Day 3B
Simulation: Bayesian Methods

A. Colin Cameron
Univ. of Calif. - Davis

... for

Center of Labor Economics
Norwegian School of Economics
Advanced Microeconometrics

Aug 28 - Sep 1, 2017

1. Introduction

- Bayesian methods provide an alternative method of computation and statistical inference to ML estimation.
  - Some researchers use a fully Bayesian approach to inference.
  - Other researchers use Bayesian computation methods (with a diffuse or uninformative prior) as a tool to obtain the MLE and then interpret results as they would classical ML results.
- The slides give generally theory and probit example done three ways
  - estimation using command `bayesmh`
  - manual implementation of Metropolis-Hastings algorithm
  - harder: manual implementation of Gibbs sampler with data augmentation.
- We focus on topics 1-5 below.

Outline

2. Bayesian Probit Example

- Generated data from probit model with
- $\Pr[y=1|x]=\Phi(0.5+1\times x)$, $x\sim N(0,1)$, $N=100$.

```
. * Generate data N = 100  Pr[y=1|x] = PHI(0 + 0.5*x)
. clear

. set obs 100
number of observations (_N) was 0, now 100

. set seed 1234567

. gen x = rnormal(0,1)

. gen ystar = 0.5 + 1*x + rnormal(0,1)

. gen y = (ystar > 0)

. gen cons = 1

. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
x	100	-.1477064	1.003931	-2.583632	2.350792
ystar	100	.2901163	1.46373	-3.372719	3.316435
y	100	.59	.4943111	0	1
cons	100	1	0	1	1

Maximum Likelihood Estimates

- MLE is $(\widehat{\beta}_1, \widehat{\beta}_2) = (0.481, 1.138)$ compared to d.g.p. values of $(0.5, 1.0)$.

```
. * Estimate model by MLE
. probit y x

Iteration 0:   log likelihood = -67.685855
Iteration 1:   log likelihood = -46.554132
Iteration 2:   log likelihood = -46.350487
Iteration 3:   log likelihood = -46.350193
Iteration 4:   log likelihood = -46.350193

Probit regression                               Number of obs     =        100
                                                LR chi2(1)        =      42.67
                                                Prob > chi2       =     0.0000
Log likelihood = -46.350193                     Pseudo R2         =     0.3152
```

y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
x	1.137895	.2236915	5.09	0.000	.6994677	1.576322
_cons	.4810185	.1591173	3.02	0.003	.1691543	.7928827

Bayesian Estimates

```
. * Following the same as version 15 command bayes, rseed(10101): probit y x
. bayesmh y x, likelihood(probit) prior({y: }, normal(0,10000)) rseed(10101)

Burn-in ...
Simulation ...

Model summary
------------------------------------------------------------------------------
Likelihood:
  y ~ probit(xb_y)

Prior:
  {y:x _cons} ~ normal(0,10000)                                           (1)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_y.

Bayesian probit regression                       MCMC iterations  =     12,500
Random-walk Metropolis-Hastings sampling         Burn-in          =      2,500
                                                 MCMC sample size =     10,000
                                                 Number of obs    =        100
                                                 Acceptance rate  =      .2081
                                                 Efficiency:  min =     .09261
                                                              avg =       .104
Log marginal likelihood = -58.903331                          max =      .1154

------------------------------------------------------------------------------
             |                                                Equal-tailed
           y |      Mean   Std. Dev.     MCSE     Median  [95% Cred. Interval]
-------------+----------------------------------------------------------------
           x |   1.17248   .2315757   .006817   1.155512   .7693411   1.644085
       _cons |  .4912772   .1649861   .005421   .4913285   .1694713   .8135924
------------------------------------------------------------------------------
```

First Output

- Bayesian analysis treats $\boldsymbol{\beta}$ as a parameter and combines
  - knowledge on $\boldsymbol{\beta}$ gained from the data - the likelihood function
  - prior knowledge on the distribution of $\boldsymbol{\beta}$ - the prior.
- Here the likelihood is that for the probit model.
- And the prior is $\beta_1 \sim N(0, 100^2)$ and $\beta_2 \sim N(0, 100^2)$.

```
Model summary
------------------------------------------------------------------------------
Likelihood:
  y ~ probit(xb_y)

Prior:
  {y:x _cons} ~ normal(0,10000)                                            (1)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_y.
```

Second Output

- This provides the Markov chain Monte Carlo details.

