# Bayesian Methods: Part 1

A. Colin Cameron
Univ. of Calif. - Davis
. .

May 2021

# 1. Introduction

- Bayesian methods provide an alternative method of computation and statistical inference to ML estimation.
  - Some researchers use a fully Bayesian approach to inference.
  - Other researchers use Bayesian computation methods (with a diffuse or uninformative prior) as a tool to obtain the MLE and then interpret results as they would classical ML results.

# Outline

# 2. Bayesian Methods: Basic Idea

- Bayesian methods for inference on $\boldsymbol{\theta}$ obtain information on $\boldsymbol{\theta}$ from two sources
  - Data - the **likelihood** function
    - for regression this is usually $L(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})$
  - Prior beliefs on $\boldsymbol{\theta}$
    - the **prior density** $\pi(\boldsymbol{\theta})$
    - this bit is new.

# Bayesian Methods: The posterior density

- Recall Bayes Theorem that $\Pr[A|B] = \Pr[A \cap B] / \Pr[B]$.
- Applying Bayes here, the **posterior density** for $\boldsymbol{\theta}$ given data $\mathbf{y}, \mathbf{X}$ is

$$p(\boldsymbol{\theta}|\mathbf{y},\mathbf{X}) = \frac{p(\boldsymbol{\theta},\mathbf{y},\mathbf{X})}{p(\mathbf{y},\mathbf{X})}$$

- So the **posterior density** of $\boldsymbol{\theta}$ is

$$p(\boldsymbol{\theta}|\mathbf{y},\mathbf{X}) = \frac{L(\mathbf{y}|\boldsymbol{\theta},\mathbf{X}) \times \pi(\boldsymbol{\theta})}{m(\mathbf{y}|\mathbf{X})}$$

  - $m(\mathbf{y}|\mathbf{X}) = \int L(\mathbf{y}|\boldsymbol{\theta},\mathbf{X}) \times \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is called the **marginal likelihood**
    - **problem: there is usually no tractable expression** for $m(\mathbf{y}|\mathbf{X})$.
- In general

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

# Bayesian Methods: The prior density

- The prior can be **informative** so it does effect $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})$
  - do this if have strong prior information on $\boldsymbol{\theta}$.
- In some simple settings such as a doctor interpreting a medical test
  - $\theta$ is scalar
  - there are no regressors so the likelihood is $L(\mathbf{y}|\theta)$
  - there can be strong prior beliefs $\pi(\theta)$.
- The prior can be **uninformative** so it has little effect on $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})$
  - e.g. $\boldsymbol{\theta}$ can take a very wide range of values (large variance)
- For econometrics regressions prior beliefs are typically uninformative over all parameters, or over all but a subset of the parameters.
- As $N \rightarrow \infty$ the prior has little effect as the likelihood dominates.

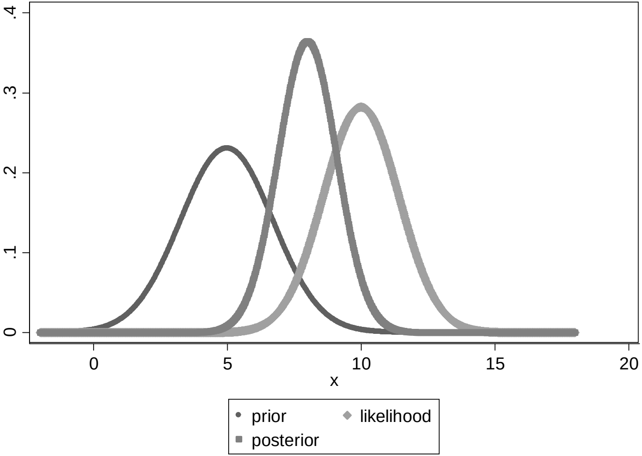
# Bayesian Methods: Inference

- Bayesian analysis bases inference on the posterior distribution.
  - the **best estimate** of $\boldsymbol{\theta}$ is the mean or the mode of the posterior distribution.
  - a **95% credible interval** (or "Bayesian confidence interval") for $\boldsymbol{\theta}$ is from the 2.5 to 97.5 percentiles of the posterior distribution
  - no need for asymptotic theory!
- Classical statisticians interpret results in the usual MLE way
  - the mode or mean of the posterior is viewed as estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$.
- Until recently only very simple Bayesian models could be computed
  - due to inability to compute $m(\mathbf{y}|\mathbf{X}) = \int L(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X}) \times \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$
    - including Bayes (1765) original example
  - MCMC methods have changed this.

