SCOTTISH GRADUATE PROGRAMME IN ECONOMICS

ECNM11060
Bayesian Econometrics

A Bayesian linear regression model

Bayesian analysis of the linear regression model

⇒ Now see how general Bayesian theory of overview lecture works in familiar regression model

⇒ In lecture, we will focus on multiple regression under classical assumptions (independent errors, homoskedasticity, etc.)

⇒ Bayesian methods for freeing up classical assumptions are relatively straightforward (some of them we will discuss in the next lectures)

A regression model

⇒ Assume K explanatory variables, $x_{i1}, \ldots, x_{iK}$, for $i = 1, .., N$, and a regression model:

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i$$

⇒ Note x_{i1} is implicitly set to 1 to allow for an intercept

⇒ Matrix notation:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

⇒ ε is $N \times 1$ vector stacked in same way as $\boldsymbol{y}$

A regression model

$\Rightarrow$ β is $K \times 1$ vector of regression coefficients

$\Rightarrow$ $\boldsymbol{X}$ is $N \times K$ matrix of regressors

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{12} & \dots & x_{1K} \\ 1 & x_{22} & \dots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N2} & \dots & x_{NK} \end{bmatrix}$$

$\Rightarrow$ Regression model can be written as:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$$

$\Rightarrow$ Classical assumptions imply: ε is $\mathcal{N}(\mathbf{0}_N, h^{-1}\boldsymbol{I}_N)$ where $h^{-1} = \sigma^2$

How to approach a regression problem as a Bayesian?

⇒ Suppose we are interested in estimating the following AR(1) model:

$$y_t = \rho y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, h^{-1})$$

⇒ We have data $\boldsymbol{y}$ and two unknown model parameters (ρ and σ^2)

⇒ We know that y_t is a stationary time series (e.g., a growth rate)

⇒ What else do we know about the data? (e.g., how is the data distributed?)

⇒ Given this information, what prior beliefs can we formulate about the AR(1) parameter ρ and the precision parameter h?

⇒ In other words, we wish to specify the prior distributions $p(\rho)$ and $p(h)$

- → What do we know about ρ for certain?
- → What do we know about h for certain?

The likelihood function

- ⇒ Likelihood can be derived under the classical assumptions
- ⇒ All elements of $\boldsymbol{X}$ are fixed (i.e., not random variables)
- ⇒ OLS quantities given by:

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}, \quad \nu = N - K, \quad s^2 = \frac{\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)}{\nu}$$

Likelihood function can be written in terms of OLS quantities (see Exercise 10.1, Bayesian Econometric Methods):

$$p(\boldsymbol{y}|\boldsymbol{\beta}, h) = \frac{1}{(2\pi)^{\frac{N}{2}}} \exp\left[-\frac{h}{2}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)'\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)\right] \times h^{\frac{N}{2}} \exp\left[-\frac{h}{2}\nu s^2\right]$$

The prior

⇒ Common starting point is natural conjugate prior

⇒ β conditional on h is now **multivariate Normal**:

$$\beta \mid h \sim \mathcal{N}(\underline{\beta}, h^{-1}\underline{\boldsymbol{V}})$$

⇒ Prior for error precision h is **Gamma**:

$$h \sim \mathcal{G}(\underline{a}, \underline{b})$$

⇒ $\underline{\beta}$, $\underline{\boldsymbol{V}}$, $\underline{a}$, and $\underline{b}$ are prior hyperparameter values chosen by the researcher

The joint posterior of β and h

$\Rightarrow$ Multiplying likelihood by prior and collecting terms, the posterior can be shown to have the same form as the prior (see Exercise 10.1)

$\Rightarrow$ The conditional posterior for β (conditional on h) is **multivariate Normal**:

$$\beta \mid \boldsymbol{y}, h \sim \mathcal{N}\left(\overline{\beta}, h^{-1}\overline{\boldsymbol{V}}\right)$$

with

$$\overline{\boldsymbol{V}} = \left(\underline{\boldsymbol{V}}^{-1} + \boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$
$$\overline{\beta} = \overline{\boldsymbol{V}}\left(\underline{\boldsymbol{V}}^{-1}\underline{\beta} + \boldsymbol{X}'\boldsymbol{X}\widehat{\beta}\right)$$

The joint posterior of β and h (cont.)

$\Rightarrow$ The (marginal) posterior for error precision h is **Gamma**:

$$h \mid \boldsymbol{y} \sim \mathcal{G}(\overline{a}, \overline{b})$$

with

$$\begin{aligned}
\overline{a} &= \underline{a} + \frac{N}{2} \\
\overline{b} &= \underline{b} + \frac{1}{2}\left(\nu s^2 + \left(\widehat{\beta} - \underline{\beta}\right)' \left[\underline{\boldsymbol{V}} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\right]^{-1} \left(\widehat{\beta} - \underline{\beta}\right)\right) \\
&= \underline{b} + \frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{y} + \underline{\beta}'\underline{\boldsymbol{V}}^{-1}\underline{\beta} - \overline{\beta}'\overline{\boldsymbol{V}}^{-1}\overline{\beta}\right)
\end{aligned}$$

The marginal posterior of β

⇒ The marginal posterior for β is obtained by integrating out h

⇒ It turns out that β is multivariate t-distributed:

$$\beta \mid \boldsymbol{y} \sim t\left(\overline{\beta}, \frac{\overline{b}\,\overline{\boldsymbol{V}}}{\overline{a}}, 2\overline{a}\right)$$

⇒ Useful results for estimation:

$$\mathbb{E}(\beta \mid \boldsymbol{y}) = \overline{\beta}$$

$$\mathbb{V}(\beta \mid \boldsymbol{y}) = \frac{\overline{b}}{\overline{a}-1}\overline{\boldsymbol{V}}$$

⇒ **Intuition:** Posterior mean and variance are weighted average of information in the prior and the data

What does a prior do?

