Model Estimation With Inequality Constraints: A Bayesian Approach Using SAS/IML® Software

Charlie Hallahan
Economic Research Service, USDA

Abstract

Economic models derived from economic theory are generally accompanied with various sets of equality and inequality restrictions. For example, in an economic relationship between consumption, $c$, and income, $y$, and $c = a + \beta y + e$, the coefficient $\beta$ represents the marginal propensity to consume and should lie in the interval $[0, 1]$. Treating the parameters as fixed and using constrained least squares, estimates can be obtained, but the complex distributional properties of the estimates complicate inference. The Bayesian approach treats the parameters as random variables whose prior distribution incorporates the inequality restrictions. The steps necessary to obtain Bayesian estimates via Monte Carlo integration will be discussed and illustrated with IML programs.

Bayesian Estimation of Regression Models

Suppose that the parameters, $\beta$, in the regression model $y = X\beta + \varepsilon$, must satisfy certain inequality restrictions. Following the classical frequentist approach, a constrained least squares estimator could be obtained by solving a quadratic programming problem. A drawback to this estimator is that its distributional properties are not straightforward, making inference difficult.

An alternative approach is to treat $\beta$ as a random variable and apply Bayesian techniques. The article by Griffiths [1988] is a good intuitive discussion of applying Bayesian methods to impose inequality restrictions on regression models. A simple example is given there of a consumption function, $y = \beta x + \varepsilon$, where $y$ is consumption and $x$ is income. The parameter $\beta$ is called the long-run marginal propensity to consume (or mpc), and economic theory states that $\beta$ should be between 0 and 1.

More complicated models, yet similar in principle, are treated by Chalfant, Gray and White [1991], Fernandez-Cornejo [1992], and Geweke [1986, 1988, 1989]. These examples will be discussed further below.

Let $n(\beta)$ be the prior probability distribution density of $\beta$ and $L(\beta/y)$ the likelihood function given the sample $y$. Then by Bayes’ Theorem,

$$p(\beta/y) = \frac{L(\beta/y) \pi(\beta)}{f(y)}$$

or $p(\beta/y) \propto L(\beta/y) n(\beta)$, where $f(y)$ is the marginal density of $y$ and $p(\beta/y)$ is the posterior density of $\beta$. The uncertainty in $\beta$ before the sample is drawn is summarized by $n(\beta)$, and $p(\beta/y)$ represents the uncertainty in $\beta$ after the sample is drawn. A Bayesian point estimate of $\beta$ can be obtained by defining a quadratic loss function and finding the value of $\beta$ that minimizes expected loss (under the posterior distribution). The optimal estimate under a quadratic loss function is the mean of the posterior distribution.

For the simple consumption function example, assuming $\varepsilon \sim N(0, I)$, the likelihood will have the form

$$L(\beta/y) = \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (y_i - x_i \beta)^2\right)$$

Writing $y_i - x \beta = y_i - x \beta^* + x \beta - x \beta$, expanding and ignoring terms not involving $\beta$, we get

$$L(\beta/y) = \exp\left(-\frac{1}{2} \sum_{i=1}^{n} x_i^2 (\beta - \beta^*)^2\right)$$

where $\beta^*$ is the OLS estimate of $\beta$. The posterior density for the mpc is then

$$p(\beta/y) \propto \pi(\beta) \exp\left(-\frac{1}{2} \sum_{i=1}^{n} x_i^2 (\beta - \beta^*)^2\right)$$

The specific form of the posterior depends on the prior, $n(\beta)$. The prior could be categorized as follows:
(i) informative: e.g., \( \beta \sim N(.85, .004) \), which is roughly equivalent to saying that \( P(.75 < \beta < .95) = 0.9 \).

(ii) non-informative: e.g., \( n(\beta) = 1 \) for \(-\infty < \beta < \infty \). This is an improper density.

(iii) inequality restriction: e.g., \( n(\beta) = 1 \) for \( 0 \leq \beta \leq 1 \), uniform density.