# 3. Normal-normal Bayesian example

- Suppose $y|\theta \sim \mathcal{N}[\theta, 100]$ ($\sigma^2$ is known from other studies)
  And we have independent sample of size $N = 50$ with $\bar{y} = 10$.
- Classical analysis uses $\bar{y}|\theta \sim \mathcal{N}[\theta, 100/N] \sim \mathcal{N}[\theta, 2]$
  Reinterpret as likelihood $\theta|\mathbf{y} \sim \mathcal{N}[\theta, 2]$.
  Then MLE $\widehat{\theta} = \bar{y} = 10$.
- Bayesian analysis introduces prior, say $\theta \sim \mathcal{N}[5, 3]$.
  We combine likelihood and prior to get posterior.
- We expect
  - posterior mean: between prior mean 5 and sample mean 10
  - posterior variance: less than 2 as prior info reduces noise
  - posterior distribution: ? Generally intractable.
- But here can show that the posterior for $\theta$ is $\mathcal{N}[8, 1.2]$.

- Prior $\mathcal{N}[5,3]$ and likelihood $\mathcal{N}[10,2]$ and yields posterior $\mathcal{N}[8,1.2]$ for $\theta$



- Classical inference: $\widehat{\theta} = \bar{y} = 10 \sim \mathcal{N}[10, 2]$
  - A 95% confidence interval for $\theta$ is $10 \pm 1.96 \times \sqrt{2} = (7.23, 12.77)$
  - i.e. if we sampled many times then 95% of the time a similarly constructed confidence interval will include the unknown constant $\theta$.
- Bayesian inference: Posterior $\theta \sim \mathcal{N}[8, 1.2]$
  - A 95% posterior interval for $\theta$ is $8 \pm 1.96 \times \sqrt{1.2} = (5.85, 10.15)$
  - i.e. with probability 0.95 the true value of $\theta$ lies in this interval.

# Role of the prior and the sample size

- For normal-normal if $y_i|\mu \sim \mathcal{N}[\mu, \sigma^2]$ with $\sigma^2$ known and prior $\mu \sim \mathcal{N}[\underline{\mu}, \underline{s}^2]$ then the posterior $\mu|\mathbf{y} \sim \mathcal{N}[\overline{\mu}, \overline{s}^2]$
  - $\overline{\mu} = \overline{s}^2 \times [(\frac{\sigma^2}{N})^{-1}\bar{y} + (\underline{s}^2)^{-1}\underline{\mu}]$ is the posterior mean
  - and $\overline{s}^2 = [(\frac{\sigma^2}{N})^{-1} + (\underline{s}^2)^{-1}]^{-1}$ is the posterior variance
    - ★ the inverse of the variance is called the precision.
- Consider variations of the preceding example with $\mu \sim \mathcal{N}[8, 1.2]$.
  - with a "diffuse" prior Bayesian gives similar numerical result to classical
    - ★ if prior is $\mu \sim \mathcal{N}[5, 100]$ then posterior is $\mu \sim \mathcal{N}[9.903, 1.961]$.
  - with a large sample we get result close to the classical result
    - ★ if $N = 5,000$ then $\bar{y} = 10 \sim \mathcal{N}[10, 0.02]$ and posterior is $\mu \sim \mathcal{N}[9.961, 0.01987]$.

# Tractable results are rare

- The tractable result for normal-normal (known variance) carries over to **exponential family using a conjugate prior**

| **Likelihood** | **Prior** | **Posterior** |
| --- | --- | --- |
| Normal (mean $\mu$) | Normal | Normal |
| Normal (precision $\frac{1}{\sigma^2}$) | Gamma | Gamma |
| Binomial $(p)$ | Beta | Beta |
| Poisson $(\mu)$ | Gamma | Gamma |

  - using conjugate prior is like augmenting data with a sample from the same distribution
  - for Normal with precision matrix $\Sigma^{-1}$ gamma generalizes to Wishart.
- But in general tractable results are not available
  - so use numerical methods, notably MCMC.
  - using tractable results in subcomponents of MCMC can speed up computation.

# 4. MCMC Example using Stata command bayes:

- Consider a linear regression log earnings - schooling example
  - men and women full-time workers in 2010.

```
. * Read in and summarize earnings - schooling data
. qui use mus229acs.dta, clear

. describe earnings lnearnings age education

Variable      Storage   Display    Value
    name         type    format    label      Variable label
-----------------------------------------------------------------------------------
earnings        float   %9.0g                 Annual earnings in $
lnearnings      float   %9.0g                 Natural logarighm of earnings
age             int     %36.0g                Age in years
education       float   %9.0g                 Educational attainment: years of
                                                schooling

. qui keep if _n <= 100

. summarize earnings lnearnings age education

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
    earnings |        100       60244    46513.19       4000     318000
  lnearnings |        100    10.76058    .7273709   8.294049   12.66981
         age |        100       43.33     10.9342         25         65
   education |        100       13.69    3.158106          0         20
```

