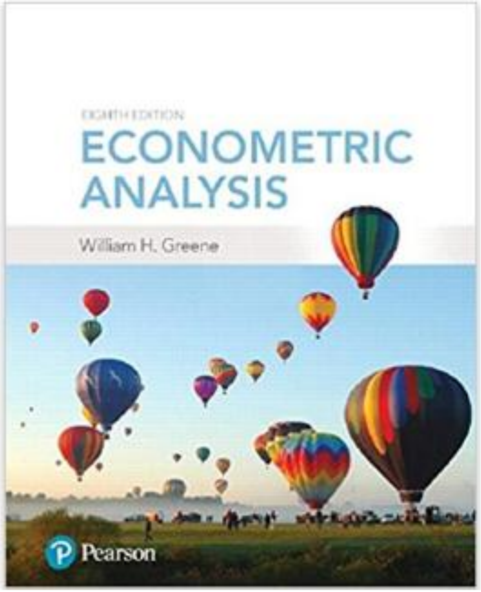
Econometrics I

Professor William Greene
Stern School of Business
Department of Economics



Econometrics I

Part 24 – Bayesian Estimation

Bayesian Estimators

- "Random Parameters" vs. Randomly Distributed Parameters
- Models of Individual Heterogeneity
  - Random Effects: Consumer Brand Choice
  - Fixed Effects: Hospital Costs

Bayesian Estimation

- Specification of conditional likelihood: f(data | parameters)
- Specification of priors: g(parameters)
- Posterior density of parameters:

$$f\,(\text{parameters}|\text{data}) = \frac{f\,(\text{data}\,|\,\text{parameters})\,g\,(\text{parameters})}{f\,(\text{data})}$$

- Posterior mean = E[parameters|data]

The Marginal Density for the Data is Irrelevant

$$f(\beta\,|\,\text{data}) = \frac{f(\text{data}\,|\,\beta)p(\beta)}{f(\text{data})} = \frac{L(\text{data}\,|\,\beta)p(\beta)}{f(\text{data})}$$

Joint density of β and data is $f(\text{data},\beta) = L(\text{data}\,|\,\beta)p(\beta)$

Marginal density of the data is

$$f(\text{data}) = \int_\beta f(\text{data},\beta)d\beta = \int_\beta L(\text{data}\,|\,\beta)p(\beta)d\beta$$

Thus, $f(\beta\,|\,\text{data}) = \dfrac{L(\text{data}\,|\,\beta)p(\beta)}{\int_\beta L(\text{data}\,|\,\beta)p(\beta)d\beta}$

$$\text{Posterior Mean} = \int_\beta p(\beta\,|\,\text{data})d\beta = \frac{\int_\beta \beta\, L(\text{data}\,|\,\beta)p(\beta)d\beta}{\int_\beta L(\text{data}\,|\,\beta)p(\beta)d\beta}$$

Requires specification of the likeihood and the prior.

Computing Bayesian Estimators

- First generation: Do the integration (math)

$$E(\beta \mid \text{data}) = \int_\beta \beta \frac{f(\text{data} \mid \beta) g(\beta)}{f(\text{data})} d\beta$$

- Contemporary - Simulation:
  - (1) Deduce the posterior
  - (2) Draw random samples of draws from the posterior and compute the sample means and variances of the samples. (Relies on the law of large numbers.)

Modeling Issues

- As n →∞, the likelihood dominates and the prior disappears ➔ Bayesian and Classical MLE converge. (Needs the mode of the posterior to converge to the mean.)
- Priors
  - Diffuse → large variances imply little prior information. (NONINFORMATIVE)
  - INFORMATIVE priors – finite variances that appear in the posterior. “Taints” any final results.

A Practical Problem

Sampling from the joint posterior may be impossible.

E.g., linear regression.

$$f(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \mathbf{X}) \propto \frac{[vs^2]^{v+2}}{\Gamma(v+2)} \left[\frac{1}{\sigma^2}\right]^{v+1} e^{-vs^2(1/\sigma^2)} [2\pi]^{-K/2} \mid \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \mid^{-1/2}$$

$$\times \exp(-(1/2)(\boldsymbol{\beta} - \mathbf{b})'[\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta} - \mathbf{b}))$$

What is this???

To do 'simulation based estimation' here, we need joint observations on $(\boldsymbol{\beta}, \sigma^2)$.

A Solution to the Sampling Problem

The joint posterior, $p(\boldsymbol{\beta},\sigma^2|\text{data})$ is intractable. But,

For inference about $\boldsymbol{\beta}$, a sample from the marginal posterior, $p(\boldsymbol{\beta}|\text{data})$ would suffice.

