Bayesian Inference in the Normal Linear Regression Model

Bayesian Analysis of the Normal Linear Regression Model

- Now see how general Bayesian theory of overview lecture works in familiar regression model
- In lecture, I will focus on multiple regression under classical assumptions (independent errors, homoskedasticity, etc.)
- Bayesian methods for freeing up classical assumptions exist (see Chapter 6 of my textbook)

The Regression Model

- Assume k explanatory variables, x_{i1},..,x_{ik} for $i = 1, .., N$ and regression model:

$$y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i.$$

- Note x_{i1} is implicitly set to 1 to allow for an intercept.
- Matrix notation:

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

- ε is $N \times 1$ vector stacked in same way as y

- β is $k \times 1$ vector
- X is $N \times k$ matrix

$$X = \begin{bmatrix} 1 & x_{12} & . & . & x_{1k} \\ 1 & x_{22} & . & . & x_{2k} \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_{N2} & . & . & x_{Nk} \end{bmatrix}$$

- Regression model can be written as:

$$y = X\beta + \varepsilon.$$

The Likelihood Function

- Likelihood can be derived under the classical assumptions:
- ε is $N(0_N, h^{-1}I_N)$ where $h = \sigma^{-2}$.
- All elements of X are either fixed (i.e. not random variables).
- Exercise 10.1, Bayesian Econometric Methods shows that likelihood function can be written in terms of OLS quantities:

$$\nu = N - k,$$

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

$$s^2 = \frac{\left(y - X\widehat{\beta}\right)' \left(y - X\widehat{\beta}\right)}{\nu}$$

- Likelihood function:

$$p(y|\beta, h) = \frac{1}{(2\pi)^{\frac{N}{2}}}$$

$$\left\{ h^{\frac{k}{2}} \exp\left[-\frac{h}{2}\left(\beta - \widehat{\beta}\right)' X'X \left(\beta - \widehat{\beta}\right)\right]\right\} \left\{ h^{\frac{\nu}{2}} \exp\left[-\frac{h\nu}{2s^{-2}}\right]\right\}$$

The Prior

- Common starting point is natural conjugate Normal-Gamma prior
- β conditional on h is now multivariate Normal:

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

- Prior for error precision h is Gamma

$$h \sim G(\underline{s}^{-2}, \underline{\nu})$$

- $\underline{\beta}, \underline{V}, \underline{s}^{-2}$ and $\underline{\nu}$ are prior hyperparameter values chosen by the researcher
- Notation: Normal-Gamma distribution

$$\beta, h \sim NG\left(\underline{\beta}, \underline{V}, \underline{s}^{-2}, \underline{\nu}\right).$$

The Posterior

- Multiply likelihood by prior and collecting terms (see Bayesian Econometrics Methods Exercise 10.1).
- Posterior is

$$\beta, h|y \sim NG\left(\overline{\beta}, \overline{V}, \overline{s}^{-2}, \overline{\nu}\right)$$

- where

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

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

$$\overline{\nu} = \underline{\nu} + N$$

and $\overline{s}^{-2}$ is defined implicitly through

$$\overline{\nu}\overline{s}^2 = \underline{\nu}\underline{s}^2 + \nu s^2 + \left(\widehat{\beta} - \underline{\beta}\right)' \left[\underline{V} + \left(X'X\right)^{-1}\right]^{-1} \left(\widehat{\beta} - \underline{\beta}\right).$$

- Marginal posterior for β: multivariate t distribution:

$$\beta|y \sim t\left(\overline{\beta}, \overline{s}^2\overline{V}, \overline{\nu}\right),$$

- Useful results for estimation:

$$E(\beta|y) = \overline{\beta}$$

- $$var(\beta|y) = \frac{\overline{\nu}\overline{s}^2}{\overline{\nu}-2}\overline{V}.$$

- 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 (for now) β is a scalar, $h = 1$ and its prior mean is zero
- Prior shrinkage: Posterior mean is pulled towards zero ("shrinkage")
- Commonly done to avoid over-fitting/over-parameterization problems
- Strength of prior shrinkage controlled through prior variance:
- If $\underline{V}$ is small, then strong prior information β is near 0.
- E.g. If $\underline{V} = 0.0001$ then $\Pr\left(-0.0196 \leq \beta \leq 0.0196\right) = 0.95$
- If $\underline{V}$ is big then prior becomes more non-informative
- If $\underline{V} = 100$ then $\Pr\left(-19.6 \leq \beta \leq 19.6\right) = 0.95$
- Note: exactly what "small" and "large" means depends on the empirical application and units of measurement of data