```
Bayesian probit regression                            MCMC iterations  =     12,500
Random-walk Metropolis-Hastings sampling              Burn-in          =      2,500
                                                      MCMC sample size =     10,000
                                                      Number of obs    =        100
                                                      Acceptance rate  =      .2081
                                                      Efficiency:  min =     .09261
                                                                   avg =       .104
Log marginal likelihood = -58.903331                               max =      .1154
```

- There were 12,500 MCMC draws
  - the first 2,500 were discarded to let the chain hopefully converge
  - and the next 10,000 were retained.
- Not all draws led to an updated value of β
  - in fact only 2,081 did
  - the 10,000 correlated draws were equivalent to 926 independent draws.

Third Output

- This provides the posterior distribution of β_1 and β_2

y	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred.	Interval]
x	1.17248	.2315757	.006817	1.155512	.7693411	1.644085
_cons	.4912772	.1649861	.005421	.4913285	.1694713	.8135924

- The posterior distribution of β_2 has mean 1.172 (average of the 10,000 draws), standard deviation 0.232, and the 2.5 to 97.5 percentiles were (0.769, 1.644).
- The results are similar to the MLE as the prior of $N(0, 100^2)$ had very large standard deviation so has little effect
  - the likelihood dominates and the MLE uses this.

y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
x	1.137895	.2236915	5.09	0.000	.6994677	1.576322
_cons	.4810185	.1591173	3.02	0.003	.1691543	.7928827

3. Bayesian Methods: Basic Idea

- Bayesian methods begin with
  - Likelihood: $L(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})$
  - Prior on $\boldsymbol{\theta}$: $\pi(\boldsymbol{\theta})$
- This yields the posterior distribution for $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) = \frac{L(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X}) \times \pi(\boldsymbol{\theta})}{f(\mathbf{y}|\mathbf{X})}$$

  - where $f(\mathbf{y}|\mathbf{X}) = \int L(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X}) \times \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is called the marginal likelihood.
- This uses the result that

$$\begin{aligned} \Pr[A|B] &= \Pr[A \cap B] / \Pr[B] \\ &= \{\Pr[B|A] \times \Pr[A]\} / \Pr[B] \\ p(\boldsymbol{\theta}|\mathbf{y}) &= \{L(\mathbf{y}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})\} / f(\mathbf{y})). \end{aligned}$$

- Bayesian analysis then bases inference on the posterior distribution.
- Estimate $\boldsymbol{\theta}$ by the mean or the mode of the posterior distribution.
- A 95% credible interval (or "Bayesian confidence interval") for $\boldsymbol{\theta}$ is from the 2.5 to 97.5 percentiles of the posterior distribution
- No need for asymptotic theory!

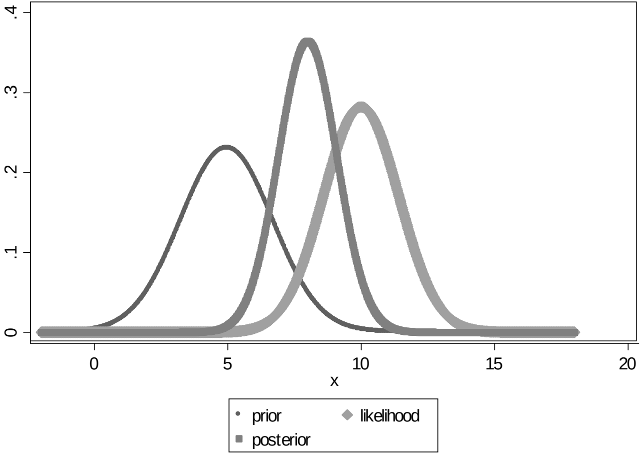
Normal-normal example

- Suppose $y|\theta \sim \mathcal{N}[\theta, 100]$ (σ^2 is known from other studies)
  And we have independent sample of size $N = 50$ with $\bar{y} = 10$.
- Classical analysis uses $\bar{y}|\theta \sim \mathcal{N}[\theta, 100/N] \sim \mathcal{N}[\theta, 2]$
  Reinterpret as likelihood $\theta|\mathbf{y} \sim \mathcal{N}[\theta, 2]$.
  Then MLE $\widehat{\theta} = \bar{y} = 10$.
- Bayesian analysis introduces prior, say $\theta \sim \mathcal{N}[5, 3]$.
  We combine likelihood and prior to get posterior.
- We expect
  - posterior mean: between prior mean 5 and sample mean 10
  - posterior variance: less than 2 as prior info reduces noise
  - posterior distribution: ? Generally intractable.
- But here can show posterior for θ is $\mathcal{N}[8, 1.2]$

Normal-normal example (continued)

- Classical inference: $\widehat{\theta} = \bar{y} = 10 \sim \mathcal{N}[10, 2]$
  - A 95% confidence interval for θ is $10 \pm 1.96 \times \sqrt{2} = (7.23, 12.77)$
  - i.e. 95% of the time this conf. interval will include the unknown constant θ.
- Bayesian inference: Posterior $\widehat{\theta} \sim \mathcal{N}[8, 1.2]$
  - A 95% posterior interval for θ is $8 \pm 1.96 \times \sqrt{1.2} = (5.85, 10.15)$