For inference about σ^2, a sample from the marginal posterior of σ^2, $p(\sigma^2|\text{data})$ would suffice.

Can we deduce these? For this problem, we do have conditionals:

$$p(\boldsymbol{\beta}|\sigma^2,\text{data}) = N[\mathbf{b},\sigma^2(\mathbf{X'X})^{-1}]$$

$$p(\sigma^2|\boldsymbol{\beta},\text{data}) = K \times \frac{\Sigma_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{\sigma^2} = \text{a gamma distribution}$$

Can we use this information to sample from $p(\boldsymbol{\beta}|\text{data})$ and $p(\sigma^2|\text{data})$?

The Gibbs Sampler

- Target: Sample from marginals of $f(x_1, x_2)$ = joint distribution
- Joint distribution is unknown or it is not possible to sample from the joint distribution.
- Assumed: $f(x_1|x_2)$ and $f(x_2|x_1)$ both known and samples can be drawn from both.
- Gibbs sampling: Obtain one draw from x_1,x_2 by many cycles between $x_1|x_2$ and $x_2|x_1$.
  - Start $x_{1,0}$ anywhere in the right range.
  - Draw $x_{2,0}$ from $x_2|x_{1,0}$.
  - Return to $x_{1,1}$ from $x_1|x_{2,0}$ and so on.
  - Several thousand cycles produces the draws
  - Discard the first several thousand to avoid initial conditions. (Burn in)
- Average the draws to estimate the marginal means.

Bivariate Normal Sampling

Draw a random sample from bivariate normal $\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right]$

(1) Direct approach: $\begin{pmatrix}v_1\\v_2\end{pmatrix}_r = \Gamma\begin{pmatrix}u_1\\u_2\end{pmatrix}_r$ where $\begin{pmatrix}u_1\\u_2\end{pmatrix}$ are two

independent standard normal draws (easy) and $\Gamma=\begin{pmatrix}1 & 0\\ \theta_1 & \theta_2\end{pmatrix}$

such that $\Gamma\Gamma'=\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}$. $\theta_1 = \rho$, $\theta_2 = \sqrt{1-\rho^2}$.

(2) Gibbs sampler: $v_1 \mid v_2 \sim N\left[\rho v_2, \sqrt{1-\rho^2}\right]$

$v_2 \mid v_1 \sim N\left[\rho v_1, \sqrt{1-\rho^2}\right]$

Gibbs Sampling for the Linear Regression Model

$$p(\boldsymbol{\beta}|\sigma^2,\text{data}) = N[\mathbf{b},\sigma^2(\mathbf{X'X})^{-1}]$$

$$p(\sigma^2|\boldsymbol{\beta},\text{data}) = K \times \frac{\Sigma_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{\sigma^{\mathbf{2}}}$$

$$= \text{ a gamma distribution}$$

Iterate back and forth between these two distributions

Application – the Probit Model

(a) $y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \qquad \varepsilon_i \sim N[0,1]$

(b) $y_i = 1$ if $y_i^* > 0$, 0 otherwise

Consider estimation of $\boldsymbol{\beta}$ and y_i^* (data augmentation)

(1) If y* were observed, this would be a linear regression

(y_i would not be useful since it is just sgn(y_i^*).)

We saw in the linear model before, $p(\boldsymbol{\beta} \mid y_i^*, y_i)$

(2) If (only) $\boldsymbol{\beta}$ were observed, y_i^* would be a draw from

the normal distribution with mean $\mathbf{x}_i'\boldsymbol{\beta}$ and variance 1.

But, y_i gives the sign of y_i^*. $y_i^* \mid \boldsymbol{\beta}, y_i$ is a draw from

the truncated normal (above if y=0, below if y=1)

Gibbs Sampling for the Probit Model

(1) Choose an initial value for $\boldsymbol{\beta}$ (maybe the MLE)

(2) Generate y_i * by sampling N observations from the truncated normal with mean $\mathbf{x}_i'\boldsymbol{\beta}$ and variance 1, truncated above 0 if $y_i = 0$, from below if $y_i = 1$.

(3) Generate $\boldsymbol{\beta}$ by drawing a random normal vector with mean vector $(\mathbf{X'X})^{-1}\mathbf{X'y}*$ and variance matrix $(\mathbf{X'X})^{-1}$

(4) Return to 2 10,000 times, retaining the last 5,000 draws - first 5,000 are the 'burn in.'

(5) Estimate the posterior mean of $\boldsymbol{\beta}$ by averaging the last 5,000 draws.

(This corresponds to a uniform prior over $\boldsymbol{\beta}$.)

