Bayesian Inference in the Linear Regression Model

Econ 690

Purdue University

Outline

The inverse Gamma distribution (again!)

- We denote the *inverted Gamma* density as $Y \sim IG(\alpha, \beta)$. Though different parameterizations exist (particularly for how β enters the density), we utilize the following form here:

$$Y \sim IG(\alpha, \beta) \Rightarrow p(y) = [\Gamma(\alpha)\beta^{\alpha}]^{-1} y^{-(\alpha+1)} \exp(-1/[y\beta]), \quad y > 0.$$

- The mean of this inverse Gamma is $E(Y) = [\beta(\alpha - 1)]^{-1}$.

▸ Jump to Prior 1, σ^2 posterior ▸ Jump to Prior 1, β posterior

The student-t distribution (again)

A continuous $k-$dimensional random vector, $Y = (Y_1, .., Y_k)'$, has a *t distribution* with mean μ (a $k-$vector), scale matrix Σ (a $k \times k$ positive definite matrix) and ν (a positive scalar referred to as a *degrees of freedom* parameter), denoted $Y \sim t(\mu, \Sigma, \nu)$, if its p.d.f. is given by:

$$f_t(y|\mu, \Sigma, \nu) = \frac{1}{c_t}|\Sigma|^{-\frac{1}{2}} \left[\nu + (y-\mu)' \Sigma^{-1} (y-\mu)\right]^{-\frac{\nu+k}{2}},$$

Jump to Prior 1, β posterior

The Linear Regression Model

- The linear regression model is the workhorse of econometrics.
- We will describe Bayesian inference in this model under 2 different priors. The "default" non-informative prior, and a conjugate prior.
- Though this is a standard model, and analysis here is reasonably straightforward, the results derived will be quite useful for later analyses of linear and nonlinear models via MCMC methods.
- We will also obtain results under a Gaussian sampling model. Later we will show how this assumption can be relaxed in practice.

The Model

The model we consider is:

$$y_i = x_i\beta + \epsilon_i, \quad i = 1, 2, \dots, n, \quad \epsilon_i | X \overset{iid}{\sim} N(0, \sigma^2).$$

Stacking quantities over i, we write

$$y = X\beta + \epsilon,$$

where

$$y_{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X_{n\times k} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \epsilon_{n\times 1} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

The Likelihood

$$y_i = x_i\beta + \epsilon_i, \quad i = 1, 2, \ldots, n, \quad \epsilon_i | X \overset{iid}{\sim} N(0, \sigma^2).$$

- The Jacobian of the transformation from ϵ to y is unity.
- Thus, the likelihood function is given as

$$\begin{aligned} L(\beta, \sigma^2) &= (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right] \\ &\propto (\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right] \end{aligned}$$

- This is one-half of what is needed to obtain the posterior $p(\beta, \sigma^2 | y)$.

Prior #1

- A standard "default" procedure is to place a non-informative (improper) prior on (β, σ^2).
- The first step in this regard is to assume prior independence between these quantities:
  -
- For the marginal prior for β, this is often specified as the "flat" (improper) prior:
  -

  for some constant c_1.

Prior #1

$$p(\beta, \sigma^2) = p(\beta)p(\sigma^2).$$

- For the variance parameter σ^2, we note that it must be positive. A common practice in this situation, dating to the pioneering work of Jeffreys, is to employ a uniform (improper) prior for the log of σ^2.
- Let $\psi = \log \sigma^2$. Then,
  -

  for some constant c_2.
- Note that the Jacobian of the transformation from ψ to σ^2 is σ^{-2}. Thus, we have the implied prior
  -

Prior #1

- Putting these together, we obtain the prior

$$p(\beta, \sigma^2) \propto \sigma^{-2}.$$

- We combine this with the likelihood

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

to obtain

  -

Prior #1

Now, consider the least-squares quantity:

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

and also define

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

Prior #1

The quadratic form in the exponential kernel of the likelihood can be manipulated as follows:

-

where the last line follows from the well-known orthogonality condition associated with the least-squares residuals.