For the MPC, a prior of type (iii) seems appropriate. The posterior becomes a truncated normal distribution in this case. In general, priors reflecting inequality restrictions lead to truncated distributions for the posterior. The next step is to find the mean of the posterior distribution. We can also ask such questions as: What is the probability that \( \beta > .95 \)? Each of these questions leads to integrals of the form:

\[
E[\beta / \mathcal{X}] = \int_0^1 \beta \, p(\beta / \mathcal{X}) \, d\beta
\]

\[
P(0 \leq \beta \leq .95) = \int_0^{.95} p(\beta / \mathcal{X}) \, d\beta
\]

In most higher dimensional cases, such integrals cannot be evaluated analytically or approximated numerically and must be estimated using Monte Carlo integration.

**Monte Carlo Integration**

The principle behind Monte Carlo integration is as follows:

To evaluate the integral \( \int g(x)f(x) \, dx \), where \( f(x) \) is a (possibly multivariate) density function, we note that \( \int g(x)f(x) \, dx = E_x[g(x)] \), the expectation of \( g(x) \) under the \( f \)-distribution. (A subscript on the expectation operator refers to the distribution with respect to which the expectation is being taken.) If we can make \( n \) draws \( \{x_i\} \) from the \( f \)-distribution, then

\[
1 \, \sum_{i=1}^{n} \, g(x_i) \approx \int g(x) \, f(x) \, dx
\]

To apply these ideas in the Bayesian estimation setting, let \( \theta \) be the parameter vector for a model with likelihood function \( L(\theta / \mathcal{Y}) \) and prior density \( n(\theta) \) and let \( g(\theta) \) be any function of interest. Then if \( \{\theta_i\}, i = 1 \) to \( n \), is a sequence of independent draws from the posterior density, \( p(\theta / \mathcal{Y}) = L(\theta / \mathcal{Y})n(\theta) / \int L(\theta / \mathcal{Y})n(\theta) \, d\theta \), it follows by the strong law of large numbers that

\[
\frac{1}{n} \sum_{i=1}^{n} \, g(\theta_i) \approx E_{\theta}[g(\theta)] = \bar{g}
\]

where \( \rightarrow \) is almost sure convergence.

When \( g(\theta) = \theta_i \), one of the components of \( \theta \), then we are just estimating \( E_{\theta}[\theta_i] \) by the sample mean of a random sample from the posterior distribution.

Suppose we want to estimate the probability that \( \theta \in D \), some region of the parameter space. Let \( g(\theta) \) be an indicator function equal to 1 if \( \theta \in D \), and 0 otherwise. Then the integral

\[
\int_D g(\theta) \, p(\theta / \mathcal{X}) \, d\theta = \int_D p(\theta / \mathcal{X}) \, d\theta
\]

is \( \text{Prob}(\theta \in D) \).

By making random draws from the posterior and counting the number of draws that satisfy \( \theta \in D \), say \( n_1 \), we can use \( n_1/n \) as the estimate.

Equation (7) states that the Monte Carlo estimate will converge to the correct limit, but doesn't indicate the rate of convergence or how accurate the approximation is. To do that, a central limit theorem is needed.

If \( \text{var}[g(\theta)] \), the variance of \( g(\theta) \) under the posterior distribution, exists, then a central limit theorem applies to show that

\[
\sqrt{n} \left( \frac{1}{n} \sum_{i=1}^{n} \, g(\theta_i) - \bar{g} \right) \rightarrow N(0, \text{var}[g(\theta)])
\]

where \( \Rightarrow \) is convergence in distribution.

Since \( \text{var}[g(\theta)] = E_{\theta}[g(\theta)^2] - E_{\theta}[g(\theta)]^2 \), Monte Carlo integration can also be used to estimate the variance.

