Bayesian Simulation Routines

By Wendy Olsen, July 2021

Contents of this Presentation

- Preliminary readings and informal sources
- Notes on Bayesian Estimation
- The Alternative Methods of MCMC
- Commands and Results with Quick Quizzes from a Child Labour Example Using GLMER and BRMS and STAN in R
- References

Chapters to read as introductory material:

- Gelman and Hill, 2007, chapters 16 and 18.
- Your prep for chapter 16 is to read chapter 15 on multilevel models too.

Gelman and Hill, 2007, *Data Analysis Using Regression and Multilevel/Hierarchical Models,* Cambridge University Press (Basic introduction to how simulation would solve a linear regression model, chapters 7-9 only; Multilevel models are around chapter 15.)

- Cowles' entire book (2015) is an alternative source of the same material.

- Albert, Jim (2009) *Bayesian Computation With R*. London: Springer. **This book is somewhat iconoclastic but has fun R code for you.**
- Cowles, Mary (2015*). Applied Bayesian Statistics*. 1st ed. Springer. **Mary's book introduces the Bayesian theory and is suitable for undergraduate students.**

Online preparatory sources

- A classic journal article on the method of Markov chain Monte carlo simulations:

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics & Computing*, **10**, 325–337.

- And lastly a long slide set to work through in your own time (6 hours' work)

Bayesian Hierarchical Modelling using WinBUGS, an advanced online training course, by Nicky Best, Alexina Mason, and Philip Li, 2011, http://www.bias-project.org.uk/WB2011Man/BHM-2011-slides.pdf. Contains Introduction to Bayesian Hierarchical Models, Priors, Model Criticism and Comparison, Longitudinal Models, and Cross-Classified Model Construction

Exemplars for MCMC

- Raymer, James, & Arkadiusz Wiśniowski (2018) Applying and testing a forecasting model for age and sex patterns of immigration and emigration, Population Studies, 72:3, 339-355, DOI: 10.1080/00324728.2018.1469784

Country data: Sweden, South Korea, and Australia, model based on UK, data sources multiple even for a single flow.

Chapter 11 of Demographic Forecasting by Bryant and Zhang, and tutorial corresponding to that. Olsen's tutorial guide helps you and you could ask me for 'tutorial 4' of our coaching series.

One Exemplar for Today's Q&A

- Suppose children report their time-use in a survey, and the children's reports for ages 5 to 16 are gathered into a matrix. Time thresholds are set by age group, 0 for 5-11 years, 24 for 12-14, and 34 hours per week for age 15-16.
- A binary outcome is set up 0 = not in harmful excessive-hours child labour, 1= harmful child labour.
- The child labour in market work includes farming, family helper in enterprise, and producing goods likefish at home.
- Using classic ILO terminology, based on the UN system, child labour in domestic work has to be counted outside productive work. This includes home based childcare services, cooking and washing.

Child Labour Logit Exemplar

- Factors affecting children getting into harmful child labour include:
  - Oversize household with an internal division of labour;
  - Having a poor mother who lacks a husband present (separated, deserted or divorced);
  - Cultural minority groups;
  - Being low in assets [not recorded in the data used here];
  - Growing toward maturity;
  - Others.

  - We do not expect gender of child to matter, but it does: boys in south Asia get more of the work that is ‘market work’ so girls are ‘less at risk’ = gender stereotyping of earning at a very young age.