# MLE (equals OLS) for Comparison

- Concentrate on coefficient of education
  - MLE is 0.0852 with se 0.0221 and 95% CI (0.041, 0.129)

```
. * ML linear regression (same as OLS with iid errors)
. regress lnearnings education age, noheader
```

| lnearnings | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| education | .0852959 | .0221804 | 3.85 | 0.000 | .0412739 | .1293178 |
| age | .0079952 | .0064063 | 1.25 | 0.215 | -.0047195 | .02071 |
| _cons | 9.246449 | .4546021 | 20.34 | 0.000 | 8.34419 | 10.14871 |

# MCMC Simple overview

- Markov chain Monte Carlo methods (MCMC) are a way to make draws of $\boldsymbol{\theta}$ from the posterior given the previous draw of $\boldsymbol{\theta}$.
- Metropolis-Hastings iterative procedure
  - at round $s$ draw $\boldsymbol{\theta}^*$ from a candidate distribution that depends on $\boldsymbol{\theta}^{(s-1)}$ and possibly the data $\mathbf{y}$, $\mathbf{X}$
  - use a rule (Metropolis or Metropolis-Hastings) to
    - ★ either set $\boldsymbol{\theta}^{(s)} = \boldsymbol{\theta}^*$ or set $\boldsymbol{\theta}^{(s)} = \boldsymbol{\theta}^{(s-1)}$.
  - thus some draws from the candidate distribution are accepted and some are not.
- The initial resulting $\boldsymbol{\theta}^{(s)}$ draws are not draws from the posterior
  - so discard the first several thousand draws.
- Hopefully after that we have (correlated) draws from the posterior.
- Given the draws from the posterior we can do almost anything.

# MCMC Example: Linear Regression

- Stata command `bayes`: prefix command is simple
  - e.g. `bayes: regress y x1 x2`
- The default sets the following priors
  - $\beta_j$ are independently $N(0, 100^2)$
  - $\sigma^2$ is inverse gamma (0.01, 0.01)
    - ★ so $1/\sigma^2$ is gamma (0.01, 0.01).
- The default sets
  - 12,500 MCMC iterations
  - first 2,500 are not used ("burn-in")
- The defaults can be changed.
- The command `bayesmh` is more flexible
  - e.g. for nonstandard models you can provide the likelihood.

# MCMC Example

- First part of output

```
. * Bayesian linear regression with uninformative prior and Stata defaults
. bayes, rseed(10101): regress lnearnings education age

Burn-in ...
Simulation ...

Model summary
------------------------------------------------------------------------------
Likelihood:
  lnearnings ~ regress(xb_lnearnings,{sigma2})

Priors:
  {lnearnings:education age _cons} ~ normal(0,10000)                       (1)
                         {sigma2} ~ igamma(.01,.01)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_lnearnings.
```

# MCMC Example (continued)

- Second part of output
  - Efficiency: the 10,000 correlated draws are equivalent to on average 929.9 independent draws
  - Acceptance rate: 3,071 of the 10,000 draws were accepted.

```
Bayesian linear regression                     MCMC iterations  =     12,500
Random-walk Metropolis-Hastings sampling       Burn-in          =      2,500
                                               MCMC sample size =     10,000
                                               Number of obs    =        100
                                               Acceptance rate  =      .3071
                                               Efficiency:  min =     .07066
                                                            avg =     .09299
Log marginal-likelihood = -133.37046                        max =      .1512
```

# MCMC Example (continued)

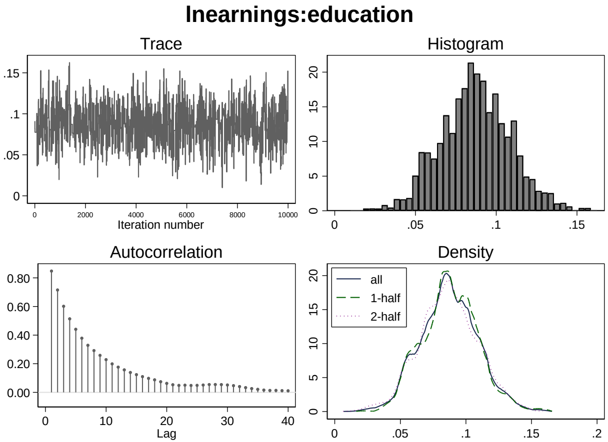
- Third part of output for regressor education
  - Posterior mean is 0.0872 with sd 0.0218 and 95% credible region (0.047, 0.131)
  - MLE is 0.0852 with se 0.0221 and 95% CI (0.041, 0.129)

| | Mean | Std. dev. | MCSE | Median | Equal-tailed [95% cred. interval] | |
|---|---|---|---|---|---|---|
| lnearnings | | | | | | |
| education | .0871874 | .0217776 | .000819 | .0868041 | .0471493 | .1312628 |
| age | .008496 | .0062873 | .000231 | .0089316 | -.0037933 | .0208249 |
| _cons | 9.198406 | .4482471 | .016292 | 9.196124 | 8.319206 | 10.09851 |
| sigma2 | .4774248 | .0711248 | .001829 | .4702676 | .3587335 | .6308758 |

Note: Default priors are used for model parameters.
Note: Adaptation tolerance is not met in at least one of the blocks.

