EC 607: Bayesian Econometrics

Bayesian Computation II

Prof. Jeremy M. Piger

Department of Economics

University of Oregon

Last Revised: March 22, 2019

1 The Gibbs Sampler

1.1 Method

- The Gibbs Sampler is a popular MCMC technique. It can be motivated as a special case of the Metropolis-Hastings algorithm. It can't always be implemented, but when it can, it is often a very efficient sampler.

- Suppose we have a target distribution of interest. To work with a specific example, suppose it is a Bayesian posterior density $p\left(\theta|Y\right)$. Suppose θ is a vector, and partition its elements into two vectors, θ_1 and θ_2, such that $\theta = \left(\theta_1', \theta_2'\right)'$. The target distribution can then be written as:

$$p\left(\theta|Y\right) = p\left(\theta_1, \theta_2|Y\right)$$

- We saw before that we could use a Metropolis-Hastings algorithm implemented in two blocks to sample from $p\left(\theta|Y\right) = p\left(\theta_1, \theta_2|Y\right)$. Suppose we do this, but with a very particular choice for the proposal densities. Specifically, suppose we set the two proposal densities as:

$$q_1\left(\theta_1|\theta_1^{[g]}, \theta_2^{[g]}\right) = p\left(\theta_1|\theta_2^{[g]}, Y\right)$$

$$q_2\left(\theta_2|\theta_1^{[g+1]}, \theta_2^{[g]}\right) = p\left(\theta_2|\theta_1^{[g+1]}, Y\right)$$

- If we plug these proposal densities into the formula for the acceptance probability, it is easy to see that the acceptance probability for proposed draws $\theta_1^{\star}$ or $\theta_2^{\star}$ will always be 1. In other words, with these proposal densities, we will always accept the proposed draws.

- Recognizing this, the Gibbs sampler is then implemented through the following algorithm:

  1. Generate a random value of $\theta_1^{[g+1]}$ from $p\left(\theta_1|\theta_2^{[g]}, Y\right)$

2. Generate a random value of $\theta_2^{[g+1]}$ from $p\left(\theta_2|\theta_1^{[g+1]}, Y\right)$

3. Increment g and go to 1.

The sampler is initiated with an arbitrary initial values, $\theta_2^{[0]}$.

- Note that a key requirement to be able to use the Gibbs Sampler is that $p\left(\theta_1|\theta_2, Y\right)$ and $p\left(\theta_2|\theta_1, Y\right)$ are known conditional densities that we can directly sample from. This isn't always the case, which is why the Gibbs Sampler can't always be used.

- By virtue of this being an example of an MH algorithm, we know that draws from the Gibbs Sampler will converge to draws from a stationary distribution that is the joint posterior density $p\left(\theta_1, \theta_2|Y\right)$. We could also see that $p\left(\theta_1, \theta_2|Y\right)$ is a stationary distribution for the Gibbs Sampler by direct inspection of the transition kernel for the Gibbs Sampler. We will do this next.

- We wish to sample from $p\left(\theta_1, \theta_2|Y\right)$. To implement this using MCMC techniques, we require a transition kernel:

$$K\left(\theta^{[g]}, \theta^{[g+1]}\right) = K\left(\left(\theta_1^{[g]}, \theta_2^{[g]}\right), \left(\theta_1^{[g+1]}, \theta_2^{[g+1]}\right)\right)$$

  such that the Markov chain converges to the target distribution $p\left(\theta_1, \theta_2|Y\right)$.

- The **Gibbs Sampler** defines a transition kernel as follows:

$$K\left(\theta^{[g]}, \theta^{[g+1]}\right) = p\left(\theta_2^{[g+1]}|\theta_1^{[g]}, Y\right) p\left(\theta_1^{[g+1]}|\theta_2^{[g+1]}, Y\right)$$

  In other words, the transition kernel is the the product of two conditional densities. The first gives the posterior probability density for $\theta_2^{[g+1]}$, conditional on the value of $\theta_1^{[g]}$. The second gives the posterior probability density for $\theta_1^{[g+1]}$, conditional on $\theta_2^{[g+1]}$.