⇒ To show main ideas assume:

- → $h = 1$ is known and β is a scalar
- → $\boldsymbol{X} = (1, \ldots, 1)'$, such that $\boldsymbol{X}'\boldsymbol{X} = N$
- → $\underline{V} = \tau$ and $\underline{\beta} = 0$

⇒ In this case the posterior is:

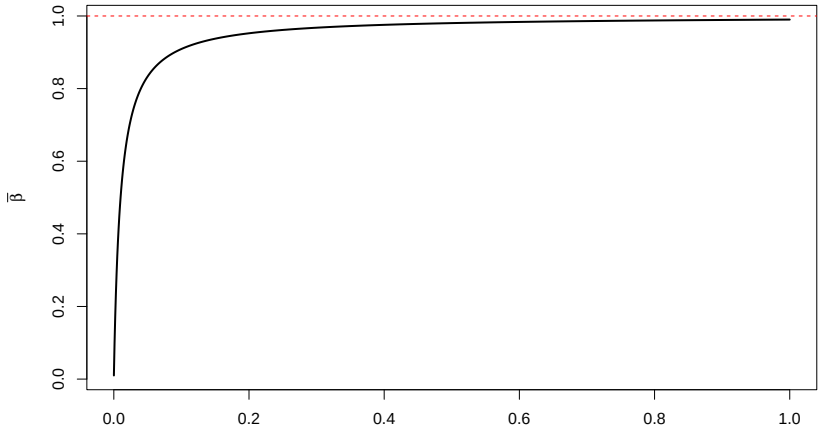
$$\overline{\beta} = \frac{N}{\frac{1}{\tau} + N}\widehat{\beta}$$

⇒ Think of the extreme cases ($\tau \to \infty$ and $\tau \to 0$)

- → If $\tau \to \infty$, then $\frac{1}{\tau} \to 0$ and $\frac{N}{\frac{1}{\tau}+N} \to 1$, so $\overline{\beta} \to \widehat{\beta}$
- → If $\tau \to 0$, then $\frac{1}{\tau} \to \infty$ and $\frac{N}{\frac{1}{\tau}+N} \to 0$, so $\overline{\beta} \to 0$ (the prior mean)

Implied prior shrinkage by varying τ

$$\bar{\beta} = \frac{N}{1/\tau + N}\hat{\beta}, \text{ for N = 100 and } \hat{\beta} = 1$$



Prior shrinkage

⇒ Posterior mean is pulled towards zero (“shrinkage”)

⇒ Commonly done to avoid over-fitting/over-parameterisation problems

⇒ Strength of prior shrinkage controlled through prior variance

- → If τ is small, then strong prior information β is near 0
- → If τ is big, then prior becomes more noninformative

⇒ Note: exactly what “small” and “large” means depends on the empirical application and units of measurement of data

Dummy observation view of conjugate shrinkage prior

- ⇒ In macroeconomics, we typically have a maximum of 700/800 observations available
- ⇒ Likely a sufficient amount of information to estimate a single parameter, but what about a large VAR model with thousands of parameters?
- ⇒ In such a case, we typically need additional information! (see Lecture 4)
- ⇒ Why not simply use a **theoretical model** to simulate data and combine it with real observed data?
- ⇒ Bayesians refer to this theoretical model as a **prior model** (our initial belief), as we can simulate observations from it without seeing any data

Dummy observation view of conjugate prior (cont.)

- $\Rightarrow$ Conjugate prior amounts to pooling real observations, $\boldsymbol{y}$ and $\boldsymbol{X}$, with dummy observations, $\underline{\boldsymbol{y}}$ and $\underline{\boldsymbol{X}}$, simulated from a theoretical model
- $\Rightarrow$ Defining $\overline{\boldsymbol{y}} = (\boldsymbol{y}', \underline{\boldsymbol{y}}')'$ and $\overline{\boldsymbol{X}} = (\boldsymbol{X}', \underline{\boldsymbol{X}}')'$, the posterior estimator (via OLS) is given by:

$$\overline{\boldsymbol{\beta}} = \left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}\overline{\boldsymbol{X}}'\overline{\boldsymbol{y}}$$

$$\overline{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X} + \underbrace{\underline{\boldsymbol{X}}'\underline{\boldsymbol{X}}}_{=\underline{\boldsymbol{V}}^{-1}}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{y} + \underline{\boldsymbol{X}}'\underline{\boldsymbol{y}}\right)$$

$$\overline{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X} + \underline{\boldsymbol{V}}^{-1}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} + \underline{\boldsymbol{X}}'\underline{\boldsymbol{X}}(\underline{\boldsymbol{X}}'\underline{\boldsymbol{X}})^{-1}\underline{\boldsymbol{X}}'\underline{\boldsymbol{y}}\right)$$

$$\overline{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X} + \underline{\boldsymbol{V}}^{-1}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \underline{\boldsymbol{V}}^{-1}\underline{\boldsymbol{\beta}}\right)$$

A noninformative prior

- $\Rightarrow$ Noninformative prior sets $\underline{a} = 0$ and $\underline{\boldsymbol{V}}$ is big (implies large prior uncertainty)
- $\Rightarrow$ But there is not a unique way of doing the latter (see Exercise 10.4 in Bayesian Econometric Methods)
- $\Rightarrow$ A common way: $\underline{\boldsymbol{V}}^{-1} = \tau^{-1}\boldsymbol{I}_K$ where τ^{-1} is a scalar and let τ^{-1} go to zero
- $\Rightarrow$ This noninformative prior is improper and becomes:

$$p(\boldsymbol{\beta}, h) \propto \frac{1}{h}$$

- $\Rightarrow$ With this choice we get OLS results:

$$\boldsymbol{\beta} \mid h, \boldsymbol{y} \sim \mathcal{N}\left(\overline{\boldsymbol{\beta}}, h^{-1}\overline{\boldsymbol{V}}\right)$$

with $\overline{\boldsymbol{V}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $\overline{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}$

$$h \mid \boldsymbol{y} \sim \mathcal{G}(\overline{a}, \overline{b})$$

with $\overline{a} = \frac{N}{2}$ and $\overline{b} = \frac{1}{2}\left(\nu \boldsymbol{s}^2\right)$

Model comparison

⇒ **Case 1:** M_1 imposes a linear restriction and M_2 does not (nested)

⇒ **Case 2:** $M_1 : \boldsymbol{y} = \boldsymbol{X}^{(1)}\beta^{(1)} + \varepsilon^{(1)}$ and $M_2 : \boldsymbol{y} = \boldsymbol{X}^{(2)}\beta^{(2)} + \varepsilon^{(2)}$, where $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ contain different explanatory variables (non-nested)

⇒ Both cases can be handled by defining models as (for $j = 1, 2$):

$$M_j : \boldsymbol{y} = \boldsymbol{X}^{(j)}\beta^{(j)} + \varepsilon^{(j)}$$

⇒ Nested model comparison defines M_2 as unrestricted regression

⇒ M_1 imposes the restriction can involve a redefinition of explanatory and dependent variable

Example: Nested model comparison

⇒ M_2 is unrestricted model

$$\boldsymbol{y} = \beta_1 + \beta_2 \boldsymbol{x}_2 + \beta_3 \boldsymbol{x}_3 + \varepsilon$$

⇒ M_1 restricts $\beta_3 = 1$, can be written:

$$\begin{aligned} \boldsymbol{y} &= \beta_1 + \beta_2 \boldsymbol{x}_2 + 1\boldsymbol{x}_3 + \varepsilon \\ \boldsymbol{y} - \boldsymbol{x}_3 &= \beta_1 + \beta_2 \boldsymbol{x}_2 + \varepsilon \end{aligned}$$

