic in μ and therefore normal:

$$p(\mu|\tau, y) \propto \prod_{j=1}^{J} N(y_j|\mu, \sigma_j^2 + \tau^2)$$

This is a normal sampling distribution with a noninformative prior density on μ.

Posterior distribution of μ given τ, y

The mean and variance are obtained by considering group means y_j as J independent estimates of μ with variance $\sigma_j^2 + \tau^2$.

Result: $\mu|\tau, y \sim N(\hat{\mu}, V_\mu)$ with

$$\hat{\mu} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2} y_j}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}} \quad \text{and} \quad V_\mu = \frac{1}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}}$$

Posterior distribution of τ given y

We could integrate $p(\mu, \tau|y)$ over μ?

It is easier to use identity $p(\tau|y) = p(\mu, \tau|y)/p(\mu|\tau, y)$ (which holds for all μ), and evaluate at $\mu = \hat{\mu}$:

$$\begin{aligned} p(\tau|y) &\propto \frac{\prod_{j=1}^{J} N(y_j|\hat{\mu}, \sigma_j^2 + \tau^2)}{N(\hat{\mu}|\hat{\mu}, V_\mu)} \\ &\propto V_\mu{}^{1/2} \prod_{j=1}^{J} (\sigma_j^2 + \tau^2)^{-1/2} \exp\left(-\frac{(y_j - \hat{\mu})^2}{2(\sigma_j^2 + \tau^2)}\right) \end{aligned}$$

Posterior distribution of τ given y

Note that V_μ and $\hat{\mu}$ are both functions of τ, and thus so is $p(\tau|y)$, so we compute $p(\tau|y)$ on a grid of values of τ.

The numerator of the first expression for $p(\tau|y)$ is the *profile* likelihood for τ given the maximum likelihood estimate of μ given τ - more details later.

Normal-normal model computation: Summary

To simulate from joint posterior distribution $p(\theta, \mu, \tau|y)$:

1. Draw τ from $p(\tau|y)$ (grid approximation)
2. Draw μ from $p(\mu|\tau, y)$ (normal distribution)
3. Draw $\theta = (\theta_1, \ldots, \theta_J)$ from $p(\theta|\mu, \tau, y)$ (independent normal distribution for each θ_j)

Apply these ideas to SAT coaching data; repeat 1000 times to obtain 1000 simulations.

SAT coaching example: post. quantiles

School	2.5%	25%	50%	75%	97.5%	y_j
A	− 2	6	10	16	32	28
B	− 5	4	8	12	20	8
C	−12	3	7	11	22	− 3
D	− 6	4	8	12	21	7
E	−10	2	6	10	17	− 1
F	− 9	2	6	10	19	1
G	− 1	6	10	15	27	18
H	− 7	4	8	13	23	12
μ	− 2	5	8	11	18	
τ	0.3	2.3	5.1	8.8	21.0	

SAT coaching example: Results

We can address more complicated questions:

$\Pr(\text{school A's effect is the max}) = 0.25$
$\Pr(\text{school B's effect is the max}) = 0.10$
$\Pr(\text{school C's effect is the max}) = 0.10$
$\Pr(\text{school A's effect is the min}) = 0.07$
$\Pr(\text{school B's effect is the min}) = 0.09$
$\Pr(\text{school C's effect is the min}) = 0.17$
$\Pr(\text{school A's effect} > \text{school C's effect}) = 0.67$

Hierarchical models: Summary

- They account for multiple levels of variability.
- There is a data-determined degree of pooling across studies.
- Classical estimates (no pooling, complete pooling) provide a starting point for analysis.
- We can draw inference about the population of schools.

APPENDIX I

Computation with hierarchical models: Overview

Conjugate case ($p(\theta|\phi)$ conjugate prior for $p(y|\theta)$)

- write $p(\theta, \phi|y) = p(\phi|y)p(\theta|\phi, y)$
- identify conditional posterior density $p(\theta|\phi, y)$ (easy for conjugate models)
- obtain marginal posterior distribution $p(\phi|y)$ (more about this step on next slide)
- draw from $p(\phi|y)$ and then $p(\theta|\phi, y)$

Approaches for obtaining $p(\phi|y)$

Integration: $p(\phi|y) = \int p(\theta, \phi|y)d\theta$

Algebra: Use $p(\phi|y) = p(\theta, \phi|y)/p(\theta|\phi, y)$ for a convenient value of θ.

Sampling from $p(\phi|y)$

- easy if it is a common distribution
- grid if ϕ is low-dimensional
- more sophisticated methods (later)

"Empirical Bayes" methods replace $p(\phi|y)$ by mode.

Statistical measures of fetal growth

Thursday 14th August 2008, morning

Lyle Gurrin

Bayesian Data Analysis
11 – 15 August 2008, Copenhagen

The fetal origins hypothesis

Fetal adaptation to an adverse intrauterine environment programs permanent change.

It is now acknowledged that their is an inverse relationship between low birthweight and subsequent elevated blood pressure (Huxley R. *Lancet* (2002)).

We'll look at quantifying fetal growth using statistical summary measures derived from serial fetal biometric data.

Western Australian Pregnancy Cohort Study

Subjects received routine ultrasound examination at 18 weeks gestation.

Additional ultrasound examination at 24, 28, 34, 38 weeks gestation as part of a randomised trial of the safety of repeated antenatal ultrasound.

Criteria for growth data

These data exclude multiples, infants born less than 37 weeks gestation, or with maternal or fetal disease.

Required agreement within 7 days between gestational age based on last menstrual period and ultrasound examination at 18 weeks.

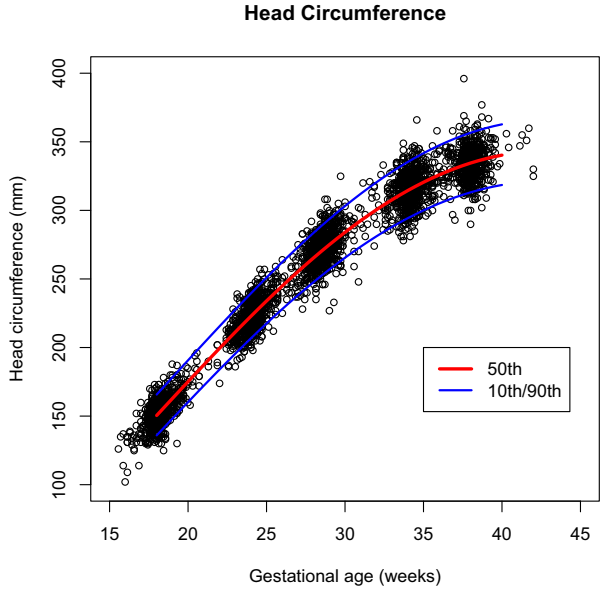
3450 ultrasound measurements on 707 fetus of five fetal dimensions (BPD, OFD, HC, AC, FL). We'll look at HC = head circumference.

Statistical modelling

Y_{ij} is the measured head circumference for the i^{th} fetus at the j^{th} timepoint $t_{ij} = 18, 24, 28, 34, 38$ weeks

The number of measurements on an individual fetus varies from 1 to 7.

The aim is to model the relationship between head circumference Y_{ij} and gestational age t_{ij}.