  - i.e. with probability 0.95 the random θ lies in this interval
- Not that with a "diffuse" prior Bayesian gives similar numerical result to classical
  - if prior is $\theta \sim \mathcal{N}[5, 100]$ then posterior is $\widehat{\theta} \sim \mathcal{N}[9.90, 0.51]$

- Prior $\mathcal{N}[5, 3]$ and likelihood $\mathcal{N}[10, 2]$ and yields posterior $\mathcal{N}[8, 1.2]$ for θ



Rare Tractable results

- The tractable result for normal-normal (known variance) carries over to exponential family using a conjugate prior

Likelihood	**Prior**	**Posterior**
Normal (mean μ)	Normal	Normal
Normal (precision $\frac{1}{\sigma^2}$)	Gamma	Gamma
Binomial (p)	Beta	Beta
Poisson (μ)	Gamma	Gamma

  - using conjugate prior is like augmenting data with a sample from the same distribution
  - for Normal with precision matrix Σ^{-1} gamma generalizes to Wishart.
- But in general tractable results not available
  - so use numerical methods, notably MCMC.
  - using tractable results in subcomponents of MCMC can speed up computation.

4. Markov chain Monte Carlo (MCMC)

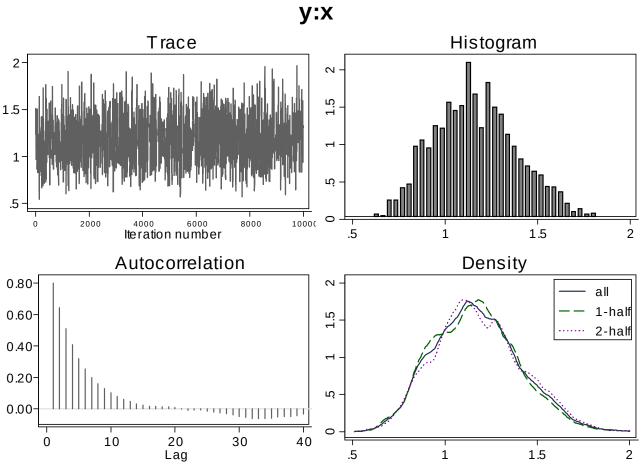
- The challenge is to compute the posterior $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})$
  - analytical results are only available in special cases.
- Instead use Markov chain Monte Carlo methods:
  - Make sequential random draws $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, ...$
  - where $\boldsymbol{\theta}^{(s)}$ depends in part on $\boldsymbol{\theta}^{(s-1)}$
    - ★ but not on $\boldsymbol{\theta}^{(s-2)}$ once we condition on $\boldsymbol{\theta}^{(s-1)}$ (Markov chain)
  - in such a way that after an initial burn-in (discard these draws) $\boldsymbol{\theta}^{(s)}$ are (correlated) draws from the posterior $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})$
    - ★ the Markov chain converges to a stationary marginal distribution which is the posterior.
- MCMC methods include
  - Metropolis algorithm
  - Metropolis-Hastings algorithm
  - Gibbs sampler

Checking Convergence of the Chain

- Once the chain has converged the draws are draws from the posterior.
- There is no way to be 100% sure that the chain has converged!
- First thing is to throw out initial draws e.g. first 2,500.
- But it has not converged if it fails some simple tests
  - if sequential draws are highly correlated
  - if sequential draws are very weakly correlated
  - if the second half of the draws have quite different distribution from the first draws
  - for MH (but not Gibbs sampler) if few draws are accepted or if almost all draws are accepted
  - if posterior distributions are multimodal (unless there is reason to expect this).

Diagnostics for Bayesian Probit Example

- `bayesgraph diagnostics {y:x}` gives diagnostics for β_2



Diagnostics (continued)

- These diagnostics suggest that the chain has converged.
- The trace shows the 10,000 draws of β_2 and shows that the value changes.
- The histogram is unimodal, fairly symmetric, and appears normally distributed
  - this is not always be the case, especially in small samples.
- The sequential draws of β_2 are correlated (like AR(1) with $\rho \simeq 0.8$).
- The first 5,000 draws have similar density to the second 5,000 draws.

Metropolis-Hastings Algorithm: Metropolis Algorithm

- We want to draw from posterior $p(\cdot)$ but cannot directly do so.
- Metropolis draws from a candidate distribution $g(\boldsymbol{\theta}^{(s)}|\boldsymbol{\theta}^{(s-1)})$
  - these draws are sometimes accepted and some times not
  - like accept-reject method but do not require $p(\cdot) \leq kg(\cdot)$
- Metropolis algorithm at the s^{th} round
  - draw candidate $\boldsymbol{\theta}^*$ from candidate distribution $g(\cdot)$
  - the candidate distribution $g(\boldsymbol{\theta}^{(s)}|\boldsymbol{\theta}^{(s-1)})$ needs to be symmetric
    - so $g(\boldsymbol{\theta}^a|\boldsymbol{\theta}^b) = g(\boldsymbol{\theta}^b|\boldsymbol{\theta}^a)$
  - set $\boldsymbol{\theta}^{(s)} = \boldsymbol{\theta}^*$ if $u < \frac{p(\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(s-1)})}$ where u is draw from uniform$[0, 1]$
    - note: normalizing constants in $p(\cdot)$ cancel out