Prior #1

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

Using our previous result, we can write this as

$$L(\beta,\sigma^2) \propto (\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}[SSE + (\beta - \widehat{\beta})'X'X(\beta - \widehat{\beta})]\right].$$

and thus

-

Prior #1

$$p(\beta, \sigma^2 | y) \propto (\sigma^2)^{-(n+2)/2} \exp\left[-\frac{1}{2\sigma^2}[SSE + (\beta - \widehat{\beta})'X'X(\beta - \widehat{\beta})]\right].$$

We can express this posterior as

$$\begin{aligned} p(\beta, \sigma^2 | y) &\propto (\sigma^2)^{-(n+2)/2} \exp\left[-\frac{1}{2\sigma^2} SSE\right] \times \\ & \exp\left[-\frac{1}{2\sigma^2}(\beta - \widehat{\beta})'X'X(\beta - \widehat{\beta})]\right] \\ &= (\sigma^2)^{-[(n-k)/2]-1} \exp\left[-\frac{1}{2\sigma^2} SSE\right] \times \\ & (\sigma^2)^{-k/2} \exp\left[-\frac{1}{2\sigma^2}(\beta - \widehat{\beta})'X'X(\beta - \widehat{\beta})]\right] \end{aligned}$$

Prior # 1

In this form, it is straightforward to see that

-

and

-

Jump to Inverse Gamma Density

Prior # 1

As for the variance parameter note that, as a property of the inverse gamma distribution,

-

Therefore, the posterior mean of the variance parameter is not the typical frequentist estimator, s^2, but approaches s^2 as $n \rightarrow \infty$ (and collapses around this value).

Prior # 1

As for the marginal posterior for β, note that

-

Note that the integrand above is the kernel of an

-

density. Thus, the desired integral is simply the reciprocal of the normalizing constant of this density.

Prior #1

For an $IG(\alpha, \beta)$ density, the reciprocal of the normalizing constant is $\Gamma(\alpha)\beta^{\alpha}$ Jump to Inverse Gamma Density.

In our case, therefore, the integrand sets

-

and

-

Prior #1

It follows that [letting $\nu = n - k, \quad s^2 = (y - X\hat{\beta})'(y - X\hat{\beta})/\nu$]:

-

In this form, it is seen that the marginal posterior for β is multivariate student-t.
Specifically,

-

Example with Log Wage Data

- We illustrate how to carry out Bayesian inference in the regression model with a simple example.
- The data set used contains 1,217 observations on three variables: hourly wages, education and a standardized test score.
- We consider the model

$$y_i = \beta_0 + \beta_1 Ed_i + \epsilon_i,$$

  where y is the log of the hourly wage.
- We also employ the flat prior

$$p(\beta_0, \beta_1, \sigma^2) \propto \sigma^{-2}.$$

Example with Log Wage Data

- Since the focus of such studies is usually on the "return to education" parameter β_1, we confine our discussion to that parameter.
- From our previous derivations, we know that

$$E(\beta|y) = \hat{\beta} = [1.18 \ .091]'$$

so that (using the posterior mean as an estimate), an additional year of education increases wages by about 9.1 percent.

Example with Log Wage Data

- In addition (see properties of the Student-t in the distributional catalog)

$$\text{Var}(\beta|y) = \frac{\nu}{\nu - 2} s^2 (X'X)^{-1},$$

with $\nu = n - k = 1,217 - 2 = 1,215$.

- Calculating the above using our data, and taking the square roots of the diagonal elements, we find

$$\text{Std}(\beta_1|y) \approx .0066.$$

Example with Log Wage Data

- Like the normal distribution, marginals and conditionals from the multivariate Student-t are also of the Student-t form. (See distributional catalog).
- In particular,

$$\beta_1|y \sim t\left(\widehat{\beta}_1, s^2(X'X)^{-1}_{(2,2)}, \nu\right).$$

- Thus, putting these pieces together, we obtain

$$\beta_1|y \sim t(.0910, [.0066]^2, 1, 215).$$

This could be plotted in Matlab to provide a picture of the marginal posterior. (Of course, in this case, the posterior is virtually identical to the normal distribution with the given mean and variance).