⇒ M_1 has dependent variable $\boldsymbol{y} - \boldsymbol{x}_3$ and intercept and $\boldsymbol{x}_2$ are explanatory variables

Example: Marginal likelihood computation

$\Rightarrow$ Marginal likelihood is (for $j = 1, 2$):

$$p(\boldsymbol{y} \mid M_j) = c^{(j)} \left(\frac{|\overline{\boldsymbol{V}}^{(j)}|}{|\underline{\boldsymbol{V}}^{(j)}|} \right)^{\frac{1}{2}} \left(\overline{b}^{(j)} \right)^{-\frac{\overline{a}^{(j)}}{2}}$$

$\Rightarrow$ $c^{(j)}$ is constant depending on prior hyperparameters, etc.

$$PO_{12} = \frac{c^{(1)} \left(\frac{|\overline{\boldsymbol{V}}^{(1)}|}{|\underline{\boldsymbol{V}}^{(1)}|} \right)^{\frac{1}{2}} \left(\overline{b}^{(1)} \right)^{-\frac{\overline{a}^{(1)}}{2}}}{c^{(2)} \left(\frac{|\overline{\boldsymbol{V}}^{(2)}|}{|\underline{\boldsymbol{V}}^{(2)}|} \right)^{\frac{1}{2}} \left(\overline{b}^{(2)} \right)^{-\frac{\overline{a}^{(2)}}{2}}} \times \frac{p(M_1)}{p(M_2)}$$

$\Rightarrow$ Posterior odds ratio depends on the prior odds ratio and contains rewards for model fit, coherency between prior and data information and parsimony

Model comparison with noninformative priors

Important Rule

- $\Rightarrow$ **Noninformative priors** acceptable for parameters *common to all models*
- $\Rightarrow$ **Informative, proper priors** required for all *model-specific* parameters

- $\Rightarrow$ Prior for $h^{(1)}, h^{(2)}$: noninformative is **fine** (common to both models); $\underline{a}^{(1)} = \underline{a}^{(2)} = 0$ still yields sensible PO_{12}
- $\Rightarrow$ Noninformative priors for $\beta^{(j)}$ **cause problems** when $K_1 \neq K_2$; e.g., $\underline{\boldsymbol{V}}^{(j)} = \tau \boldsymbol{I}_{K_j}$, $\tau \to \infty$, since $|\underline{\boldsymbol{V}}^{(j)}|$ depends on K_j, it does not cancel across models
- $\Rightarrow$ Consequence:

Condition		Outcome
$K_1 < K_2$	$PO_{12} \to \infty$	(**always favours** M_1)
$K_1 > K_2$	$PO_{12} \to 0$	(**always favours** M_2)

Prediction

⇒ Want to predict:

$$y^* = \boldsymbol{x}^{*\prime}\boldsymbol{\beta} + \varepsilon^*$$

⇒ Remember, prediction is based on:

$$p\left(y^* \mid \boldsymbol{y}\right) = \int \int p\left(y^* \mid \boldsymbol{y}, \boldsymbol{\beta}, h\right) p(\boldsymbol{\beta}, h \mid \boldsymbol{y}) d\boldsymbol{\beta} dh$$

⇒ The resulting predictive:

$$y^* \mid \boldsymbol{y} \sim t\left(\boldsymbol{x}^{*\prime}\overline{\boldsymbol{\beta}}, \frac{\overline{b}}{\overline{a}}\left\{1 + \boldsymbol{x}^{*\prime}\overline{\boldsymbol{V}}\boldsymbol{x}^*\right\}, 2\overline{a}\right)$$

⇒ Model comparison, prediction and posterior inference about $\boldsymbol{\beta}$ can all be done analytically

⇒ So no need for posterior simulation in this model

⇒ However, let us illustrate Monte Carlo integration in this model

Monte Carlo integration

⇒ Remember the basic LLN we used for Monte Carlo integration

⇒ Let $\beta^{(s)}$, for $s = 1, .., S$, be a random sample from $p(\beta \mid \boldsymbol{y})$ and $g(\bullet)$ be any function and define

$$\widehat{g}_S = \frac{1}{S} \sum_{s=1}^{S} g\left(\beta^{(s)}\right)$$

⇒ then $\widehat{g}_S$ converges to $\mathbb{E}\left[g(\beta) \mid \boldsymbol{y}\right]$ as S goes to infinity

⇒ How would you write a computer program which did this?

Monte Carlo integration in action

- ⇒ *Step 1:* Take a random draw, $\beta^{(s)}$ from the posterior for β using a random number generator for the multivariate t distribution
- ⇒ *Step 2:* Calculate $g\left(\beta^{(s)}\right)$ and keep this result
- ⇒ *Step 3:* Repeat Steps 1 and 2 S times
- ⇒ *Step 4:* Take the average of the S draws $g\left(\beta^{(1)}\right), \ldots, g\left(\beta^{(S)}\right)$
- ⇒ These steps will yield an estimate of $\mathbb{E}\left[g(\beta) \mid \boldsymbol{y}\right]$ for any function of interest
- ⇒ Remember: Monte Carlo integration yields only an approximation for $\mathbb{E}\left[g(\beta) \mid \boldsymbol{y}\right]$ (since you cannot set $S = \infty$)
- ⇒ By choosing S, can control the degree of approximation error

Monte Carlo integration in action (cont.)

In **R** this could look like this (you will do something similar in Matlab)

```
tXX <- crossprod(X)                    # Compute X'X
tXy <- crossprod(X,y)                  # Compute X'y
tyy <- crossprod(y)                    # Compute y'y

# Posterior moments
V_po.inv     <- (tXX + V_pr.inv)       # Posterior precision
V_po         <- solve(V_po.inv)        # Posterior variance-covariance
Vchol_po     <- t(chol(V_po))          # Cholesky factor of posterior VC
beta_po      <- V_po%*%(tXy + V_pr.inv%*%b_pr) # Posterior mean

a_po         <- s_pr + N/2                      # Posterior DoF
b_po         <- S_pr + (tyy + t(beta_pr)%*%V_pr.inv%*%beta_pr
                  -  t(beta_po)%*%V_po.inv%*%beta_po)/2 # Posterior scaling
```

Monte Carlo integration in action (cont.)