Generating Random Draws from f(X)

The inverse probability method of sampling random draws:

If F(x) is the CDF of random variable x, then a random draw on x may be obtained as $F^{-1}(u)$ where u is a draw from the standard uniform (0,1).

Examples:

Exponential: $f(x)=\theta\exp(-\theta x)$; $F(x)=1-\exp(-\theta x)$

$x = -(1/\theta)\log(1-u)$

Normal: $F(x) = \Phi(x)$; $x = \Phi^{-1}(u)$

Truncated Normal: $x=\mu_i + \Phi^{-1}[1-(1-u)*\Phi(\mu_i)]$ for y=1;

$x= \mu_i + \Phi^{-1}[u\Phi(-\mu_i)]$ for y=0.

```
? Generate raw data
Calc     ; Ran(13579) $
Sample   ; 1 - 250 $
Create   ; x1  = rnn(0,1) ; x2 = rnn(0,1) $
Create   ; ys  = .2 + .5*x1 - .5*x2 + rnn(0,1) ; y = ys > 0 $
Namelist; x    = one,x1,x2$
Matrix   ; xxi = <x'x> $
Calc     ; Rep = 200 ; Ri = 1/(Rep-25)$
? Starting values and accumulate mean and variance matrices
Matrix   ; beta=[0/0/0] ; bbar=init(3,1,0);bv=init(3,3,0)$$
Proc     = gibbs $  Markov Chain – Monte Carlo iterations
Do for   ; simulate ; r =1,Rep $
? ------- [ Sample y* | beta ] --------------------------
Create   ; mui = x'beta ; f = rnu(0,1)
         ; if(y=1)  ysg = mui + inp(1-(1-f)*phi( mui));
             (else) ysg = mui + inp(     f *phi(-mui)) $
? ------- [ Sample beta | y*] ---------------------------
Matrix   ; mb = xxi*x'ysg ; beta = rndm(mb,xxi) $
? ------- [ Sum posterior mean and variance. Discard burn in. ]
Matrix   ; if[r > 25] ; bbar=bbar+beta ; bv=bv+beta*beta'$
Enddo    ; simulate $
Endproc $
Execute  ; Proc = Gibbs $
Matrix   ; bbar=ri*bbar ; bv=ri*bv-bbar*bbar' $
Probit   ; lhs = y ; rhs = x $
Matrix   ; Stat(bbar,bv,x) $
```

Example: Probit MLE vs. Gibbs

```
--> Matrix  ; Stat(bbar,bv); Stat(b,varb) $
+---------------------------------------------------+
|Number of observations in current sample =    1000 |
|Number of parameters computed here        =       3 |
|Number of degrees of freedom              =     997 |
+---------------------------------------------------+
+---------+--------------+----------------+--------+---------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] |
+---------+--------------+----------------+--------+---------+
 BBAR_1         .21483281      .05076663     4.232    .0000
 BBAR_2         .40815611      .04779292     8.540    .0000
 BBAR_3        -.49692480      .04508507   -11.022    .0000
+---------+--------------+----------------+--------+---------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] |
+---------+--------------+----------------+--------+---------+
 B_1            .22696546      .04276520     5.307    .0000
 B_2            .40038880      .04671773     8.570    .0000
 B_3           -.50012787      .04705345   -10.629    .0000
```

A Random Effects Approach

- Allenby and Rossi, "Marketing Models of Consumer Heterogeneity"
  - Discrete Choice Model – Brand Choice
  - "Hierarchical Bayes"
  - Multinomial Probit
- Panel Data: Purchases of 4 brands of Ketchup

Structure

Conditional data generation mechanism

$$y_{it,j}{}^* = \boldsymbol{\beta}_i \mathbf{x}_{it,j} + \varepsilon_{it,j}, \ \textit{Utility for consumer } i, \textit{choice } t, \textit{brand } j.$$

$$Y_{it,j} \quad = 1[y_{it,j}{}^* = \textit{maximum utility among the } J \textit{ choices}]$$

$$\mathbf{x}_{it,j} = (\text{constant, log price, "availability," "featured"})$$

$$\varepsilon_{it,j} \sim N[0, \lambda_j], \lambda_1 = 1$$

Implies a J outcome multinomial probit model.