This is a valid transition kernel for a Markov chain, as the Markov property holds. The probability of state $\theta^{[g+1]}$ depends only on the past states through $\theta^{[g]}$.

- The Markov chain associated with the Gibbs Sampler transition kernel will have a stationary distribution equal to the target distribution $p\left(\theta|Y\right)$ if the following is true:

$$p\left(\theta_1^{[g+1]},\theta_2^{[g+1]}|Y\right)=\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}p\left(\theta_2^{[g+1]}|\theta_1^{[g]},Y\right)p\left(\theta_1^{[g+1]}|\theta_2^{[g+1]},Y\right)p\left(\theta_1^{[g]},\theta_2^{[g]}|Y\right)d\theta_1^{[g]}d\theta_2^{[g]}$$

- Is this true?

$$p\left(\theta_1^{[g+1]},\theta_2^{[g+1]}|Y\right)=\int\limits_{-\infty}^{\infty}p\left(\theta_2^{[g+1]}|\theta_1^{[g]},Y\right)p\left(\theta_1^{[g+1]}|\theta_2^{[g+1]},Y\right)p\left(\theta_1^{[g]}|Y\right)d\theta_1^{[g]}$$

$$p\left(\theta_1^{[g+1]},\theta_2^{[g+1]}|Y\right)=p\left(\theta_1^{[g+1]}|\theta_2^{[g+1]},Y\right)\int\limits_{-\infty}^{\infty}p\left(\theta_2^{[g+1]}|\theta_1^{[g]},Y\right)p\left(\theta_1^{[g]}|Y\right)d\theta_1^{[g]}$$

$$p\left(\theta_1^{[g+1]},\theta_2^{[g+1]}|Y\right)=p\left(\theta_1^{[g+1]}|\theta_2^{[g+1]},Y\right)\int\limits_{-\infty}^{\infty}p\left(\theta_2^{[g+1]},\theta_1^{[g]}|Y\right)d\theta_1^{[g]}$$

$$p\left(\theta_1^{[g+1]},\theta_2^{[g+1]}|Y\right)=p\left(\theta_1^{[g+1]}|\theta_2^{[g+1]},Y\right)p\left(\theta_2^{[g+1]}|Y\right)$$

$$p\left(\theta_1^{[g+1]},\theta_2^{[g+1]}|Y\right)=p\left(\theta_1^{[g+1]},\theta_2^{[g+1]}|Y\right)$$

Thus, the transitional kernel for the Gibbs Sampler has a stationary distribution equal to the target distribution.

- Note that there was nothing special about the ordering in the transition kernel in this derivation. We could have alternatively used the transition kernel:

$$K\left(\theta^{[g]},\theta^{[g+1]}\right)=p\left(\theta_1^{[g+1]}|\theta_2^{[g]}|Y\right)p\left(\theta_2^{[g+1]}|\theta_1^{[g+1]}|Y\right)$$

- This derivation shows that the Gibbs sampler has the right transition kernel in terms of the existence of a stationary distribution. Again, this means that if we were lucky enough to have our initial value for the chain, $\theta^{[0]}$ be from $p\left(\theta^{[0]}\right)$, then all subsequent draws, $\theta^{[g]}, g = 1, 2, \cdots$ would also be draws from $p\left(\theta^{[g]}\right)$.

- However, if we are starting from some arbitrary $\theta^{[0]}$, we need the Gibbs transition kernel to satisfy irreducibility, recurrence and aperiodicity to guarantee convergence to a unique stationary distribution. Conditions on the Gibbs transition kernel necessary to demonstrate convergence are provided in Geweke (2005, *Contemporary Bayesian Econometrics and Statistics*). However, these are rarely checked in practice for specific models. For most models used in econometric practice, concerns about the applicability of convergence theorems are not substantial.