In **R** this could look like this (you will do something similar in Matlab)

```
# Monte carlo integration
nsave <- 10000 # No. of draws

# Produce draws for h from the Gamma distribution
h_store <- rgamma(nsave,a_po,b_po)

# Produce draws for beta from the conditional Normal distribution beta|h
beta_store <- matrix(NA,nsave,K)
for(irep in 1:nsave)
{
  beta_store[irep,] <- beta_po +
              sqrt(1/h_store[irep])*Vchol_po %*% rnorm(K)
}

# Compute posterior mean of each element in beta
beta_pm <- apply(beta_store, 2, mean)
```

$\Rightarrow$ Note: By integrating over h, which is Gamma-distributed, β will be t-distributed

An empirical illustration

Empirical illustration

⇒ Dataset: $N = 546$ houses sold in Windsor, Canada (1987)

⇒ Dependent variable: y_i = sales price (Canadian dollars)

⇒ Regressors:

- x_{i2}: lot size (square feet)
- x_{i3}: number of bedrooms
- x_{i4}: number of bathrooms
- x_{i5}: number of storeys

Setup of empirical illustration

⇒ Results shown for both **informative** and **noninformative** priors

⇒ Our prior implies statements of the form "if we compare two houses which are identical except the first house has one bedroom more than the second, then we expect the first house to be worth $\$5,000$ more than the second"; this yields prior mean, then choose large prior variance to indicate prior uncertainty

⇒ Tables on the next few slides present some empirical results (textbook has lots of discussion of how you would interpret them)

⇒ **95% HPDI** = Highest Posterior Density Interval: shortest interval $[a, b]$ such that

$$p(a \leq \beta_j \leq b \mid \boldsymbol{y}) = 0.95$$

Comparing prior and posterior means for β

	Prior	Posterior	
	Informative	Noninformative prior	Informative prior
β_1	0 (10,000)	−4,009.55 (3,593.16)	−4,035.05 (3,530.16)
β_2	10 (5)	5.43 (0.37)	5.43 (0.37)
β_3	5,000 (2,500)	2,824.61 (1,211.45)	2,886.81 (1,184.93)
β_4	10,000 (5,000)	17,105.17 (1,729.65)	16,965.24 (1,708.02)
β_5	10,000 (5,000)	7,634.90 (1,005.19)	7,641.23 (997.02)

Standard deviations in parentheses

Model comparison involving β

	$p(\beta_j > 0 \mid \boldsymbol{y})$	95% HPDI	PO for $\beta_j = 0$
Informative Prior			
β_1	0.13	$[-10{,}957; 2{,}887]$	4.14
β_2	1.00	$[4.71; 6.15]$	2.25×10^{-39}
β_3	0.99	$[563.5; 5{,}210.1]$	0.39
β_4	1.00	$[13{,}616; 20{,}314]$	1.72×10^{-19}
β_5	1.00	$[5{,}686; 9{,}596]$	1.22×10^{-11}
Noninformative Prior			
β_1	0.13	$[-11{,}055; 3{,}036]$	—
β_2	1.00	$[4.71; 6.15]$	—
β_3	0.99	$[449.3; 5{,}200]$	—
β_4	1.00	$[13{,}714; 20{,}497]$	—
β_5	1.00	$[5{,}664; 9{,}606]$	—

PO = posterior odds; — not defined under noninformative prior

Posterior results for β_2 (analytical versus Monte Carlo)

Method	Mean	Std. dev.	Numerical std. error
Analytical	5.4316	0.3662	
Monte Carlo (S replications)			
$S = 10$	5.3234	0.2889	0.0913
$S = 100$	5.4877	0.4011	0.0401
$S = 1{,}000$	5.4209	0.3727	0.0118
$S = 10{,}000$	5.4330	0.3677	0.0037
$S = 100{,}000$	5.4323	0.3664	0.0012

As S increases, MC estimates converge to the analytical solution and the numerical standard error shrinks

Summary so far

- ⇒ So far: Normal linear regression with **natural conjugate prior** ⇒ posterior, marginal likelihood and predictive distributions all available in **closed form**
- ⇒ Other priors / more complex models ⇒ **no analytical results**
- ⇒ Next: Extensions of the regression model requiring **Gibbs sampler** for posterior computation
  - → Gibbs Sampler is a special case of **Markov Chain Monte Carlo (MCMC)**
  - → Keep the Normal linear regression model (under the classical assumptions) as before
  - → Likelihood function same as before
  - → Parameters of model will still be β and h

The independent Normal-Gamma prior

⇒ Before we had conjugate prior where $p(\boldsymbol{\beta} \mid h)$ was Normal density and $p(h)$ Gamma density

⇒ Now use similar prior, but assume prior **independence** between $\boldsymbol{\beta}$ and h

$$p(\boldsymbol{\beta}, h) = p(\boldsymbol{\beta})\, p(h)$$

⇒ $p(\boldsymbol{\beta})$ is Normal and $p(h)$ is Gamma:

$$\boldsymbol{\beta} \sim \mathcal{N}\left(\underline{\boldsymbol{\beta}}, \underline{\boldsymbol{V}}\right)$$
$$h \sim \mathcal{G}(\underline{a}, \underline{b})$$

⇒ Key difference: now $\underline{\boldsymbol{V}}$ is now the prior covariance matrix of $\boldsymbol{\beta}$, with conjugate prior we had $\mathbb{V}(\boldsymbol{\beta} \mid h) = h^{-1}\underline{\boldsymbol{V}}$

The posterior

- $\Rightarrow$ Posterior $\propto$ prior $\times$ likelihood
- $\Rightarrow$ Joint posterior for β and h: **no closed form** $\Rightarrow$ cannot be used directly for inference
- $\Rightarrow$ But **conditional posterior** for β given h takes a simple form:

$$\beta \mid \boldsymbol{y}, h \sim \mathcal{N}\left(\overline{\beta}, \overline{\boldsymbol{V}}\right)$$

with

$$\overline{\boldsymbol{V}} = \left(\underline{\boldsymbol{V}}^{-1} + h\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

$$\overline{\beta} = \overline{\boldsymbol{V}}\left(\underline{\boldsymbol{V}}^{-1}\underline{\beta} + h\boldsymbol{X}'\boldsymbol{y}\right)$$

The posterior (cont.)

⇒ **Conditional posterior** for h given β is Gamma:

$$h \mid \boldsymbol{y}, \beta \sim \mathcal{G}\left(\overline{a}, \overline{b}\right)$$

with

$$\overline{a} = \underline{a} + \frac{N}{2}, \qquad \overline{b} = \underline{b} + (\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)$$

⇒ Econometrician is interested in **joint posterior** $p(\beta, h \mid \boldsymbol{y})$ or marginal $p(\beta \mid \boldsymbol{y})$, **not** in the conditional posteriors

⇒ **Key point:** Since

$$p(\beta, h \mid \boldsymbol{y}) \neq p(\beta \mid \boldsymbol{y}, h)\, p(h \mid \boldsymbol{y}, \beta)$$

the conditionals **do not directly characterize** the joint posterior

Popular MCMC algorithms

The Gibbs sampler

- ⇒ The **Gibbs sampler** uses conditional posteriors to draw $\beta^{(s)}$ and $h^{(s)}$ for $s = 1, \ldots, S$ ⇒ averages yield posterior estimates, as with Monte Carlo integration
- ⇒ Powerful and widely-used tool for posterior simulation in econometric models
- ⇒ **General setup:** $\boldsymbol{\theta}$ is a p-vector of parameters with likelihood $p(\boldsymbol{y} \mid \boldsymbol{\theta})$, prior $p(\boldsymbol{\theta})$ and posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$
- ⇒ Partition $\boldsymbol{\theta}$ into B **blocks**:

$$\boldsymbol{\theta} = \left(\boldsymbol{\theta}'_{(1)}, \boldsymbol{\theta}'_{(2)}, \ldots, \boldsymbol{\theta}'_{(B)}\right)'$$