    - equivalently set $\boldsymbol{\theta}^{(s)} = \boldsymbol{\theta}^*$ if $\ln u < \ln p(\boldsymbol{\theta}^*) - \ln p(\boldsymbol{\theta}^{(s-1)})$
  - otherwise set $\boldsymbol{\theta}^{(s)} = \boldsymbol{\theta}^{(s-1)}$
- Random walk Metropolis uses $\boldsymbol{\theta}^{(s)} \sim \mathcal{N}[\boldsymbol{\theta}^{(s-1)}, \mathbf{V}]$ for fixed $\mathbf{V}$
  - ideally $\mathbf{V}$ such that 25-50% of candidate draws are accepted.

Metropolis-Hastings Algorithm

- Metropolis-Hastings is a generalization
  - the candidate distribution $g(\boldsymbol{\theta}^{(s)}|\boldsymbol{\theta}^{(s-1)})$ need not be symmetric
  - the acceptance rule is then $u < \frac{p(\boldsymbol{\theta}^{*})\times g(\boldsymbol{\theta}^{*}|\boldsymbol{\theta}^{(s-1)})}{p(\boldsymbol{\theta}^{(s-1)})\times g(\boldsymbol{\theta}^{(s-1)}|\boldsymbol{\theta}^{*})}$
  - Metropolis algorithm itself is often called Metropolis-Hastings.
- Independence chain MH uses $g(\boldsymbol{\theta}^{(s)})$ not depending on $\boldsymbol{\theta}^{(s-1)}$ where $g(\cdot)$ is a good approximation to $p(\cdot)$
  - e.g. Do ML for $p(\boldsymbol{\theta})$ and then $g(\boldsymbol{\theta})$ is multivariate T with mean $\widehat{\boldsymbol{\theta}}$, variance $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}}]$.
  - multivariate rather than normal as has fatter tails.
- M and MH called Markov chain Monte Carlo
  - because $\boldsymbol{\theta}^{(s)}$ given $\boldsymbol{\theta}^{(s-1)}$ is a first-order Markov chain
  - Markov chain theory proves convergence to draws from $p(\cdot)$ as $s \rightarrow \infty$
  - poor choice of candidate distribution leads to chain stuck in place.

Probit with random walk Metropolis

- Consider probit model $\Pr[y_i = 1|\mathbf{x}_i, \boldsymbol{\beta}] = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$.
- The likelihood is

$$L(\mathbf{y}|\boldsymbol{\beta}, \mathbf{X}) = \prod_{i=1}^{N} \Phi(\mathbf{x}_i'\boldsymbol{\beta})^{y_i}(1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta}))^{1-y_i}$$

- Use an uninformative prior (all values of $\boldsymbol{\beta}$ equally likely)

$$\pi(\boldsymbol{\beta}) \propto 1$$

  - even though prior is improper the posterior will be proper
- The posterior is

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) &\propto L(\mathbf{y}|\boldsymbol{\beta}, \mathbf{X}) \times \pi(\boldsymbol{\beta}) \\ &\propto \prod_{i=1}^{N} \Phi(\mathbf{x}_i'\boldsymbol{\beta})^{y_i}(1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta}))^{1-y_i} \times 1 \\ &\propto \prod_{i=1}^{N} \Phi(\mathbf{x}_i'\boldsymbol{\beta})^{y_i}(1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta}))^{1-y_i} \end{aligned}$$

  - Note: we know $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$ only up to a scale factor
- We use Metropolis algorithm to make draws from this posterior.

Random walk Metropolis draws

- The random walk MH uses a draw from $\mathcal{N}[\boldsymbol{\beta}^{(s-1)}, c\mathbf{I}]$ where c is set.
  - So we draw $\boldsymbol{\beta}^{*} = \boldsymbol{\beta}^{(s-1)} + \mathbf{v}$ where $\mathbf{v}$ is draw from $\mathcal{N}[\mathbf{0}, c\mathbf{I}]$
- For $u \sim \text{uniform}[0,1]$ draw and acceptance probability $paccept = p(\boldsymbol{\beta}^{*})/p(\boldsymbol{\beta}^{(s-1)})$
  - set $\boldsymbol{\beta}^{(s)} = \boldsymbol{\beta}^{*}$ if $u < paccept$
  - set $\boldsymbol{\beta}^{(s)} = \boldsymbol{\beta}^{(s-1)}$ if $u > paccept$
- Taking logs, equivalent to $\boldsymbol{\beta}^{(s)} = \boldsymbol{\beta}^{*}$ if $\ln u < \ln(paccept)$ where
  - $\ln(paccept) = [\sum_i y_i \ln \Phi(\mathbf{x}_i'\boldsymbol{\beta}^{*}) + (1-y_i)\ln(1-\Phi(\mathbf{x}_i'\boldsymbol{\beta}^{*}))]$ $-[\sum_i y_i \ln \Phi(\mathbf{x}_i'\boldsymbol{\beta}^{(s-1)}) + (1-y_i)\ln(1-\Phi(\mathbf{x}_i'\boldsymbol{\beta}^{(s-1)}))]$

Numerical example

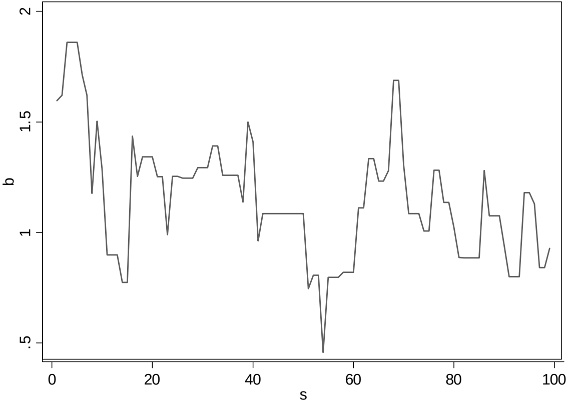
- Do Bayesian
  - uninformative prior so $\pi(\boldsymbol{\beta}) = 1$
    - ★ an improper prior here is okay.
  - random walk MH with $\boldsymbol{\beta}^* = \boldsymbol{\beta}^{(s-1)} + \mathbf{v}$ where $\mathbf{v}$ is draw from $\mathcal{N}[\mathbf{0}, 0.25\mathbf{I}]$
    - ★ $c = 0.25$ chosen after some trial and error
  - First $10,000$ MH draws were discarded (burn-in)
  - Next $10,000$ draws were kept.

Mata code