- Assuming this is a convergent chain, we can then begin to collect the draws of $\theta^{[g]} = \left(\theta_1^{[g]}, \theta_2^{[g]}\right)$ after g is sufficiently large to ensure convergence has occurred. These draws will then be valid draws from $p\left(\theta|Y\right)$.

- Note that while each post-convergence draw of $\theta^{[g]}$ considered by itself will be a draw from $p\left(\theta|Y\right)$, the draws will not be independent draws. Indeed, the dependence between the draws can be quite high. For this reason, some have argued for using only every j^{th} draw to reduce the dependence in the draws. However, the ergodic theorem tells us this is unnecessary. The general rule is that the more dependence there is in your draws, the longer you need to sample to obtain convergence, and the longer you need to sample post-convergence to adequately estimate your objects of interest from the posterior distribution.

- The choice of how θ is split into θ_1 and θ_2 is called “blocking”. This choice is not important for the theoretical justification of the Gibbs Sampler. In practice the blocking choice is made in a way to make construction of the required conditional posterior distributions easier / feasible.

- Finally, although we have presented it this way, there is nothing about the Gibbs

Sampler that requires we have two blocks. Suppose we block θ into J blocks, $\theta = (\theta_1', \theta_2', \cdots, \theta_J')'$. Then we can implement the Gibbs Sampler via the following algorithm:

1. Generate a random value of $\theta_J^{[g+1]}$ from $p\left(\theta_J | \theta_{j<J}^{[g]}, Y\right)$

2. Generate a random value of $\theta_{J-1}^{[g+1]}$ from $p\left(\theta_{J-1} | \theta_J^{[g+1]}, \theta_{j<J-1}^{[g]}, Y\right)$

3. Generate a random value of $\theta_{J-2}^{[g+1]}$ from $p\left(\theta_{J-2} | \theta_J^{[g+1]}, \theta_{J-1}^{[g+1]}, \theta_{j<J-2}^{[g]}, Y\right)$

.

.

J. Generate a random value of $\theta_1^{[g+1]}$ from $p\left(\theta_1 | \theta_{j>1}^{[g+1]}, Y\right)$

Repeat steps 1 through J.

The sampler can be initialized with arbitrary initial value $\theta_{j<J}^{[0]}$.

- Finally, it is sometimes the case that we are implementing a Gibbs Sampler, but one (or more) of the conditional densities $p\left(\theta_j | \theta_{\neq j}, Y\right)$ are unknown. In this case, we can substitute a draw from what is often called a "Metropolis step," meaning that we draw a proposal for θ_j from a proposal density and accept or reject this proposal using a Metropolis-Hastings acceptance probability designed to produce a chain with convergent distribution $p\left(\theta_j | \theta_{\neq j}, Y\right)$. This produces a so-called "Metropolis-within-Gibbs" sampler.

2 Sampling the Posterior Predictive Density

- Recall, the posterior predictive density is the Bayesian object of interest for forecasting:

$$p\left(y^* | Y\right) = \int_\theta p\left(y^*, \theta | Y\right) d\theta$$

- For most models we care about, we won't be able to calculate this integral directly. However, recall from our discussion of sampling that we can generate a draw from $p\left(y^*,\theta|Y\right)$ by first generating a draw from the posterior distribution for θ, $p\left(\theta|Y\right)$, and then generating a draw from $p\left(y^*|\theta,Y\right)$. This second draw can usually be generated easily given knowledge of the likelihood function. Recall then that the draws of y^* from $p\left(y^*,\theta|Y\right)$ will be valid draws from the posterior predictive density, $p\left(y^*|Y\right)$.

- Using these draws we can then estimate objects of interest from the posterior predictive density, or the density itself, to an arbitrary degree of accuracy.