Modelling strategy

We follow the methodology in Royston P. *Stat. Med.* (1995):

Transform both sides of the regression equation to establish approximately linear relationship between transformed outcome $Y_{ij}^{(\lambda)}$ and timescale $g(t_{ij})$

The longitudinal design suggests the use of *linear mixed model*.

Transformation of Y_{ij}

Use the familiar **Box-Cox** transformation:

$$\begin{aligned} Y_{ij}^{(\lambda)} &= (Y_{ij}^{\lambda} - 1)/\lambda \quad \text{if} \quad \lambda \neq 0 \\ &= \log(Y_{ij}) \quad \text{if} \quad \lambda = 0 \end{aligned}$$

We account for gestational age when choosing λ by fitting a cubic polynomial in time t_{ij}.

Transformed t_{ij}

We assume that $Y_{ij}^{(\lambda)}$ is linear in a second degree *fractional polynomial* in t_{ij} (Royston P, Altman DG *Appl. Stat.* (1994)).

$$\begin{aligned} g(t_{ij}) &= \xi_0 + \xi_1 t_{ij}^{(p_1)} + \xi_2 t_{ij}^{(p_2)} \\ t_{ij}^{(p_1)} &= t_{ij}^{p_1} \quad \text{if} \quad p_1 \neq 0 \\ &= \log(t_{ij}) \quad \text{if} \quad p_1 = 0 \end{aligned}$$

If $p_1 = p_2$ then let $t_{ij}^{(p_2)} = t_{ij}^{(p_1)} \log(t_{ij})$.

Fractional polynomials in t_{ij}

So $g(t_{ij}) = \xi_0 + \xi_1 t_{ij}^{(p_1)} + \xi_2 t_{ij}^{(p_2)}$.

Select p_1, p_2 from $\{-3, -2, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, 2, 3\}$.

Use a grid search to find p_1, p_2 providing the best fit to $Y_{ij}^{(\lambda)}$.

Estimate ξ_1 and ξ_2 using maximum likelihood, with separate intercepts for each subject.

Let $X_{ij} = t_{ij}^{(p_1)} + (\xi_2/\xi_1)\ t_{ij}^{(p_2)}$.

Transformations for head circumference

For head circumference we find that $\hat{\lambda} = 0.56 \approx 0.50$, equivalent to the square root transformation.

We use a quadratic transformation of gestational age:

$$X_{ij} = t_{ij} - 0.0116638 t_{ij}^2.$$

Simple linear model

The simplest linear model would be:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij},$$

where

$$\varepsilon_{ij} \sim \mathsf{N}(0, \sigma_\varepsilon^2).$$

Mixed linear model

We can extend this to a *mixed* linear model by allowing subject-specific intercepts and gradients:

$$Y_{ij} = (\beta_0 + u_{0i}) + (\beta_1 + u_{1i})X_{ij} + \varepsilon_{ij},$$

where

$$\varepsilon_{ij} \sim \mathsf{N}(0, \sigma^2_\varepsilon).$$

and

$$\begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix} = \mathsf{N}_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma^2_0 & \sigma_{01} \\ \sigma_{01} & \sigma^2_1 \end{bmatrix} \right]$$

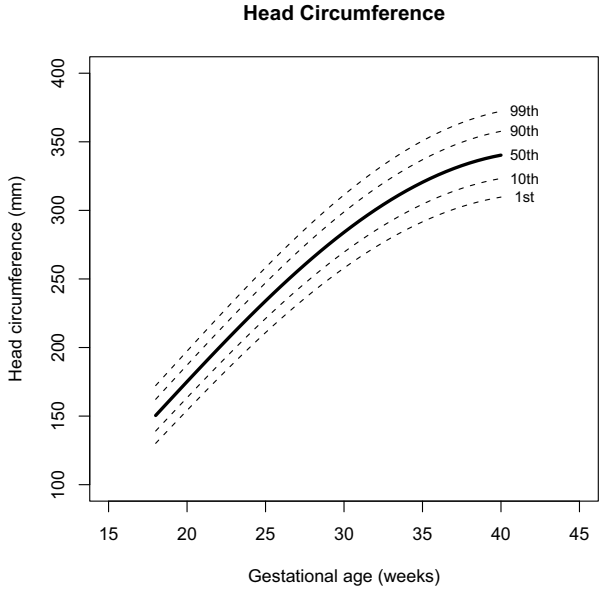
with $\mathrm{cov}(\varepsilon_{ij}, u_i) = 0$.

References centiles

The models can be used to derive reference centiles:

$$\begin{aligned} \mathrm{var}(Y_{ij}) &= \mathrm{var}(u_{0i}) + 2\mathrm{cov}(u_{0i}, u_{1i})X_{ij} \\ &\quad + \mathrm{var}(u_{1i})X^2_{ij} + \mathrm{var}(\varepsilon_{ij}) \\ &= \sigma^2_0 + 2\sigma_{01}X_{ij} + \sigma^2_1 X^2_{ij} + \sigma^2_\varepsilon \end{aligned}$$

The variance of Y_{ij} (square root of head circumference) is quadratic in gestational age.



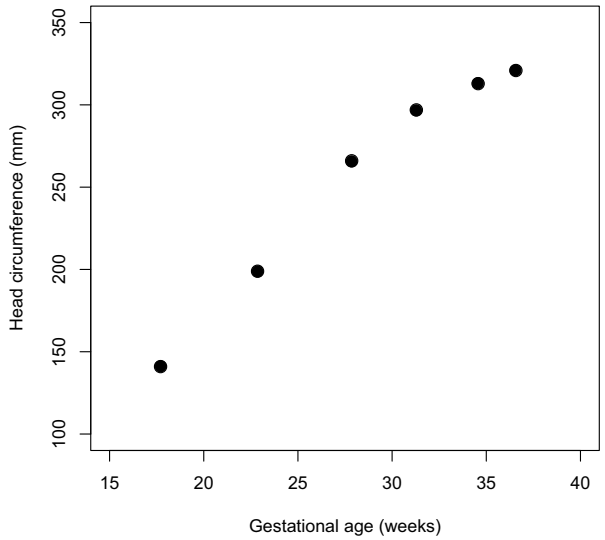
Measure of fetal growth

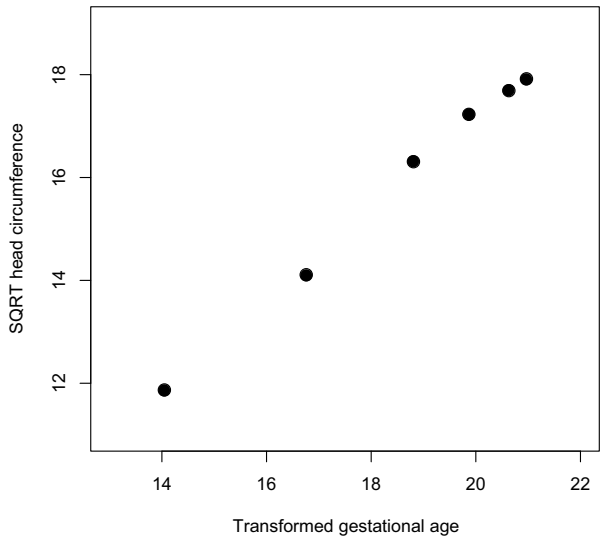
We call also use models to derive measures of growth or change in size:

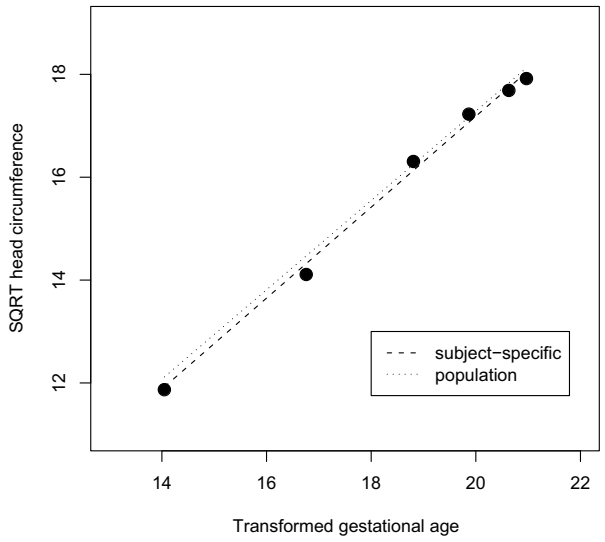
$Y_{i1}^{(\lambda)}$ and $Y_{i2}^{(\lambda)}$ are bivariate normal, so $Y_{i2}^{(\lambda)}$ given $Y_{i1}^{(\lambda)}$ is univariate normal. The "conditional Z-score" is

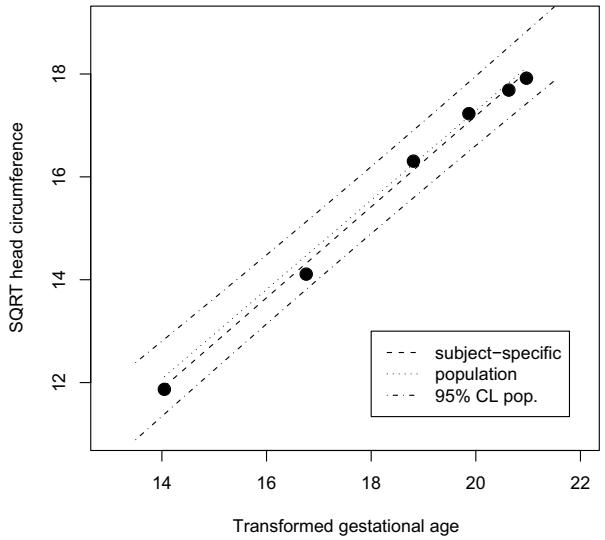
$$Z_{2|1} = \frac{Y_{i2}^{(\lambda)} - \mathrm{E}(Y_{i2}^{(\lambda)}|Y_{i1}^{(\lambda)})}{\sqrt{\mathrm{var}(Y_{i2}^{(\lambda)}|Y_{i1}^{(\lambda)})}}$$

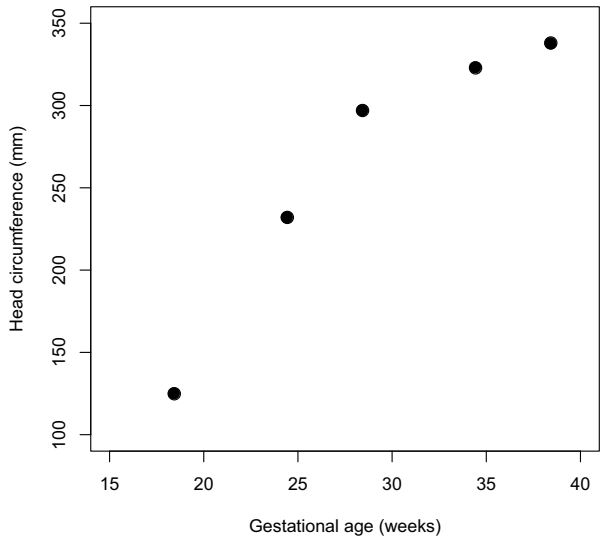
We used measures at 38 weeks gestation conditional on value at 18 weeks gestation and relating these to birthweight and subsequent blood pressure in childhood.

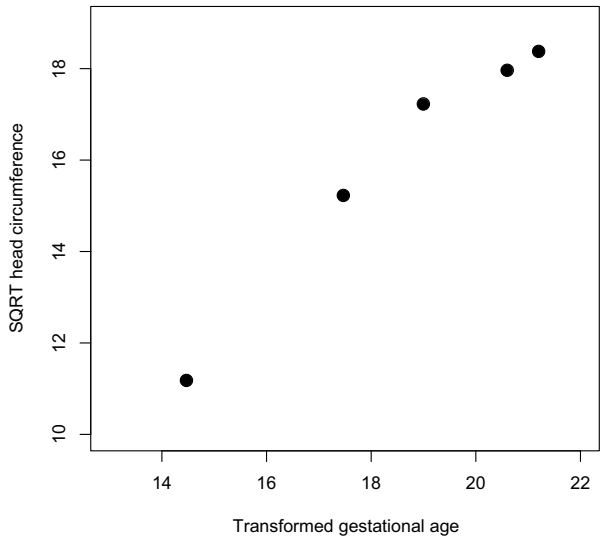


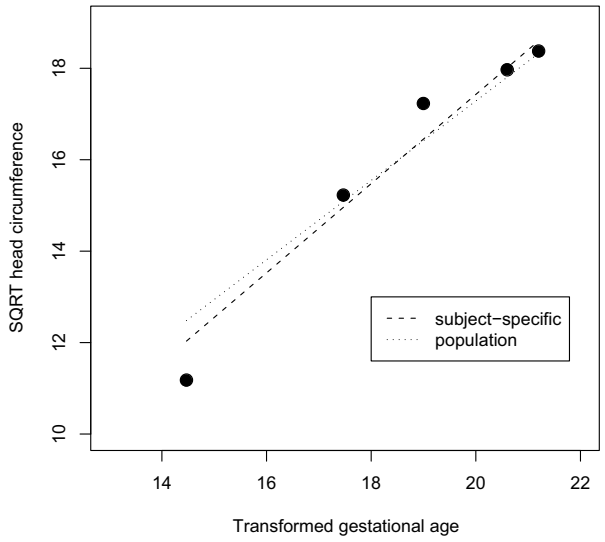


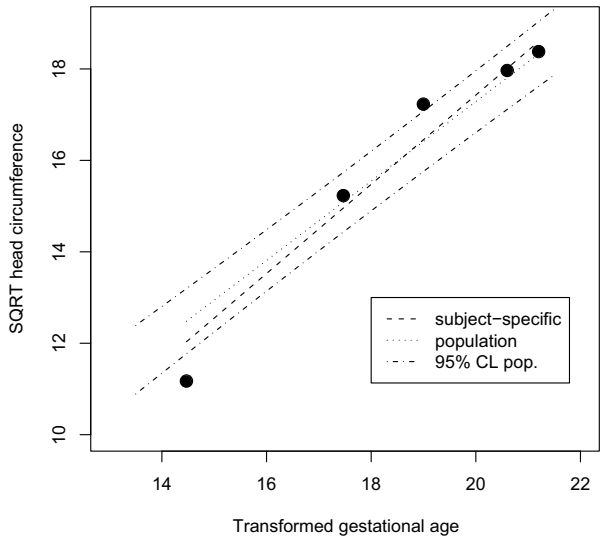

SQRT head circumference
12
14
16
18
14
16
18
20
22
Transformed gestational age
subject−specific
population


SQRT head circumference
12
14
16
18
14
16
18
20
22
Transformed gestational age
subject−specific
population
95% CL pop.