```
for (irep=1; irep<=20000; irep++) {
   bcandidate = bdraw + 0.25*rnormal(k,1,0,1)   // bdraw is previous value of b
   phixb = normal(X*bcandidate)
   lpostcandidate = e'( y:*ln(phixb) + (e-y):*ln(e-phixb)   // e = J(n,1,1)
   laccprob = lpostcandidate - lpostdraw   // lpostdraw post. prob. from last round
   if ( ln(runiform(1,1)) < laccprob ) {
      lpostdraw = lpostcandidate
      bdraw = bcandidate
   }
   // Store the draws after burn-in of b
   if (irep>10000) {
   j = irep-10000
   b_all[.,j] = bdraw // These are the posterior draws
   }
}
```

Correlated draws

- The first 100 draws (after burn-in) from the posterior density of β_2
- Flat sections are where the candidate draw was not accepted.

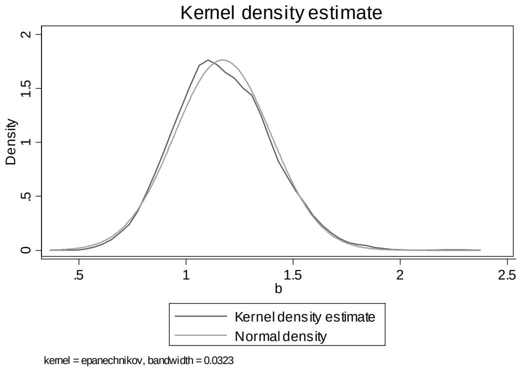


- Correlations of the 10,000 draws of β_2 die out reasonably quickly
  - This varies a lot with choice of c in $\boldsymbol{\beta}^* = \boldsymbol{\beta}^{(s-1)} + \mathcal{N}[\mathbf{0}, c\mathbf{I}]$
- The acceptance rate for 10,000 draws was 0.4286 - very high.

