The Normal Linear Regression Model with Natural Conjugate Prior

March 7, 2016
The Normal Linear Regression Model with Natural Conjugate Prior

The plan

- Estimate simple regression model using Bayesian methods
- Formulate prior
- Combine prior and likelihood to compute posterior
- Model comparison

Main reading: Ch.2 in Gary Koop’s *Bayesian Econometrics*
Recap
Bayes Rule

Deriving Bayes’ Rule

\[ p(A, B) = p(A | B)p(B) \]

Symmetrically

\[ p(A, B) = p(B | A)p(A) \]

Implying

\[ p(A | B)p(B) = p(B | A)p(A) \]

or

\[ p(A | B) = \frac{p(B | A)p(A)}{p(B)} \]

which is known as Bayes Rule.
Data and parameters

The purpose of Bayesian analysis is to use the data $y$ to learn about the “parameters” $\theta$

- Parameters of a statistical model
- Or anything not directly observed

Replace $A$ and $B$ in Bayes rule with $\theta$ and $y$ to get

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

The probability density $p(\theta | y)$ then describes what do we know about $\theta$, given the data.
The posterior density

The posterior density is often the object of fundamental interest in a Bayesian estimation.

- The posterior is the “result”.
- We are interested in the entire conditional distribution of the parameters $\theta$

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}$$

Today we will start filling in blanks and specify a prior and data generating process (i.e. a likelihood function) in order to find the posterior.
The linear regression model
The linear regression model

We will start with something very basic

- The linear regression model

\[ y_i = \beta x_i + \varepsilon_i \sim N(0, \sigma^2) \]

\[ i = 1, 2, \ldots N \]

The variable \( x_i \) is fixed (i.e. not random) or independent of \( \varepsilon_i \) and with a probability density function \( p(x_i \mid \lambda) \) where the parameter(s) \( \lambda \) does not include \( \beta \) or \( \sigma^2 \)

- Walk before we run etc…
Figure 2.1: Marginal Prior and Posteriors for $\beta$

- **Prior**
- **Posterior**
- **Likelihood**
The likelihood function
Defining the likelihood function

The assumptions made about the linear regression model implies that $p(y_i \mid \beta, \sigma^2)$ is Normal.

$$p(y_i \mid \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(y_i - \beta x_i)^2}{2\sigma^2} \right]$$

or that

$$p(y \mid \beta, \sigma^2) = \frac{1}{(2\pi)^{\frac{N}{2}} \sigma^N} \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \beta x_i)^2 \right]$$

where

$$y \equiv \begin{bmatrix} y_1 & y_2 & \cdots & y_N \end{bmatrix}^\prime$$
Reformulating the likelihood function

It will be convenient to use that

\[
\sum_{i=1}^{N} (y_i - \beta x_i)^2 = v s^2 + \left(\beta - \hat{\beta}\right)^2 \sum_{i=1}^{N} x_i^2
\]

where

\[
\begin{align*}
 v &= N - 1 \\
 \hat{\beta} &= \frac{\sum x_i y_i}{\sum x_i^2} \\
 s^2 &= \frac{\sum (y_i - \beta x_i)^2}{v}
\end{align*}
\]
Reformulating the likelihood function

The likelihood function can then be written as

\[ p (y \mid \beta, \sigma^2) = \frac{1}{(2\pi)^{N/2}} \left\{ h^{1/2} \exp \left[ - \frac{h}{2} \left( \beta - \hat{\beta} \right)^2 \sum_{i=1}^{N} x_i^2 \right] \right\} \times \left\{ h^{v/2} \exp \left[ - \frac{hv}{2s^{-2}} \right] \right\} \]

where \( h \equiv \sigma^{-2} \).
Formulating a prior
The Prior density

The prior density should reflect information about $\theta$ that we have prior to observing $y$.

Priors should preferably

- be easy to interpret
- be of a form that facilitates computing the posterior

Natural conjugate priors have both of these advantages.
Conjugate priors

- Conjugate priors, when combined with the likelihood function, result in posteriors that are of the same family of distributions.
- Natural conjugate priors has the same functional form as the likelihood function.