Head circumference (mm)
100
150
200
250
300
350
15
20
25
30
35
40
Gestational age (weeks)


SQRT head circumference
10
12
14
16
18
14
16
18
20
22
Transformed gestational age


SQRT head circumference
10
12
14
16
18
14
16
18
20
22
Transformed gestational age
subject−specific
population


SQRT head circumference
10
12
14
16
18
14
16
18
20
22
Transformed gestational age
subject−specific
population
95% CL pop.

Fitting the model in BUGS

Subject-specific index runs from `i = 1 to 707`

`u[i,1:2] ~ dmnorm((0,0),Omega.beta[,])`

`u[i,1]` ($= u_{0i}$) and `u[i,2]` ($= u_{1i}$) are multivariate normally distributed.

`mu.beta[1]` is the fixed effects intercept β_0.
`mu.beta[2]` is the fixed effects gradient β_1.

`Omega.beta[,]` is the *inverse* of the random effects variance-covariance matrix.

Fitting the model in BUGS

The observation-specific index runs from `j = 1 to 3097`

```
mu[j] <-   (mu.beta[1] + u[id[j],1])
         + (mu.beta[2] + u[id[j],2])*X[j]
```

The observation-specific mean `mu[j]` uses `id[j]` as the first index of the random effect array `u[1:707,1:2]` to pick up the `id` number of the j^{th} observation and select the correct row of the random effect array.

The final step adds the error term to the mean:

`Y[j] ~ dnorm(mu[j],tau.e)`

Prior distributions for model parameters

We use the "standard" noninformative prior distributions for the fixed effects β_0, β_1 and σ_e.

We use a *Wishart* prior distribution for the precision matrix `Omega.beta`, which requires

- A degrees of freedom parameter (we use the smallest possible value, 2, the rank of the precision matrix).
- A scale matrix representing our prior guess at the order of magnitude of the covariance matrix of the random effects (we assume $\sigma_0^2 = 0.5$, $\sigma_1^2 = 0.1$ and $\sigma_{01} = 0$.)

Analysis twin data and GLMMs

Thursday 14th August 2008, afternoon

Lyle Gurrin

Bayesian Data Analysis
11 – 15 August 2008, Copenhagen

This lecture

- A model for paired data.
- Extend this to a genetic model for twin data.
- Fitting these models in BUGS.
- An introduction to the example on mammographic (breast) density.

Paired data

Paired data provide an excellent example of Bayesian methods in BUGS since these data are

- the simplest type of correlated data structure;
- naturally represented as a hierarchical model.

We begin with the basic model to capture correlation in paired data, and then extend this to accommodate different correlation in monozygous (MZ) and dizygous (DZ) twin pairs.

Notation for paired data

Suppose we have a continuously valued outcome y measured in each individual of n twin pairs.

Let y_{ij} denote the measurement of the j^{th} individual in the i^{th} twin pair, where $j = 1, 2$ and $i = 1, 2, \ldots, n$.

We do not consider additional measurements on exposure variables at this stage.

A model for a single pair

We assume that, for the i^{th} pair, that y_{i1} and y_{i2} have common mean

$$\mathrm{E}(y_{i1}) = \mathrm{E}(y_{i2}) = a_i,$$

where (for now) a_i is assumed to be fixed.

It is also assumed that the two measurements have common variance

$$\mathrm{var}(y_{i1}) = \mathrm{var}(y_{i2}) = \sigma_e^2.$$

Conditional on the value of a_i, y_{i1} and y_{i2} are uncorrelated: $\mathrm{cov}(y_{i1}, y_{i2}) = 0$.

A model for a single pair

We can write this model as

$$\begin{aligned} y_{i1} &= a_i + \varepsilon_{i1} \\ y_{i2} &= a_i + \varepsilon_{i2} \end{aligned}$$

where

$$\mathrm{var}(\varepsilon_{i1}) = \mathrm{var}(\varepsilon_{i2}) = \sigma_e^2$$

and

$$\mathrm{cov}(\varepsilon_{i1}, \varepsilon_{i2}) = 0,$$

with an implicit assumption that

$$\varepsilon_{ij} \sim \mathsf{N}(0, \sigma_e^2).$$

A hierarchical model for paired data

We can extend this simple structure for paired data to a hierarchical model by assuming a normal population distribution for the pair-specific means a_i

$$a_i \sim \mathsf{N}(\mu, \sigma_a^2).$$

A hierarchical model for paired data

This is an example of the hierarchical normal-normal model studied in lectures earlier.

The sampling model is (for $j = 1, 2$)

$$y_{ij}|a_i, \sigma_e^2 \sim \mathsf{N}(a_i, \sigma_e^2)$$

The pair-specific mean model is (for $j = 1, 2$)

$$a_i|\mu, \sigma_a^2 \sim \mathsf{N}(\mu, \sigma_a^2)$$

Unconditional mean and variance of y_{ij}

We can use the iterative expectation and variance formulae:

$$\begin{aligned}
\mathrm{E}(y_{ij}) &= \mathrm{E}(\mathrm{E}(y_{ij}|a_i)) \\
&= \mathrm{E}(a_i) \\
&= \mu. \\
\mathrm{var}(y_{ij}) &= \mathrm{var}(\mathrm{E}(y_{ij}|a_i)) + \mathrm{E}(\mathrm{var}(y_{ij}|a_i)) \\
&= \mathrm{var}(a_i) + \mathrm{E}(\sigma_e^2) \\
&= \sigma_a^2 + \sigma_e^2.
\end{aligned}$$

So $y_{ij} \sim \mathsf{N}(\mu, \sigma_a^2 + \sigma_e^2)$.

Bivariate distribution of (y_{i1}, y_{i2})

In fact the joint distribution of y_{i1} and y_{i2} is bivariate normal:

$$\begin{pmatrix} y_{i1} \\ y_{i2} \end{pmatrix} = \mathsf{N}_2 \left[\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{bmatrix} \sigma_a^2 + \sigma_e^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 + \sigma_e^2 \end{bmatrix} \right]$$

BUGS does allow vector nodes to have a multivariate normal (MVN) distribution, but it requires careful specification of the parameters defining the MVN distribution.

Introduction to twin data

Measurements on twins are a special case of paired data and provides a natural matched design.

Outcomes (continuous or binary) are *possibly* influenced by both genetic and environmental factors.

There has been extensive development of statistical methods to deal with quantitative traits, concordance/discordance, affected sib-pairs etc.

Studying basic models for twins provide a good introduction to analysing family data.

Introduction to twin data

We will now extend the simple paired model to accommodate monozygous (MZ) and dizygous (DZ) pairs.

MZ twins are "identical" and share all of their genes.

DZ twins are siblings and share on average half of their genes (same relationship as siblings like brother and sister).