```
. * Give the corrleations and the acceptance rate in the random walk chain MH
. corrgram b, lags(10)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.8330   0.8331   6940.9  0.0000
2        0.6956   0.0056    11781  0.0000
3        0.5848   0.0140    15203  0.0000
4        0.4889  -0.0089    17595  0.0000
5        0.4089   0.0010    19268  0.0000
6        0.3369  -0.0172    20404  0.0000
7        0.2798   0.0075    21188  0.0000
8        0.2287  -0.0132    21712  0.0000
9        0.1896   0.0104    22071  0.0000
10       0.1558  -0.0054    22314  0.0000

. quietly summarize accept

. display "MH acceptance rate = " r(mean) "
MH acceptance rate = .4286
```

Posterior density

- Kernel density estimate of the 10,000 draws of β_2
  - centered around approx. 0.4 with standard deviation of 0.1-0.2.



- More precisely
  - Posterior mean of β_2 is 1.171 and standard deviation is 0.226
  - A 95% percent Bayesian credible interval for β_2 is $(0.754, 1.633)$.

```
. summarize b

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
           b |     10,000    1.171479    .2263332    .396735   2.341014

. centile b, centile(2.5, 97.5)

                                                       -- Binom. Interp. --
    Variable |       Obs  Percentile    Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
           b |    10,000        2.5    .7540872        .7451204    .7699984
             |                 97.5    1.633189        1.622456    1.652172
```

- Whereas probit MLE was 1.137 with standard error 0.226
  - and 95% confidence interval $(0.699, 1.576)$.

6. Gibbs sampler and Data Augmentation: Gibbs Sampler

- Gibbs sampler
  - case where posterior is partitioned e.g. $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$
  - and make alternating draws from $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)$ and $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$
  - gives draws from $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ even though

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) \times p(\boldsymbol{\theta}_2) \neq p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) \times p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1).$$

- Gibbs is special case of MH
  - usually quicker than usual MH
  - if need MH to draw from $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)$ and/or $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$ called MH within Gibbs.
  - extends to e.g. $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$ make sequential draws from $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$, $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \boldsymbol{\theta}_3)$ and $p(\boldsymbol{\theta}_3|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$
  - requires knowledge of all of the full conditionals.
- M, MH and Gibbs yield correlated draws of $\boldsymbol{\theta}^{(s)}$
  - but still give correct estimate of marginal posterior distribution of θ (once discard burn-in draws)
  - e.g. estimate posterior mean by $\frac{1}{S}\sum_{s=1}^{S} \boldsymbol{\theta}^{(s)}$.

Data Augmentation: Summary

- Latent variable models (probit, Tobit, ...) observe $y_1, ..., y_N$ based on latent variables $y_1^*, ..., y_N^*$.
- Bayesian data augmentation introduces $y_1^*, ..., y_N^*$ as additional parameters
  - then posterior is $p(y_1^*, ..., y_N^*, \boldsymbol{\theta})$.
- Use Gibbs sampler
  - alternating draws between $p(\boldsymbol{\theta}|y_1^*, ..., y_N^*)$ and $p(y_1^*, ..., y_N^*|\boldsymbol{\theta})$.
- Draws of $\boldsymbol{\theta}|y_1^*, ..., y_N^*$ can use known results for linear regression
  - since regular regression once $y_1^*, ..., y_N^*$ are known
- Draws from $p(y_1^*, ..., y_N^*|\boldsymbol{\theta})$ are called data augmentation
  - since we augment observed $y_1, ..., y_N$ with unobserved $y_1^*, ..., y_N^*$.

Probit example: algorithm

- Likelihood: Probit model with latent variable formulation
  - $y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}[0,1]$.
  - $y_i = \begin{cases} 1 & y_i^* > 0 \\ 0 & y_i^* \leq 0 \end{cases}$
- Prior: uniform prior (all values equally likely)
  - $\pi(\boldsymbol{\beta}) = 1$
- $\boldsymbol{\beta}|\mathbf{y}^*$: Tractable result for $\mathbf{y}^*|\boldsymbol{\beta}, \mathbf{X} \sim \mathcal{N}[\mathbf{X}\boldsymbol{\beta}, \mathbf{I}]$ and uniform prior on $\boldsymbol{\beta}$
  - $p(\boldsymbol{\beta}|\mathbf{y}^*, \mathbf{X})$ is $\mathcal{N}[\widehat{\boldsymbol{\beta}}, (\mathbf{X}'\mathbf{X})^{-1}]$ where $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^*$.
- $\mathbf{y}^*|\boldsymbol{\beta}$: Data augmentation draws $y_1^*, ..., y_N^*$ as parameters.
  - $p(y_1^*, ..., y_N^*|\boldsymbol{\beta}, \mathbf{y}, \mathbf{X})$ is truncated normal so
    - ★ If $y_i = 1$ draw from $\mathcal{N}[\mathbf{x}_i'\boldsymbol{\beta}, 1]$ left truncated at 0
    - ★ If $y_i = 0$ draw from $\mathcal{N}[\mathbf{x}_i'\boldsymbol{\beta}, 1]$ right truncated at 0
- So draw $\boldsymbol{\beta}^{(s)}$ from $p(\boldsymbol{\beta}|y_1^{*(s-1)}, ..., y_N^{*(s-1)}, \mathbf{y}, \mathbf{X})$
  then draw $y_1^{*(s)}, ..., y_N^{*(s)}$ from $p(y_1^*, ..., y_N^*|\boldsymbol{\beta}^{(s)}, \mathbf{y}, \mathbf{X})$.

Numerical example

- Consider the same probit example as used for random walk MH
- Code is given in file **bayes2017.do**
- All draws are accepted for the Gibbs sampler.
- Correlations of the 10,000 draws of β_2 die out quite quickly

```
. corrgram b, lags(10)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.7980   0.7984   6369.5  0.0000
2        0.6387   0.0055    10450  0.0000
3        0.5074  -0.0105    13026  0.0000
4        0.4016  -0.0042    14640  0.0000
5        0.3147  -0.0088    15631  0.0000
6        0.2475   0.0032    16244  0.0000
7        0.1912  -0.0085    16610  0.0000
8        0.1470  -0.0022    16827  0.0000
9        0.1161   0.0092    16961  0.0000
10       0.0905  -0.0030    17043  0.0000
```

Posterior distribution

- Similar to other methods.