The posterior can $p(\theta \mid y)$ then be thought of as being the result of two "sub-samples"

$$p(\theta \mid y) \propto p(y \mid \theta)p(\theta)$$

- The sufficient statistic about $\theta$ of the first "sub-sample" is described by the hyper-parameters of the prior.
- The information about $\theta$ in second "sub-sample" can be described by the sufficient statistic for the actual observed sample (i.e. $y$).

This feature makes natural conjugate priors easy to interpret.
Specifying a prior density: It’s your choice

For the Normal regression model we need to ”elicit”, i.e. formulate, a prior $p(\theta)$ for $\beta$ and $h$ which we can denote $p(\beta, h)$.

- It will be convenient to use that
  \[ p(\beta, h) = p(\beta \mid h)p(h) \]
  and let $\beta \mid h \sim N(\beta, h^{-1}V)$ and let $h \sim G(s^{-2}, v)$.

- The natural conjugate prior for $\beta$ and $h$ is then denoted
  \[ \beta, h \sim NG(\underline{\beta}, V, s^{-2}, v) \]

Notation: Koop uses bars below parameters to denote hyper-parameters of a prior, and bars above to denote hyper-parameters of a posterior.
What makes a prior conjugate?

Short answer: Convenient functional forms
Long answer: Remember the likelihood function

\[ p(y \mid \beta, \sigma^2) = \frac{1}{(2\pi)^{N/2}} \left\{ h^{1/2} \exp \left[ -\frac{h}{2} \left( \beta - \hat{\beta} \right)^2 \sum_{i=1}^{N} x_i^2 \right] \right\} \times \left\{ h^{\nu/2} \exp \left[ -\frac{hv}{2s^{-2}} \right] \right\} \]

and consider the normal-gamma prior distribution \( p(\beta \mid h) p(h) \) where \( \beta \mid h \sim N(\beta, h^{-1}V) \) and \( h \sim G(s^{-2}, \nu) \). We then have

\[ p(\beta \mid h) p(h) = \left\{ \frac{1}{\sqrt{2\pi h^{-1}V}} \exp \left[ -\frac{(\beta - \hat{\beta})^2}{2h^{-1}V} \right] \right\} \times \left\{ c_G^{-1} h^{\frac{\nu-2}{2}} \exp \left( -\frac{hv}{2s^{-2}} \right) \right\} \]

Note the similar functional forms.
The posterior
The Normal-Gamma posterior

It is possible to use Bayes rule

\[
p(\beta, h \mid y) = \frac{p(y \mid \beta, h)p(\beta \mid h)p(h)}{p(y)}
\]

to find an analytical expression for the posterior that is Normal-Gamma

\[
p(\beta, h \mid y) = \left\{ \frac{1}{\sqrt{2\pi V}} \exp \left[ -\frac{(\beta - \bar{\beta})^2}{2V} \right] \right\} \left\{ c_G^{-1} h^{\frac{v-2}{2}} \exp \left( -\frac{h\bar{V}}{2\bar{s}^{-2}} \right) \right\}
\]

i.e. the same form as the prior

- \( \beta, h \mid y \sim NG(\bar{\beta}, \bar{V}, \bar{s}^{-2}, \bar{v}) \)
- \( \beta, h \sim NG(\bar{\beta}, \bar{V}, \bar{s}^{-2}, \bar{v}) \)
The posterior

The (hyper) parameters of the posterior

$$\beta, h \mid y \sim NG(\overline{\beta}, \overline{V}, \overline{s^{-2}}, \overline{v})$$

are a combination of the (hyper) parameters of the prior and sample information

$$\overline{V} = \frac{1}{V^{-1} + \sum x_i^2}$$

$$\overline{\beta} = \overline{V} \left( V^{-1} \beta + \hat{\beta} \sum x_i^2 \right)$$

$$\overline{v} = v + N$$

$$\overline{vs^2} = vs^2 + vs^2 + \frac{(\hat{\beta} - \beta)^2}{V + \sum x_i^2}$$
Posterior marginal distribution of $\beta$

In regression analysis, we are often interested in the mean and the variance of the coefficient $\beta$. The mean can be found by integrating the posterior w.r.t. $h$ and $\beta$

$$E(\beta | y) = \int \int \beta p(\beta, h | y) dh d\beta = \int \beta p(\beta | y) d\beta$$

The density $p(\beta | y)$ is called the marginal distribution of $\beta$. It can be shown that

$$\beta | y \sim t(\bar{\beta}, \bar{s}^2 \bar{V}, \bar{v})$$

so that

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

$$\text{var}(\beta | y) = \frac{\bar{v}s^2}{\bar{v} - 2\bar{V}}$$
Combining prior and sample information

The posterior mean of $\beta$ is given by

$$\bar{\beta} = \bar{V} \left( \frac{1}{\bar{V}^{-1}} \beta + \hat{\beta} \sum x_i^2 \right)$$

$$\bar{V} = \frac{1}{\frac{1}{\bar{V}^{-1}} + \sum x_i^2}$$

Do these expressions make sense?