A Noninformative Prior

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

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

- With this choice we get OLS results.

$$\beta, h|y \sim NG\left(\overline{\beta}, \overline{V}, \overline{s}^{-2}, \overline{\nu}\right)$$

- where

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

$$\overline{\beta} = \widehat{\beta}$$

$$\overline{\nu} = N$$

$$\overline{\nu}\overline{s}^2 = \nu s^2.$$

Model Comparison

- Case 1: M_1 imposes a linear restriction and M_2 does not (nested).
- Case 2: $M_1 : y = X_1\beta_{(1)} + \varepsilon_1$ and $M_2 : y = X_2\beta_{(2)} + \varepsilon_2$, where X_1 and X_2 contain different explanatory variables (non-nested).
- Both cases can be handled by defining models as (for $j = 1, 2$):

$$M_j : y_j = X_j\beta_{(j)} + \varepsilon_j$$

- Non-nested model comparison involves $y_1 = y_2$.
- 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

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

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

$$y - x_3 = \beta_1 + \beta_2 x_2 + \varepsilon$$

- M_1 has dependent variable $y - x_3$ and intercept and x_2 are explanatory variables

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

$$p(y_j|M_j) = c_j \left(\frac{|\overline{V}_j|}{|\underline{V}_j|} \right)^{\frac{1}{2}} \left(\overline{\nu}_j \overline{s}_j^2 \right)^{-\frac{\overline{\nu}_j}{2}}$$

- c_j is constant depending on prior hyperparameters, etc.
-

$$PO_{12} = \frac{c_1 \left(\frac{|\overline{V}_1|}{|\underline{V}_1|} \right)^{\frac{1}{2}} \left(\overline{\nu}_1 \overline{s}_1^2 \right)^{-\frac{\overline{\nu}_1}{2}} p(M_1)}{c_2 \left(\frac{|\overline{V}_2|}{|\underline{V}_2|} \right)^{\frac{1}{2}} \left(\overline{\nu}_2 \overline{s}_2^2 \right)^{-\frac{\overline{\nu}_2}{2}} p(M_2)}$$

- 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: *When comparing models using posterior odds ratios, it is acceptable to use noninformative priors over parameters which are common to all models. However, informative, proper priors should be used over all other parameters.*
- If we set $\underline{\nu}_1 = \underline{\nu}_2 = 0$. Posterior odds ratio still has a sensible interpretation.
- Noninformative prior for h_1 and h_2 is fine (these parameters common to both models)
- But noninformative priors for $\beta_{(j)}$'s cause problems which occur largely when $k_1 \neq k_2$. (Exercise 10.4 of Bayesian Econometric Methods)
- E.g. noninformative prior for $\beta_{(j)}$ based on $\underline{V}_j^{-1} = cI_{k_j}$ and letting $c \to 0$. Since $|\underline{V}_j| = \frac{1}{c^{k_j}}$ terms involving k_j do not cancel out.
- If $k_1 < k_2$, PO_{12} becomes infinite, while if $k_1 > k_2$, PO_{12} goes to zero.

Prediction

- Want to predict:

$$y^* = X^*\beta + \varepsilon^*$$

- Remember, prediction is based on:

$$p\left(y^*|y\right) = \int\int p\left(y^*|y,\beta,h\right)p(\beta,h|y)d\beta dh.$$

- The resulting predictive:

$$y^*|y \sim t\left(X^*\overline{\beta},\overline{s}^2\left\{I_T + X^*\overline{V}X^{*\prime}\right\},\overline{\nu}\right)$$

- Model comparison, prediction and posterior inference about β 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 $p(\beta|y)$ and $g\left(.\right)$ be any function and define

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

- then $\widehat{g}_S$ converges to $E\left[g(\beta)|y\right]$ as S goes to infinity.
- How would you write a computer program which did this?

- *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 $g\left(\beta^{(1)}\right), ..., g\left(\beta^{(S)}\right)$.
- These steps will yield an estimate of $E\left[g(\beta)|y\right]$ for any function of interest.
- Remember: Monte Carlo integration yields only an approximation for $E\left[g(\beta)|y\right]$ (since you cannot set $S = \infty$).
- By choosing S, can control the degree of approximation error.
- Using a CLT we can obtain 95% confidence interval for $E[g(\beta)|y]$
- Or a numerical standard error can be reported.

