A Tutorial on Bayesian Estimation of the Normal Linear Model*

Donald J. Lacombe

February 7, 2022

Abstract

The normal linear model is one of the workhorses of applied modeling across many different academic disciplines. This tutorial will illustrate the mathematical details of Bayesian estimation of the normal linear regression model. Particular attention is paid to the mathematical derivations required to obtain the full conditional distributions required for Gibbs sampling. The models are derived using diffuse as well as natural conjugate priors for the parameters.

Introduction

The normal linear model is one of the workhorses of applied modeling across many different academic disciplines, such as economics, political science, geography, criminology, and a host of others. Linear regression is usually one of the first regression models applied researchers study, due to its simplicity in regards to mathematical derivation as well as computation. However, in recent years, the Bayesian paradigm has reached maturity, due to the availability of easy–to–use software packages. The Bayesian paradigm is conceptually fairly easy to understand, however, the mathematical details surrounding estimation issues are difficult to comprehend unless one has some help. Once a few mathematical ideas are known and applied, the derivation of such models becomes much easier to understand and apply. In particular, this tutorial will focus on the mathematical details required to implement various Markov Chain Monte Carlo techniques, such as the popular Gibbs sampler, and give the reader complete and full mathematical details so that they may understand the underlying mathematical derivations. Additionally, Python code will be introduced to see how these mathematical ideas as conceptualized in code.

The normal linear regression model is mathematically represented as follows:

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

where y is a $N \times 1$ vector of observations on the dependent variable, X is an $N \times k$ matrix of explanatory variables, and ε represents the i.i.d. error term that is assumed to be normally distributed, with mean zero and covariance matrix $\sigma^2 I_n$.

*School of Financial Planning, Texas Tech University, Lubbock, TX

The normal linear regression model, or linear regression model, can be estimated via maximum likelihood or via the so–called "normal equations" that leads to a closed–form solution for the coefficients of the model, which is computationally convenient.

However convenient the normal linear model is to estimate using maximum likelihood or the closed–form normal equations, there are distinct advantages to using Bayesian methods. We now focus on the Bayesian paradigm in the next section.

A Brief Introduction to the Bayesian Paradigm

The Bayesian paradigm consists of three entities: the posterior distribution, the prior distribution, and the likelihood function. The likelihood function is probably the most familiar to those working in the frequentist domain and indeed it is the same entity in the Bayesian paradigm. In terms of normal linear model, the likelihood function is as follows:

$$L(\beta, \sigma, y, X) = \frac{1}{(2\pi)^{\frac{N}{2}}\sigma^2} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right\}$$

We will revisit the likelihood function in a subsequent section.

The next entity to be discussed in the Bayesian paradigm is the role of the prior distribution. The prior distribution is designed to capture the prior beliefs of the researcher and to formalize those beliefs in a probability distribution. Although the designation of the prior distribution is of utmost importance, in this introduction, we are interested only in how the specification of certain priors changes the mathematical derivations of the full conditional distributions. As a starting point, so-called "non informative","diffuse", or "ignorant" priors for the parameters in the model, namely β, and σ be specified. Mathematically, this can be accomplished as follows:

$$p(\beta) \propto constant$$
$$p(\sigma) \propto \frac{1}{\sigma}$$

In other words, the prior for the β term is a constant (usually set to one), and the prior on the σ term is a standard diffuse prior for this parameter that is used frequently in the literature. In fact, the use of this diffuse parameter for σ allows one to easily recognize the form of the conditional distribution required for Gibbs sampling, which will be illustrated later.

The final entity, and perhaps the most important from a Bayesian perspective, is the specification of the posterior distribution. The posterior distribution summarizes all of the information about the parameters of the model and is the focus of all Bayesian inference. The posterior distribution is derived from Bayes Rule, which can be summarized as follows:

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

Frequently, the relationship between the posterior, prior, and likelihood is summarized in the phrase "the posterior is proportional to the prior times the likelihood" and thus can be written

$$p(\theta|y) \propto p(y|\theta)\,p(\theta)$$

This relationship holds because the denominator in the previous equation above does not involve the parameters, so the relationship is a proportional one[1].

Now that the basic elements of the Bayesian paradigm are in order, it is time to start assembling the constituent parts in order to derive the full conditional distributions required for Gibbs sampling.