Child Labour Logit Multilevel Model in GLMER

- `glmer(childlab ~ female + hhsize+ rural + fhh + (1 | country) + (1 | ageyear), mysamp, family = binomial("logit"))`
- ## Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [glmerMod]
  ## Family: binomial (logit) Formula: childlab ~ female + hhsize + rural + fhh + (1 | country) + (1 | ageyear)
  ## Data: mysamp
- ## AIC BIC logLik deviance df.resid
  ## 2933.284 2983.757 -1459.642 2919.284 9993
  ## Random effects:
  ## Groups Name Std.Dev.
  ## ageyear (Intercept) 0.7104
  ## country (Intercept) 1.0561
  ## Number of obs: 10000, groups: ageyear, 13; country, 6
  ## Fixed Effects:
  ## (Intercept) female hhsize rural fhh
  ## -3.32666 -0.42979 0.02289 0.27317 0.19181

Quick Quiz

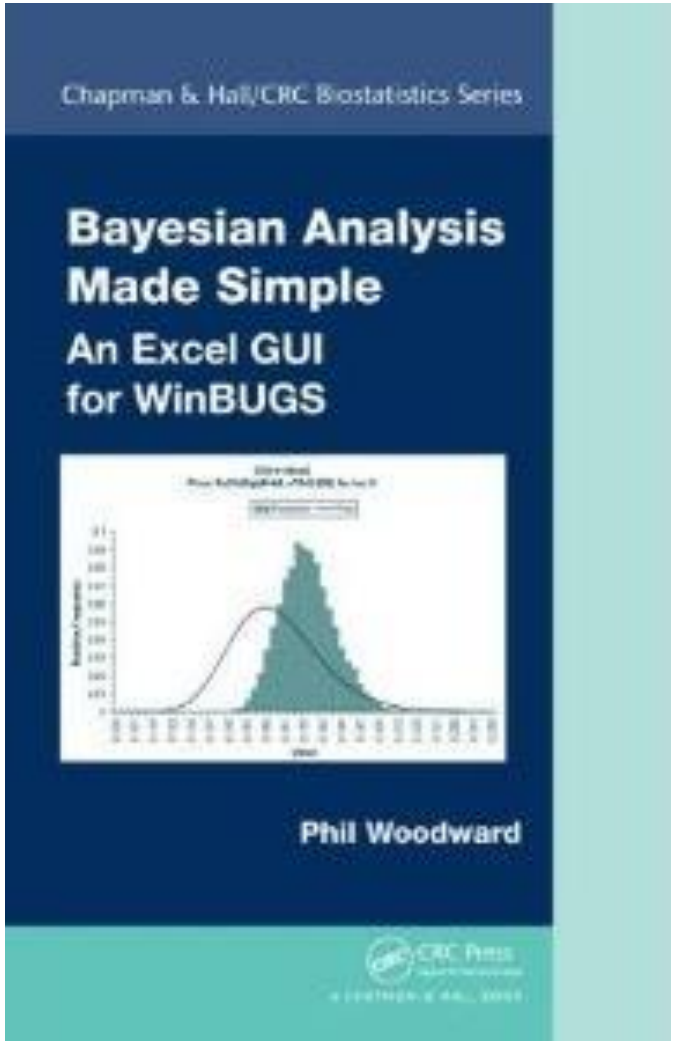
- What are the glmer commands for getting the confidence interval boundaries in a table format? Hint a list can be subsetted using $... So you name your results as fitglmer then str(fitglmer) and lastly *** command *** fitglmer$&&&& [name the correct item in the list]

- What is the loglik used for?

- What is the item name within the ‘fit’ object that holds the Level 2 random intercepts for age? How do you list these?

Another reading, if you get time: Using Excel!!

- Glance at this simple book if you wish:



Child Labour Logit Multilevel Model in Stan

- We use package brms to run rstan for a m.l. model.
- The 2nd level here is Country. It can be a factor, but even as a character mode variable it is set up as a factor in brms via the 'forcing' or 'coercion' that occurs when we express (1 | country) for the random intercepts.