Bayesian Priors

Prior Densities

$$\boldsymbol{\beta}_i \sim N[\bar{\boldsymbol{\beta}}, \mathbf{V}_\beta],$$

$$\text{Implies } \boldsymbol{\beta}_i = \bar{\boldsymbol{\beta}} + \mathbf{w}_i, \mathbf{w}_i \sim N[\mathbf{0}, \mathbf{V}_\beta]$$

$$\lambda_j \sim \text{Inverse Gamma}[v, s_j] \text{ (looks like chi-squared), } v{=}3,\ s_i = 1$$

Priors over model parameters

$$\bar{\boldsymbol{\beta}} \sim N[\bar{\bar{\boldsymbol{\beta}}}, a\mathbf{V}_\beta],\ \bar{\bar{\boldsymbol{\beta}}} = \mathbf{0}$$

$$\mathbf{V}_\beta^{-1} \sim \text{Wishart}[v_0, \mathbf{V}_0], v_0 = 8, \mathbf{V}_0 = 8\mathbf{I}$$

Bayesian Estimator

- Joint Posterior= $E[\beta_1,...,\beta_N,\bar{\beta},V_\beta,\lambda_1,...,\lambda_J \mid data]$
- Integral does not exist in closed form.
- Estimate by random samples from the joint posterior.
- Full joint posterior is not known, so not possible to sample from the joint posterior.

Gibbs Cycles for the MNP Model

- Samples from the marginal posteriors

Marginal posterior for the individual parameters

(Known and can be sampled)

$$\boldsymbol{\beta}_i \mid \bar{\boldsymbol{\beta}}, \mathbf{V}_\beta, \boldsymbol{\lambda}, data$$

Marginal posterior for the common parameters

(Each known and each can be sampled)

$$\bar{\boldsymbol{\beta}} \mid \mathbf{V}_\beta, \boldsymbol{\lambda}, data$$

$$\mathbf{V}_\beta \mid \bar{\boldsymbol{\beta}}, \boldsymbol{\lambda}, data$$

$$\boldsymbol{\lambda} \mid \bar{\boldsymbol{\beta}}, \mathbf{V}_\beta,, data$$

Results

- Individual parameter vectors and disturbance variances
- Individual estimates of choice probabilities
- The same as the "random parameters model" with slightly different weights.
- Allenby and Rossi call the classical method an "approximate Bayesian" approach.
  - (Greene calls the Bayesian estimator an "approximate random parameters model")
  - Who's right?
    - Bayesian layers on implausible uninformative priors and calls the maximum likelihood results "exact" Bayesian estimators
    - Classical is strongly parametric and a slave to the distributional assumptions.
    - Bayesian is even more strongly parametric than classical.
    - Neither is right – Both are right.

Comparison of Maximum Simulated Likelihood and Hierarchical Bayes

- Ken Train: "A Comparison of Hierarchical Bayes and Maximum Simulated Likelihood for Mixed Logit"
- Mixed Logit

$$U(i,t,j) = \boldsymbol{\beta}_i{}'\mathbf{x}(i,t,j) + \varepsilon(i,t,j),$$

$$i = 1,\ldots,N \text{ individuals,}$$

$$t = 1,\ldots,T_i \text{ choice situations}$$

$$j = 1,\ldots,J \text{ alternatives (may also vary)}$$

Stochastic Structure – Conditional Likelihood

$$\text{Prob}(i,j,t) = \frac{\exp(\boldsymbol{\beta}_i'\mathbf{x}_{i,j,t})}{\sum_{j=1}^{\mathbf{J}}\exp(\boldsymbol{\beta}_i'\mathbf{x}_{i,j,t})}$$

$$Likelihood = \prod_{t=1}^{T}\frac{\exp(\boldsymbol{\beta}_i'\mathbf{x}_{i,j*,t})}{\sum_{j=1}^{\mathbf{J}}\exp(\boldsymbol{\beta}_i'\mathbf{x}_{i,j*,t})}$$

$j* =$ indicator for the specific choice made by i at time t.

Note individual specific parameter vector, $\boldsymbol{\beta}_i$

Classical Approach

$$\boldsymbol{\beta}_i \sim N[\mathbf{b},\boldsymbol{\Omega}];\ \text{write } \boldsymbol{\Omega}=\boldsymbol{\Gamma}\boldsymbol{\Gamma}'$$

$$\boldsymbol{\beta}_i = \mathbf{b}+\mathbf{w}_i$$

$$= \mathbf{b}+\boldsymbol{\Gamma}\mathbf{v}_i \ \text{ where } \boldsymbol{\Gamma} = diag(\gamma_j^{1/2})\ (uncorrelated)$$

$$Log-likelihood = \sum_{i=1}^{N}\log\int_{\mathbf{w}}\prod_{t=1}^{T}\frac{\exp[(\mathbf{b}+\mathbf{w}_i)'\mathbf{x}_{i,j*,t}]}{\sum_{j=1}^{\mathbf{J}}\exp[(\mathbf{b}+\mathbf{w}_i)_i'\mathbf{x}_{i,j,t}]}d\mathbf{w}_i$$