Within-pair covariation in twin data

If a quantitative trait (continuous outcome variable) is under the influence of *shared* genetic factors then we expect the within-pair covariation to be smaller in DZ pairs than in MZ pairs:

$$\text{cov}_{DZ}(y_{i1}, y_{i2}) = \rho_{DZ:MZ}\text{cov}_{MZ}(y_{i1}, y_{i2})$$

So $\rho_{DZ:MZ}$ is the ratio of covariances between DZ and MZ pairs. We assume that $\text{var}(y_{ij})$ is the same in MZ and DZ twins and so $\rho_{DZ:MZ}$ is also the ratio of the within-pair correlation in DZ and MZ pairs.

Interpretation of $\rho_{DZ:MZ}$

- If $\rho_{DZ:MZ} = 1$ then the outcome is no more correlated between individuals in MZ pairs than in DZ pairs, so no evidence of genetic influence.
- If $\rho_{DZ:MZ} = \frac{1}{2}$ then we have an *additive* genetic model, also known as the "classical twin model".
- If $0 < \rho_{DZ:MZ} < 1$ and $\rho_{DZ:MZ} \neq \frac{1}{2}$ then any genetic model will need to be non-additive (*eg* gene-gene interaction) or incorporate a contribution to variation from shared environment.

Specifying the additive model in BUGS

Recall our original model:

$$\begin{aligned} y_{i1} &= a_i + \varepsilon_{i1} \\ y_{i2} &= a_i + \varepsilon_{i2} \end{aligned}$$

where

$$\begin{aligned} \text{var}(\varepsilon_{ij}) &= \sigma_e^2 \\ \text{cov}(\varepsilon_{i1}, \varepsilon_{i2}) &= 0 \\ a_i &\sim \text{N}(\mu, \sigma_a^2) \end{aligned}$$

Random effect sharing in MZ and DZ pairs

Now instead of just one random effect a_i per pair, extend the model to incorporate three random effects per pair a_{i1}, a_{i2} and a_{i3}, all i.i.d. N(μ, σ_a^2).

Regardless of zygosity both individuals in a twin-pair share the first random effect a_{i1}.

Individuals in MZ pairs share the second random effect a_{i2}.

For DZ pairs, one member of the pair receives a_{i2} and the other member receives a_{i3}.

So DZ pairs share less than MZ pairs. Scaling appropriately by $\rho = \rho_{DZ:MZ}$ we have...

Model equations

For MZ pairs

$$\begin{aligned} y_{i1} &= \sqrt{\rho}\, a_{i1} + \sqrt{1-\rho}\, a_{i2} + \varepsilon_{i1} \\ y_{i2} &= \sqrt{\rho}\, a_{i1} + \sqrt{1-\rho}\, a_{i2} + \varepsilon_{i2} \end{aligned}$$

For DZ pairs

$$\begin{aligned} y_{i1} &= \sqrt{\rho}\, a_{i1} + \sqrt{1-\rho}\, a_{i2} + \varepsilon_{i1} \\ y_{i2} &= \sqrt{\rho}\, a_{i1} + \sqrt{1-\rho}\, a_{i3} + \varepsilon_{i2} \end{aligned}$$

Covariance and variance for MZ and DZ pairs

Since MZ pairs share both a_{i1} and a_{i2}, the within-pair covariance of y_{i1} and y_{i2} is $\sqrt{\rho}^2\sigma_a^2 + \sqrt{1-\rho}^2\sigma_a^2 = \sigma_a^2$.

DZ pairs share only a_{i1} so the corresponding within-pair covariance is $\sqrt{\rho}^2\sigma_a^2 = \rho\sigma_a^2$.

The variance of y_{ij} is, however, $\sigma_a^2 + \sigma_e^2$ for both MZ and DZ pairs.

Model equations

Clearly the model for MZ pairs

$$\begin{aligned} y_{i1} &= \sqrt{\rho}\, a_{i1} + \sqrt{1-\rho}\, a_{i2} + \varepsilon_{i1} \\ y_{i2} &= \sqrt{\rho}\, a_{i1} + \sqrt{1-\rho}\, a_{i2} + \varepsilon_{i2} \end{aligned}$$

is identical to the original model

$$\begin{aligned} y_{i1} &= a_{i1} + \varepsilon_{i1} \\ y_{i2} &= a_{i1} + \varepsilon_{i2} \end{aligned}$$

but computations in BUGS are easier if both members of the pair share some random effects but not others if they are DZ.

Application: Mammographic density

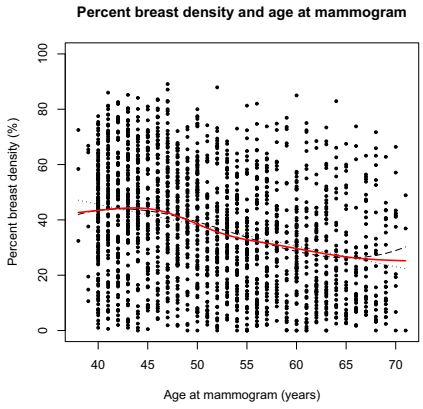
Female twin pairs (571 MZ 380 DZ) aged 40 to 70 were recruited in Australia and North America.

The outcome is percent mammographic density, the ratio of dense tissue to non-dense tissue from mammographic scan.

Age-adjusted mammographic density is a risk factor for breast cancer.

The correlation for percent density was 0.63 in MZ pairs and 0.27 in DZ pairs adjusted for age and location.

Other risk factors for breast density: height, weight, reproductive, diet.



Generalised Linear Mixed Models

Melbourne Sexual Health Clinic (MSHC)

consults	PID	warts
80	1	1
816	41	46
726	12	37
2891	38	137
79	4	4
1876	34	73
469	8	27
1124	13	76
210	10	6
539	8	28
1950	22	101
1697	24	86
811	13	56
908	52	48
944	19	65
832	10	33
1482	10	62
456	0	20
420	0	8
1258	3	58
1101	1	22
109	0	3
1006	2	62

23 sexual health physicians diagnosing each patient with either:

- pelvic inflammatory disease (PID)
- genital warts

Are there differences between physicians in the proportion diagnosed with PID or warts?

Diagnosis frequency in MSHC

consults	PID	warts
80	1.25	1.25
816	5.02	5.64
726	1.65	5.10
2891	1.31	4.74
79	5.06	5.06
1876	1.81	3.89
469	1.71	5.76
1124	1.16	6.76
210	4.76	2.86
539	1.48	5.19
1950	1.13	5.18
1697	1.41	5.07
811	1.60	6.91
908	5.73	5.29
944	2.01	6.89
832	1.20	3.97
1482	0.67	4.18
456	0.00	4.39
420	0.00	1.90
1258	0.24	4.61
1101	0.09	2.00
109	0.00	2.75
1006	0.20	6.16

The proportion of patients diagnosed varies:

- PID 0% – 5.73%
  1.49% (weighted)
  1.72% (unweighted)
- Warts 1.25% – 6.91%
  4.86% (weighted)
  4.59% (unweighted)

Are the warts percentages "better spread" than the PID percentages (most less 2% but four are around 5%)?

Data Structure

For each outcome (PID or warts) the data are of the form

$$(n_i, y_i); \qquad i = 1, \ldots, 23$$

where

n_i is the number of consultations (patients seen) by physician i.

y_i is the number of patients diagnosed with the condition.