- Command

```
fitbrms0 <- brm(childlab ~ 1+ (1|country), data=mysamp,
                family = bernoulli(link="logit"),
                chains = 4, cores = 4, inits="0")
```

```
## Compiling Stan program...
```

```
## Start sampling
```

```
mcmc_plot(fitbrms0)
```

```
 or fitbrms7 <-brm(childlab ~  Gender + hhsize+ minority + rural + fhh +agec+agec2+agec3+  (1 |
country) , bernoulli(link = "logit"), data=mysampallten, chains=4, cores= 4, iter=1200)
```

Child Labour Logit Brms Results

```
##    Data: mysampallten (Number of observations: 10000)
## Samples: 4 chains, each with iter = 1200; warmup = 600; thin = 1;
##          total post-warmup samples = 2400
##
## Group-Level Effects:
## ~country (Number of levels: 6)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     1.35      0.53     0.66     2.75 1.01      613     1161
##
## Population-Level Effects:
##             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept      -2.97      0.57    -4.07    -1.74 1.01      780      943
## GenderFemale   -0.41      0.11    -0.61    -0.21 1.00     2360     1907
## hhsize          0.00      0.02    -0.04     0.03 1.00     2875     1494
## minority        0.29      0.16    -0.04     0.60 1.00     2618     1763
## rural           0.34      0.12     0.11     0.57 1.00     2814     1715
## fhh            -0.27      0.21    -0.70     0.14 1.00     2457     1581
## agec           -0.02      0.04    -0.09     0.05 1.00     2387     1637
## agec2          -0.01      0.01    -0.02    -0.00 1.01     3665     1975
## agec3           0.01      0.00     0.00     0.01 1.00     2546     1728
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

- mcmc_plot(fitbrms7) etc.

Child Labour Logit Brms Quick Quiz

1 Why did I take just 'Number of observations: 10000' rather than 422K ?

2 What advantage to using 4 chains, each with iter = 1200?

3 What is throwing out the warmup session of 600 iterations?

Note: total post-warmup samples = 2400
4 What does the abbreviation CI stand for, below? (BE CAREFUL)_____

```
##                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)      1.35      0.53      0.66      2.75 1.01      613     1161