# MCMC Example: Diagnostics

- For $\beta_{educ}$ shows several graphical diagnostics
  - use `bayesgraph diagnostics {lnearnings:education}`



# Convergence of Chain

- There is no formal test.
- Can do multiple independent chains and see if the variability of the posterior mean of $\theta$ across chains is small, relative to the variation of draws of $\theta$ within each chain.
- Consider the *jth* of *m* chains
  - $\widehat{\theta}_j$ = posterior mean and $s_j$ = posterior variance
- $B$ measures variation between chains
  - $B = \frac{1}{m-1}\sum_{j=1}^{m}(\widehat{\theta}_j - \overline{\widehat{\theta}})^2$ where $\overline{\widehat{\theta}} = \frac{1}{m}\sum_{j=1}^{m}\widehat{\theta}_j$.
- $W$ measures variation in $\theta$ within chains
  - $W = \frac{1}{m}\sum_{j=1}^{m} s_j^2$.
- The Gelman-Rubin Rc statistic Rc $\simeq \frac{W+B}{W}$
  - Actually uses an adjustment for finite number of chains
  - A common threshold is Rc$< 1.1$ (equivalently $\frac{B}{W} < 0.1$).

# Convergence of Chain (continued)

- * Check convergence using multiple chains
- bayes, rseed(10101) nchains(5): regress lnearnings education age

```
Bayesian linear regression                    Number of chains    =          5
Random-walk Metropolis-Hastings sampling      Per MCMC chain:
                                                  Iterations      =     12,500
                                                  Burn-in         =      2,500
                                                  Sample size     =     10,000
                                              Number of obs       =        100
                                              Avg acceptance rate =      .3402
                                              Avg efficiency: min =     .07201
                                                              avg =      .1053
                                                              max =      .1815
Avg log marginal-likelihood = -133.35288      Max Gelman-Rubin Rc =      1.002
```

| | Mean | Std. dev. | MCSE | Median | Equal-tailed [95% cred. | interval] |
|---|---|---|---|---|---|---|
| lnearnings | | | | | | |
| education | .085597 | .0222416 | .000371 | .0855127 | .0416117 | .12877 |
| age | .0079981 | .0063156 | .000096 | .0081201 | -.0044435 | .0202879 |
| _cons | 9.241303 | .4537841 | .007116 | 9.23721 | 8.355778 | 10.14552 |
| sigma2 | .4763385 | .0699901 | .000735 | .4693347 | .3578036 | .6313855 |

# Convergence of Chain (continued)

- Preceding gave average Rc across the four parameters of $1.002 < 1.1$.
- Now get Rc for each parameter.

```
. * Give Gelman-Rubin Rc statistic for each parameter
. bayesstats grubin

Gelman-Rubin convergence diagnostic

Number of chains     =          5
MCMC size, per chain =     10,000
Max Gelman-Rubin Rc  =   1.002092

---------------------------
             |         Rc
-------------+-------------
lnearnings   |
   education |    1.00161
         age |   1.001305
       _cons |   1.002092
-------------+-------------
      sigma2 |   1.000309
---------------------------

Convergence rule: Rc < 1.1
```

# MCMC Example: Some bayes: code

```
* Estimation
bayes rseed(10101): regress y x
* Summary statistics for model parameters
bayesstats summary {y:x}
* Probability that slope is in range 0.4 to 0.6
bayestest interval {y:x}, lower(0.4) upper(0.6)
* Effective sample size
bayesstats ess
* Graphical Diagnostics
bayesgraph diagnostics {y:x}
* Convergence diagnostics
bayes, rseed(10101) nchains(5): regress y x
bayesstats grubin
```

# 5. Markov chain Monte Carlo (MCMC)

- The challenge is to compute the posterior $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})$
  - analytical results are only available in special cases.
  - early numerical methods used importance sampling to estimate posterior moments.
- Instead use Markov chain Monte Carlo methods:
  - Make sequential random draws $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$.
  - where $\boldsymbol{\theta}^{(s)}$ depends in part on $\boldsymbol{\theta}^{(s-1)}$
    - but not on $\boldsymbol{\theta}^{(s-2)}$ once we condition on $\boldsymbol{\theta}^{(s-1)}$ (so a Markov chain)
  - in such a way that after an initial burn-in (discard these draws) $\boldsymbol{\theta}^{(s)}$ are (correlated) draws from the posterior $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})$
    - the Markov chain converges to a stationary marginal distribution which is the posterior.