Deriving the Full Conditional Distributions

The first order of business in deriving the full conditional distributions required for Gibbs sampling is to mathematically derive the posterior distribution. This is accomplished by multiplying the likelihood and the priors together to form the posterior distribution. The priors in this example are assumed to be independent, i.e. $p(\beta,\sigma) \propto p(\beta)\,p(\sigma)$, so that they may be multiplied by the likelihood without any problem. The likelihood can be conveniently expressed mathematically as follows:

$$L(\beta,\sigma,y,X) = (2\pi)^{-\frac{N}{2}}\sigma^{-N}\exp\left\{-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)\right\}$$

which is simply taking the leading terms in the original likelihood function above and inverting. In a Bayesian analysis when formulating the likelihood, we can write the likelihood as follows:

$$L(\beta,\sigma,y,X) \propto \sigma^{-N}\exp\left\{-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)\right\}$$

One may wonder what happened to the $(2\pi)^{-\frac{N}{2}}$ term. In Bayesian analysis, since we are working with a proportional relationship, we can ignore any constants that do not involve the variable of interest, which is why the above term "disappears".

A Brief Introduction to Gibbs Sampling

The common objective in all Bayesian exercises is to derive the marginal distributions that summarize knowledge about the parameters, θ , conditional on the data, y. The posterior distribution, however, is "difficult to work with" meaning that some of the required posterior marginal distributions, namely $p(\beta|y)$ and

[1]The denominator is referred to as the marginal likelihood. The marginal likelihood will not play a role in the derivation of the full conditional distributions required for Gibbs sampling, however it does play a role in model choice.

$p(\sigma|y)$, are not available in closed form. In other words, one cannot integrate out the parameter of interest and derive the marginal distribution using pencil and paper. In general settings, this integration is difficult or impossible, which is a major impediment to analysis. The solution is to Gibbs sample.

The necessary conditions for Gibbs sampling the linear regression model, or any model, for that matter, are two. First, the fully conditional distributions comprising the joint posterior must be available in closed form. Second, these forms must be tractable in the sense that it is easy to draw samples from them.

A simplified example may help in explaining the idea behind the Gibbs sampler. Suppose one has a joint density of the form $p(x, y, z)$ and that the marginal distributions of interest, $p(x)$, $p(y)$, and $p(z)$ are not available in closed form. Normally, this would lead to an intractable situation. However, we can use Gibbs sampling to achieve the desired goal. Essentially, Gibbs sampling requires that we obtain random draws from each of the component full conditional distributions i.e., $p(x|y)$ and $p(x|y)$ derived from the joint posterior distribution. Under weak regularity conditions, this chain converges in distribution to the true marginal quantities (the marginal posterior PDF's) that we seek. Therefore, Gibbs sampling is a very convenient alternative that allows inference to take place.

The next order of business is deriving the full conditional distributions for the linear regression model. When deriving the full conditional distributions, there are three mathematical ideas that one should keep in mind. First, all non-essential constants can be ignored because they are subsumed into the constant of proportionality. Second, only the terms involved in the conditional distribution need be collected when deriving the conditional distribution. Lastly, "completing the square" will play a role in deriving conditional distributions as well.

Obtaining the posterior requires we must multiply the likelihood by the priors.The prior on β is diffuse and equal to one, so there is no change to the likelihood. When we multiply the likelihood by the prior for σ, we obtain the following posterior distribution:

$$L(\beta, \sigma, y, X) \propto \sigma^{-(N+1)} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right\}$$

The $\sigma^{-(N+1)}$ term is obtained as follows: $\sigma^{-N} \times \sigma^{-1} = \sigma^{-(N+1)}$. We will see later on that specifying the prior in this manner will enable us to recognize the form of the conditional distribution for σ.The posterior is now fully defined and explained so we are ready to derive the full conditional distributions required for Gibbs sampling.

The full conditional distribution that we seek are the following:

$$p(\beta|\sigma)$$
$$p(\sigma|\beta)$$

We will start by examining the full conditional distribution for σ. In order to define the full conditional distribution for σ, or any other parameter, we simply pick out the terms in the posterior distribution that contain σ. This leads to the following conditional distribution for σ:

$$p(\sigma \,|\beta\,) \propto \sigma^{-(N+1)} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)'\,(y - X\beta)\right\}$$