Empirical Illustration

- Data set on $N = 546$ houses sold in Windsor, Canada in 1987.
- $y_i =$ sales price of the i^{th} house measured in Canadian dollars,
- $x_{i2} =$ the lot size of the i^{th} house measured in square feet,
- $x_{i3} =$ the number of bedrooms in the i^{th} house,
- $x_{i4} =$ the number of bathrooms in the i^{th} house,
- $x_{i5} =$ the number of storeys in the i^{th} house.

- Example uses informative and noninformative priors.
- Textbook discusses how you might elicit a prior.
- 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.
- The following tables 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:

$$p\left(a \leq \beta_j \leq b \middle| y\right) = 0.95.$$

Prior and Posterior Means for β
(standard deviations in parentheses)

	Prior	Posterior	
	Informative	Using Noninf Prior	Using Inf 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)

Model Comparison involving β			
Informative Prior			
	$p\left(\beta_j > 0 \vert y\right)$	95% HPDI	Posterior Odds for $\beta_j = 0$
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			
	$p\left(\beta_j > 0 \vert y\right)$	95% HPDI	Posterior Odds for $\beta_j = 0$
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	$[5,664, 9,606]$	—

Posterior Results for β_2 Calculated Various Ways

	Mean	Standard Deviation	Numerical St. Error
Analytical	5.4316	0.3662	—
Number of Reps			
$S = 10$	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

Summary

- So far we have worked with Normal linear regression model using natural conjugate prior
- This meant posterior, marginal likelihood and predictive distributions had analytical forms
- But with other priors and more complicated models do not get analytical results.
- Next we will present some popular extensions of the regression model to introduce another tool for posterior computation: the Gibbs sampler.
- The Gibbs sampler is a special type of Markov Chain Monte Carlo (MCMC) algorithm.

Normal Linear Regression Model with Independent Normal-Gamma Prior

- Keep the Normal linear regression model (under the classical assumptions) as before.
- Likelihood function presented above
- Parameters of model are β and h.

The Prior

- Before we had conjugate prior where $p\left(\beta|h\right)$ was Normal density and $p\left(h\right)$ Gamma density.
- Now use similar prior, but assume prior independence between β and h.
- $p\left(\beta, h\right) = p\left(\beta\right) p\left(h\right)$ with $p\left(\beta\right)$ being Normal and $p\left(h\right)$ being Gamma:

$$\beta \sim N\left(\underline{\beta}, \underline{V}\right)$$

and

$$h \sim G(\underline{s}^{-2}, \underline{\nu})$$

Key difference: now $\underline{V}$ is now the prior covariance matrix of β, with conjugate prior we had $var(\beta|h) = h^{-1}\underline{V}$.

The Posterior

- The posterior is proportional to prior times the likelihood.
- The joint posterior density for β and h does not take form of any well-known and understood density – cannot be directly used for posterior inference.
- However, conditional posterior for β (i.e. conditional on h) takes a simple form:

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

- where

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

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

- Conditional posterior for h takes simple form:

$$h|y,\beta \sim G(\overline{s}^{-2},\overline{\nu})$$

where

$$\overline{\nu} = N + \underline{\nu}$$

and

$$\overline{s}^2 = \frac{(y - X\beta)'(y - X\beta) + \underline{\nu}\underline{s}^2}{\overline{\nu}}$$

- Econometrician is interested in $p(\beta, h|y)$ (or $p(\beta|y)$), NOT the posterior conditionals, $p(\beta|y,h)$ and $p(h|y,\beta)$.
- Since $p(\beta, h|y) \neq p(\beta|y,h)\,p(h|y,\beta)$, the conditional posteriors do not directly tell us about $p(\beta, h|y)$.
- But, there is a posterior simulator, called the *Gibbs sampler*, which uses conditional posteriors to produce random draws, $\beta^{(s)}$ and $h^{(s)}$ for $s = 1, .., S$, which can be averaged to produce estimates of posterior properties just as with Monte Carlo integration.

The Gibbs Sampler

- Gibbs sampler is powerful tool for posterior simulation used in many econometric models.
- We will motivate general ideas before returning to regression model
- General notation: θ is a $p-$vector of parameters and $p\left(y|\theta\right), p\left(\theta\right)$ and $p\left(\theta|y\right)$ are the likelihood, prior and posterior, respectively.
- Let θ be partitioned into *blocks* as $\theta = \left(\theta'_{(1)}, \theta'_{(2)}, .., \theta'_{(B)}\right)'$. E.g. in regression model set $B = 2$ with $\theta_{(1)} = \beta$ and $\theta_{(2)} = h$.