# Markov Chains

- A Markov chain is a stochastic sequence of possible events in which the probability of each event depends only on the state attained in the previous event
- Under suitable assumptions the chain converges to a stationary marginal distribution.
- Here the MCMC method is set up so that this stationary distribution is the desired posterior.
- The one caveat is that while in theory the chain converges
  - in practice it can take many rounds to converge
  - and there is no formal test of whether convergence has occurred.

# Leading MCMC methods

- **1.** Metropolis algorithm
  - Nicholas Metropolis, Arianna W. Rosenbluth, Marshall Rosenbluth, Augusta H. Teller and Edward Teller (1953), "Equation of State Calculations by Fast Computing Machines", *Journal of Chemical Physics*.
- **2.** Metropolis-Hastings algorithm
  - Relax the metropolis requirement that the candidate distribution is symmetric
  - W.K. Hastings (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications ", *Biometrika*.
- **3.** Gibbs sampler
  - special case where conditional posteriors are known
  - A.E. Gelfand and A.F.M. Smith (1990), *JASA*, is a key statistical paper for Gibbs sampler and more generally use of MCMC methods in statistics.

# Metropolis Algorithm

- We want to draw from **posterior** $p(\cdot)$ but usually cannot directly do so.
- Metropolis draws from a **candidate** distribution $g(\boldsymbol{\theta}^{(s)}|\boldsymbol{\theta}^{(s-1)})$
  - these draws are sometimes accepted and some times not
  - like accept-reject method but do not require $p(\cdot) \leq kg(\cdot)$
- Metropolis algorithm at the $s^{th}$ round
  - draw candidate $\boldsymbol{\theta}^*$ from candidate distribution $g(\cdot)$
  - the candidate distribution $g(\boldsymbol{\theta}^{(s)}|\boldsymbol{\theta}^{(s-1)})$ needs to be symmetric
    - ★ so it must satisfy $g(\boldsymbol{\theta}^a|\boldsymbol{\theta}^b) = g(\boldsymbol{\theta}^b|\boldsymbol{\theta}^a)$
  - draw $u$ from uniform$[0,1]$

$$\begin{aligned}\boldsymbol{\theta}^{(s)} &= \boldsymbol{\theta}^* \text{ if } u < \frac{p(\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(s-1)})} \\ &= \boldsymbol{\theta}^{(s-1)} \text{ otherwise.}\end{aligned}$$

# Metropolis Algorithm (continued)

- Because we only use a ratio of posteriors the difficult normalizing constant (the marginal likelihood) does not have to be computed

$$\frac{p(\boldsymbol{\theta}^{*}|\mathbf{y},\mathbf{X})}{p(\boldsymbol{\theta}^{(s-1)}|\mathbf{y},\mathbf{X})} = \frac{\frac{L(\mathbf{y}|\boldsymbol{\theta}^{*},\mathbf{X})\times\pi(\boldsymbol{\theta}^{*})}{m(\mathbf{y}|\mathbf{X})}}{\frac{L(\mathbf{y}|\boldsymbol{\theta}^{(s-1)},\mathbf{X})\times\pi(\boldsymbol{\theta}^{(s-1)})}{m(\mathbf{y}|\mathbf{X})}} = \frac{L(\mathbf{y}|\boldsymbol{\theta}^{*},\mathbf{X})\times\pi(\boldsymbol{\theta}^{*})}{L(\mathbf{y}|\boldsymbol{\theta}^{(s-1)},\mathbf{X})\times\pi(\boldsymbol{\theta}^{(s-1)})}$$

- For proof that the Markov chain converges to the desired distribution see, for example, Cameron and Trivedi (2005), p.451
  - the proof requires that the candidate distribution is symmetric.
- Taking logs

$$\begin{aligned} \boldsymbol{\theta}^{(s)} &= \boldsymbol{\theta}^{*} \text{ if } \ln u < \ln p(\boldsymbol{\theta}^{*}) - \ln p(\boldsymbol{\theta}^{(s-1)}) \\ &= \boldsymbol{\theta}^{(s-1)} \text{ otherwise.} \end{aligned}$$

- Random walk Metropolis draws from $\boldsymbol{\theta}^{(s)} \sim \mathcal{N}[\boldsymbol{\theta}^{(s-1)}, \mathbf{V}]$ for fixed $\mathbf{V}$
  - ideally $\mathbf{V}$ such that 25-50% of candidate draws are accepted.