## Population-Level Effects:
##              Estimate Est.Error l-95% CI u-95% CI Rhat
## fhh  -0.27    0.21  -0.70  0.14  1.00
```

5 Is the effect of Female Household Head, after controlling for age and country and rural+other, different from zero, and please use a sentence?

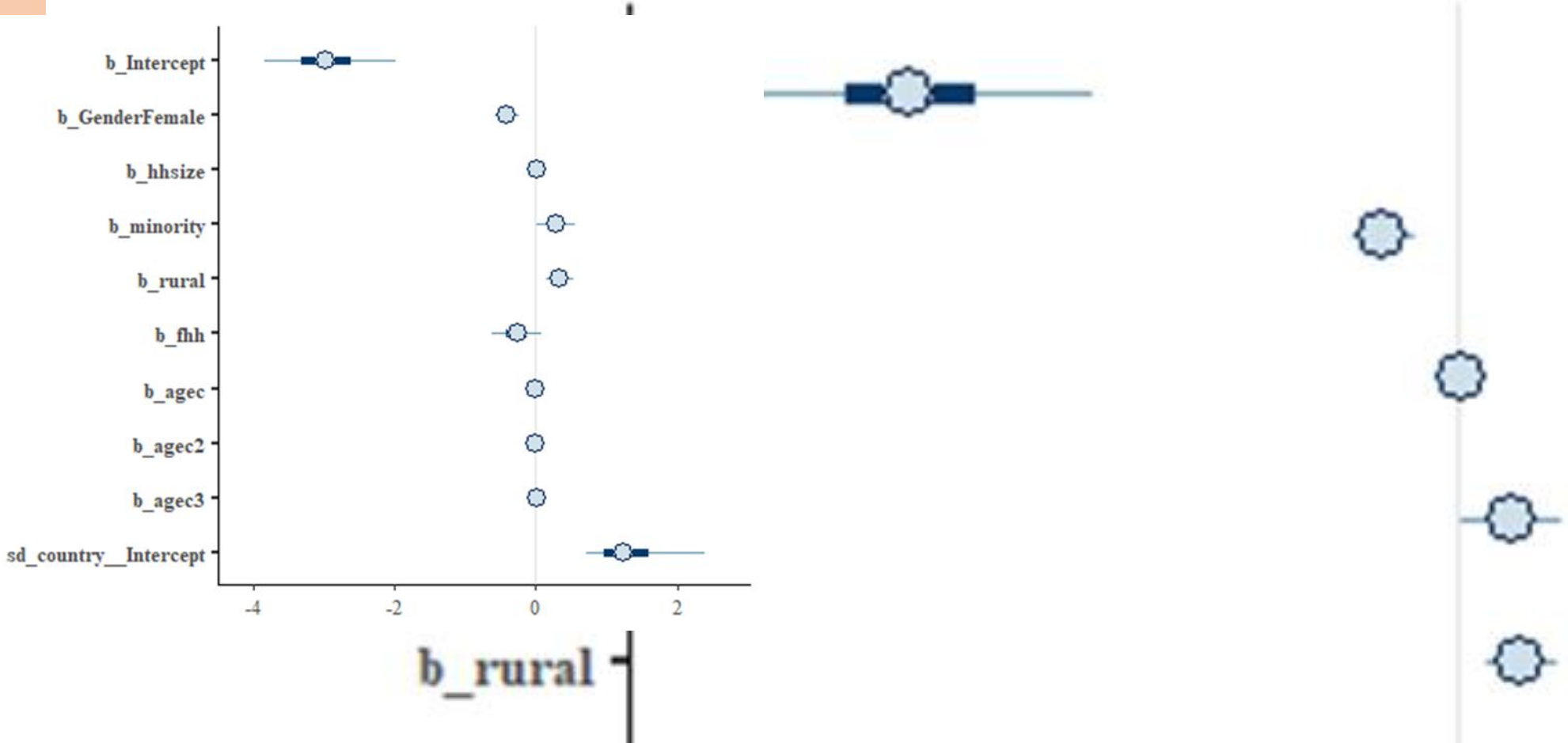
6 ***Why*** were **samples were drawn using sampling(NUTS)? [simply, why did R do that?]**

- mcmc_plot(fitbrms7) etc. 7. If **NUTS** is the sampler, then what is 'mcmc' here?

The coefficients are shown here.

7. Why is sd_country_intercept so important?
8. What do the default brms blue lines mean? It's important to be sure about this if you use the mcmc_plot defaults in brms.

9. Is minority status across six South Asian countries non-zero? >0?



Bayes' Theorem, written for a regression

D=Data matrix, P is probability density function, and theta ($\boldsymbol{\theta}$) is a vector of all parameters.

- $\mathbf{P(\theta|D) = \frac{P(D|\theta)\cdot P(\theta)}{P(D)}}$
- **So therefore, $\mathbf{P(\theta|D) \propto P(D|\theta)\cdot P(\theta)}$** ***The symbol*** $\propto$ **means 'is proportional to'.**
- **And P(θ|reality) ←→P($\hat{\theta}|D$) [Data are a trace of the reality]**
- **And P($\hat{\theta}|D$) $\propto$ Likelihood · Prior [$\hat{\theta}$ is the set of estimates of θ]**

Quiz questions.
1 Is the 'data' a set of unknowns? Yes/no.
2 Is the set of all parameters going to include y-hat (subscript i)? Yes/no.
3 If theta has just 2 parameters, and one is estimated wrongly or badly, will that affect the estimate of the other one? Yes/no.
4 If we imagine iterating around the dataset 10,000 times, (S=samples, with replacement, S=10,000), trying out different theta values, how would we choose which is a better vector θ (Theta)? _____

Another quiz, this one is harder.

$\hat{\theta}$ is the set of estimates of θ

Suppose we write the posterior as:

- **P($\hat{\theta}|\boldsymbol{D}$) $\propto$ Likelihood $\cdot$ Prior**

Then if we take logs, what is the effect of the prior on the log posterior?

& Does the prior have both direct and indirect effects on the log posterior?

We have a likelihood function on the right-hand side. It is a multiplicative function of its terms.

Bayesian Model Comparison

Under the general model, the predictive density of y is given by the integral

$$f(y) = \int \prod_{j=1}^{N} f(y_j|\theta)g(\theta)d\theta.$$

In chapter 18 of Gelman and Hill, 2007, we can see example of likelihood functions for each generalised linear modelling option (the linear model; the binary logistic model; and the Poisson model).

Sources for how MCMC works

- 1. NUTS is the no – U – Turn sampler
- 2 BUGS is the Bayesian Gibbs Sampler
- 3 Jags package in R uses Gibbs Sampler, it's Just Another Gibbs Sampler.
- 4 Stan uses Metropolis Hastings.
- Readings BUGS = Lund et al., Gibbs, read Gelman and Hill, Chapters 15-16; JAGS read the rJags package documentation; Metropolis-Hastings, read Gelman, et al., 2013, 3rd ed (URL free pdf), chapters

NUTS and BUGS

- The BUGS doodle diagram is used to illustrate the hierarchical nature of main parameters versus hyperparameters.
- BUGS was described by Kery and Schaub (2012). In Chapter 2, they explain that a closed-form formula can be used if the Posterior function is derivable from the prior and the likelihood. The algebraic requirement is called conjugacy. As explained in Gelman, et al., 2013, conjugacy is a limiting requirement but is important for grasping how the posterior is scaled.
- We use simulations to estimate the posterior function if we do not have a simpler, algebraic solution involving conjugate distributions.

The No-U-Turn Sampler avoids the ‘snags’ of a model arriving at a cycle that is not the correct parameter set

- See the free online document by Hoffman and Gelman (2011) if you want to see the technical equations of the Hamiltonian Monte Carlo algorithm.
- The Hamiltonian Monte Carlo was a step forward compared with older routines, which could get ‘hung’ or stuck in a cycle [the random walk problem, or inability to escape one zone of the parameter space], with a resulting report of a parameter set θ that was not the best one.