- $\bar{V}^{-1}$ is the precision of the prior
- $\sum x_i^2$ is the ”precision” of the data

What happens when $\bar{V}^{-1} = 0$? when $\bar{V} \to 0$?
Example
Simple “empirical” example

Consider a sample of 50 observations generated from the model

\[ y_i = \beta x_i + \varepsilon_i \sim N(0, \sigma^2) \]

with \( \beta = 2, \sigma^2 = 1 \) and \( x_i \sim N(0, 1) \)

Set the prior hyper parameters in \( \beta, h \sim NG(\beta, V, s^{-2}, v) \) to

\[
\begin{align*}
\beta &= 1.5 \\
V &= 0.25 \\
s^{-2} &= 1 \\
v &= 10
\end{align*}
\]
The data
Figure 2.1: Marginal Prior and Posteriors for $\beta$

- **Prior**
- **Posterior**
- **Likelihood**
Change the prior mean

\[
\begin{align*}
\beta & = 2.5 \\
\mathcal{V} & = 0.25 \\
\mathcal{S}^{-2} & = 1 \\
\nu & = 10
\end{align*}
\]
Figure 2.1: Marginal Prior and Posteriors for $\beta$
Change the prior precision

\[
\begin{align*}
\beta &= 1.5 \\
V &= 0.5 \\
S^{-2} &= 1 \\
\nu &= 10
\end{align*}
\]
Figure 2.1: Marginal Prior and Posteriors for $\beta$

<table>
<thead>
<tr>
<th>Prior</th>
<th>Posterior</th>
<th>Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The graph shows the marginal prior and posteriors for $\beta$ as probability density functions. The x-axis represents $\beta$, and the y-axis represents probability density.
Change the prior mean of variance

\[
\begin{align*}
\beta &= 1.5 \\
V &= 0.25 \\
\sigma^{-2} &= 10 \\
v &= 10
\end{align*}
\]
Figure 2.1: Marginal Prior and Posteriors for $\beta$

- **Prior**
- **Posterior**
- **Likelihood**
Change the prior degrees of freedom

\[ \beta = 1.5 \]
\[ V = 0.25 \]
\[ s^{-2} = 1 \]
\[ v = 100 \]
Figure 2.1: Marginal Prior and Posteriors for $\beta$
More data: \( N=100 \)
Figure 2.1: Marginal Prior and Posteriors for $\beta$

- **Prior**
- **Posterior**
- **Likelihood**

![Graph showing marginal prior and posteriors for $\beta$.](image-url)
Even more data: $N=500$
Figure 2.1: Marginal Prior and Posteriors for $\beta$ probability density

<table>
<thead>
<tr>
<th>Prior</th>
<th>Posterior</th>
<th>Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>0.5</td>
<td>0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>1</td>
<td>0.05</td>
<td>0.06</td>
</tr>
</tbody>
</table>
Prior precise but with $\beta$ far from $\hat{\beta}$
Figure 2.1: Marginal Prior and Posteriors for $\beta$

<table>
<thead>
<tr>
<th>Prior</th>
<th>Posterior</th>
<th>Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0$</td>
<td>$0.005$</td>
<td>$0.01$</td>
</tr>
<tr>
<td>$0.5$</td>
<td>$0.01$</td>
<td>$0.015$</td>
</tr>
<tr>
<td>$1$</td>
<td>$0.02$</td>
<td>$0.025$</td>
</tr>
<tr>
<td>$1.5$</td>
<td>$0.015$</td>
<td></td>
</tr>
<tr>
<td>$2$</td>
<td>$0.02$</td>
<td></td>
</tr>
<tr>
<td>$2.5$</td>
<td>$0.015$</td>
<td></td>
</tr>
<tr>
<td>$3$</td>
<td>$0.005$</td>
<td></td>
</tr>
<tr>
<td>$3.5$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$4$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Noninformative prior

Non-informative prior

\[
\begin{align*}
V^{-1} &= 0 \\
\nu &= 0
\end{align*}
\]

\[
\begin{align*}
\overline{V} &= \frac{1}{\sum x_i^2} \\
\bar{\beta} &= \hat{\beta} \\
\nu &= N \\
\overline{\nu s^2} &= \nu s^2
\end{align*}
\]

which are the same as the OLS estimates.