# Metropolis-Hastings Algorithm

- Metropolis-Hastings is a generalization
  - the candidate distribution $g(\boldsymbol{\theta}^{(s)}|\boldsymbol{\theta}^{(s-1)})$ need not be symmetric
  - the acceptance rule is then $u < \frac{p(\boldsymbol{\theta}^{*})\times g(\boldsymbol{\theta}^{*}|\boldsymbol{\theta}^{(s-1)})}{p(\boldsymbol{\theta}^{(s-1)})\times g(\boldsymbol{\theta}^{(s-1)}|\boldsymbol{\theta}^{*})}$
  - Metropolis algorithm itself is often called Metropolis-Hastings.
- Independence chain MH uses $g(\boldsymbol{\theta}^{(s)})$ not depending on $\boldsymbol{\theta}^{(s-1)}$ where $g(\cdot)$ is a good approximation to $p(\cdot)$
  - e.g. Do ML for $p(\boldsymbol{\theta})$ and then $g(\boldsymbol{\theta})$ is multivariate $T$ with mean $\widehat{\boldsymbol{\theta}}$, variance $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}}]$.
  - multivariate rather than normal as has fatter tails.
- M and MH called Markov chain Monte Carlo
  - because $\boldsymbol{\theta}^{(s)}$ given $\boldsymbol{\theta}^{(s-1)}$ is a first-order Markov chain
  - Markov chain theory proves convergence to draws from $p(\cdot)$ as $s \rightarrow \infty$
  - poor choice of candidate distribution leads to chain stuck in place.

# Gibbs sampler

- Gibbs sampler (a general method for making draws)
  - draw $(\mathbf{Y}_1, \mathbf{Y}_2)$ by alternating draws from $f(\mathbf{y}_1|\mathbf{y}_2)$ and $f(\mathbf{y}_2|\mathbf{y}_1)$
  - after many draws gives draws from $f(\mathbf{y}_1, \mathbf{y}_2)$ even though

$$f(\mathbf{y}_1, \mathbf{y}_2) = f(\mathbf{y}_1|\mathbf{y}_2) \times f(\mathbf{y}_2) \neq f(\mathbf{y}_1|\mathbf{y}_2) \times f(\mathbf{y}_2|\mathbf{y}_1).$$

- Suppose posterior is partitioned e.g. $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$
  - and we can make draws from $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)$ and $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$.
- Gibbs is special case of MH
  - usually quicker than usual MH
  - if need MH to draw from $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)$ and/or $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$ called MH within Gibbs.
  - extends to e.g. $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$ make sequential draws from $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$, $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \boldsymbol{\theta}_3)$ and $p(\boldsymbol{\theta}_3|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$
  - requires knowledge of all of the full conditionals.

# Correlated Draws

- M, MH and Gibbs yield correlated draws of $\boldsymbol{\theta}^{(s)}$.
- But it still gives correct estimate of marginal posterior distribution of $\boldsymbol{\theta}$ (once discard burn-in draws)
  - e.g. estimate posterior mean by $\frac{1}{S}\sum_{s=1}^{S}\boldsymbol{\theta}^{(s)}$.
- The precision of this estimate will, however, decline with greater correlation of the draws
  - the efficiency statistic measures this
  - if the efficiency statistic is low then make more draws (after the burn-in).

# Stata bayes: and bayesmh commands

- The `bayes:` prefix command can be applied to over 50 estimation commands including `regress`, `xtreg`, `logit`, `mlogit`, `ologit` and `xtlogit`. Defaults such as priors can be changed.
- The `bayesmh` command is more flexible and allows one to program ones own models.
- The default version of `bayesmh` can give somewhat different results to `bayes:` because `bayes:` takes advantage of the knowledge of the particular model used, such as blocking of model parameters to improve the efficiency of the sampling algorithm.

# bayesmh command equal to earlier bayes: regress command

- The following command gives exactly the same results as the earlier bayes, rseed(10101): regress lnearnings education age
- `bayesmh` command example

```
bayesmh lnearnings education age, likelihood(normal({sigma2})) ///
prior({lnearnings:education}, normal(0,10000)) ///
prior({lnearnings:age}, normal(0,10000)) ///
prior({lnearnings:_cons},normal(0,10000)) ///
prior({sigma2},igamma(0.01,0.01)) rseed(10101) ///
block({lnearnings: education age _cons}) block({sigma2})
```

- If the last line (blocking) is dropped the results differ
  - blocking can really speed up computation.

# 6. Further discussion: Specification of prior

- As $N \rightarrow \infty$ data dominates the prior $\pi(\boldsymbol{\theta})$
  and then posterior $\boldsymbol{\theta}|\mathbf{y} \overset{a}{\sim} \mathcal{N}[\widehat{\boldsymbol{\theta}}_{\text{ML}}, I(\widehat{\boldsymbol{\theta}}_{\text{ML}})^{-1}]$
    - but in finite samples prior can make a difference.
- Noninformative and improper prior
    - has little effect on posterior
    - a uniform or flat prior (with all values equally likely) is frequent choice
    - this is an improper prior if $\boldsymbol{\theta}$ is unbounded
    - but usually the posterior is still proper
        - ★ if $\pi(\boldsymbol{\theta}) = c$ we need $\int L(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} = c\int L(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})d\boldsymbol{\theta}$ to be finite
    - not invariant to transformation of $\boldsymbol{\theta}$ (e.g. $\theta \rightarrow e^{\theta}$).
- Jeffreys prior sets $\pi(\boldsymbol{\theta}) \propto \det[I(\boldsymbol{\theta})^{-1}]$, $I(\boldsymbol{\theta}) = -\partial^2 \ln L/\partial\theta\partial\theta'$
    - invariant to transformation
    - for linear regression under normality this is uniform prior for $\boldsymbol{\beta}$
    - also an improper prior.