```
. summarize b
```

Variable	Obs	Mean	Std. Dev.	Min	Max
b	10,000	1.163722	.2227863	.43323	2.311867

```
. centile b, centile(2.5, 97.5)
```

Variable	Obs	Percentile	Centile	— Binom. Interp. — [95% Conf.	Interval]
b	10,000	2.5	.7625044	.7494316	.7674681
		97.5	1.623944	1.608732	1.639934

More complicated example: Multinomial probit

- Likelihood: Multinomial probit model (MLE has high-dimensional integral)
  - $U_{ij}^* = \mathbf{x}_{ij}'\boldsymbol{\beta} + \varepsilon_{ij}$, $\boldsymbol{\varepsilon}_i \sim \mathcal{N}[\mathbf{0}, \Sigma_\varepsilon]$
  - $y_{ij} = 1$ if $U_{ij}^* > U_{ik}^*$ all $k \neq j$
- Prior for $\boldsymbol{\beta}$ and Σ_ε^{-1} may be normal and Wishart
- Data augmentation
  - Latent utilities $\mathbf{U}_i = (U_{i1}, ..., U_{im})$ are introduced as auxiliary variables
  - Let $\mathbf{U} = (\mathbf{U}_1, ..., \mathbf{U}_N)$ and $\mathbf{y} = (y_1, ..., y_N)$
- Gibbs sampler for joint posterior $p(\boldsymbol{\beta}, \mathbf{U}, \Sigma_\varepsilon | \mathbf{y}, \mathbf{X})$ cycles between
  - 1. Conditional posterior for $\boldsymbol{\beta}|\mathbf{U}, \Sigma_\varepsilon, \mathbf{y}, \mathbf{X}$
  - 2. Conditional posterior for $\Sigma_\varepsilon|\boldsymbol{\beta}, \mathbf{U}, \mathbf{y}, \mathbf{X}$, and
  - 3. Conditional posterior for $\mathbf{U}_i|\boldsymbol{\beta}, \Sigma_\varepsilon, \mathbf{y}, \mathbf{X}$.
- Albert and Chib (1993) provide a quite general treatment.
- McCulloch and Rossi (1994) provide a substantive MNP application.

7. Further discussion: Specification of prior

- As $N \rightarrow \infty$ data dominates the prior $\pi(\boldsymbol{\theta})$ and then posterior $\boldsymbol{\theta}|\mathbf{y} \overset{a}{\sim} \mathcal{N}[\widehat{\boldsymbol{\theta}}_{\text{ML}}, I(\widehat{\boldsymbol{\theta}}_{\text{ML}})^{-1}]$
  - but in finite samples prior can make a difference.
- Noninformative and improper prior
  - has little effect on posterior
  - uniform prior (all values equally likely) is obvious choice
    - ★ improper prior if $\boldsymbol{\theta}$ unbounded usually causes no problem
    - ★ not invariant to transformation (e.g. $\theta \rightarrow e^{\theta}$)
  - Jeffreys prior sets $\pi(\boldsymbol{\theta}) \propto \det[I(\boldsymbol{\theta})^{-1}]$, $I(\boldsymbol{\theta}) = \partial^2 \ln L / \partial\theta\partial\theta'$
    - ★ invariant to transformation
    - ★ for linear regression under normality this is uniform prior for $\boldsymbol{\beta}$
    - ★ also an improper prior.

- Proper prior (informative or uninformative)
  - informative becomes uninformative as prior variance becomes large.
  - use conjugate prior if available as it is tractable
  - hierarchical (multi-level) priors are often used
    - ★ Bayesian analog of random coefficients
    - ★ let $\pi(\boldsymbol{\theta})$ depend on unknown parameters τ which in turn have a completely specified distribution
    - ★ $p(\boldsymbol{\theta}, \boldsymbol{\tau}|\mathbf{y}) \propto L(\mathbf{y}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}|\boldsymbol{\tau}) \times \pi(\boldsymbol{\tau})$ so $p(\boldsymbol{\theta}|\mathbf{y}) \propto \int p(\boldsymbol{\theta}, \boldsymbol{\tau}|\mathbf{y}) d\boldsymbol{\tau}$
- Poisson example with y_i Poisson$[\mu_i = \exp(\mathbf{x}_i, \boldsymbol{\beta})]$
  - $p(\boldsymbol{\beta}, \boldsymbol{\mu}, |\mathbf{y}, \mathbf{X}) \propto L(\mathbf{y}|\boldsymbol{\mu}) \times \pi(\boldsymbol{\mu}|\boldsymbol{\beta}) \times \pi(\boldsymbol{\beta})$
  - where $\pi(\boldsymbol{\mu}_i|\boldsymbol{\beta})$ is gamma with mean $\exp(\mathbf{x}_i'\boldsymbol{\beta})$
  - and $\pi(\boldsymbol{\beta})$ is $\boldsymbol{\beta} \sim \mathcal{N}[\underline{\boldsymbol{\beta}}, \underline{\mathbf{V}}]$.

Convergence of MCMC

- Theory says chain converges as $s \longrightarrow \infty$
  - could still have a problem with one million draws.
- Checks for convergence of the chain (after discarding burn-in)
  - graphical: plot $\theta^{(s)}$ to see that $\theta^{(s)}$ is moving around
  - correlations: of $\theta^{(s)}$ and $\theta^{(s-k)}$ should $\longrightarrow 0$ as k gets large
  - plot posterior density: multimodality could indicate problem
  - break into pieces: expect each 1,000 draws to have similar properties
  - run several independent chains with different starting values.
- But it is not possible to be 100% sure that chain has converged.

Bayesian model selection

- Bayesians use the marginal likelihood
  - $f(\mathbf{y}|\mathbf{X}) = \int L(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X}) \times \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$
  - this weights the likelihood (used in ML analysis) by the prior.
- Bayes factor is analog of likelihood ratio

$$B = \frac{f_1(\mathbf{y}|\mathbf{X})}{f_2(\mathbf{y}|\mathbf{X})} = \frac{\text{marginal likelihood model 1}}{\text{marginal likelihood model 2}}$$

  - one rule of thumb is that the evidence against model 2 is
    - weak if $1 < B < 3$ (or approximately $0 < 2 \ln B < 2$)
    - positive if $1 < B < 3$ (or approximately $2 < 2 \ln B < 6$)
    - strong if $20 < B < 150$ (or approximately $6 < 2 \ln B < 10$)
    - very strong if $B > 150$ (or approximately $2 \ln B > 10$).
- Can use to "test" $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$ against $H_a : \boldsymbol{\theta} = \boldsymbol{\theta}_2$.
- The posterior odds ratio weights B by priors on models 1 and 2.

- Problem: MCMC methods to obtain the posterior avoid computing the marginal likelihood
  - computing the marginal likelihood can be difficult
  - see Chib (1995), JASA, and Chib and Jeliazkov (2001), JASA.
- An asymptotic approximation to the Bayes factor is

$$B_{12} = \frac{L_1(\mathbf{y}|\widehat{\theta}, \mathbf{X})}{L_2(\mathbf{y}|\widehat{\theta}, \mathbf{X})} N^{(k_2 - k_1)/2}$$

  - This is the Bayesian information criterion (BIC) or Schwarz criterion.

What does it mean to be a Bayesian?

- Bayesian inference is a different inference method
  - treats $\boldsymbol{\theta}$ as intrinsically random
  - whereas classical inference treats $\boldsymbol{\theta}$ as fixed and $\widehat{\boldsymbol{\theta}}$ as random.
- Modern Bayesian methods (Markov chain Monte Carlo)
  - make it much easier to compute the posterior distribution than to maximize the log-likelihood.
- So classical statisticians:
  - use Bayesian methods to compute the posterior