- Intuition: i) Monte Carlo integration takes draws from $p\left(\theta|y\right)$ and averages them to produce estimates of $E\left[g\left(\theta\right)|y\right]$ for any function of interest $g\left(\theta\right)$.
- ii) In many models, it is not easy to draw from $p\left(\theta|y\right)$. However, it often is easy to draw from $p\left(\theta_{(1)}|y,\theta_{(2)},..,\theta_{(B)}\right)$, $p\left(\theta_{(2)}|y,\theta_{(1)},\theta_{(3)}..,\theta_{(B)}\right)$, ..., $p\left(\theta_{(B)}|y,\theta_{(1)},..,\theta_{(B-1)}\right)$.
- Note: Preceding distributions are *full conditional posterior distributions* since they define a posterior for each block conditional on all other blocks.
- iii) Drawing from the full conditionals will yield a sequence $\theta^{(1)},\theta^{(2)},..,\theta^{(s)}$ which can be averaged to produce estimates of $E\left[g\left(\theta\right)|y\right]$ in the same manner as Monte Carlo integration.
- This is called Gibbs sampling

More motivation for the Gibbs sampler

- Regression model with $B = 2$: β and h
- Suppose that you have one random draw from $p\left(\beta|y\right)$. Call this draw $\beta^{(0)}$.
- Since $p\left(\beta, h|y\right) = p\left(h|y, \beta\right) p\left(\beta|y\right)$, a draw from $p\left(h|y, \beta^{(0)}\right)$ is a valid draw of h. Call this $h^{(1)}$.
- Since $p\left(\beta, h|y\right) = p\left(\beta|y, h\right) p\left(h|y\right)$, a random draw from $p\left(\beta|y, h^{(1)}\right)$ is a valid draw of β. Call this $\beta^{(1)}$
- Hence, $\left(\beta^{(1)}, h^{(1)}\right)$ is a valid draw from $p\left(\beta, h|y\right)$.
- You can continue this reasoning indefinitely producing $\left(\beta^{(s)}, h^{(s)}\right)$ for $s = 1, .., S$

- Hence, if you can successfully find $\beta^{(0)}$, then sequentially drawing $p\left(h|y,\beta\right)$ and $p\left(\beta|y,h\right)$ will give valid draws from posterior.
- Problem with above strategy is that it is not possible to find such an initial draw $\beta^{(0)}$.
- If we knew how to easily take random draws from $p\left(\beta|y\right)$, we could use this and $p\left(h|\beta,y\right)$ to do Monte Carlo integration and have no need for Gibbs sampling.
- 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 $p\left(\beta,h|y\right)$.
- In practice, choose $\beta^{(0)}$ in some manner and then run the Gibbs sampler for S replications.
- Discard S_0 initial draws ("the *burn-in")* and remaining S_1 used to estimate $E\left[g\left(\theta\right)|y\right]$

Why is Gibbs sampling so useful?

- In Normal linear regression model with independent Normal-Gamma prior Gibbs sampler is easy
- $p\left(\beta|y,h\right)$ is Normal and $p\left(h|y,\beta\right)$ and Gamma (easy to draw from)
- Huge number of other models have hard joint posterior, but easy posterior conditionals
- 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 etc etc.
- What if the full posterior conditionals do not have simple form?
- Many other algorithms exist for handling general cases, Metropolis-Hastings algorithm is most popular

The Metropolis-Hastings Algorithm

- This is another popular class of algorithms useful when Gibbs sampling is not easy
- For now, I leave the regression model and return to our general notation:
- θ is a vector of parameters and $p\left(y|\theta\right), p\left(\theta\right)$ and $p\left(\theta|y\right)$ are the likelihood, prior and posterior, respectively.
- Metropolis-Hastings algorithm takes draws from a convenient *candidate generating density*.
- Let θ^* indicate a draw taken from this density which we denote as $q\left(\theta^{(s-1)};\theta\right)$.
- Notation: θ^* is a draw taken of the random variable θ whose density depends on $\theta^{(s-1)}$.

- We are drawing the wrong distribution, $q\left(\theta^{(s-1)};\theta\right)$, instead of $p\left(\theta|y\right)$
- We have to correct for this somehow.
- Metropolis-Hastings algorithm corrects for this via an acceptance probability
- Takes candidate draws, but only some of these candidate draws are accepted.