- ⇒ For example, regression model: $B = 2$ with $\boldsymbol{\theta}_{(1)} = \beta$ and $\boldsymbol{\theta}_{(2)} = h$

Intuition of the Gibbs sampler

- ⇒ **Monte Carlo integration**: Draw from $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ and average to estimate $\mathbb{E}[g(\boldsymbol{\theta}) \mid \boldsymbol{y}]$ for any function $g(\boldsymbol{\theta})$
- ⇒ **Problem:** Drawing directly from $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ is often **not possible** (e.g., not available in closed form)
- ⇒ But it often is easy to draw from **full conditional posteriors**:

$$p(\boldsymbol{\theta}_{(b)} \mid \boldsymbol{y}, \{\boldsymbol{\theta}_{(j)}\}_{j \neq b}), \quad b = 1, \ldots, B$$

- ⇒ **Gibbs sampler**: Cycle through draws from each full conditional; in turn, produces a sequence $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(S)}$ that can be averaged to estimate $\mathbb{E}[g(\boldsymbol{\theta}) \mid \boldsymbol{y}]$, just as with Monte Carlo integration

More motivation for the Gibbs sampler

- $\Rightarrow$ Regression model: $B = 2$ blocks, $\boldsymbol{\theta} = \{\boldsymbol{\beta}, h\}$
- $\Rightarrow$ Suppose we have one draw $\boldsymbol{\beta}^{(0)}$ from $p(\boldsymbol{\beta} \mid \boldsymbol{y})$
- $\Rightarrow$ Then since

$$p(\boldsymbol{\beta}, h \mid \boldsymbol{y}) = p(h \mid \boldsymbol{y}, \boldsymbol{\beta})\, p(\boldsymbol{\beta} \mid \boldsymbol{y})$$

  a draw from $p(h \mid \boldsymbol{y}, \boldsymbol{\beta}^{(0)})$ is a valid draw of h; call this $h^{(1)}$
- $\Rightarrow$ Similarly, since

$$p(\boldsymbol{\beta}, h \mid \boldsymbol{y}) = p(\boldsymbol{\beta} \mid \boldsymbol{y}, h)\, p(h \mid \boldsymbol{y})$$

  a draw from $p(\boldsymbol{\beta} \mid \boldsymbol{y}, h^{(1)})$ is a valid draw of $\boldsymbol{\beta}$; call this $\boldsymbol{\beta}^{(1)}$
- $\Rightarrow$ Hence $\left\{\boldsymbol{\beta}^{(1)}, h^{(1)}\right\}$ is a valid draw from $p(\boldsymbol{\beta}, h \mid \boldsymbol{y})$ and continuing yields $\left\{\boldsymbol{\beta}^{(s)}, h^{(s)}\right\}$ for $s = 1, \ldots, S$

More motivation for the Gibbs sampler (cont.)

- $\Rightarrow$ Hence, if you can successfully find $\beta^{(0)}$, then sequentially drawing $p(h \mid \boldsymbol{y}, \beta)$ and $p(\beta \mid \boldsymbol{y}, h)$ will give valid draws from posterior
- $\Rightarrow$ Problem with above strategy is that it is not possible to find such initial $\beta^{(0)}$
- $\Rightarrow$ If we knew how to easily take random draws from $p(\beta \mid \boldsymbol{y})$, we could use this and $p(h \mid \beta, \boldsymbol{y})$ to do Monte Carlo integration and have no need for Gibbs sampling
- $\Rightarrow$ However, it can be shown that subject to weak conditions, the initial draw $\beta^{(0)}$ does not matter: Gibbs sampler will converge to a sequence of draws from $p(\beta, h \mid \boldsymbol{y})$
- $\Rightarrow$ In practice, choose $\beta^{(0)}$ in some manner and then run the Gibbs sampler for S replications
- $\Rightarrow$ Discard S_0 initial draws (the so-called "**burn-in**") and remaining S_1 used to estimate $\mathbb{E}[g(\boldsymbol{\theta}) \mid \boldsymbol{y}]$

Why is Gibbs sampling so useful?

- $\Rightarrow$ In Normal linear regression model with independent Normal-Gamma prior, Gibbs sampler is easy
- $\Rightarrow$ $p(\beta \mid \boldsymbol{y}, h)$ is Normal and $p(h \mid \boldsymbol{y}, \beta)$ is Gamma (easy to draw from)
- $\Rightarrow$ Huge number of other models have hard joint posterior, but easy posterior conditionals
- $\Rightarrow$ Tobit, probit, stochastic frontier model, Markov switching model, threshold autoregressive, smooth transition threshold autoregressive, other regime switching models, state space models, some semiparametric regression models, etc.
- $\Rightarrow$ What if the full posterior conditionals do not have simple form?
- $\Rightarrow$ Many other algorithms exist for handling general cases, Metropolis-Hastings algorithm is most popular

The Metropolis-Hastings (MH) algorithm

- ⇒ Another popular MCMC algorithm, useful when **Gibbs sampling is not feasible**
- ⇒ General notation: $\boldsymbol{\theta}$ is a vector of parameters with likelihood $p(\boldsymbol{y} \mid \boldsymbol{\theta})$, prior $p(\boldsymbol{\theta})$ and posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$
- ⇒ Key idea: Draw from a convenient **candidate generating density** $q(\boldsymbol{\theta}^{(s-1)}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}^*$ denotes a candidate draw whose density depends on $\boldsymbol{\theta}^{(s-1)}$
- ⇒ Since we draw from $q(\boldsymbol{\theta}^{(s-1)}; \boldsymbol{\theta})$ instead of $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, we have to **correct for this** via an **acceptance probability**, with only some candidate draws being accepted

MH algorithm: Steps

1. Choose starting value $\boldsymbol{\theta}^{(0)}$
2. Draw candidate $\boldsymbol{\theta}^*$ from $q(\boldsymbol{\theta}^{(s-1)}; \boldsymbol{\theta})$
3. Compute acceptance probability $\alpha(\boldsymbol{\theta}^{(s-1)}, \boldsymbol{\theta}^*)$
4. Set $\boldsymbol{\theta}^{(s)} = \begin{cases} \boldsymbol{\theta}^* & \text{with probability } \alpha(\boldsymbol{\theta}^{(s-1)}, \boldsymbol{\theta}^*) \\ \boldsymbol{\theta}^{(s-1)} & \text{with probability } 1 - \alpha(\boldsymbol{\theta}^{(s-1)}, \boldsymbol{\theta}^*) \end{cases}$
5. Repeat Steps 2 to 4 for $s = 1, \ldots, S (= S_0 + S_1)$; discard S_0 burn-in draws
6. Estimate $\mathbb{E}[g(\boldsymbol{\theta}) \mid \boldsymbol{y}]$ by averaging $g(\boldsymbol{\theta}^{(S_0+1)}), \ldots, g(\boldsymbol{\theta}^{(S)})$