Example with Log Wage Data

- Again, like the normal distribution, one can convert the more general location-scale version of the t distribution to its standardized form by noting

$$\frac{\beta_1 - \widehat{\beta}_1}{\sqrt{s^2(X'X)^{-1}_{(2,2)}}} \Big| y \sim t(0, 1, \nu).$$

- Thus, using the command "tcdf" in Matlab, we can calculate quantities of interest like

$$\Pr(\beta_1 < .10|y) = T_\nu \left(\frac{.10 - \widehat{\beta}_1}{\sqrt{s^2(X'X)^{-1}_{(2,2)}}} \right) \approx .9135,$$

with T_ν denoting the cdf of the standardized t distribution with ν degrees of freedom.

Prior #2

This time, suppose you employ the conjugate prior:

-

Before deriving the posterior results under this prior, we must first review the following completion of the square formula:

$$(x-\underline{\mu}_1)'A(x-\underline{\mu}_1)+(x-\underline{\mu}_2)'B(x-\underline{\mu}_2)=$$
$$(x-\overline{\mu})'C(x-\overline{\mu})+(\underline{\mu}_1-\underline{\mu}_2)'D(\underline{\mu}_1-\underline{\mu}_2),$$

where

$$\begin{aligned} C &= A+B \\ \overline{\mu} &= C^{-1}(A\underline{\mu}_1+B\underline{\mu}_2) \\ D &= \left(A^{-1}+B^{-1}\right)^{-1} \end{aligned}$$

Jump to Prior #2, $\beta|\sigma^2$ posterior

Prior #2

The prior can therefore be written as

$$p(\beta|\sigma^2) \propto [\sigma^2]^{-k/2} \exp\left[-\frac{1}{2\sigma^2}(\beta-\mu)' V_\beta^{-1}(\beta-\mu)\right],$$

$$p(\sigma^2) \propto [\sigma^2]^{-(a+1)} \exp\left[-\frac{1}{b\sigma^2}\right].$$

The posterior is obtained by combining these priors with the likelihood:

$$L(\beta,\sigma^2) \propto (\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}[SSE + (\beta-\widehat{\beta})'X'X(\beta-\widehat{\beta})]\right].$$

Prior #2

Putting these pieces together, we obtain

$$p(\beta, \sigma^2|y) \propto [\sigma^2]^{-\left(\frac{n+k}{2}+a+1\right)} \times$$
$$\exp\left[-\frac{1}{2\sigma^2}\left((\beta-\mu)'V_\beta^{-1}(\beta-\mu)+(\beta-\widehat{\beta})'X'X(\beta-\widehat{\beta})\right)\right] \times$$
$$\exp\left[-\frac{1}{2\sigma^2}SSE\right]\exp\left(-\frac{1}{b\sigma^2}\right).$$

Prior #2

Using our completion of the square formula ▸ Jump to formula we can write

$$
\begin{aligned}
& (\beta-\mu)' V_\beta^{-1}(\beta-\mu) \quad + \quad (\beta-\widehat{\beta})' X' X(\beta-\widehat{\beta}) \\
= & (\beta-\overline{\beta})' \overline{V}_\beta(\beta-\overline{\beta}) \quad + \quad (\mu-\widehat{\beta})' \tilde{V}_\beta(\mu-\widehat{\beta}),
\end{aligned}
$$

where

$$
\begin{aligned}
\overline{V}_\beta &= V_\beta^{-1}+X'X \\
\overline{\beta} &= \overline{V}_\beta^{-1}\left[V_\beta^{-1}\mu+X'X\hat{\beta}\right] \\
\tilde{V}_\beta &= \left[V_\beta+(X'X)^{-1}\right]^{-1}
\end{aligned}
$$

Prior #2

- Note that the second quadratic form does not involve β and thus is absorbed in the normalizing constant of the posterior conditional $\beta|\sigma^2, y$.