- The Metropolis-Hastings algorithm takes following form:
- *Step 1:* Choose a starting value, $\theta^{(0)}$.
- *Step 2:* Take a candidate draw, θ^* from the candidate generating density, $q\left(\theta^{(s-1)}; \theta\right)$.
- *Step 3:* Calculate an acceptance probability, $\alpha\left(\theta^{(s-1)}, \theta^*\right)$.
- *Step 4:* Set $\theta^{(s)} = \theta^*$ with probability $\alpha\left(\theta^{(s-1)}, \theta^*\right)$ and set $\theta^{(s)} = \theta^{(s-1)}$ with probability $1 - \alpha\left(\theta^{(s-1)}, \theta^*\right)$.
- *Step 5:* Repeat Steps 1, 2 and 3 S times.
- *Step 6:* Take the average of the S draws $g\left(\theta^{(1)}\right), ..., g\left(\theta^{(S)}\right)$.

- These steps will yield an estimate of $E\left[g(\theta)|y\right]$ for any function of interest.
- Note: As with Gibbs sampling, Metropolis-Hastings algorithm requires the choice of a starting value, $\theta^{(0)}$. To make sure that the effect of this starting value has vanished, wise to discard S_0 initial draws.
- Intuition for acceptance probability, $\alpha\left(\theta^{(s-1)}, \theta^*\right)$, given in textbook (pages 93-94).

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

Choosing a Candidate Generating Density

- Independence Chain Metropolis-Hastings Algorithm
- Uses a candidate generating density which is independent across draws.
- That is, $q\left(\theta^{(s-1)};\theta\right)=q^{*}\left(\theta\right)$ and the candidate generating density does not depend on $\theta^{(s-1)}$.
- Useful in cases where a convenient approximation exists to the posterior. This convenient approximation can be used as a candidate generating density.
- Acceptance probability simplifies to:

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

Choosing a Candidate Generating Density

- Random Walk Chain Metropolis-Hastings Algorithm
- Popular with DSGE – useful when you cannot find a good approximating density for the posterior.
- No attempt made to approximate posterior, rather candidate generating density is chosen to wander widely, taking draws proportionately in various regions of the posterior.
- Generates candidate draws according to:

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

where w is called the *increment random variable*.

- Acceptance probability simplifies to:

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

- Choice of density for w determines form of candidate generating density.
- Common choice is Normal:

$$q\left(\theta^{(s-1)}; \theta\right) = f_N(\theta|\theta^{(s-1)}, \Sigma).$$

- Researcher must select Σ. Should be selected so that the acceptance probability tends to be neither too high nor too low.
- There is no general rule which gives the optimal acceptance rate. A rule of thumb is that the acceptance probability should be roughly 0.5.
- A common approach sets $\Sigma = c\Omega$ where c is a scalar and Ω is an estimate of posterior covariance matrix of θ (e.g. the inverse of the Hessian evaluated at the posterior mode)

Bayesian Model Averaging

Overview

- BMA can be used with any set of models
- Here use it with 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 $p(M_r|y)$
- Let M_r for $r = 1, .., R$ denote R models.
- If ϕ is a parameter to be estimated (or a function of parameters) or a variable to be forecast, then the rules of probability imply:

$$p\left(\phi|y\right) = \sum_{r=1}^{R} p\left(\phi|y, M_r\right) p\left(M_r|y\right)$$

- Allows for a formal treatment of model uncertainty.
- Model selection: choose a single model and act as though it were true
- BMA incorporates uncertainty about which model generated the data.

The Model Space

- Let X_r be a $N \times k_r$ matrix containing some (or all) columns of X, then each model is
$$y = \alpha \iota_N + X_r \beta_r + \varepsilon$$
- ι_N is a $N \times 1$ vector of ones so as to say each model contains an intercept
- Other assumptions as for Normal linear regression model under classical assumptions.
- 2^K possible choices for X_r and, thus, the number of models, $R = 2^K$.
- Computational concerns: estimating every model will be impossible if K is large
- BMA empirical example will have $K = 41$
- If each model could be estimated in 0.001 seconds, over 100 years to estimate them all
- Use natural conjugate prior to make estimation of each model as fast as possible