3 Gibbs Sampler Example 1: Normal Linear Regression with Independent Normal and Gamma Priors

- Recall the Gaussian linear regression model with likelihood:

$$p\left(Y|\beta,h,\mathrm{X}\right)=\left(2\pi\right)^{-\frac{N}{2}}h^{\frac{N}{2}}\exp\left[-\frac{h}{2}\left(\beta-\widehat{\beta}_{OLS}\right)'\mathrm{X}'\mathrm{X}\left(\beta-\widehat{\beta}_{OLS}\right)\right]\exp\left[-\frac{h}{2}e'_{OLS}e_{OLS}\right]$$

The priors are:

$$p\left(\beta,h\right)=p\left(\beta\right)p\left(h\right)$$

where:

$$\beta\sim N\left(\mu,V\right)$$

$$h\sim\text{Gamma}\left(m,v\right)$$

The full equation for these prior probability distribution functions are:

$$p\left(\beta\right)=\left(2\pi\right)^{-\frac{k}{2}}\left|V\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\left(\beta-\mu\right)'V^{-1}\left(\beta-\mu\right)\right]$$

$$p\left(h\right)=\frac{1}{\left(\frac{2m}{v}\right)^{v/2}}\frac{1}{\Gamma\left(\frac{v}{2}\right)}h^{\frac{v-2}{2}}\exp\left[-\frac{hv}{2m}\right]$$

- To implement the Gibbs sampler, we will use the blocking β and h. We then require both conditional posterior distributions, $p\left(\beta|h,X,Y\right)$ and $p\left(h|\beta,X,Y\right)$.
- Consider first $p\left(\beta|h,X,Y\right)$. From Bayes Rule:

$$p\left(\beta|h,Y,\mathrm{X}\right)\propto p\left(Y|\beta,h,\mathrm{X}\right)p\left(\beta|h,\mathrm{X}\right)$$

Because of the prior independence of β and h, as well as the exogeneity of X, $p\left(\beta|h,\mathrm{X}\right)$ in this equation is equivalent to our prior, $p\left(\beta\right)$.

$$p\left(\beta|h,Y,\mathrm{X}\right)\propto p\left(Y|\beta,h,\mathrm{X}\right)p\left(\beta\right)$$

Now, plugging in the equations for these components and eliminating terms that don't depend on β we have:

$$p\left(\beta|h,Y,\mathrm{X}\right)\propto\exp\left[-\frac{1}{2}\left(\left(\beta-\widehat{\beta}_{OLS}\right)'h\mathrm{X}'\mathrm{X}\left(\beta-\widehat{\beta}_{OLS}\right)+\left(\beta-\mu\right)'V^{-1}\left(\beta-\mu\right)\right)\right]$$

If we use our completion of the square formula and again eliminate terms that do not depend on β, we obtain:

$$p\left(\beta|h,Y,\mathrm{X}\right)\propto\exp\left[-\frac{1}{2}\left[\left(\beta-\overline{\mu}\right)'\overline{V}\left(\beta-\overline{\mu}\right)\right]\right],$$

where:

$$\begin{aligned}
\overline{V} &= h\mathrm{X}'\mathrm{X} + V^{-1} \\
\overline{\mu} &= \overline{V}^{-1}\left(h\mathrm{X}'\mathrm{X}\widehat{\beta}_{OLS} + V^{-1}\mu\right) \\
&= \overline{V}^{-1}\left(h\mathrm{X}'Y + V^{-1}\mu\right)
\end{aligned}$$

This is recognized as the kernel of a multivariate normal distribution with mean vector $\overline{\mu}$ and variance-covariance matrix $\overline{V}^{-1}$. Thus:

$$\beta|h, Y, \mathrm{X} \sim N\left(\overline{\mu}, \overline{V}^{-1}\right)$$

- To complete the Gibbs sampler we then need $p\left(h|\beta, Y, \mathrm{X}\right)$. Applying Bayes Rule as before, we have:

$$p\left(h|\beta, Y, \mathrm{X}\right) \propto p\left(Y|\beta, h, \mathrm{X}\right)p\left(h\right)$$

