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:

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

- 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 $c \rightarrow 0$. 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:

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

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)

<table>
<tr><th colspan="4">Model Comparison involving β</th></tr>
<tr><th colspan="4">Informative Prior</th></tr>
<tr><th></th><th>$p\left(\beta_j > 0 | y\right)$</th><th>95% HPDI</th><th>Posterior Odds
for $\beta_j = 0$</th></tr>
<tr><td>β_1</td><td>0.13</td><td>$[-10,957,2,887]$</td><td>4.14</td></tr>
<tr><td>β_2</td><td>1.00</td><td>$[4.71,6.15]$</td><td>2.25×10^{-39}</td></tr>
<tr><td>β_3</td><td>0.99</td><td>$[563.5,5,210.1]$</td><td>0.39</td></tr>
<tr><td>β_4</td><td>1.00</td><td>$[13,616,20,314]$</td><td>1.72×10^{-19}</td></tr>
<tr><td>β_5</td><td>1.00</td><td>$[5,686,9,596]$</td><td>1.22×10^{-11}</td></tr>
<tr><th colspan="4">Noninformative Prior</th></tr>
<tr><th></th><th>$p\left(\beta_j > 0 | y\right)$</th><th>95% HPDI</th><th>Posterior Odds
for $\beta_j = 0$</th></tr>
<tr><td>β_1</td><td>0.13</td><td>$[-11,055,3,036]$</td><td>—</td></tr>
<tr><td>β_2</td><td>1.00</td><td>$[4.71,6.15]$</td><td>—</td></tr>
<tr><td>β_3</td><td>0.99</td><td>$[449.3,5,200]$</td><td>—</td></tr>
<tr><td>β_4</td><td>1.00</td><td>$[13,714,20,497]$</td><td>—</td></tr>
<tr><td>β_5</td><td>1.00</td><td>$[5,664,9,606]$</td><td>—</td></tr>
</table>

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:

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

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 $p\left(\beta, h|y\right)$ (or $p\left(\beta|y\right)$), NOT the posterior conditionals, $p\left(\beta|y, h\right)$ and $p\left(h|y, \beta\right)$.
- Since $p\left(\beta, h|y\right) \neq p\left(\beta|y, h\right) p\left(h|y, \beta\right)$, the conditional posteriors do not directly tell us about $p\left(\beta, h|y\right)$.
- 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.

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