Sampling model at each dose level

For each physician, the patients (i.e. their outcomes) are assumed to be exchangeable (there is no information to distinguish one patient from another).

We model the outcomes within-physician as independent given a physician-specific probability of death θ_i, which leads to the familiar binomial sampling model:

$$y_i|\theta_i \sim \text{Bin}(n_i, \theta_i)$$

Setting up a model across physicians

Typical assumption is that each θ_i is an independent parameter, that is, a fixed effect.

We can re-express this model using *logistic regression*:

$$\text{logit}(\theta_i) = \log(\theta_i/(1-\theta_i)) = \alpha_i$$

where α_i is the physician-specific log odds of diagnosis with the condition.

Estimating the fixed effects

MLE: Estimates are:

$$\begin{aligned}\hat{\theta}_i &= y_i/n_i \\ \hat{\alpha}_i &= \log(y_i/(n_i - y_i))\end{aligned}$$

Bayes: Estimate θ as the posterior mean using some (beta?) prior distribution (see Lecture 1).

An even more extreme assumption is that $\alpha_i = \alpha$ for some common log-odds of diagnosis α.

Also this provides no way of quantifying variability in frequency of diagnosis - is there a compromise?

A hierarchical model

Replace independence with exchangeability - allow the α_i to be drawn from a "population" distribution of diagnosis frequencies:

$$\alpha_i \sim \mathsf{N}(\mu, \tau^2).$$

Then assume the y_i are *conditionally* independent binomial random variables given the α_i:

$$y_i|\alpha_i \sim \mathrm{Bin}(n_i, \mathsf{expit}(\alpha_i))$$

where

$$\mathsf{expit}(\alpha_i) = \mathsf{logit}^{-1}(\alpha_i) = \exp(\alpha_i)/(1+\exp(\alpha_i)) = \theta_i.$$

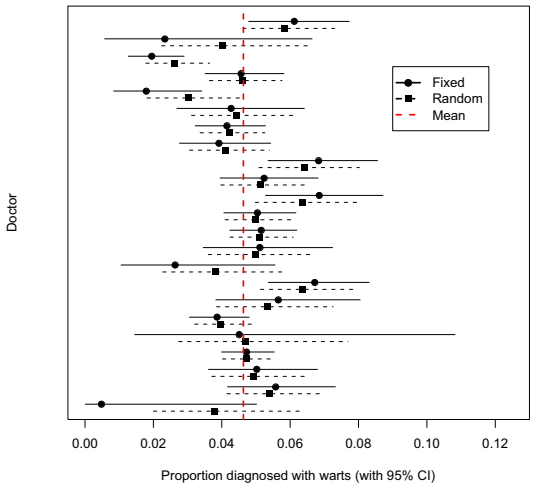
Logistic-normal GLMM

This logistic-normal model, where the α_i's are given a distribution, and so are *random* rather than *fixed* effects, is an example of a *generalised linear mixed model* (GLMM).

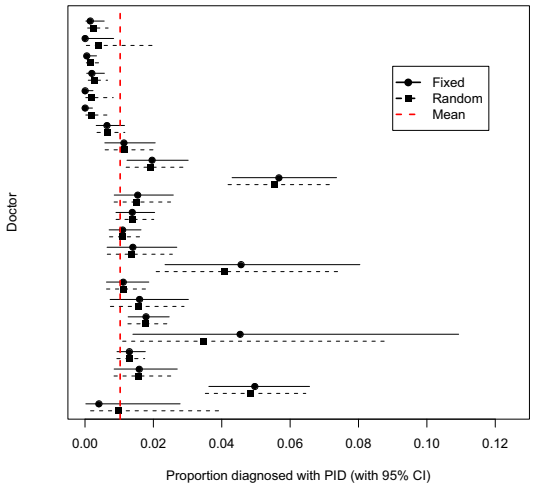
The model is easily implemented in `BUGS` to generate posterior distributions for α_i's and hyperparameters μ and τ.

This simple model is available in Stata and SAS but without much flexibility to extend to random coefficients.

Fixed and random effects for warts



Fixed and random effects for PID



Summary output from BUGS

Warts:

	mean	sd	2.5%	50%	97.5%	Rhat	n.eff
mu	-3.02	0.08	-3.19	-3.02	-2.87	1	5700
tau2	0.10	0.06	0.03	0.09	0.24	1	15000
tau	0.31	0.08	0.18	0.30	0.49	1	15000

PID:

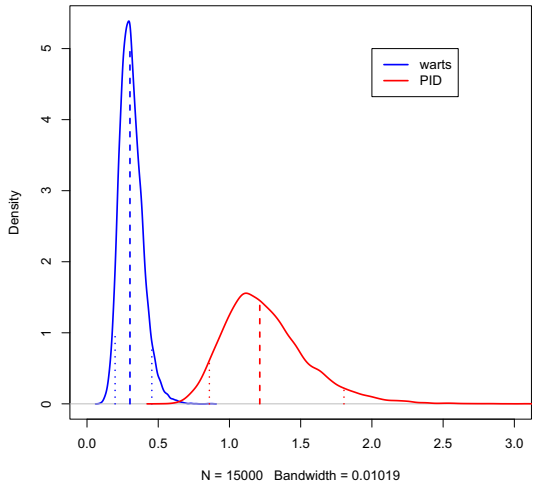
	mean	sd	2.5%	50%	97.5%	Rhat	n.eff
mu	-4.57	0.32	-5.27	-4.55	-3.98	1	1000
tau2	1.67	0.84	0.65	1.47	3.82	1	15000
tau	1.26	0.29	0.81	1.21	1.95	1	15000

Summary output from BUGS

PID after removing those physicians with PID frequency $> 2.5\%$:

	mean	sd	2.5%	50%	97.5%	Rhat	n.eff
mu	-4.86	0.27	-5.50	-4.83	-4.39	1	4500
tau2	1.07	0.70	0.29	0.90	2.93	1	4100
tau	0.99	0.30	0.54	0.95	1.71	1	4100

Posterior densities for τ



Penalised loss functions for model comparison and the DIC

Friday 15th August 2008, morning

Lyle Gurrin

Bayesian Data Analysis
11 – 15 August 2008, Copenhagen

The need to compare models

Model choice is an important part of data analysis.

This lecture: Present the Deviance Information Criterion (DIC) and relate it to a cross-validation procedure for model checking.

When should we use the DIC?

Posterior predictive checking

One approach to Bayesian model checking is based on hypothetical replicates of the same process that generated the data.

In posterior predictive checking (Chapter 6 of BDA), replicate datasets are simulated using draws from the posterior distribution of the model parameters.

The adequacy of the model is assessed by the faithfulness with which the replicates reproduce key features of the original data.

Two independent datasets

Suppose that we have training data $\mathbf{Z}$ and test data $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_n\}$.

We assess model adequacy using a loss function $L(\mathbf{Y}, \mathbf{Z})$ which measures the ability to make prediction of $\mathbf{Y}$ from $\mathbf{Z}$.

Suitable loss functions are derived from *scoring rules* which measure the utility of a probabilistic forecast of $\mathbf{Y}$ represented by a probability distribution $p(\mathbf{y})$.

The log scoring rule

One sensible scoring rule is the log scoring rule

$$A \log \{(p(y))\} + B(y)$$

for essentially arbitrary constant A and function $B(.)$ of the data $\mathbf{Y}$.