BMA Priors

- We want a prior for model r that is:
- Informative (so as to provide valid marginal likelihoods for model comparison)
- Objective (requiring minimal subjective input)
- Automatic (does not have to be individually chosen for each of the many models)
- g-prior is commonly used:
- Prior mean shrinks coefficients towards zero:

$$\underline{\beta}_r = 0$$

- Prior covariance matrix is $h^{-1}\underline{V}_r$ where

$$\underline{V}_r = \left(gX_r'X_r\right)^{-1}$$

- g is a scalar

The g-prior

- The g-prior was suggested in Zellner (1986)
- Justification:
- Under non-informative prior $h^{-1}\left(X_r'X_r\right)^{-1}$ is posterior covariance matrix
- Amount of information in data for estimating β_r (information matrix)
- Prior covariance matrix $h^{-1}\left(gX_r'X_r\right)^{-1}$ says:
- Prior information that $\beta_r = 0$ takes same form as data information
- g controls relative strengths of the prior and data information.
- $g = 1$: prior and data are given equal weight.
- $g = 0.01$: prior information receives one per cent of the weight as data
- There exist commonly-used rules of thumb for choosing g
- Or g can be treated as unknown parameter with own prior and estimated
- Noninformative prior for h typically used

BMA Posterior

- With natural conjugate prior, analytical results for M_r
- Posterior is Normal-Gamma
- Marginal likelihood (for producing posterior model probs) analytical
- Predictive density is t-distribution
- Key thing: for each model, everything we need can be calculated quickly
- But even with this, doing BMA with 2^K models for $K > 20$ or so too computationally demanding

BMA Computation

- Previously we talked about posterior simulation as tool for learning about complicated posteriors
- For BMA can do model simulation
- A popular algorithm is Markov Chain Monte Carlo Model Composition (MC^3)
- Similar to a random walk Metropolis-Hastings algorithm, but models are drawn instead of parameters

MC-cubed

- $M^{(s)}$ for $s = 1, .., S$ are drawn models
- Averaging estimates/forecasts over drawn models will converge to the true BMA posterior or predictive estimates as $S \rightarrow \infty$.
- if ϕ is parameter of interest, then

$$\widehat{\phi} = \frac{1}{S} \sum_{s=1}^{S} E\left(\phi | y, M^{(s)}\right)$$

- will converge to $E(\phi|y)$.
- Frequencies with which models are drawn can be used to calculate Bayes factors.
- If MC^3 algorithm draws M_i A times and M_j B times, then $\frac{A}{B}$converges to Bayes factor comparing M_i to M_j.
- In practice, discard initial draws as burn-in

MC-cubed: How are models drawn?

- Want to draw $s = 1, .., S$ and suppose you have drawn $M^{(s-1)}$
- Candidate model, M^{*}, is proposed drawn randomly (with equal probability) from a set of models including:
- i) $M^{(s-1)}$
- ii) all models which delete one explanatory variable from $M^{(s-1)}$
- iii) all models which add one explanatory variable to $M^{(s-1)}$.
- Candidate model accepted with probability:

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

- If M^{*} is accepted then $M^{(s)} = M^{*}$, else $M^{(s)} = M^{(s-1)}$.
- Can prove MC-cubed will converge to true BMA posterior/predictive

BMA Application: The Determinants of Economic Growth

- To illustrate BMA use a classic cross-country growth regression data set
- Why do some countries grow faster than others?
- Numerous potential explanations (e.g. education, investment, governance, institutions, trade, colonialism, etc. etc.)
- Dependent variable: average growth in GDP per capita from 1960-1992
- $K = 41$ explanatory variables (all normalized by subtracting of mean and dividing by st. dev.)
- This is Big Data
- But data set has only $N = 72$ countries
- Note: will use this data set in machine learning lecture

BMA Application

- Cross-country growth regression data set with $N = 72$ and $K = 41$
- Use common recommendation to set $g = \frac{1}{N}$ if $N > K^2$ or $g = \frac{1}{K^2}$ if $N \leq K^2$
- Run MC-cubed 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
- Next table indicates convergence
- Note that best model receives less than 1% of posterior model
- Model selection puts all weight on this single model — ignoring huge amount of model uncertainty

Posterior Model Probabilities for Top 10 Models		
	$p\left(M_r\middle|y\right)$ Analytical	$p\left(M_r\middle|y\right)$ MC^3 estimate
1	0.0087	0.0089
2	0.0076	0.0077
3	0.0051	0.0050
4	0.0034	0.0035
5	0.0031	0.0032
6	0.0029	0.0029
7	0.0027	0.0025
8	0.0027	0.0027
9	0.0027	0.0026
10	0.0024	0.0022

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 st. dev.)
- BMA ensures parsimony by averaging over many small models
- Average number of exp. vars in a model drawn by MC^3 is 11.4

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