Maximize over $\mathbf{b},\boldsymbol{\Gamma}$ using maximum simulated likelihood

(random parameters model)

Bayesian Approach – Gibbs Sampling and Metropolis-Hastings

$$Posterior = \prod_{i=1}^{N} L(data \mid \beta_i, \mathbf{\Omega}) \times priors$$

$$Prior = N(\beta_1,...,\beta_N \mid \mathbf{b}, \mathbf{\Omega})\ (normal)$$

$$\times IG(\gamma_1,...,\gamma_N \mid parameters)\ (Inverse\ gamma)$$

$$\times g(\mathbf{b} \mid assumed\ \ parameters)\ (Normal\ with\ large\ variance)$$

Gibbs Sampling from Posteriors: **b**

$$p(\mathbf{b} \mid \beta_1,...,\beta_N,\Omega) = Normal[\bar{\beta},(1/N)\Omega]$$

$$\bar{\beta} = (1/N)\sum_{i=1}^{N}\beta_i$$

Easy to sample from Normal with known mean and variance by transforming a set of draws from standard normal.

Gibbs Sampling from Posteriors: Ω

$$p(\gamma_k \mid \mathbf{b}, \beta_1, ..., \beta_N) \sim \textit{Inverse Gamma}[1+N, 1+N\bar{V}_k]$$

$$\bar{V}_k = (1/N)\sum_{i=1}^{N}(\beta_{k,i} - b_k)^2 \text{ for each k=1,...,K}$$

Draw from inverse gamma for each k:

Draw 1+N draws from N[0,1] = $h_{r,k}$,

then the draw is $\dfrac{(1+N\bar{V}_k)}{\sum_{r=1}^{R} h_{r,k}^2}$

Gibbs Sampling from Posteriors: β_i

$$p(\beta_i \mid \mathbf{b},\mathbf{\Omega}) = M \times L(data \mid \beta_i) \times g(\beta_i \mid \mathbf{b},\mathbf{\Omega})$$

M=a constant, L=likelihood, g=prior

(This is the definition of the posterior.)

Not clear how to sample.

Use Metropolis Hastings algorithm.

Metropolis – Hastings Method

$Define:$

$\beta_{i,0}$ = an 'old' draw (vector)

$\beta_{i,1}$ = the 'new' draw (vector)

$\text{d}_\text{r} = \sigma\, \boldsymbol{\Gamma}\, \mathbf{v}_\mathbf{r}$,

σ=a constant (see below)

$\boldsymbol{\Gamma}$ = the diagonal matrix of standard deviations

$\mathbf{v}_\mathbf{r}$ =a vector of K draws from standard normal

Metropolis Hastings: A Draw of β_i

$$Trial\ value: \tilde{\beta}_{i,1} = \beta_{i,0} + d_r$$

$$R = \frac{Posterior(\tilde{\beta}_{i,1})}{Posterior(\beta_{i,0})} (Ms\ cancel)$$

U = a random draw from U(0,1)

If U < R, use $\tilde{\beta}_{i,1}$, *else keep* $\beta_{i,0}$

During Gibbs iterations, draw $\beta_{i,1}$

σ controls acceptance rate. Try for .4.

Application: Energy Suppliers

- N=361 individuals, 2 to 12 hypothetical suppliers
- X= (1) fixed rates,
  (2) contract length,
  (3) local (0,1),
  (4) well known company (0,1),
  (5) offer TOD rates (0,1),
  (6) offer seasonal rates (0,1).

Estimates: Mean of Individual β_i

	MSL Estimate	Bayes Posterior Mean
Price	-1.04 (0.396)	-1.04 (0.0374)
Contract	-0.208 (0.0240)	-0.194 (0.0224)
Local	2.40 (0.127)	2.41 (0.140)
Well Known	1.74 (0.0927)	1.71 (0.100)
TOD	-9.94 (0.337)	-10.0 (0.315)
Seasonal	-10.2 (0.333)	-10.2 (0.310)

Reconciliation: A Theorem (Bernstein-Von Mises)

- The posterior distribution converges to normal with covariance matrix equal to 1/n times the information matrix (same as classical MLE). (The distribution that is converging is the posterior, not the sampling distribution of the estimator of the posterior mean.)
- The posterior mean (empirical) converges to the mode of the likelihood function. Same as the MLE. A proper prior disappears asymptotically.
- Asymptotic sampling distribution of the posterior mean is the same as that of the MLE.