When random sampling from the posterior is not possible, the next best alternative may be to sample from the likelihood density, \( L(\theta / \mathcal{Y}) / \int L(\theta / \mathcal{Y}) \, d\theta \). Let \( c = \int L(\theta / \mathcal{Y}) \, d\theta \), the
normalizing constant for the likelihood kernel. Then we can write

\[ E_p[g(\theta)] = \int g(\theta) \pi(\theta) \, d\theta \]

\[ = \frac{\int g(\theta) \pi(\theta) L(\theta|x) \, d\theta}{\int \pi(\theta) L(\theta|x) \, d\theta} \]

\[ = \frac{\int g(\theta) \pi(\theta) L(\theta|x) \, d\theta}{\int \pi(\theta) L(\theta|x) \, d\theta} \]

\[ = \frac{\int g(\theta) \pi(\theta) L(\theta|x) \, d\theta}{\int \pi(\theta) L(\theta|x) \, d\theta} \]

\[ = \frac{E_L[g(\theta) \pi(\theta)]}{E_L[\pi(\theta)]} \]

Note that as soon as we sample from a distribution other than the posterior, we must estimate a ratio of integrals.

Given \{\theta_i\}, \( i = 1 \) to \( n \), a sequence of independent draws from the likelihood density, it can be shown that

\[ \frac{1}{n} \sum_{i=1}^{n} \frac{g(\theta_i) \pi(\theta_i)}{\pi(\theta_i)} = E_p[g(\theta)] \equiv \bar{g} \]

and that

\[ \sqrt{n}(\bar{g} - g) \sim N(0, \text{var}(g(\theta) \pi(\theta))) \]

where \( \text{var}(g(\theta) \pi(\theta)) \) represents the variance of \( g(\theta)\pi(\theta) \) under the likelihood distribution.

Importance Sampling

When it is not possible to draw from either the likelihood or posterior distributions, draws are made from another density called the importance density. This works because of the simple identity

\[ \int h(x) f(x) \, dx = \int \frac{h(x) f(x)}{f(x)} I(x) \, dx \]

\[ = \int w(x) I(x) \, dx \]

\[ = E_L[w(x)] \]

Therefore,

\[ \int h(x) f(x) \, dx = \frac{1}{n} \sum_{i=1}^{n} \frac{h(x_i) f(x_i)}{f(x_i)} \]

where the \( x_i \) are random draws from the \( \pi \)-distribution.

Let \( l(\theta) \) be the density from which the random draws \{\theta_i\} are made and define the weight function \( w(\theta) = L(\theta|x)\pi(\theta)/l(\theta) \). Then the Monte Carlo approximation to \( E_p[g(\theta)] \) is

\[ \bar{g} = \frac{\sum_{i=1}^{n} g(\theta_i) w(\theta_i)}{\sum_{i=1}^{n} w(\theta_i)} \]

Simple Monte Carlo integration, as described above, has either \( I(\pi) \sim L(\pi) \) with \( w(\pi) = k \cdot \pi(\theta) \) or \( I(\pi) \sim L(\pi)\pi(\theta) \) with \( w(\pi) = k \).

The main theorem in Geweke (1989) provides a way to measure the numerical accuracy of the Monte Carlo approximation.

Theorem (Geweke) Suppose \( E_p[w(\theta)] < \infty \) and \( E_p[g(\theta)w(\theta)] < \infty \). Let

\[ \sigma^2 = E_p[(g(\theta) - \bar{g})^2w(\theta)] \]

\[ \theta_n^2 = \frac{\sum_{i=1}^{n} (g(\theta_i) - \bar{g})^2 w(\theta_i)^2}{\sum_{i=1}^{n} w(\theta_i)^2} \]

Then

\[ \pi^\text{nse}(\bar{g} - \theta) = N(0, \sigma^2) \]

\[ \pi \theta_n^2 = \sigma^2 \]

\[ \theta_n = (\theta_n^2)^{1/2} \]

is called the numerical standard error (nse) of \( \bar{g} \) and measures the
numerical accuracy of the estimate $\bar{g}_n$. The definition of the nse shows that it is sensitive to $\text{var}[g(\theta)]$ under the posterior (which is not in the control of the statistician) and the size of the weights, $w(\theta)$ (which is controllable by choice of an importance density).