The acceptance probability

⇒ The acceptance probability is:

$$\alpha\left(\boldsymbol{\theta}^{(s-1)}, \boldsymbol{\theta}^{*}\right) = \min\left[\frac{p(\boldsymbol{\theta}^{*} \mid \boldsymbol{y})\; q(\boldsymbol{\theta}^{*}; \boldsymbol{\theta}^{(s-1)})}{p(\boldsymbol{\theta}^{(s-1)} \mid \boldsymbol{y})\; q(\boldsymbol{\theta}^{(s-1)}; \boldsymbol{\theta}^{*})},\; 1\right]$$

⇒ Intuition: Candidate $\boldsymbol{\theta}^{*}$ is **always accepted** if it has higher posterior density than $\boldsymbol{\theta}^{(s-1)}$; otherwise accepted with probability < 1 (see textbook pp. 93–94)

⇒ As with Gibbs sampling, discard S_0 **burn-in** draws to remove dependence on starting value $\boldsymbol{\theta}^{(0)}$

⇒ Note: The **Gibbs sampler** is a special case of MH where the **full conditionals** serve as candidate generating densities for each block $\boldsymbol{\theta}_{(b)}$, resulting in $\alpha(\boldsymbol{\theta}_{(b)}^{(s-1)}, \boldsymbol{\theta}_{(b)}^{*}) = 1$, with every candidate draw being accepted

Candidate generating density: Independence chain

$\Rightarrow$ Candidate density does not depend on $\boldsymbol{\theta}^{(s-1)}$:

$$q(\boldsymbol{\theta}^{(s-1)}; \boldsymbol{\theta}) = q(\boldsymbol{\theta})$$

$\Rightarrow$ Useful when a convenient **approximation to the posterior** exists

$\Rightarrow$ Acceptance probability simplifies to:

$$\alpha\left(\boldsymbol{\theta}^{(s-1)}, \boldsymbol{\theta}^{*}\right) = \min\left[\frac{p(\boldsymbol{\theta}^{*} \mid \boldsymbol{y})\; q(\boldsymbol{\theta}^{(s-1)})}{p(\boldsymbol{\theta}^{(s-1)} \mid \boldsymbol{y})\; q(\boldsymbol{\theta}^{*})},\; 1\right]$$

Candidate generating density: Random walk chain

- ⇒ **Random walk chain** popular in DSGE models; useful when no good posterior approximation exists
- ⇒ Candidate draw:

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^{(s-1)} + \boldsymbol{w}$$

  where $\boldsymbol{w}$ is the **increment random variable**
- ⇒ Acceptance probability simplifies to:

$$\alpha\left(\boldsymbol{\theta}^{(s-1)}, \boldsymbol{\theta}^*\right) = \min\left[\frac{p(\boldsymbol{\theta}^* \mid \boldsymbol{y})}{p(\boldsymbol{\theta}^{(s-1)} \mid \boldsymbol{y})},\ 1\right]$$

- ⇒ Common choice: $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, so $q(\boldsymbol{\theta}^{(s-1)}; \boldsymbol{\theta}) = f_N(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(s-1)}, \Sigma)$
- ⇒ Researcher must choose $\Sigma = c\Omega$, where Ω is an estimate of the posterior covariance (e.g., inverse Hessian at posterior mode) and c is a scalar tuned so that the acceptance rate is **neither too high nor too low** (rule of thumb: between 0.2 and 0.4)

Example of independence chain MH

⇒ Consider a simple AR(1) model:

$$y_t = \beta y_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, h^{-1})$$

⇒ We have prior knowledge that y_t is **persistent but stationary**: $\beta \in (0, 1)$

⇒ This may motivate choosing a **Beta prior**: $\beta \sim \mathcal{B}(\underline{c}, \underline{d})$

⇒ **Problem**: Gaussian likelihood $\times$ Beta prior results in a posterior $p(\beta \mid \boldsymbol{y})$ with **no closed form**

⇒ In addition, no well-known conditionals available; Gibbs sampling not possible

⇒ Use independence chain MH with a **Uniform candidate density** $q(\beta) = \mathcal{U}(0, 1)$

Sketch of independence chain MH implementation in R

```r
# Step 1: Propose candidate draw from U(0,1)
beta.prop <- runif(1, 0, 1)

# Step 2: Log likelihood at proposed and current draw
lik.prop <- sum(dnorm(y, X%*%beta.prop, sqrt(1/h.draw), log=TRUE))
lik.acc  <- sum(dnorm(y, X%*%beta.acc,  sqrt(1/h.draw), log=TRUE))

# Step 3: Log posterior: log p(beta|y) = log p(y|beta) + log p(beta)
cond.post.prop <- lik.prop + dbeta(beta.prop, c0, d0, log=TRUE)
cond.post.acc  <- lik.acc  + dbeta(beta.acc,  c0, d0, log=TRUE)

# Step 4: Log acceptance probability and coin flip
log.alpha <- cond.post.prop - cond.post.acc

if(log.alpha > log(runif(1))){ # Simulate a coin flip with runif(1)
  beta.acc <- beta.prop        # Accept: move to candidate beta.prop
  accept   <- accept + 1
}                              # Reject: stay at beta.acc (implicit)
```

Bayesian model averaging (BMA)

BMA in a nutshell

- ⇒ BMA can be used with any set of models; here use it in a Big Data regression (many explanatory variables)
- ⇒ **Model selection:** Choose a single model and present estimates or forecasts based on it
- ⇒ **Model averaging:** Take a weighted average of estimates or forecasts from all models with weights given by $p(M_r \mid \boldsymbol{y})$, incorporating uncertainty about which model generated the data
- ⇒ Allows for a formal treatment of **model uncertainty**
- ⇒ Let M_r, for $r = 1, \ldots, R$, denote R competing models, for any parameter ϕ or forecast of interest, the rules of probability imply:

$$p(\phi \mid \boldsymbol{y}) = \sum_{r=1}^{R} p(\phi \mid \boldsymbol{y}, M_r)\, p(M_r \mid \boldsymbol{y})$$

Model space of a big data regression

⇒ Let $\boldsymbol{X}^{(r)}$ be $N \times K_r$ submatrix of X; each model takes the form:

$$\boldsymbol{y} = \alpha\, \boldsymbol{\iota}_N + \boldsymbol{X}^{(r)}\beta^{(r)} + \varepsilon$$

with ι_N being $N \times 1$ vector of ones (intercept included in all models)