The next step is to identify the distributional form of the conditional distribution for σ so that we may take random draws from this distribution. A standard reference for distributional forms is Zellner (1996, p. 371).The conditional distribution for σ is in the form of an inverse Gamma distribution, which is equation A.37b and has the following form:

$$f^{IG}\,(\sigma \,|\upsilon, s\,) = \frac{2}{\Gamma\,(\nu/2)}\left(\frac{\upsilon s^2}{2}\right)^{\frac{v}{2}} \frac{1}{\sigma^{\upsilon+1}} \exp^{\frac{-\upsilon s^2}{2\sigma^2}}$$

We can now use the first mathematical idea in eliminating the constant in the inverse Gamma distribution. The inverse Gamma distribution is defined in terms of σ so we can eliminate any terms that do not involve σ, which leads to the following functional form:

$$f^{IG}\,(\sigma \,|\upsilon, s\,) \propto \frac{1}{\sigma^{\upsilon+1}} \exp^{\frac{-\upsilon s^2}{2\sigma^2}}$$

One further simplification to the standard formula for the inverse Gamma distribution will bring this formula into a form that is similar to the form of the conditional distribution for σ. Simply invert the first term (i.e. the $\frac{1}{\sigma^{\upsilon+1}}$ term) and the functional form for the inverse Gamma now looks like the following:

$$f^{IG}\,(\sigma \,|\upsilon, s\,) \propto \sigma^{-(\upsilon+1)} \exp\left(\frac{-\upsilon s^2}{2\sigma^2}\right)$$

We can now let $\nu = N$ and $\upsilon s^2 = (y - X\beta)'\,(y - X\beta)$ and the functional form of the posterior conditional distribution for σ is identical to the inverted Gamma distribution and the first step in the Gibbs sampling algorithm is to take a random draw from the inverted Gamma distribution with the appropriate parameters, i.e., ν and νs^2.

The next conditional distribution that needs explanation is the conditional distribution for β, which is $p\,(\beta \,|y\,)$. Only the β terms in the posterior distribution need to be used, and this results in the following conditional distribution for β:

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

We can now use the second mathematical idea, completing the square, to find the conditional distribution for β.

The process of completing the square usually occurs when one is working with the kernel of the multivariate normal density[2]. The multivariate normal density has the following form (Zellner 1996, p. 379)

$$f^{MVN}(x\,|\theta,\Sigma) = (2\pi)^{-\frac{N}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(x-\theta)'\Sigma^{-1}(x-\theta)\right\}$$

We can once again ignore the leading constants and this will result is the following normal kernel:

$$f^{MVN}(x\,|\theta,\Sigma) \propto \exp\left\{-\frac{1}{2}(x-\theta)'\Sigma^{-1}(x-\theta)\right\}$$

We begin completing the square by expanding out the following term form the normal kernel:

$$(x-\theta)'\Sigma^{-1}(x-\theta)$$

This expansion is a straightforward application of matrix algebra rules (and the standard FOIL rule from algebra) and results in the following:

$$x'\Sigma^{-1}x - x'\Sigma^{-1}\theta - \theta'\Sigma^{-1}x + \theta'\Sigma^{-1}\theta$$

We are interested in completing the square in θ because that represents our parameter vector in the linear regression model, i.e., the β's. We can ignore any term in the above expansion that does not involve θ because those terms are constants. The resulting mathematical formula is the result of choosing only the relevant terms:

$$\theta'\Sigma^{-1}\theta - \theta'\Sigma^{-1}x - x'\Sigma^{-1}\theta + constant$$

Completing the square in β is a similar exercise. After expansion of the terms in the normal kernel, we obtain the following general form:

$$\beta'(term_1)\beta - \beta'(term_2) - (term_2)\beta$$

We can now examine how this is applied in the context of the normal linear regression model. The conditional distribution for β can be expressed as follows:

$$p(\beta\,|\sigma) \propto \exp\left\{-\frac{1}{2}(y-X\beta)'\Sigma_\beta^{-1}(y-X\beta)\right\}$$

The portion of the normal kernel that we are interested in is the $(y-X\beta)'\Sigma_\beta^{-1}(y-X\beta)$ term. We also need to define a covariance matrix, which can be expressed as $\Sigma_\beta^{-1} = (\sigma^2 I_N)^{-1}$.

The expression that we wish to expand to complete the square is the following:

$$(y-X\beta)'\Sigma_\beta^{-1}(y-X\beta)$$

[2]The kernel of any probability density is the part excluding any integrating constants.