As a standard of reference, we could use the nse that would occur from using the posterior itself as the Monte Carlo sampling distribution. When $l(\theta) = p(\theta|y)$, then $\sigma^2 = \text{var}[g(\theta)]$ and the nse from $n_p$ Monte Carlo draws would be $\text{var}[g(\theta)]/n_p$. For example, 10,000 draws would reduce the nse to 1% of the variance of $g(\theta)$ under the posterior.

We could ask the question: if another density $l(\theta)$ is used as the sampling density, how many iterations, $n_\alpha$, would it take to achieve the same nse, i.e., $\text{var}[g(\theta)]/n_\alpha = \sigma^2/n_p$, or $\text{var}[g(\theta)]/n_\alpha = \sigma^2/n_p$. Geweke calls this ratio the relative numeric efficiency, RNE. The RNE can be interpreted as the ratio of the number of Monte Carlo draws from the posterior to the number of draws from an importance density to achieve the same nse. A RNE close to 1 indicates an importance density as efficient as the posterior. Geweke has shown that non-random sampling (as discussed in the next section) can result in RNE’s close to 100.

**Antithetic Sampling**

So far we have talked about drawing independent samples from a distribution. When we sample from the posterior and our estimate is $\bar{g}_n$, as in equation (7), then $\text{var}[\bar{g}_n] = \text{var}[g(\theta)]/n$. Antithetic sampling consists of drawing in correlated pairs with the goal of decreasing $\text{var}[\bar{g}_n]$. To see how this would work, suppose that $X$ and $X'$ are draws from a distribution with mean $\mu$ and variance $\sigma^2$. The usual estimate of $\mu$ is $\bar{\mu} = \frac{1}{2}(X + X')$. If $X$ and $X'$ are independent, then $\text{var}(\bar{\mu}) = \frac{1}{2}\sigma^2$. In general, $\text{var}(\bar{\mu}) = \frac{1}{4}\text{var}(X) + 2\text{cov}(X,X') + \text{var}(X') = \frac{1}{2}\sigma^2(1 + \text{corr}(X,X'))$. Thus if $X$ and $X'$ are negatively correlated, the $\text{var}(\bar{\mu})$ can be reduced.

In many Bayesian estimation problems the likelihood is symmetric about an MLE estimate $\widehat{\theta}$ and a random draw is typically of the form $\widehat{\theta} + Z$, where $Z$ is a draw from, say, a multivariate normal or t distribution. The corresponding antithetic draw is then $\widehat{\theta} - Z$.

**Some Examples**

The series of papers by Geweke contain a number of interesting examples applying the above techniques. For example:

Geweke (1986): Three regression examples are presented with various kinds of inequality restrictions ranging from straightforward sign constraints, such as $\beta_2 > \beta_3$ or $\beta_4 > 0$, to nonlinear restrictions on an autoregressive model. The restriction that an AR model be stable requires that the roots of the associated lag polynomial have all its roots outside the unit circle. The SAS/IML function polyroot is exactly what is needed to check each draw to see if the restriction holds.

Geweke (1988): The means of predictive densities in a vector autoregression are calculated.

Geweke (1989): Markov chain and ARCH models are studied.

The paper by Chalfant, Gray and White (1991) uses the above described techniques to estimate a system of demand equations. Economic theory imposes a set of equality (symmetry, homogeneity) and inequality (monotonicity, concavity and substitutability) restrictions on the parameters in the system. The equality restrictions are easily handled by direct substitution. The discussion below will describe in general terms the kinds of quantities that arise in their model and prior and to show how SAS/IML can do the necessary calculations.

Let $g$ be the vector of parameters in the system and suppose $g \in \mathbb{R}^n$, where $\mathbb{R} = \{\text{real numbers}\}$. The inequality restrictions can be expressed as $g \in \mathbb{D} \subseteq \mathbb{R}^n$, for some subset $D$ of $\mathbb{R}^n$. The (inequality restriction) prior would be $\pi(g) \propto c$, $\forall g \in D$, $c$ constant, i.e. the prior information is that $g$ must satisfy the inequality restrictions, but is otherwise uninformative.

