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
- Reading: textbook chapters 2, 3 and 6
- Chapter 2 presents theory for simple regression model (no matrix algebra)
- Chapter 3 does multiple regression
- In lecture, I will go straight to multiple regression
- Begin with regression model under classical assumptions (independent errors, homoskedasticity, etc.)
- Chapter 6 frees up classical assumptions in several ways
- Lecture will cover one way: Bayesian treatment of a particular type of heteroskedasticity

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.

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 causes 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 $c \longrightarrow 0$. 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 $g(.)$ 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 $E[g(\beta)|y]$ 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:

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

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			
	$p\left(\beta_j > 0 \mid y\right)$	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			
	$p\left(\beta_j > 0 \mid y\right)$	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 $p(\beta|h)$ was Normal density and $p(h)$ Gamma density.
- Now use similar prior, but assume prior independence between β and h.
- $p(\beta, h) = p(\beta)\, p(h)$ with $p(\beta)$ being Normal and $p(h)$ 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

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

- 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 $p(\theta|y)$ and averages them to produce estimates of $E[g(\theta)|y]$ for any function of interest $g(\theta)$.
- ii) In many models, it is not easy to draw from $p(\theta|y)$. 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 $E[g(\theta)|y]$ 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 $p(\beta|y)$. Call this draw $\beta^{(0)}$.
- Since $p(\beta, h|y) = p(h|y, \beta)\, p(\beta|y)$, a draw from $p\left(h|y, \beta^{(0)}\right)$ is a valid draw of h. Call this $h^{(1)}$.
- Since $p(\beta, h|y) = p(\beta|y, h)\, p(h|y)$, 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 $p(\beta, h|y)$.
- 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 $p(h|y,\beta)$ and $p(\beta|y,h)$ 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 $p(\beta|y)$, we could use this and $p(h|\beta,y)$ 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 $p(\beta,h|y)$.
- 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 $E[g(\theta)|y]$

Why is Gibbs sampling so useful?

- In Normal linear regression model with independent Normal-Gamma prior Gibbs sampler is easy
- $p(\beta|y,h)$ is Normal and $p(h|y,\beta)$ 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.
- Also models of form I will now discuss

Regression Models with General Error Covariance

- $$y = X\beta + \varepsilon.$$
- Before assumed ε was $N(0_N, h^{-1} I_N)$.
- Many other models involve

$$\varepsilon \sim N(0_N, h^{-1}\Omega)$$

- for some positive definite Ω.
- E.g. heteroskedasticity, autocorrelated errors, Student-t errors, random effects panel data models, SUR models, ARMA models, etc.

- Standard theorem in matrix algebra:
- An $N \times N$ matrix P exists with the property that $P\Omega P' = I_N$.
- Multiply both sides of regression model by P:

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

- where $y^\dagger = Py$, $X^\dagger = PX$ and $\varepsilon^\dagger = P\varepsilon$.
- It can be verified that $\varepsilon^\dagger$ is $N\left(0_N, h^{-1} I_N\right)$.
- Hence, transformed model is identical to Normal linear regression model.

- If Ω is known, Bayesian analysis of regression model with general error covariance matrix is straightforward (simply work with transformed model).
- If Ω is unknown, often can use Gibbs sampling
- Gibbs sampler could draw from $p(\beta|y, h, \Omega)$, $p(h|y, \beta, \Omega)$ and $p(\Omega|y, \beta, h)$
- Note: what if $p(\Omega|y, \beta, h)$ does not have a convenient form to draw from?
- Metropolis-Hastings algorithms are popular (to be discussed below, see pages 92-99 of textbook)
- "Metropolis-within-Gibbs" algorithms popular.

- Example: use an independent Normal-Gamma prior for β and h
- At this stage use general notation, $p(\Omega)$, to indicate the prior for Ω.
- Thus prior used is

$$p(\beta, h, \Omega) = p(\beta)\, p(h)\, p(\Omega)$$

- where:

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

and

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

- Exercise 13.1 of Bayesian Econometric Methods shows:

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

- where

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

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

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

- where $\widehat{\beta}\left(\Omega\right)$ is the GLS estimator

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

- and

$$\overline{s}^2 = \frac{\left(y - X\beta\right)'\Omega^{-1}\left(y - X\beta\right) + \underline{\nu}\underline{s}^2}{\overline{\nu}}$$

- Posterior for Ω conditional on β and h:

$$p(\Omega|y,\beta,h) \propto$$
$$p(\Omega)\left|\Omega\right|^{-\frac{1}{2}}\left\{\exp\left[-\frac{h}{2}(y-X\beta)'\Omega^{-1}(y-X\beta)\right]\right\}$$

- Often $p(\Omega|y,\beta,h)$ take an easy form (e.g. with autocorrelated errors).
- Gibbs sampler: $p(\beta|y,h,\Omega)$ is Normal, $p(h|y,\beta,\Omega)$ is Gamma and $p(\Omega|y,\beta,h)$
- We will use Gibbs samplers for VARs and state space models shortly

Prediction Using the Gibbs Sampler

- Want to predict T unobserved values $y^* = (y_1^*, .., y_T^*)'$, which are generated as:

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

- ε^* is $N\left(0, h^{-1}\Omega\right)$
- We want $p\left(y^*|y\right)$ but cannot be derived analytically.
- But we do know y^* is $N\left(X^*\beta, h^{-1}\Omega\right)$
- Predictive features of interest can be written as $E\left[g\left(y^*\right)|y\right]$ for some function $g\left(.\right)$.
- E.g. Predictive mean of y_i^* implies $g\left(y^*\right) = y_i^*$