  - use an uninformative prior so $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) \simeq L(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})$
  - so $\boldsymbol{\theta}$ that maximizes the posterior is also the MLE.
- Others go all the way and be Bayesian:
  - give Bayesian interpretation to e.g. use credible intervals
  - if possible use an informative prior that embodies previous knowledge.

8. Appendix: Analytically obtaining the Posterior

- Bayesian methods
  - Combine likelihood: $L(\mathbf{y}|\boldsymbol{\theta},\mathbf{X})$
  - and prior on $\boldsymbol{\theta}$: $\boldsymbol{\pi}(\boldsymbol{\theta})$
  - to yield the posterior $p(\boldsymbol{\theta}|\mathbf{y},\mathbf{X})$
- Suppress $\mathbf{X}$ for simplicity
  - $p(\boldsymbol{\theta}|\mathbf{y}) = p(\boldsymbol{\theta},\mathbf{y})/p(\mathbf{y})$ using $\Pr[A|B] = \Pr[A \cap B]/\Pr[B]$
  - and $p(\boldsymbol{\theta},\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})$ using $\Pr[A \cap B] = \Pr[B|A] \times \Pr[A]$
  - So $p(\boldsymbol{\theta}|\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})/p(\mathbf{y})$
- This yields the posterior distribution for $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta}|\mathbf{y},\mathbf{X}) = \frac{L(\mathbf{y}|\boldsymbol{\theta},\mathbf{X}) \times \pi(\boldsymbol{\theta})}{f(\mathbf{y}|\mathbf{X})}$$

  - $f(\mathbf{y}|\mathbf{X}) = \int L(\mathbf{y}|\boldsymbol{\theta},\mathbf{X}) \times \boldsymbol{\pi}(\boldsymbol{\theta})d\boldsymbol{\theta}$ is a normalizing constant called the marginal likelihood.

Example: Scalar normal (known variance) and normal prior

- $y_i|\theta \sim \mathcal{N}[\theta, \sigma^2]$ where σ^2 is known.
- Likelihood: $\mathbf{y} = (y_1, ..., y_N)$ for independent data has likelihood

$$\begin{aligned} L(\mathbf{y}|\theta) &= \prod_{i=1}^{N} \{ \tfrac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\tfrac{1}{2\sigma^2}(y_i - \theta)^2\} \\ &= (2\pi\sigma^2)^{-N/2} \exp\{-\tfrac{1}{2\sigma^2} \textstyle\sum_{i=1}^{N}(y_i - \theta)^2\} \\ &\propto \exp\{-\tfrac{1}{2\sigma^2} \textstyle\sum_{i=1}^{N}(y_i - \theta)^2\} \end{aligned}$$

- Prior: $\theta \sim \mathcal{N}[\mu, \tau^2]$ where μ and τ^2 are specified

$$\begin{aligned} \pi(\theta) &= \tfrac{1}{\sqrt{2\pi\tau^2}} \exp\{-\tfrac{1}{2\tau^2}(\theta - \mu)^2\} \\ &\propto \exp\{-\tfrac{1}{2\tau^2}(\theta - \mu)^2\} \end{aligned}$$

- Note: $\propto$ means "is proportional to"
  - We can drop a normalizing constant that does not depend on θ.

- Normal-normal posterior

$$
\begin{aligned}
p(\theta|\mathbf{y}) &= \frac{L(\mathbf{y}|\theta)\times\pi(\theta)}{\int L(\mathbf{y}|\theta)\times\pi(\theta)dy} \\
&\propto L(\mathbf{y}|\theta)\times\pi(\theta) \\
&\propto \exp\{-\tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^N(y_i-\theta)^2\}\times\exp\{-\tfrac{1}{2\tau}(\theta-\mu)^2\} \\
&\propto \exp\{-\tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^N(y_i-\theta)^2-\tfrac{1}{2\tau^2}(\theta-\mu)^2\} \\
&\propto \exp\{-\tfrac{N}{2\sigma^2}(\theta-\bar{y})^2-\tfrac{1}{2\tau^2}(\theta-\mu)^2\}\ (*) \\
&\propto \exp\{-\tfrac{1}{2}[\tfrac{(\theta-\mu)^2}{\tau^2}+\tfrac{(\theta-\bar{y})^2}{\sigma^2/N}]\} \\
&\propto \exp\{-\tfrac{1}{2}[\tfrac{(\theta-b)^2}{a^2}]\}\ \text{completing the square} \\
&\sim \mathcal{N}[b,a^2]
\end{aligned}
$$

  - $a^2=[(\frac{\sigma^2}{N})^{-1}+(\tau^2)^{-1}]^{-1}$ and $b=a^2\times[(\frac{\sigma^2}{N})^{-1}\bar{y}+(\tau^2)^{-1}\mu]$
  - step (*) uses $\sum_i(y_i-\theta)^2=\sum_i(y_i-\bar{y})^2+N(\bar{y}-\theta)^2$ and can ignore first sum as does not depend on θ
  - $c_1(z-a_1)^2+c_2(z-a_2)^2=(z-\frac{c_1a_1+c_2a_2}{(c_1+c_2)})^2+\frac{c_1c_2}{(c_1+c_2)}(a_1-a_2)^2$.

- Posterior density = normal.
- Posterior variance = inverse of the sum of the precisions
  - precision is the inverse of the variance

$$\text{Posterior variance: } a^2 = [(\tfrac{\sigma^2}{N})^{-1} + (\tau^2)^{-1}]^{-1}$$

$$= [\text{sample precision of } \bar{y} + \text{ prior precision of } \theta]^{-1}$$

- Posterior mean = weighted sum of $\bar{y}$ and prior mean μ
  - where the weights are the precisions

$$\text{Posterior mean: } b = a^2[(\tfrac{\sigma^2}{N})^{-1}\bar{y} + (\tau^2)^{-1}\mu]$$

- Bayesian analysis works with the precision and not the variance.
- More generally σ^2 is unknown
  - then use a gamma prior for the precision $1/\sigma^2$.

Linear regression under normality with normal prior

- Result for i.i.d. case extends to linear regression with $\text{Var}[\mathbf{y}] = \sigma^2\mathbf{I}$ and σ^2 known
  - Likelihood: $\mathbf{y}|\boldsymbol{\beta}, \mathbf{X} \sim \mathcal{N}[\mathbf{X}\boldsymbol{\beta}, \, \sigma^2\mathbf{I}]$
  - Prior: $\boldsymbol{\beta} \sim \mathcal{N}[\underline{\boldsymbol{\beta}}, \, \underline{\mathbf{V}}]$
  - Posterior: $\boldsymbol{\beta}|\mathbf{y}, \mathbf{X} \sim \mathcal{N}[\overline{\boldsymbol{\beta}}, \, \overline{\mathbf{V}}]$ where
    - ★ $\overline{\mathbf{V}} = [\text{sample precision of } \widehat{\boldsymbol{\beta}} + \text{ prior precision of } \boldsymbol{\beta}]^{-1}$
    - ★ $\overline{\mathbf{V}} = [(\sigma^2(\mathbf{X}'\mathbf{X})^{-1})^{-1} + \underline{\mathbf{V}}^{-1}]^{-1}$
      $= [\frac{1}{\sigma^2}(\mathbf{X}'\mathbf{X})^{-1})^{-1} + \underline{\mathbf{V}}^{-1}]$
    - ★ $\overline{\boldsymbol{\beta}} = \overline{\mathbf{V}}[(\sigma^2(\mathbf{X}'\mathbf{X})^{-1})^{-1}\widehat{\boldsymbol{\beta}}_{\text{OLS}} + \underline{\mathbf{V}}^{-1}\underline{\boldsymbol{\beta}}]$
      $= \overline{\mathbf{V}}[\frac{1}{\sigma^2}(\mathbf{X}'\mathbf{y}) + \underline{\mathbf{V}}^{-1}\underline{\boldsymbol{\beta}}]$
- When σ^2 is unknown use a gamma prior for the precision $1/\sigma^2$.
- When $\text{Var}[\mathbf{y}] = \Sigma$ and Σ is unknown use a Wishart prior for Σ^{-1}.

9. Some References

- The material is covered in
  - CT(2005) MMA chapter 13
- Bayesian books by econometricians that feature MCMC are
  - Geweke, J. (2003), *Contemporary Bayesian Econometrics and Statistics*, Wiley.
  - Koop, G., Poirier, D.J., and J.L. Tobias (2007), *Bayesian Econometric Methods*, Cambridge University Press.
  - Koop, G. (2003), *Bayesian Econometrics*, Wiley.
  - Lancaster, T. (2004), *Introduction to Modern Bayesian Econometrics*, Wiley.
- Most useful (for me) book by statisticians
  - Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin (2003), *Bayesian Data Analysis*, Second Edition, Chapman & Hall/CRC.