The system of equations is estimated using Zellner’s Seemingly Unrelated Regression (SUR) technique, where the error vector $\varepsilon$ is
assumed to be distributed as a multivariate normal, \( \mathbf{\theta} \sim N(\mathbf{Q}, \mathbf{I}) \). A normal likelihood, diffuse prior on \( \Sigma \), and inequality prior on \( \mathbf{\theta} \) results in a (marginal) posterior density for \( \mathbf{\theta} \) of the form:

\[
p(\mathbf{\theta} | y) = \pi(\mathbf{\theta}) | A |^{-T/2}
\]

where \( T \) = the number of observations and \( A = (a_i) \),

\[
a_i = g_i(\mathbf{\theta}) g_i(\mathbf{\theta})
\]

and \( g_i(\mathbf{\theta}) \) is the \( i \)th vector of residuals for the \( i \)th equation in the system. Note that \( p(\mathbf{\theta} | y) \) is a truncated distribution since \( n(\mathbf{\theta}) = 0 \) for \( \mathbf{\theta} \notin D \). Since \( p(\mathbf{\theta} | y) \) is not a "familiar" density, a random sample cannot be drawn from the posterior distribution and a truncated multivariate t-distribution is used as an importance density.

The following two questions can now be addressed:

1. What is the probability of \( \mathbf{\theta} \) satisfying the inequality restrictions? i.e., evaluate the integral

\[
P(\mathbf{\theta} \in D) = \int_{\mathbf{\theta} \in D} \frac{L(\mathbf{\theta} | y)}{I(\mathbf{\theta})} I(\mathbf{\theta}) \, d\mathbf{\theta}
\]

assuming a diffuse prior on \( \mathbf{\theta} \).

2. What is the mean of the posterior distribution? i.e., evaluate the integral

\[
E_p(\mathbf{\theta} | y, \mathbf{\theta}) = \frac{\int \mathbf{\theta} \pi(\mathbf{\theta}) \frac{L(\mathbf{\theta} | y)}{I(\mathbf{\theta})} I(\mathbf{\theta}) \, d\mathbf{\theta}}{\int_{\mathbf{\theta} \in D} \frac{L(\mathbf{\theta} | y)}{I(\mathbf{\theta})} I(\mathbf{\theta}) \, d\mathbf{\theta}}
\]

The last equation follows since \( n(\mathbf{\theta}) \) is just the indicator function for the set \( D \).

An outline of the algorithm to answer these questions, along with the accompanying IML code, is given next:

**Step 1:** Estimate the unrestricted model by iterated SUR to obtain the maximum likelihood estimates \( \hat{\mathbf{\theta}} \) and \( V(\hat{\mathbf{\theta}}) \). Since there is no need to reinvent the wheel in SAS/IML, this can be done with Proc Model in SAS/ETS software.

**Step 2:** Construct the multivariate t distribution function to serve as the importance density. In the following IML code, the SAS data set 'ets.theta' contains the output from Proc Model. The first observation is \( \mathbf{\theta} \) and the remaining observations are the rows of \( V(\hat{\mathbf{\theta}}) \). Since \( \mathbf{\theta} \) has 15 elements in this example, 'ets.theta' has 16 observations. The first section of code brings the output from Proc Model into IML.

```plaintext
* read estimates saved from Proc Model; library ets 'mydir';
+ 1st read theta and transpose;
use mydir.theta var{p1 p2 ... p15};
read point 1 into theta;
theta = theta';
+ next read covariance matrix;
read point (2: 16) into cov;
+ find Cholesky factor of cov;
H = root(cov);
+ save matrices in storage catalog;
reset storage = 'mylib.unconstr';
store theta H;
```

Now that we have a matrix \( H \) such that \( V(\hat{\mathbf{\theta}}) = H' H \), we can construct the multivariate t with \( \lambda \) degrees of freedom. Antithetic draws are used as suggested by Geweke [1988].

```plaintext
* set degrees of freedom for this problem;
df = 4;
+ 1st draw multivariate normal, size of theta;
nparms = nrow(theta);
norm1 = normal(J(nparms,1,0));
+ next draw multivariate normal, size of df;
norm2 = normal(J(df,1,0));
+ construct chi-square using norm2;
chisqr = (norm2' * norm2);
+ now put pieces together;
add = H'norm1/(sqrt(chisqr/df));
* get random draw from importance distn;
theta1 = theta + add;
* get antithetic draw;
theta2 = theta - add;
```

The above statements would appear in a loop where a total of \( N \) draws would be made.