- The use of a ‘momentum variable’ complements the use of the posterior likelihood itself. In each round of simulation, one draw is taken from the prior of one parameter within the overall posterior function formula, based on the priors and the likelihoods within it. After this draw, a judgement is made using some ‘rule’ whether to accept or reject this new value, compared with the initial or previous value of that parameter. Then it goes to the next simulation round, or draw.

Effective Sample Size

Hoffman and Gelman (2011) explain:

The ESS of a set of M correlated samples θ (of dimensionality 1:**M**) with respect to any function f(θ) is the number of independent draws from the target posterior distribution p(θ) that would give a Monte Carlo estimate of the mean under p of f(θ) with the same level of precision as the estimate given by the mean of the function f for the correlated samples θ, of dimensionality 1:**M**.

The ESS of a sample is a measure of how many **independent samples** a set of correlated samples is worth for the purposes of estimating the mean of some function.

They give the routine for a ‘naïve’ and a ‘more efficient’ Hamiltonian sampler. The latter gives a larger ESS for less computation.

In the current scene, we use ESS to indicate the success of an estimate of a single parameter rather than looking at the goodness of fit overall straight away. Larger ESS is considered better.

Declaring a simple multilevel model (Gelman and Hill, Chapter 16)

- In a simple model,
- y_i ~ N($\alpha_j + \beta x_i, \sigma_y^2$) and α_j ~ N($\mu_\alpha, \sigma_\alpha^2$)

(page 346)

The second normal distribution is acting as a prior for the main model intercepts, which are distributed across Level 2 units whose subscript is j. The Level 1 units have subscript i.

Therefore, μ_α and σ_α are the hyperparameters here. The hyperparameters by default are given a uniform distribution each, or a half-t-distribution for variance which is non-negative. *Ibid.*

Quick quiz

- In a simple model,
- y_i ~ N($\alpha_j + \beta x_i, \sigma_y^2$) and α_j ~ N($\mu_\alpha, \sigma_\alpha^2$)

What are the two main parameters then?

What prior distribution would you suggest for Beta if you want only a weakly informative prior?

We write the posterior thus:

- $p(\alpha, \beta, \mu_\alpha, \sigma_y, \sigma_\alpha)$

- The posterior is proportional to the RHS as shown in this formula:
- $p(\alpha, \beta, \mu_\alpha, \sigma_y, \sigma_\alpha) \propto \prod_{j=1}^{J} \mathrm{N}(\alpha_j | \mu_\alpha, \sigma_\alpha^2)$

-- this particular equation represents only random intercepts, not random slopes, in a simplified multilevel model.

Each parameter such as β has its own values, within each chain, and this can be 'traced' over time through the S iterations, e.g. 2,000 iterations. If there are 2 chains, they should reach similar levels for β.

The different algorithms help the machine reach well-mixed chains.

They merge on a posterior that is close to the reality which generated the data.

The posterior has 4 dimensions for 4 parameters, e.g. α, μ, σ_y, σ_α, in a simpler empty multi-level model

- See Gelman and Hill, 2007, chapter 18, section 18.4 for the R code of a Gibbs sampler for a multilevel model. Here are the steps.
- 1. Choose how many chains of iterations to have.
- Decide on the domain of each key parameter, restrict these to ≥ 0 where possible.
- Set up initial values for each parameter. Usually random numbers.
- 2. Decide how many iterations to have, within each chain.
- For each parameter, in one iteration, it may be amended by taking a random simulation draw from that parameter's distribution. (We use the prior distribution here.)
- The formula to use is posterior=likelihood*prior, see example the likelihood formulas in equation 18.10, page 393, combined with simple prior distributions. The example is very simplified.
- 3. Now the chain mixing involves deciding when the chains have resolved.

The Metropolis Algorithm