Plugging the equations into this formula and eliminating all terms that don't rely on h we have:

$$\begin{aligned}
p\left(h|\beta, Y, \mathrm{X}\right) &\propto h^{\frac{N}{2}}\exp\left[-\frac{h}{2}\left(\beta - \widehat{\beta}_{OLS}\right)'\mathrm{X}'\mathrm{X}\left(\beta - \widehat{\beta}_{OLS}\right)\right]\exp\left[-\frac{h}{2}e'_{OLS}e_{OLS}\right]h^{\frac{v-2}{2}}\exp\left[-\frac{hv}{2m}\right] \\
p\left(h|\beta, Y, \mathrm{X}\right) &\propto h^{\frac{N+v-2}{2}}\exp\left[-\frac{h}{2}\left[\left(\beta - \widehat{\beta}_{OLS}\right)'\mathrm{X}'\mathrm{X}\left(\beta - \widehat{\beta}_{OLS}\right) + e'_{OLS}e_{OLS} + \frac{v}{m}\right]\right]
\end{aligned}$$

This is the kernel of a Gamma distribution with parameters $\overline{m}$ and $\overline{v}$:

$$\begin{aligned}
\overline{v} &= N + v \\
\overline{m} &= \frac{\overline{v}}{\left(\beta - \widehat{\beta}_{OLS}\right)'\mathrm{X}'\mathrm{X}\left(\beta - \widehat{\beta}_{OLS}\right) + e'_{OLS}e_{OLS} + \frac{v}{m}}
\end{aligned}$$

This can be simplified somewhat by recalling that:

$$(Y - X\beta)' (Y - X\beta) = \left(\beta - \widehat{\beta}_{OLS}\right)' \mathrm{X}'\mathrm{X} \left(\beta - \widehat{\beta}_{OLS}\right) + e_{OLS}' e_{OLS}$$

So:

$$\overline{m} = \frac{\overline{v}}{(Y - X\beta)' (Y - X\beta) + \frac{v}{m}}$$

Thus:

$$h|\beta, Y, \mathrm{X} \sim \text{Gamma}\left(\overline{m}, \overline{v}\right)$$

- The Gibbs sampler for the linear regression model with independent normal prior for slope coefficients and Gamma prior for the precision parameter is then implemented as follows:

  1. Generate a random value of $h^{[g+1]}$ from Gamma $\left(\overline{m}, \overline{v}\right)$
  2. Generate a random value of $\beta^{[g+1]}$ from $N\left(\overline{\mu}, \overline{V}^{-1}\right)$
  3. Increment g and go to 1.

  The sampler is initialized with an arbitrary initial value for β, denoted $\beta^{[0]}$:

4 Exercise: Gibbs Sampler for Linear Regression Model

In the zipped collection of filed titled "Linear Regression GS.zip" you will find a dataset called "TeachingRatings.xls," which holds data regarding $N = 463$ courses taught at the University of Texas at Austin over the period 2000-2002. The various series are:

- course_eval: An instructor evaluation score, on a score of 1 (very unsatisfactory) to 5 (excellent).

- beauty: Rating of instructor physical appearance by a panel of six students, averaged across the six panelists. This variable is shifted to have a sample mean of zero. Thus, a score of 0 is "average" beauty.
- female: A dummy variable that is 1 if the instructor is female.
- minority: A dummy variable that is one if is the instructor is non-White.
- nnenglish: A dummy variable that is one if the instructor is a non-native English speaker.
- intro: A dummy variable that is one if the course is "Introductory", which is mainly large freshman and sophomore classes.
- one_credit: A dummy variable that is one if the course is a single-credit elective course.
- age: The instructor's age.