**Step 3:** With each replication in the above loop, check if the random draws \( \mathbf{\theta}_1 \) and \( \mathbf{\theta}_2 \) satisfy the inequality restrictions. Using the "successful" draws, say there are \( n \) out of \( N \) such that

```plaintext
* read estimates saved from Proc Model;
library ets 'mydir';
+ 1st read theta and transpose;
use mydir.theta var{p1 p2 ... p15};
read point 1 into theta;
theta = theta';
+ next read covariance matrix;
read point (2:16) into cov;
+ find Cholesky factor of cov;
H = root(cov);
+ save matrices in storage catalog;
reset storage = 'mylib.unconstr';
store theta H;
```
total of N, we can estimate the posterior mean and probability of the restrictions holding by:

\[(18) \text{Define } w(\theta) = \frac{L(\theta \mid y)}{I(\theta)} \]

\[= \frac{\sum_{i=1}^{n} \theta_{i} w(\theta_{i})}{\sum_{i=1}^{n} w(\theta_{i})} \]

\[P(\theta \in D) = \frac{\sum_{i=1}^{n} w(\theta_{i})}{\sum_{i=1}^{n} w(\theta_{i})} \]

The denominators in the above expressions are normalizing factors to account for the fact that we did not carry along constants with the various density functions.

Step 4. The next task with each draw $\theta$ is to evaluate the posterior and the importance densities at that value of $\theta$. Assuming the $T \times q$ matrix $\text{errs}$ ($T =$ number of observations and $q =$ number of equations estimated) has already been defined to be the errors for the system of equations in the model after the drawn $\theta$ is substituted into the equations, we have:

- define matrix $A$ used for posterior;
  
  $A = \text{errs} \cdot \text{errs}$;

- marginal posterior pdf for theta;
  
  $f_{\text{post}} = 1/(\text{det}(A)) \cdot (\text{obs} / 2)$;

- multivariate $t$, importance pdf;
  
  $f_{t} = 1/(\text{df} + \text{add} \cdot \text{inv}(\text{cov}) \cdot \text{add}) \cdot (\text{df} + \text{parm} / 2)$;

Step 5. The last calculation to make for each iteration is to check whether or not the draw, $\theta$, is "successful" or not, i.e., does $\theta$ satisfy the inequality restrictions of $\theta \in D$. In this example "concavity" means that the eigenvalues of a certain matrix are negative. Assuming the matrix of interest is called $\text{submat}$ and has already been constructed, then:

- find eigenvalues of submat;
  
  eval = eigval(submat);

- check if all eigenvalues negative;
  
  concave = (eval(\langle \rangle) \leq .000000001);

- count successful draws;
  
  count = count + concave;

Two things to note about the above code. First, SAS/IML being a matrix language, there is a built-in function to find eigenvalues of (symmetric) matrices. Second, the variable concave is defined by a logical condition, eval(\langle \rangle) \leq .000000001. The expression eval(\langle \rangle) selects the maximum element of the matrix eval. If eval(\langle \rangle) is non-positive (allowing for some round-off error), then the value of concave is 1, else concave = 0. Thus the variable count is only incremented when the inequality restriction holds for the drawn $\theta$.

Conclusion

SAS/IML provides all the necessary tools to implement Monte Carlo integration, with importance and antithetic sampling, in order to estimate posterior means and various probabilities associated with Bayesian inference. The series of papers by Geweke provide a good summary of these procedures.

References


SAS, SAS/ETS, and SAS/IML are registered trademarks of SAS Institute Inc., Cary NC, USA