When we perform this expansion using the standard rules of matrix algebra (and the FOIL method from algebra), we obtain the following:

$$y'\Sigma_{\beta}^{-1}y - y'\Sigma_{\beta}^{-1}X\beta - \beta'X'\Sigma_{\beta}^{-1}y + \beta'X'\Sigma_{\beta}^{-1}X\beta$$

We can reorder the terms in this equation and ignore all terms that do not involve β so that it matches the template we observed before:

$$\beta'X'\Sigma_{\beta}^{-1}X\beta - \beta'X'\Sigma_{\beta}^{-1}y - y'\Sigma_{\beta}^{-1}X\beta$$

From the above equation, we see that by completing the square in β we put the expanded term into a form that is recognizable as a multivariate normal density. In this particular example, $term_1 = \beta'X'\Sigma_{\beta}^{-1}X\beta$ and $term_2 = \beta'X'\Sigma_{\beta}^{-1}y$. We can now place the $term_1$ and $term_2$ terms in the above formula to define the covariance and mean for the conditional distribution for β:

$$covariance = \left(X'\Sigma_{\beta}^{-1}X\right)^{-1}$$

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

We can simplify the above expressions by noting that $\Sigma_{\beta}^{-1} = \left(\sigma^2 I_N\right)^{-1}$. Now, we can write the above expression as follows:

$$covariance = \left(X'\frac{1}{\sigma^2 I_N}X\right)^{-1} = \sigma^2(X'X)^{-1}$$

$$mean = \left(X'\Sigma_{\beta}^{-1}X\right)^{-1}\left(X'\Sigma_{\beta}^{-1}y\right) = \left(X'\frac{1}{\sigma^2 I_N}X\right)^{-1}\left(X'\frac{1}{\sigma^2 I_N}y\right) = (X'X)^{-1}X'y$$

You'll note that the expressions for the covariance and mean are exactly the formulas that one obtains from the normal equations when using Ordinary Least Squares (OLS)! In our Gibbs sampling algorithm, we can take random draws from the multivariate normal distribution with mean and covariance matrix as in the above formulas.

The Full Conditional Distribution for β and σ Under a Normal and Inverse Gamma Prior

In this section, we will utilize a multivariate normal prior for the β term.Why might we use such a prior? A well-defined prior distribution for the regression parameters becomes useful in model comparisons between competing models or when the researcher has valid prior information for the regression parameters. Another reason to use a multivariate normal prior is that it is in the same family of distributions as the conditional distribution for β. This is referred to as a conjugate prior distribution. Regardless of the motivation, the specification of a multivariate normal prior for β will change the derivation of the full conditional distribution for these parameters relative to the diffuse prior case.

In many cases, you will hear of the idea of "adding prior information" to a model. In the case of our linear regression model, we will literally do just that: we will

add a prior covariance matrix and a prior mean to our conditional distribution for β.

We start with a prior distribution for the β's, which can be represented as the kernel of a multivariate normal distribution:

$$\left(\beta - \hat{\beta}\right)' \Sigma_{\hat{\beta}}^{-1} \left(\beta - \hat{\beta}\right)$$

When this term is expanded using standard rules of matrix algebra as before (i.e. using the FOIL method from algebra), we are left with the following expression:

$$\beta' \Sigma_{\hat{\beta}}^{-1} \beta - \beta' \Sigma_{\hat{\beta}}^{-1} \hat{\beta} - \hat{\beta}' \Sigma_{\hat{\beta}}^{-1} \beta + \hat{\beta}' \Sigma_{\hat{\beta}}^{-1} \hat{\beta}$$

In this situation, we would like to complete the square in β which leads to the following form, while treating the $\hat{\beta}' \Sigma_{\hat{\beta}}^{-1} \hat{\beta}$ term as a constant because it does not involve β:

$$\beta' \Sigma_{\hat{\beta}}^{-1} \beta - \beta' \Sigma_{\hat{\beta}}^{-1} \hat{\beta} - \hat{\beta}' \Sigma_{\hat{\beta}}^{-1} \beta$$