This dataset was used in Hamermesh and Parker, 2005, "Beauty in the Classroom: InstructorsÕ Pulchritude and Putative Pedagogical Productivity," *Economics of Education Review.*

We will consider a linear regression in which course_eval is the independent variable, labeled Y and beauty, female, minority, nnenglish, intro, age, and one_credit are independent variables, labeled X_2 through X_8 respectively. X_1 will be an $N \times 1$ vector of ones, which will incorporate an intercept into the regression model. Consider the following linear regression model with Gaussian errors:

$$Y = \mathrm{X}\beta + \epsilon,$$

$$\epsilon \sim N\left(\mathbf{0_N}, h^{-1} I_N\right)$$

where $Y = (y_1, y_2, \cdots, y_N)'$, $X = (X_1, X_2, \cdots, X_k)$, $X_j = (x_{j,1}, x_{j,2}, \cdots, x_{j,N})'$, $\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_N)'$, and $\beta = (\beta_1, \beta_2, \cdots, \beta_k)'$. Here $k = 8$ is the number of slope parameters in the model (including the intercept).

Assume we have independent Normal and Gamma prior distributions for β and h:

$$\begin{aligned} p(\beta, h) &= p(\beta)\, p(h) \\ \beta &\sim N(\mu, V) \\ h &\sim \text{Gamma}(m, v) \end{aligned}$$

The full equations for these prior probability distribution functions are:

$$p(\beta) = (2\pi)^{-\frac{k}{2}} |V|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\beta-\mu)' V^{-1} (\beta-\mu)\right]$$

$$p(h) = \frac{1}{\left(\frac{2m}{v}\right)^{v/2}} \frac{1}{\Gamma\left(\frac{v}{2}\right)} h^{\frac{v-2}{2}} \exp\left[-\frac{hv}{2m}\right]$$

We set the hyperparameters for the prior distribution as $\mu = 0_8$, $V = 10I_8$, $m = 1$ and $v = 3$.

The Matlab programs "Bayesian_Beauty_Regression.m", "gen_beta.m", and "gen_h.m" are set up to sample from $p(\beta, h|Y)$ using the Gibbs Sampler. Work with these programs to make sure you understand them. Here are some suggested activities:

1. Sample the posterior distribution, $p(\beta, h|Y)$ using the Gibbs Sampler. Make a table listing the posterior mean, posterior standard deviation, and 90% central posterior interval for each β parameter and for the disturbance variance, $\sigma^2 = h^{-1}$. Also in the table include the OLS estimate of each β parameter.

2. Plot an estimate of the marginal posterior density of the coefficient on the "beauty" variable, $p(\beta_2)$. You can do this by using the Matlab command **ksdensity** applied to the random draws of β_2.

3. Suppose you have a null hypothesis that perceived beauty increases course evaluations, all else equal. Using the random samples from the Gibbs Sampler, report your estimate of $\Pr(\beta_2 > 0)$ as a way of evaluating the evidence regarding this hypothesis.

4. Suppose you believe that perceived beauty affects course evaluations differently for male vs. female instructors. Modify the code to remove X_2 from the regression, and replace it with two new variables, $X_2 * X_3$ and $X_2 * (1 - X_3)$. Report your estimates of the posterior mean, standard deviation, and 90% central posterior interval of the slope parameters on these two new variables. Also, plot the marginal posterior density for each of these slope parameters on the same graph.