- Proper prior (informative or uninformative)
  - informative becomes uninformative as prior variance becomes large.
  - use conjugate prior if available as it is tractable
  - hierarchical (multi-level) priors are often used
    - ★ Bayesian analog of random coefficients
    - ★ let $\pi(\boldsymbol{\theta})$ depend on unknown parameters $\tau$ which in turn have a completely specified distribution
    - ★ $p(\boldsymbol{\theta},\boldsymbol{\tau}|\mathbf{y}) \propto L(\mathbf{y}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}|\boldsymbol{\tau}) \times \pi(\boldsymbol{\tau})$ so $p(\boldsymbol{\theta}|\mathbf{y}) \propto \int p(\boldsymbol{\theta},\boldsymbol{\tau}|\mathbf{y})d\boldsymbol{\tau}$
- Poisson example with $y_i$ Poisson$[\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})]$
  - $p(\boldsymbol{\beta},\boldsymbol{\mu},|\mathbf{y},\mathbf{X}) \propto L(\mathbf{y}|\boldsymbol{\mu}) \times \pi_1(\boldsymbol{\mu}|\mathbf{X},\boldsymbol{\beta}) \times \pi_2(\boldsymbol{\beta})$
  - where $\pi_1(\boldsymbol{\mu}_i|\boldsymbol{\beta})$ is gamma with mean $\exp(\mathbf{x}_i'\boldsymbol{\beta})$
  - and $\pi_2(\boldsymbol{\beta})$ is $\boldsymbol{\beta} \sim \mathcal{N}[\underline{\boldsymbol{\beta}}, \underline{\mathbf{V}}]$
    - ★ works better than $p(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}) \propto L(\mathbf{y}|\mathbf{X},\boldsymbol{\beta}) \times \pi(\boldsymbol{\beta})$.

# Informative Prior Example

- Consider lnearnings regressed on intercept, education and age.
- Education: $N[0.06, 0.01^2]$ means 95% sure that earnings increase proportionately by between 0.04 and 0.08 (so between 4% and 8%) with one more year of education.
- Age: $N[0.02, 0.01^2]$ means 95% sure that earnings increase by between 0% and 4% with one more year of aging.
- Intercept: Not clear so choose a diffuse $N[10, 10]$ prior
  - need to be very careful with prior for intercept
  - $N[10, 10]$ prior is very informative for earnings rather than lnearnings.
- sigma2 ($\sigma^2$): difficult to explain but choose a reasonably diffuse prior.

```
* bayesmh example with informative priors
bayesmh lnearnings education age, likelihood(normal({var})) ///
 prior({lnearnings:education}, normal(0.06,0.0001)) ///
 prior({lnearnings:age}, normal(0.02,0.0001)) ///
 prior({lnearnings:_cons},normal(10,100)) ///
 prior({var},igamma(1,0.5)) rseed(10101)
```

# Convergence of MCMC

- Theory says chain converges as $s \longrightarrow \infty$
    - could still have a problem with one million draws.
- Checks for convergence of the chain (after discarding burn-in)
    - graphical: plot $\theta^{(s)}$ to see that $\theta^{(s)}$ is moving around
    - correlations: of $\theta^{(s)}$ and $\theta^{(s-k)}$ should $\rightarrow 0$ as $k$ gets large
    - plot posterior density: multimodality could indicate problem
    - break into pieces: expect each 1,000 draws to have similar properties
    - run several independent chains with different starting values
        - ★ Gelman-Rubin statistic.
- But it is not possible to be 100% sure that chain has converged.

# Bayesian model selection

- Bayesians use the marginal likelihood
  - $m(\mathbf{y}|\mathbf{X}) = \int L(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X}) \times \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$
  - this weights the likelihood (used in ML analysis) by the prior.
- Bayes factor is analog of likelihood ratio

$$B = \frac{m_1(\mathbf{y}|\mathbf{X})}{m_2(\mathbf{y}|\mathbf{X})} = \frac{\text{marginal likelihood model 1}}{\text{marginal likelihood model 2}}$$

  - one rule of thumb is that the evidence against model 2 is
    - weak if $1 < B < 3$ (or approximately $0 < 2 \ln B < 2$)
    - positive if $1 < B < 3$ (or approximately $2 < 2 \ln B < 6$)
    - strong if $20 < B < 150$ (or approximately $6 < 2 \ln B < 10$)
    - very strong if $B > 150$ (or approximately $2 \ln B > 10$).
- Can use to "test" $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$ against $H_a : \boldsymbol{\theta} = \boldsymbol{\theta}_2$.
- The **posterior odds ratio** weights $B$ by priors on models 1 and 2
  - so now use priors on both $\boldsymbol{\theta}$ and the model.