We can now identify the $term_1$ and $term_2$ terms from the above expression in order to place it into the predetermined multivariate normal form. The two terms are as follows:

$$covariance = \left(\Sigma_{\hat{\beta}}^{-1}\right)$$

$$mean = \left(\Sigma_{\hat{\beta}}^{-1} \hat{\beta}\right)$$

As mentioned earlier, the prior covariance and mean are added to the covariance and mean terms that we derived earlier. Specifically, the new conditional distributions with the added covariance and mean terms are as follows:

$$covariance = \left(X' \Sigma_{\beta}^{-1} X + \Sigma_{\hat{\beta}}^{-1}\right)^{-1}$$

$$mean = \left(X' \Sigma_{\beta}^{-1} X\right)^{-1} \left(X' \Sigma_{\beta}^{-1} y + \Sigma_{\hat{\beta}}^{-1} \hat{\beta}\right)$$

As is evident in the above equations, we literally are adding a prior covariance to the covariance term, and a prior mean to the mean term. As before, we now take random draws from the multivariate normal distribution using the mean and covariance formulas outlines in the above two equations.

The final modification that will be explored in this introduction is the specification of a proper prior for the σ parameter. The most commonly used proper prior for this parameter is the inverse gamma distribution:

$$f^{IG}\left(\sigma \,|\upsilon, s\,\right) \propto \sigma^{-(\upsilon+1)} \exp\left\{-\frac{\upsilon s^2}{2\sigma^2}\right\}$$

where the leading integrating constant has been suppressed as before. Whenever a new prior for any parameter is proposed, we must alter the posterior before deriving the conditional distribution for that parameter[3]. In this case, we will have the following priors:

$$p(\sigma) \propto \sigma^{-(\upsilon_0+1)} \exp\left\{-\frac{\upsilon_0 {s_0}^2}{2\sigma^2}\right\}$$

$$p(\beta) \sim MVN\left(\beta \middle| \widehat{\beta}, \Sigma_{\widehat{\beta}}^{-1}\right)$$

which consists of an inverse gamma prior for σ and a multivariate normal prior for β. the next step is to multiply the likelihood by all of the priors to derive the new joint posterior distribution:

$$L(y, X, \beta, \sigma) = (2\pi)^{\frac{N}{2}} \sigma^{-N} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right\} \times p(\beta) \times p(\sigma)$$

where $p(\beta)$ and $p(\sigma)$ ore the prior distributions are as above. The conditional distribution for β does not change from the previous derivation and therefore, we will concentrate on the derivation for the conditional distribution for σ.

Mathematically, the conditional distribution for σ can be expressed as follows, after collecting only the terms in the posterior that depend on σ:

$$p(\sigma|\beta) \propto \sigma^{-N} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right\} \times \sigma^{-(\upsilon_0+1)} \exp\left\{-\frac{\upsilon_0 {s_0}^2}{2\sigma^2}\right\}$$

We now can collect terms in these two expressions to obtain the full conditional distribution for σ. First, we will collect the σ terms in each section of the equation:

$$\sigma^{-N} + \sigma^{-(\upsilon_0+1)} = \sigma^{-(N+\upsilon_0+1)}$$

which is simply each exponent associated with each σ term added together. Next, we can add together the normal kernel parts as follows:

$$\exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta) - \frac{1}{2\sigma^2}(\upsilon_0 s_0^2)\right\}$$

$$\exp\left\{-\frac{(y - X\beta)'(y - X\beta) + \upsilon_0 s_0^2}{2\sigma^2}\right\}$$

This leaves us with the following expression for the full conditional distribution for σ:

$$p(\sigma) \propto \sigma^{-(N+\upsilon_0+1)} \exp\left\{-\frac{(y - X\beta)'(y - X\beta) + \upsilon_0 s_0^2}{2\sigma^2}\right\}$$

[3]Since the posterior distribution is the prior multiplied by the likelihood, whenever a new prior is proposed, the posterior has to be modified to accommodate the new prior distribution(s).

which is in the form of an inverse gamma distribution:

$$f^{IG}\left(\sigma\,|\upsilon,s\,\right)\propto\sigma^{-(\upsilon+1)}\exp\left\{-\frac{\upsilon s^{2}}{2\sigma^{2}}\right\}$$

Now we can let $\upsilon=N+\nu_{0}+1$ and $\upsilon s^{2}=\left(y-X\beta\right)^{'}\left(y-X\beta\right)+\upsilon_{0}s_{0}^{2}$ and the Gibbs sampling algorithm would now consist of a random draw from the inverse gamma distribution given the above arguments.

Conclusion

Bayesian econometrics has seen widespread use in a variety of academic disciplines. Although there are many different software options for estimating Bayesian models, this tutorial takes a different path by showing all of the details of Bayesian estimation using Gibbs sampling for the normal linear regression model.