Parametric models

Consider the situation where all candidate models share a common vector of parameters $\boldsymbol{\theta}$ (called the *focus*) and differ only in the (prior) structure for $\boldsymbol{\theta}$.

Assume also that $\mathbf{Y}$ and $\mathbf{Z}$ are conditionally independent given $\boldsymbol{\theta}$, so that $p(\mathbf{Y}|\boldsymbol{\theta}, \mathbf{Z}) = p(\mathbf{Y}|\boldsymbol{\theta})$.

The log scoring rule then becomes the log likelihood of the data $\mathbf{Y}$ as a function of $\boldsymbol{\theta}$, or equivalently the *deviance* $-2 \log\{p(\mathbf{Y}|\boldsymbol{\theta})\}$.

Loss functions

Two suggested loss functions are the "plug-in" deviance

$$L^{\mathrm{p}}(\mathbf{Y}, \mathbf{Z}) = -2 \log[p\{\mathbf{Y}|\overline{\boldsymbol{\theta}}(\mathbf{Z})\}]$$

where $\overline{\boldsymbol{\theta}}(\mathbf{Z}) = E(\boldsymbol{\theta}|\mathbf{Z})$ and the expected deviance

$$L^{\mathrm{e}}(\mathbf{Y}, \mathbf{Z}) = -2 \int \log\{p(\mathbf{Y}|\boldsymbol{\theta})\} p(\boldsymbol{\theta}|\mathbf{Z}) d\boldsymbol{\theta}$$

where the expectation is taken over the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{Z}$ with $\mathbf{Y}$ considered fixed.

Loss functions

Both loss functions are derived from the deviance but there are important difference between L^{p} and L^{e}.

- The plug-in deviance is sensitive to reparametrisation but the expected deviance is co-ordinate free.
- The plug-in deviance gives equal loss to all models that yield the same posterior expectation of $\boldsymbol{\theta}$, regardless of precision.
- The expected deviance is a function of the full posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{Z}$ so takes precision into account.

The problem

How to proceed when there are no training data?

An obvious idea is to use the test data to estimate $\boldsymbol{\theta}$ and assess the fit of the model, that is, re-use the data to create the loss function $L(\mathbf{Y}, \mathbf{Y})$.

But this is optimistic. "In-sample" prediction will always do better than "out-of-sample" prediction.

Quantifying the optimism

We can get an idea of the degree of optimism for loss functions that decompose into a sum of contributions from each Y_i (e.g. independence)

$$L(\mathbf{Y}, \mathbf{Z}) = \sum_{i=1}^{n} L(Y_i, \mathbf{Z})$$

We can gauge the optimism by comparing $L(Y_i, \mathbf{Y})$ with $L(Y_i, \mathbf{Y}_{-\mathbf{i}})$ where $\mathbf{Y}_{-\mathbf{i}}$ is the data with Y_i removed.

Quantifying the optimism

The expected decrease in loss from using $L(Y_i, \mathbf{Y})$ in place of $L(Y_i, \mathbf{Y}_{-\mathbf{i}})$ is

$$p_{\text{opt}_i} = E\{L(Y_i, \mathbf{Y}_{-\mathbf{i}}) - L(Y_i, \mathbf{Y})|\mathbf{Y}_{-\mathbf{i}}\}$$

which is the *optimism* of $L(Y_i, \mathbf{Y})$. The loss function

$$L(Y_i, \mathbf{Y}) + p_{\text{opt}_i}$$

has the same expectation given $\mathbf{Y}_{-\mathbf{i}}$ as the cross-validation loss $L(Y_i, \mathbf{Y}_{-\mathbf{i}})$ so is equivalent for an observer who has not seen Y_i.

Penalised loss function

The same argument applies to each Y_i in turn.

Proposal: Use the sum of the penalised loss functions

$$L(\mathbf{Y}, \mathbf{Y}) + p_{\text{opt}}$$

to assess model accuracy where

$$p_{\text{opt}} = \sum_{i=1}^{n} p_{\text{opt}_i}$$

is the cost of using the data twice.

Deviance Information Criterion (DIC)

The DIC is defined as

$$\text{DIC} = \overline{D} + p_D$$

where $\overline{D} = E(D|\mathbf{Y})$ is a measure of model fit and p_D is the "effective number of parameters", a measure of model complexity defined by

$$p_D = \overline{D} - D\{\overline{\boldsymbol{\theta}}(\mathbf{Y})\}$$

The effective number of parameters

Using previous notation, $\overline{D} = L^{\mathsf{e}}(\mathbf{Y}, \mathbf{Y})$ and $D\{\overline{\boldsymbol{\theta}}(\mathbf{Y})\} = L^{\mathsf{p}}(\mathbf{Y}, \mathbf{Y})$, so p_D can be written as

$$p_D = L^{\mathsf{e}}(\mathbf{Y}, \mathbf{Y}) - L^{\mathsf{p}}(\mathbf{Y}, \mathbf{Y})$$

p_D can be decomposed into the sum of individual contributions

$$p_D = \sum_{i=1}^{n} p_{D_i} = \sum_{i=1}^{n} L^{\mathsf{e}}(Y_i, \mathbf{Y}) - L^{\mathsf{p}}(Y_i, \mathbf{Y})$$

The normal linear model

We can do the algebra explicitly for the normal linear model that assumes the Y_i are scalar.

Both the expected and plug-in deviance have a penalty term like $p_{D_i}/(1 - p_{D_i})$.

But the large sample behaviour depends on the dimension of $\boldsymbol{\theta}$.

The normal linear model

If the dimension of $\boldsymbol{\theta}$ is fixed then p_{D_i} is $O(n^{-1})$ so the penalised losses can be written as

$$\overline{D} + kp_D + O(n^{-1})$$

where $k = 1$ for the plug-in deviance and $k = 2$ for the expected deviance. So the penalised plug-in deviance is the same as the DIC for regular linear models with scalar outcomes.

Example: DIC in hierarchical models

In random effects models, however, it is quite common for the dimension of $\boldsymbol{\theta}$ to increase with n.

The behaviour of the penalised plug-in deviance may be different from DIC.

We illustrate this with the normal-normal hierarchical model, also known as the one-way random-effects analysis of variance (ANOVA).

The Normal-normal hierarchical model

The two-level hierarchical model assumes that:

$$\begin{aligned} Y_i|\theta_i &\sim \mathsf{N}(\theta_i, \sigma_i^2), \\ \theta_i|\mu,\tau &\sim \mathsf{N}(\mu, \tau^2) \end{aligned}$$

where the variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ are fixed and μ and τ are given noninformative priors.

The Normal-normal hierarchical model

Assume candidate models are indexed by τ, and the deviance is defined as

$$D(\boldsymbol{\theta}) = \sum_{i=1}^{n} [(y_i - \theta_i)/\sigma_i]^2$$

It can be shown that the contribution to the effective number of parameters from observation i is

$$p_{D_i} = \rho_i + \frac{\rho_i(1-\rho_i)}{\sum_{j=1}^{n}\rho_j}$$

where $\rho_i = \tau^2/(\tau^2 + \sigma_i^2)$ is the intra-class correlation coefficient.