⇒ With K potential regressors there are $R = 2^K$ possible models (each regressor is either included or excluded; 2 options per regressor)

⇒ **Computational challenge**: For $K = 41$, $R = 2^{41} \approx 2$ trillion models (at 0.001 seconds per model, estimation would take **over 100 years**)

⇒ Use a **natural conjugate prior** to maximize estimation speed per model (closed-form solutions), combined with MCMC model composition to explore the model space without estimating every model

BMA priors: The g-prior

- ⇒ We want a prior for model M_r that is **informative** (valid marginal likelihoods), **objective** (minimal subjective input), and **automatic** (no individual choice per model)
- ⇒ The **g-prior** (Zellner, 1986) satisfies all three
- ⇒ Prior shrinks coefficients towards zero:

$$\underline{\beta}^{(r)} = \mathbf{0}, \qquad \underline{\boldsymbol{V}}^{(r)} = \left(g\boldsymbol{X}^{(r)\prime}\boldsymbol{X}^{(r)}\right)^{-1}$$

- ⇒ Prior information that $\beta^{(r)} = \mathbf{0}$ takes the **same form** as data information
- ⇒ g is a scalar controlling the **relative weight** of prior vs data information:
  - → $g = 1$: equal weight to prior and data; $g = 0.01$: prior receives 1% of the weight of data; g can also be chosen via rules of thumb or treated as unknown
- ⇒ Noninformative prior typically used for h

BMA posterior

⇒ With natural conjugate prior, model M_r yields **analytical results**

- → Posterior: Normal-Gamma
- → Marginal likelihood: Closed form
- → Predictive density: Student-t

⇒ For each model, everything needed can be computed **quickly**

⇒ However, for $K > 20$, evaluating all 2^K models remains **computationally infeasible**

BMA computation: MC^3 algorithm

⇒ **MC^3** (Markov Chain Monte Carlo Model Composition): Draws models $M^{(s)}$, for $s = 1, \ldots, S$, instead of parameters

⇒ For any quantity of interest ϕ:

$$\hat{\phi} = \frac{1}{S} \sum_{s=1}^{S} \mathbb{E}\left(\phi \mid \boldsymbol{y}, M^{(s)}\right) \xrightarrow{S \to \infty} \mathbb{E}(\phi \mid \boldsymbol{y})$$

⇒ Frequencies with which models are drawn converge to **posterior model probabilities**: if M_i drawn A times and M_j drawn B times, A/B converges to the Bayes factor BF_{ij}

⇒ Discard initial draws as **burn-in**

MC^3: How are models drawn?

⇒ Given current model $M^{(s-1)}$, propose candidate M^* drawn randomly with equal probability from the following set of models:

- → $M^{(s-1)}$
- → All models deleting one regressor from $M^{(s-1)}$
- → All models adding one regressor to $M^{(s-1)}$

⇒ Accept M^* with probability:

$$\alpha\left(M^{(s-1)}, M^*\right) = \min\left[\frac{p(\boldsymbol{y} \mid M^*)\, p(M^*)}{p(\boldsymbol{y} \mid M^{(s-1)})\, p(M^{(s-1)})}, 1\right]$$

⇒ Set $M^{(s)} = M^*$ if accepted, else $M^{(s)} = M^{(s-1)}$

⇒ Can be shown MC^3 converges to the true **BMA posterior**

Classic BMA application

Determinants of economic growth

- ⇒ Classic cross-country growth regression: Why do some countries grow faster than others?
- ⇒ Dependent variable: Average GDP per capita growth, 1960–1992
- ⇒ $K = 41$ potential explanatory variables (education, investment, governance, institutions, trade, etc.), all standardized
- ⇒ $N = 72$ countries (cross-sectional observations)
- ⇒ This is Big Data: **big K, small N**

Classic BMA application (cont.)

Determinants of economic growth

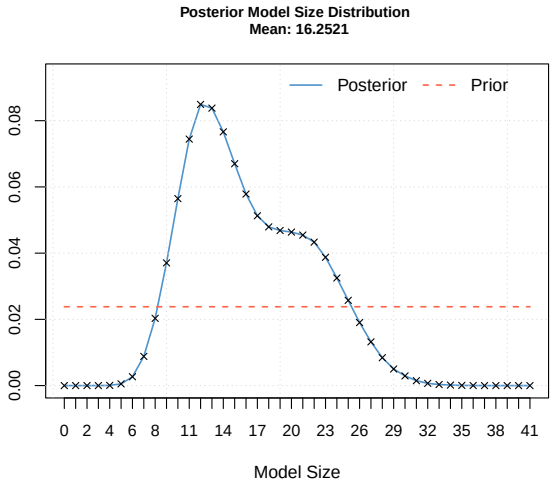
- ⇒ Cross-country growth regression data set with $N = 72$ and $K = 41$
- ⇒ Use common recommendation to set $g = \frac{1}{N}$ (`UIP`; unit information prior) and $g = \frac{1}{K^2}$ (`RIC`; risk inflation criteria)
- ⇒ Run MC^3 algorithm for $2,200,000$ draws, discarding first $200,000$ as burn-in
- ⇒ Is this enough draws?
- ⇒ Convergence diagnostic: Calculate posterior model probabilities analytically and using MC^3 and compare
- ⇒ Model selection puts all weight on this single model — ignoring huge amount of model uncertainty

Replicating the analysis in R

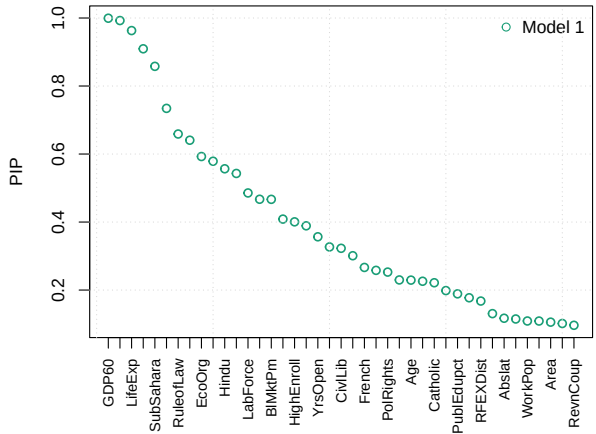
Bayesian Model Averaging can be implemented using the `BMS` package

```r
# Load package
library(BMS)

# Load data
data(datafls)

# Estimate BMA model
bma.fls <- bms(
  datafls,
  burn = 200000,      # burn-in draws
  iter = 2000000,     # posterior draws
  g    = "UIP"        # unit information prior
)
```