But this prior is so-called *improper*, i.e. the density \( p(\beta, h) \) does not integrate to 1

\[
\int \int p(\beta, h) d\beta dh = \infty \neq 1
\]
Model comparison

We may have two models that could both potentially explain $y$

$$y_i = \beta x_{ji} + \epsilon_{ji} \sim N(0, h_j^{-1})$$

$j = 1, 2$

For model comparison, it is important that the models explain the \textit{same} data

Improper priors can be used only for parameters shared by both models
Model comparison

We can formulate priors and compute posterior just as before

- Prior: $\beta_j, h_j \sim NG \left( \beta_j, V_j, s_j^{-2}, v_j \right)$
- Posterior: $\beta_j, h_j \mid y \sim NG \left( \bar{\beta}_j, \bar{V}_j, \bar{s}_j^{-2}, \bar{v}_j \right)$

Again, the (hyper) parameters of the posterior are a combination of the (hyper) parameters of the prior and sample information

\[
\begin{align*}
\bar{V}_j &= \dfrac{1}{V_j^{-1} + \sum x_{ji}^2} \\
\bar{\beta}_j &= \bar{V}_j \left( V_j^{-1} \beta_j + \hat{\beta}_j \sum x_{ji}^2 \right) \\
\bar{v}_j &= v_j + N \\
\bar{v}_j \bar{s}_j^2 &= v_j s_j^2 + v_j s_j^2 + \frac{\left( \hat{\beta}_j - \beta_j \right)^2}{\bar{V}_j + \sum x_{ji}^2}
\end{align*}
\]
The posterior odds ratio is the relative probabilities of two models conditional on the data

\[ PO_{12} \equiv \frac{p(M_1 | y)}{p(M_2 | y)} = \frac{p(y | M_1)p(M_1)}{p(y | M_2)p(M_2)} \]

- The prior \( p(M_j) \) model probabilities are formulated before seeing the data.
- The marginal likelihood \( p(y | M_j) \) is calculated as

\[ p(y | M_j) = \int \int p(y | \beta, h)p(\beta_j, h_j)dh_jd\beta_j \]
The marginal likelihood for the Normal-Gamma model

The marginal likelihood for the Normal-Gamma model is given by

\[ p(y \mid M_j) = c_j \left( \frac{\bar{V}_j}{V_j} \right)^{\frac{1}{2}} \left( \bar{v}_j s_j^2 \right)^{-\frac{v_j}{2}} \]

where

\[ c_j = \frac{\Gamma \left( \frac{\bar{v}_j}{2} \right) \left( \bar{v}_j s_j^2 \right)^{-\frac{v_j}{2}}}{\Gamma \left( \frac{v_j}{2} \right) \pi^{\frac{N}{2}}} \]

and \( \Gamma(\cdot) \) is the Gamma function.
Posterior odds ratio

The posterior odds ratio is then given by

\[ PO_{12} = \frac{c_1 \left( \frac{V_1}{V_1} \right)^{1/2} \left( \bar{V}_1 \bar{s}_1^2 \right)^{-\frac{v_1}{2}}}{c_2 \left( \frac{V_2}{V_2} \right)^{1/2} \left( \bar{V}_2 \bar{s}_2^2 \right)^{-\frac{v_2}{2}}} \]

The posterior odds ratio rewards

- fitting the data via the term \( \bar{V}_j \bar{s}_j^2 \) which contains the sum of squared errors \( v_j \bar{s}_j^2 \)
- coherence between prior and OLS estimate via the term \( \bar{V}_j \bar{s}_j^2 \) which contains \( \left( \beta_j - \hat{\beta}_j \right)^2 \)

When models have different number of parameters, the posterior odds ratio also rewards parsimony, i.e. the model with fewer parameters.
That’s it for today.