- But, using LLN, if we can find $y^{*(s)}$ for $s = 1, .., S$ which are draws from $p\left(y^{*}|y\right)$, then

$$\widehat{g}_Y = \frac{1}{S}\sum_{s=1}^{S} g\left(y^{*(s)}\right)$$

  will converge to $E\left[g\left(y^{*}\right)|y\right]$.
- The following strategy provides draws of y^{*}.
- For every $\beta^{(s)}$, $h^{(s)}, \Omega^{(s)}$ from Gibbs sampler, take a draw (or several) of $y^{*(s)}$ from $p\left(y^{*}|\beta^{(s)}, h^{(s)}, \Omega^{(s)}\right)$ (a Normal density)
- We now have draws $\beta^{(s)}$, $h^{(s)}, \Omega^{(s)}$ and $y^{*(s)}$ for $s = 1, .., S$ which we can use for posterior or predictive inference.
- Why are these the correct draws? Use rules of conditional probability (see pages 72-73 of textbook for details).

Heteroskedasticity of an Unknown Form: Student-t Errors

- We will give one example which illustrates a few general concepts.
- It turns out that heteroskedasticity of an unknown form in Normal linear regression model is, in a sense, equivalent to a regression model with Student-t errors.
- This is a simple example of a *mixture model*.
- Mixture models are very popular right now in many fields as a way of making models more flexible (e.g. non-Normal errors, "nonparametric" treatment of regression line, etc.).

Heteroskedasticity of an Unknown Form: Student-t Errors

- Heteroskedasticity occurs if:

$$\Omega = \begin{bmatrix} \omega_1 & 0 & . & . & 0 \\ 0 & \omega_2 & 0 & . & . \\ . & 0 & . & . & . \\ . & . & . & . & 0 \\ 0 & . & . & 0 & \omega_N \end{bmatrix}$$

- In other words, $var(\varepsilon_i) = h^{-1}\omega_i$ for $i = 1, .., N$.
- With N observations and $N + k + 1$ parameters to estimate (i.e. β, h and $\omega = (\omega_1, .., \omega_N)'$), treatment of heteroskedasticity of unknown form may sound like a difficult task.
- Solution: use a *hierarchical prior* (ω_is drawn from some common distribution – parameters of that distribution estimated from the data).
- Hierarchical priors are commonly used as a way of making flexible, parameter-rich models more amenable to statistical analysis.
- Allows us to free up the assumption of Normal errors

A Hierarchical Prior for the Error Variances

- We begin by eliciting $p(\omega)$.
- Work with error precisions rather than variances and, hence, we define $\lambda \equiv (\lambda_1, \lambda_2, .., \lambda_N)' \equiv \left(\omega_1^{-1}, \omega_2^{-1}, .., \omega_N^{-1}\right)'$.
- Consider the following prior for λ:

$$p(\lambda) = \prod_{i=1}^{N} f_G(\lambda_i | 1, \nu_\lambda) \qquad (**)$$

- Note f_G is the Gamma p.d.f.
- The prior for λ depends on a hyperparameter, ν_λ, and assumes each λ_i comes from the same distribution.
- In other words, λ_is are i.i.d. draws from the Gamma distribution.
- This assumption (or something similar) is necessary to deal with the problems caused by the high-dimensionality of λ.

A Hierarchical Prior for the Error Variances

- Why should the λ_is be i.i.d. draws from the Gamma distribution with mean 1.0?
- Can prove this model is *exactly the same* as the linear regression model with i.i.d. Student-t errors with ν_λ degrees of freedom (Bayesian Econometric Methods Exercise 15.1).
- In other words, if we had begun by assuming:

$$p\left(\varepsilon_i\right) = f_t\left(\varepsilon_i | 0, h^{-1}, \nu_\lambda\right)$$

- for $i = 1, .., N$, we would have ended up with exactly the same posterior.

A Hierarchical Prior for the Error Variances

- Note: we now have model with more flexible error distribution, but we are still our familiar Normal linear regression model framework.
- Note: a popular way of making models/distributions more flexible is through: *mixture of Normals* distributions.
- Our treatment here is an example of a *scale mixture of Normals*.
- If ν_λ is unknown, need a prior $p(\nu_\lambda)$.
- Note that now the prior for λ is specified in two steps, the first being (**), the other being $p(\nu_\lambda)$.
- Alternatively, the prior for λ can be written as $p(\lambda|\nu_\lambda)\, p(\nu_\lambda)$.
- Priors written in two (or more) steps in this way are referred to as hierarchical priors.

Bayesian Computation with Student-t Model

- Geweke (1993, Journal of Applied Econometrics) develops a Gibbs sampler for taking draws of the parameters in the model: β, h, λ and ν_λ.
- $p(\beta|y, h, \lambda)$ and $p(h|y, \beta, \lambda)$ are as discussed previously
- Focus on $p(\lambda|y, \beta, h, \nu_\lambda)$ and $p(\nu_\lambda|y, \beta, h, \lambda)$.
- Bayesian Econometric Methods, Exercise 15.1 derives posterior conditionals for λ_is as

$$p(\lambda_i|y, \beta, h, \nu_\lambda) = f_G\left(\lambda_i|\frac{\nu_\lambda + 1}{h\varepsilon_i^2 + \nu_\lambda}, \nu_\lambda + 1\right)$$

- $p(\nu_\lambda|y, \beta, h, \lambda)$ depends on $p(\nu_\lambda)$. Geweke uses a particular prior density and derives a method of drawing from this density (thus completing the Gibbs sampler).

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)

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
- Note: There are methods for calculating marginal likelihoods using Gibbs sampler output