In the limit as $\tau^2 \to 0$

There are two limiting situations.

In the limit as $\tau^2 \to 0$, the ANOVA model tends to a pooled model where all observations have the same prior mean μ. In this limit, $p_D = 1$ and both the DIC and penalised plug-in deviance are equal to

$$\sum_{j=1}^{n} [(Y_j - \overline{Y})/\sigma_j]^2 + 2 \tag{45}$$

where

$$\overline{Y} = \frac{\sum_{j=1}^{n} Y_j/\sigma_j^2}{\sum_{j=1}^{n} 1/\sigma_j^2}. \tag{46}$$

In the limit as $\tau^2 \to \infty$

In the limit as $\tau^2 \to \infty$, the ANOVA model tends to a saturated fixed-effects model, in which $\mathbf{Y_{-i}}$ contains no information about the mean of Y_i. In this limit, $p_D = n$ and DIC $= 2n$, but the penalised plug-in deviance tends to infinity.

So there is strong disagreement between DIC and the penalised plug-in deviance in this case.

Conclusion

For linear models with scalar outcomes, DIC is a good approximation to the penalised plug-in deviance whenever $p_{D_i} \ll 1$ for all i, which implies $p_D \ll n$.

So p_D/n may be used as an indicator of the validity of DIC in such models.

Further reading

Plummer M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, 1–17.

Spiegelhalter D, Best N, Carlin B, van der Linde A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B* **64**, 583–639.

WinBUGS website, which includes David Spiegelhalter's slides from the "IceBUGS" workshop in 2006.

Comparison of two methods of oximetry measurement

Friday 15th August 2008, afternoon

Bendix Carstensen

Bayesian Data Analysis
11 – 15 August 2008, Copenhagen

Comparing methods of measurement

We may wish to compare two different methods for measuring the same underlying (true) value.

Example 1: Lung function measurements using a *spirometer* (expensive but accurate) and a *peak flow meter* (cheap but less accurate).

Example 2: Standard or traditional method of measurement compared to a new method based on recent technology.

Two methods for oxygen saturation

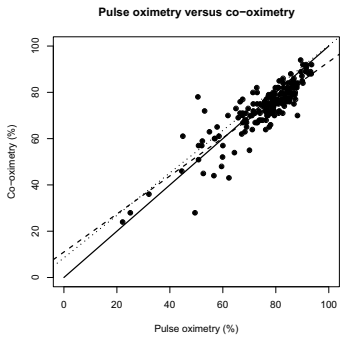
infant	co.ox	pulse.ox
1	78	71
1	76.4	72
1	77.2	73
2	68.7	68
2	67.6	67
2	68.3	68
3	82.9	82
3	80.1	77
3	80.7	77
4	62.3	43
4	65.8	69
4	67.5	77
5	75.8	76
5	73.7	72
5	76.3	68
6	78	79
6	78.8	78
6	77.3	78

177 oximetry measures (about three per infant):

- Co-oximetry using biochemistry
- Pulse oximetry using light reflectance

Is there a difference between the two methods?

Two methods for oxygen saturation



Identity line.

Regression lines:

pulse.ox on **co.ox**

co.ox on **pulse.ox**

Comparing two methods of measurement

The two methods (pulse oximetry and co-oximetry) are clearly strongly related - their correlation coefficient is 0.87.

Correlation quantifies the *association* between the measurements and not the *agreement* between them.

The correlation

- would still be high even if one method was systematically in error;
- will depend on the range of the true quantity in the sample.

Method comparison and regression?

Similarly, there is not much point in considering the hypothesis that $\beta_0 = 0$ (intercept) and $\beta_1 = 1$ (slope).

Any attempt at modelling should take into account the fact that both methods have measurement errors.

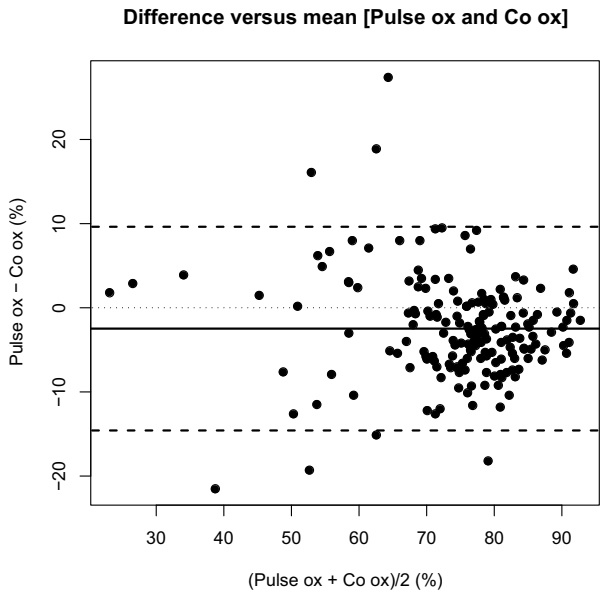
Evidence for linear relationship is of little use if there is large random variation between the methods.

Statistical measures of agreement

Bland and Altman (*Lancet* (1986) **1** 307-310) suggested that the extent of agreement could be examined by plotting the difference between pairs of measurements on the vertical axis against the mean of each pair on the horizontal axis.

If one method is known to be accurate then the mean difference will indicate systematic bias.

Avoid plotting the difference against either of the individual measurements due to the problem of *regression to the mean*.



Distribution of the difference

We work with a model for the sampling distribution of the difference, and combine this with a prior distribution for the *mean* difference.

Give a nominal 95% posterior prediction interval for the difference (incorporating uncertainty in both the mean and variance parameter):

Mean difference $\pm$ 2 $\times$ std dev(difference)

This is often called the *limits of agreement* and is based on the usual assumption that the difference are approximately normally distributed.

General conclusions

More detailed analyses would look to see if the differences vary systematically with the "level" (ie mean value) of the measurements.

Even if the mean difference is very close to zero (statistically and "clinically") the variation may still be large. Once we have the limits of agreement it is up to the investigators to decide how to proceed.

Motivation for the Bayesian approach

1. Capture explicitly uncertainty in the parameters governing our measurement comparison, in particular the mean difference δ and the measurement error σ_e^2.
2. Represent the hierarchical structure of the data using a (population) distribution of random effects at each level of the hierarchy.
3. Easy to incorporate extensions:
   - Method-specific residual variances.
   - Bias that depends on the level of measurement.

Structure of the oximetry data

Three level hierarchy:

1. Individual infant (subject-specific random effect representing their level of oxygen saturation).
2. Three separate measurement occasions (occasion-specific random effect representing the oxygen saturation on that occasion).
3. Two methods of measurement, pulse- and co-oximetry (structural term for "bias" as well as random error).

Prior distributions

We need prior distributions for the parameters representing the mean difference between the two methods of measurement, as well as the variance components representing

1. Subject-specific effects: σ_u^2
2. Measurement occasion: σ_v^2
3. Random error: σ_e^2

Further reading

Carstensen B. Comparing and predicting between several methods of measurement. *Biostatistics* (2004) **5**, 399–413.

Carstensen B, Simpson JA, Gurrin LC. Practical Aspects of Method Comparison Studies *Int. J. Biostat.* (2008)

Bland JM, Altman DG. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* (1999), **8**, 136–160.