- Each parameter in the model is one of the variables inside the posterior function. Candidates for the value of a parameter are chosen from one of several possible distribution types, or densities:
  - Normal Distribution, or Student's T Distribution,
  - Or a Uniform Density.

(Source: Congdon, 2010: 10)

Remember posterior = likelihood*prior, and this is evaluated, using old parameter values, plus this one new parameter value.

The whole thing is kept if it is better, or revert to previous if not better.

Rule: Compare Posterior_2 with Posterior_1 and decide.

Metropolis wrote about this in 1953.

Metropolis-Hastings Algorithm also considers that we can takes the derivative of the posterior in each iteration

- See Gelman and Hill, sections 18.6 in Chapter 18.
- One iteration involves raising or lowering one parameter by a quantum, depending on a random draw from that parameter's prior.

- After the draw, the posterior can be calculated, using the previous iteration's values of the data, and all other parameters.

- The hyperparameters are included in the round so for example:
- 3 parameters plus 5 hyperparameters in linear regression means 8 in total in the Theta vector.

Summary of special routines

- The use of a momentum function or another supplementary function
- Within the routine for taking S iterations
- Can help the computer to resolve either a single chain
  - Gibbs sampling in JAGS (Just Another Gibbs Sampler)
  - Or in multiple chains in Gibbs, Metropolis-Hastings or other samplers.
- The use of a momentum function is a metaphor for ‘movement’ in the parameter space, and good ones avoid doubling back, thus reducing the risk of cycling endlessly.

Quick quiz, revisited

- In a simple model,
- y_i ~ N($\alpha_j + \beta x_i, \sigma_y^2$) and α_j ~ N($\mu_\alpha, \sigma_\alpha^2$)

Can you write an expression for the expectation of the function at the left?

Here is the posterior for the above model

$p(\alpha, \beta, \mu_\alpha, \sigma_y, \sigma_\alpha) \propto \prod_{j=1}^{J} \mathrm{N}(\alpha_j | \mu_\alpha, \sigma_\alpha^2)$

Can you algebraically take a derivative of this function with respect to one of the hyperparameters? No, that would get very hard, but it can be done in theory and in practice, allowing for the part which is 'likelihood' and the part which is prior. When dealing with the former, the priors are taken as given, and when dealing with the priors, we take the rest of the likelihood as given. In derivatives, this means a constant slope non-dependent on other terms, at a given time in the running of the routine.

Concluding points

- The use of priors creates an improvement over conjugate Bayesian models because we can computerise the estimation.
    - Conjugate models formed key contents of Gelman, et al., 2013 chapters, and are algebraically huge and complex, but do offer solutions to problems
    - They also help us understand what 'prior' distributions are preferable, see Cowles (2015)
- Algorithms help us avoid getting trapped in the wrong part of the parameter space.
- Using 2+ chains helps the machine to resolve quicker, splitting work into ¼ as many iterations, for example 2 chains, each with 1,000 iterations.
- Metropolis-Hastings algorithm uses the posterior function, moving through stages, to consider small changes of each prior in turn, and each main parameter in term, through all the S iterations.
- There is a need for a rule to decide whether to keep, amend, or reject a new parameter value.
- Efficient estimation uses ESS as a measure, with one ESS value for each parameter. BIC could be used to measure the overall fit of one model.
- BIC only makes sense when comparing 2 or more models.

Child Labour Logit Brms Quick Quiz - Answers

1 Why did I take just 'Number of observations: 10000' rather than 422K ? **FASTER**

2 What advantage to using 4 chains, each with iter = 1200? **FASTER**

3 What is throwing out the warmup session of 600 iterations? **Don't use nonconverged results of chain-mixing.**

Note: total post-warmup samples = 2400

4 What does the abbreviation CI stand for, below? **(Credible Interval)**