5. Rerun the Gibbs Sampler with different initial values. Do the results change much?

6. **Extra, Extra Credit:** Suppose you believe that the conditional **volatility** of course evaluations might be different for male vs. female instructors. Suppose we believe that:

$$\epsilon_i \sim N(0, h_i) \tag{1}$$

where $h_i = h$ if $X_{3i} = 0$ and $h_i = h^*$ if $X_{3i} = 1$. The distribution for ϵ is then given by:

$$\epsilon \sim N(\mathbf{0_N}, \mathrm{diag}(h_1, h_2, h_3, \cdots, h_N))$$

where $\mathrm{diag}(\cdot)$ indicates a diagonal matrix with diagonal entries given by its argument. Here h measures the precision of the regression for a course with a male instructor, while h^* measures the precision for a course with a female instructor. You should use independent gamma priors for h and h^*:

$$\begin{aligned} p(h, h^*) &= p(h)\, p(h^*) \\ h &\sim \mathit{Gamma}(m, v) \\ h^* &\sim \mathit{Gamma}(m^*, v^*) \end{aligned}$$

where $m = m^* = 1$ and $v = v^* = 3$.

Modify the code you used from the previous problem to sample the posterior distribution, $p\left(\beta, h, h^{*}|Y\right)$ using the Gibbs Sampler. Define $\sigma = h^{-0.5}$ and $\sigma^{*} = h^{*-0.5}$. Report the posterior means of σ and σ^{*} and plot the estimated probability density function for each of these (on the same graph). Also report the posterior mean of the ratio:

$$\frac{\sigma^{*}}{\sigma}$$

and plot the estimated probability density function of the draws of this ratio.

You will need to design a modified Gibbs sampler to complete this problem. I suggest that you use a three-block Gibbs Sampler with one block being β and the other two being h and h^{*}.

5 Gibbs Sampler Example 2: Probit Model

- For our second example, we will show how the Gibbs Sampler can be implemented for a probit model. This example is interesting in its own right, and will also introduce a technique known as **Data Augmentation**.
- Suppose we have a discrete, bivariate, random variable $y_i \in \{0, 1\}$. The probit model assumes that:

$$y_i = \begin{cases} 1 \text{ if } y_i^{*} \geq 0 \\ 0 \text{ if } y_i^{*} < 0 \end{cases}$$

where:

$$y_i^{*} = x_i\beta + \epsilon_i$$
$$\epsilon_i \sim N\left(0, h^{-1}\right)$$

and $x_i = (x_{1,i}, x_{2,i}, \cdots, x_{k,i})$, where $x_{j,i}$ is the i^{th} observation of the j^{th} variable.

- In the probit model y_i^* is not observed, and is referred to as a **latent** variable or **latent** data.

- The probit model implies the following probability distribution for y_i:

$$\begin{aligned}\Pr\left(y_i = 1|\beta, h, x_i\right) &= \Pr\left(\epsilon_i \geq -x_i\beta|\beta, h, x_i\right) \\ &= \Pr\left(-h^{0.5}\epsilon_i \leq h^{0.5}x_i\beta|\beta, h, x_i\right) \\ &= \Pr\left(\epsilon_i^* \leq h^{0.5}x_i\beta|\beta, h, x_i\right)\end{aligned}$$

where $\epsilon_i^* = \left(-h^{0.5}\epsilon_i\right) \sim N\left(0, 1\right)$.

Thus, we have:

$$\Pr\left(y_i = 1|\beta, h, x_i\right) = \Phi\left(h^{0.5}x_i\beta\right)$$

where $\Phi\left(\cdot\right)$ is the standard normal cumulative density function. Using similar calculations we have:

$$\Pr\left(y_i = 0|\beta, h, x_i\right) = 1 - \Phi\left(h^{0.5}x_i\beta\right)$$

Note that since β and $h^{0.5}$ enter the probability distribution for y_i in exactly the same way, they will not be separately identified in the likelihood function. The usual practice is to normalize $h = 1$. Then:

$$\begin{aligned}\Pr\left(y_i = 1|\beta, x_i\right) &= \Phi\left(x_i\beta\right) \\ \Pr\left(y_i = 0|\beta, x_i\right) &= 1 - \Phi\left(x_i\beta\right)\end{aligned}$$

In these equations I have eliminated the conditioning on $h = 1$, but it should be remembered that all probability distributions computed below are conditional on $h = 1$.