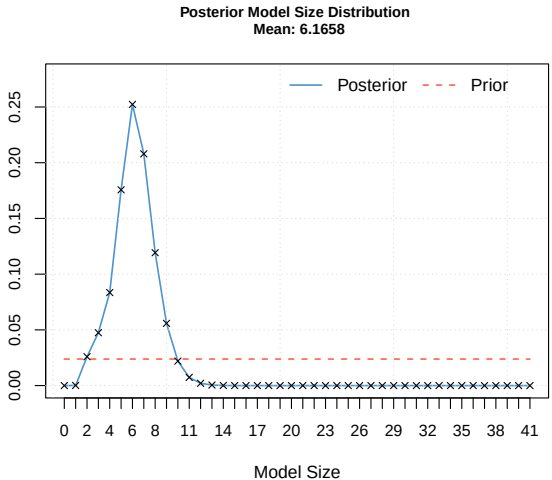
Posterior model size distribution for `g = UIP`



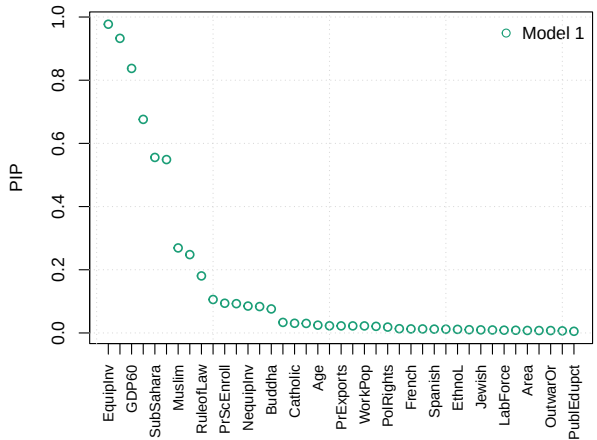
Posterior model size distribution for `g = UIP`



Posterior model size distribution for `g = RIC`



Posterior model size distribution for `g = RIC`



BMA application

- ⇒ Next table presents results:
- ⇒ Posterior mean and standard deviation for each explanatory variable using BMA and BMS
- ⇒ Rule of thumb: If an estimate (posterior mean) more than two standard deviations from zero likely to be important
- ⇒ Column labelled ”Prob.” = probability that the corresponding explanatory variable should be included (= proportion of models drawn by MC^3 which contain the corresponding explanatory variable)
- ⇒ BMS ensures parsimony by choosing 14 variables
- ⇒ By ignoring model uncertainty estimates are more precise (smaller standard dev.)
- ⇒ BMA ensures parsimony by averaging over many small models
- ⇒ Average number of explanatory variables in a model drawn by MC^3 is 16.25

Point Estimates and Standard Devs of Regression Coefficients (Mean and standard deviations multiplied by 100)					
	BMA			BMS	
Explanatory Variable	Prob.	Mean.	St. Dev.	Mean	St. Dev.
Primary School Enrolment	0.207	0.104	0.234	0.048	0.018
Life expectancy	0.933	0.961	0.392	0.090	0.020
GDP level in 1960	0.999	−1.425	0.278	−1.463	0.193
Fraction GDP in Mining	0.459	0.147	0.181	0.322	0.108
Degree of Capitalism	0.457	0.151	0.183	0.387	0.094
No. Years Open Economy	0.513	0.260	0.283	0.557	0.138
% Pop. Speaking English	0.069	−0.011	0.047	–	–
% Pop. Speak. For. Lang.	0.068	0.012	0.059	–	–
Exchange Rate Distortions	0.082	−0.017	0.070	–	–
Equipment Investment	0.923	0.552	0.236	0.548	0.128
Non-equipment Investment	0.434	0.136	0.174	0.347	0.099
St. Dev. of Black Mkt. Prem.	0.048	−0.006	0.037	–	–
Outward Orientation	0.037	−0.003	0.029	–	–

Point Estimates and Standard Devs of Regression Coefficients (Mean and standard deviations multiplied by 100)					
	BMA			BMS	
Explanatory Variable	Prob.	Mean.	St. Dev.	Mean	St. Dev.
Black Market Premium	0.179	−0.040	0.097	–	–
Area	0.030	−0.001	0.021	–	–
Latin America	0.215	−0.082	0.191	–	–
Sub-Saharan Africa	0.738	−0.473	0.347	−0.543	0.124
Higher Education Enrolment	0.046	−0.008	0.056	–	–
Public Education Share	0.032	−0.001	0.024	–	–
Revolutions and Coups	0.031	−0.001	0.023	–	–
War	0.075	−0.014	0.062	–	–

Posteror Estimates and Standard Devs of Regression Coefficients					
	Bayesian Model Averaging			Single Best Model	
Explanatory Variable	Prob.	Mean	St. Dev.	Mean	St. Dev.
Political Rights	0.094	−0.028	0.107	–	–
Civil Liberties	0.131	−0.050	0.015	−0.284	0.176
Latitude	0.041	0.001	0.052	–	–
Age	0.085	−0.015	0.058	–	–
British Colony	0.041	−0.003	0.032	–	–
Fraction Buddhist	0.196	0.047	0.109	–	–
Fraction Catholic	0.128	−0.011	0.121	–	–
Fraction Confucian	0.990	0.493	0.127	0.503	0.090
Ethnolinguistic Fractionalization	0.060	0.010	0.056	–	–
French Colony	0.049	0.007	0.040	–	–

Posteror Estimates and Standard Devs of Regression Coefficients					
	Bayesian Model Averaging			Single Best Model	
Explanatory Variable	Prob.	Mean	St. Dev.	Mean	St. Dev.
Fraction Hindu	0.126	−0.035	0.120	–	–
Fraction Jewish	0.037	−0.002	0.028	–	–
Fraction Muslim	0.640	0.025	0.023	0.295	0.093
Primary Exports	0.100	−0.029	0.105	−0.352	0.136
Fraction Protestant	0.455	−0.143	0.178	−0.277	0.098
Rule of Law	0.489	0.244	0.279	0.563	0.134
Spanish Colony	0.058	0.010	0.068	–	–
Population Growth	0.037	0.005	0.048	–	–
Ratio Workers to Population	0.045	−0.005	0.043	–	–
Size of Labor Force	0.075	0.018	0.097	–	–

Summary

- ⇒ This lecture shows how Bayesian ideas work in familiar context (regression model)
- ⇒ Occasionally analytical results are available (no need for posterior simulation)
- ⇒ Usually posterior simulation is required
- ⇒ Monte Carlo integration is simplest, but rarely possible to use it
- ⇒ Gibbs sampling (and related MCMC) methods can be used for estimation and prediction for a wide variety of models
- ⇒ Metropolis-Hastings algorithms popular and can be combined with Gibbs sampling (Metropolis-within-Gibbs)
- ⇒ Note: There are methods for calculating marginal likelihoods using Gibbs sampler output