```
##                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)      1.35      0.53      0.66      2.75 1.01      613     1161

## Population-Level Effects:
##             Estimate Est.Error l-95% CI u-95% CI Rhat
## fhh  -0.27    0.21  -0.70  0.14  1.00
```

5 Is the effect of Female Household Head, after controlling for age and country and rural+other, different from zero, and please use a sentence?

There is 95% probability that the population from which this sample was taken has a FHH effect that is >0 for the risk of a child getting into harmful market child labour. This is after controls.

6 ***Why*** were **samples were drawn using sampling(NUTS)? R knew that a default NUTS was implied.**

- 7. **NUTS** is the sampler; what is 'mcmc' here? **Markov-Chain, Monte Carlo simulation.**

References

Contact and Queries:
Wendy.olsen@Manchester.ac.uk

- Congdon, Peter D. (2010), *Applied Bayesian Hierarchical Methods*, London: CRC Press, Chapter 1.
- Gelman, Andrew, and Jennifer Hill (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge: Cambridge University Press.
- Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, A. Vehtari, and Donald B. Rubin, 2013, *Bayesian Data Analysis*, 3rd ed., London: CRC Press and Taylor & Francis, Chapman and Hall. Series: Texts in Statistical Science.
- Hoffman, and Andrew Gelman (2011), The No-U-Turn Sampler, mimeo in *Arxiv*, https://arxiv.org/pdf/1111.4246.pdf
- Kruschke, John K. (2015), 2nd ed., *Doing Bayesian Data Analysis: A Tutorial With R, JAGS, and STAN,* Amsterdam: Academic Press.
- Kéry, Marc, and Michael Schaub (2012), *Bayesian Population Analysis Using WinBUGS: A hierarchical perspective*, Elsevier, Academic Press. (Chapter 2)
- Lunn, Christopher Jackson, Nicky Best, Andrew Thomas, and David Spiegelhalter, 2013, *The BUGS Book: A Practical Introduction to Bayesian Analysis*, CRC Press, Chapman and Hall.
- Lunn, Christopher Jackson, Andrew Thomas, Nicky Best, & David Spiegelhalter (2000). WinBUGS—a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics & Computing*, **10**, 325–337.