- Problem: MCMC methods to obtain the posterior avoid computing the marginal likelihood
  - computing the marginal likelihood can be difficult
  - see Chib (1995), JASA, and Chib and Jeliazkov (2001), JASA.
- An asymptotic approximation to the Bayes factor is

$$B_{12} = \frac{L_1(\mathbf{y}|\widehat{\theta}, \mathbf{X})}{L_2(\mathbf{y}|\widehat{\theta}, \mathbf{X})} N^{(k_2 - k_1)/2}$$

  - Here model 1 is nested in model 2 and due to asymptotics the prior has no influence (so the ratio of posteriors is the ratio of likelihoods)
  - This is the Bayesian information criterion (BIC) or Schwarz criterion.

# What does it mean to be a Bayesian?

- Modern Bayesian methods (Markov chain Monte Carlo)
  - make it much easier to compute the posterior distribution than to maximize the log-likelihood.
- So classical statisticians:
  - use Bayesian methods to compute the posterior
  - use an uninformative prior so $p(\boldsymbol{\theta}|\mathbf{y},\mathbf{X}) \simeq L(\mathbf{y}|\boldsymbol{\theta},\mathbf{X})$
  - so $\boldsymbol{\theta}$ that maximizes the posterior is also the MLE.
- Others go all the way and be Bayesian:
  - give Bayesian interpretation
    - ★ e.g. use credible intervals
    - ★ e.g. given draws of $\theta$ can easily do inference on transformations of $\theta$
  - if possible use an informative prior that embodies previous knowledge.

# 7. Appendix: Accept-reject method

- There are many ways to random draws from a distribution such as inverse-transformation method.
- The accept-reject method can be used when
  - we want to draw from density $f(x)$ but this is difficult
  - we have a candidate density $g(x)$ that we can make draws from
  - for any value of $x$ we can compute $f(x)$ and $g(x)$
  - **key:** $g(x)$ covers $f(x)$ with $f(x) \leq kg(x)$ for some $r$ and all $x$
    - ★ this is often not possible, especially in tails for e.g. $-\infty < x < \infty$
    - ★ Metropolis and Metropolis-Hastings do not have this restriction.
    - ★ The accept-reject method to get draws from $f(x)$
  - draw $x$ from $g(x)$
  - draw $u$ from uniform(0,1) and accept the draw $x$ if

$$u \leq \frac{f(x)}{kg(x)}$$

# Accept-reject method proof

- $Y$ denotes the random variable generated by the accept-reject method
  $X$ denotes a random variable with density $g(x)$ and
  $U$ denotes a draw from the uniform. Then $Y$ has c.d.f.

$$
\begin{aligned}
\Pr[Y \leq y] &= \Pr\left[X \leq y | U \leq f(x)/kg(x)\right] \\
&= \frac{\Pr\left[X \leq y, U \leq f(x)/kg(x)\right]}{\Pr\left[U \leq f(x)/kg(x)\right]} \\
&= \frac{\int_{-\infty}^{y} \{\int_{0}^{f(x)/kg(x)} du\} g(x) dx}{\int_{-\infty}^{\infty} \{\int_{0}^{f(x)/kg(x)} du\} g(x) dx} \\
&= \frac{\int_{-\infty}^{y} [f(x)/kg(x)] g(x) dx}{\int_{-\infty}^{\infty} [f(x)/kg(x)] g(x) dx} \\
&= \frac{\int_{-\infty}^{y} [f(x)/k] dx}{\int_{-\infty}^{\infty} [f(x)/k] dx} \\
&= \int_{-\infty}^{y} f(x) dx
\end{aligned}
$$

# 8. Some References


- Chapter 13 "Bayesian Methods" in A. Colin Cameron and Pravin K. Trivedi, Microeconometrics: Methods and Applications, Cambridge University Press.
- Chapter 29 "Bayesian Methods: basics" in A. Colin Cameron and Pravin K. Trivedi, Microeconometrics using Stata, Second edition, forthcoming.
- Bayesian books by econometricians that feature MCMC are
  - Geweke, J. (2003), Contemporary Bayesian Econometrics and Statistics, Wiley.
  - Koop, G., Poirier, D.J., and J.L. Tobias (2007), Bayesian Econometric Methods, Cambridge University Press.
  - Koop, G. (2003), Bayesian Econometrics, Wiley.
  - Lancaster, T. (2004), Introduction to Modern Bayesian Econometrics, Wiley.
- Most useful (for me) book by statisticians
  - Gelman, A., J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari and D.B. Rubin (2013), Bayesian Data Analysis, Third Edition, Chapman & Hall/CRC.