- To derive the posterior conditional $\beta|\sigma^2, y$, we can consider only those terms in the expression for $p(\beta, \sigma^2|y)$ that involve β. This produces:

$$p(\beta|\sigma^2, y) \propto \exp\left[-\frac{1}{2\sigma^2}(\beta - \overline{\beta})'\overline{V}_\beta(\beta - \overline{\beta})\right],$$

- or equivalently,

$$\beta|\sigma^2, y \sim N(\overline{\beta}, \sigma^2\overline{V}_\beta^{-1}).$$

Prior #2

- Consider what happens when a "flat" prior for β is employed in the sense that V_β is a diagonal matrix with "large" elements on the diagonal. Then,
- $$\overline{V}_\beta = V_\beta^{-1} + X'X \approx X'X$$
- $$\overline{\beta} = \overline{V}_\beta^{-1} \left[V_\beta^{-1}\mu + X'X\hat{\beta} \right] \approx (X'X)^{-1}X'X\hat{\beta} \approx \hat{\beta}.$$
- Thus, results approach those obtained for Prior #1, as expected.

Marginal likelihood in the LRM

Consider the stacked regression model:

$$y = X\beta + \epsilon, \quad \epsilon | X \sim N(0, \sigma^2 I_n).$$

Suppose we employ the following priors:

$$\beta | \sigma^2 \sim N(\mu, \sigma^2 V_\beta)$$

$$\sigma^2 \sim IG\left(\frac{\nu}{2}, 2(\nu\lambda)^{-1}\right).$$

(To this point, we have used $IG(a, b)$ as the prior for σ^2. The above is the same thing, but simply writes the prior hyperparemeters in a slightly different way - this will simplify the resulting expressions).

Marginal likelihood in the LRM

Note that

$$p(y|\sigma^2) = \int p(y|\beta, \sigma^2)p(\beta|\sigma^2)d\beta.$$

Note that our prior for β can be written as

$$\beta = \mu + \eta, \quad \eta \sim N(0, \sigma^2 V_\beta).$$

Substituting this result into our regression model, we obtain

$$y = X\mu + [X\eta + \epsilon],$$

or equivalently,

-

Marginal likelihood in the LRM

$$y|\sigma^2 \sim N(X\mu, \sigma^2[I_n + XV_\beta X']).$$

Recall that the marginal likelihood is $p(y)$, and thus we need to determine

$$p(y) = \int_0^\infty p(y|\sigma^2)p(\sigma^2)d\sigma^2.$$

Writing this out, we obtain

$$\begin{aligned} p(y) \quad \propto \quad & \int_0^\infty [\sigma^2]^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(y - X\mu)'[I_n + XV_\beta X']^{-1}(y - X\mu)\right) \times \\ & [\sigma^2]^{-\left(\frac{\nu}{2}+1\right)} \exp\left(-\frac{1}{2[\nu\lambda]^{-1}\sigma^2}\right) \end{aligned}$$

Marginal likelihood in the LRM

Grouping together like terms, we obtain

$$
\begin{aligned}
p(y) \quad \propto \quad & \int_0^\infty [\sigma^2]^{-\left(\frac{n+\nu}{2}+1\right)} \times \\
& \exp\left(-\frac{1}{\sigma^2}\frac{1}{2}\left[(y - X\mu)'[I_n + XV_\beta X']^{-1}(y - X\mu) + (\nu\lambda)\right]\right)
\end{aligned}
$$

As before, we recognize the above as the kernel of an IG density. Specifically, it is the kernel of an

$$
IG\left(\frac{n+\nu}{2}, 2\left[(y - X\mu)'[I_n + XV_\beta X']^{-1}(y - X\mu) + (\nu\lambda)\right]^{-1}\right).
$$

density.

Marginal likelihood in the LRM

The results of our earlier exercise of finding the marginal posterior distribution of β can be re-applied here. When doing so, we find that the kernel of the marginal likelihood is given as follows:

$$p(y) \propto \left[(\nu\lambda) + (y - X\mu)'[I_n + XV_\beta X']^{-1}(y - X\mu)\right]^{-\frac{n+\nu}{2}}$$

In this form, it is seen that

-

This result can be used to test hypotheses (i.e., calculate Bayes factors) in a regression context.