- Note that y_i^* takes the form of a Normal linear regression with known precision. This will be useful in later derivations.

- Suppose we have N observations on y_i, collected in $Y = (y_1, y_2, \cdots, y_N)'$, and x_i, collected in $X = (x_1', x_2', \cdots, x_N')'$. We wish to conduct Bayesian estimation of the parameters of this model, which is β.

- To do so we need a prior for β. Here we assume a Gaussian prior:

$$\beta \sim N(\mu, V)$$

- In the Bayesian framework, latent data is treated like any other unobserved object, and incorporated into the problem as one of the unknown quantities of interest. Thus, we will approach Bayesian estimation of the probit model by constructing a Gibbs sampler to sample from the joint posterior distribution of β and $Y^* = (y_1^*, y_2^*, \cdots, y_N^*)'$:

$$p(\beta, Y^* | Y, X)$$

  To implement the Gibbs Sampler we will sample iteratively from the two conditional posterior distributions:

$$p(\beta | Y^*, Y, X)$$
$$p(Y^* | \beta, Y, X)$$

  Note that in one of these distributions we will simulate a realization of the latent data, Y^*. Then, will condition on this realization of Y^* in the second distribution to obtain a draw of β. This simulating of latent data to use as conditioning information in other simulation steps is known as **data augmentation**.

- We then need to figure out how to sample from the two required distributions.

Sampling $p\left(\beta|Y^*, Y, X\right)$:

First, note that conditional on Y^*, Y provides no additional information, as its elements simply indicate the signs of the elements of Y^*. Thus:

$$p\left(\beta|Y^*, Y, X\right) = p\left(\beta|Y^*, X\right)$$

Second, note that:

$$Y^* = X\beta + \epsilon \tag{2}$$

where $\epsilon \sim N\left(0_N, I_N\right)$. Thus, Y^* takes the form of a linear regression model with slope parameters β and known precision ($h = 1$). We also have a Gaussian prior for β that is independent of h. Using the results we had earlier for the posterior of β in a linear regression conditional on $h = 1$ we then have:

$$\beta|Y^*, \mathrm{X} \sim N\left(\overline{\mu}, \overline{V}^{-1}\right)$$

where:

$$\begin{aligned} \overline{V} &= \mathrm{X}'\mathrm{X} + V^{-1} \\ \overline{\mu} &= \overline{V}^{-1}\left(\mathrm{X}'Y^* + V^{-1}\mu\right) \end{aligned}$$

We can then obtain random draws of β from this normal distribution.

Sampling $p\left(Y^*|\beta, Y, X\right)$:

Applying Bayes Rule we have

$$p\left(Y^*|\beta, Y, X\right) \propto \Pr\left(Y|\beta, Y^*, X\right) p\left(Y^*|\beta, X\right)$$

Since Y is a deterministic function of Y^*, the first term on the right-hand side above is equivalent to:

$$\Pr\left(Y|\beta, Y^*, X\right) = \Pr\left(Y|Y^*\right)$$

This will equal 1 if Y correctly indicates the sign of Y^* and 0 otherwise. Also, it can be written as the product of N independent distributions:

$$\Pr\left(Y|Y^*\right) = \prod_{i=1}^{N} \Pr\left(y_i|y_i^*\right)$$

Each term in this product is one if y_i correctly indicates the sign of y_i^* and 0 otherwise. The second term on the right-hand side is given by:

$$Y^*|\beta, X \sim N\left(X\beta, I_N\right)$$

This distribution can also be written as the product of N independent distributions:

$$p\left(Y^*|\beta, X\right) = \prod_{i=1}^{N} p\left(y_i^*|\beta, X_i\right)$$

where

$$y_i^*|\beta, X_i \sim N\left(X_i\beta, 1\right)$$

Putting this all together we have:

$$p\left(Y^*|\beta, Y, X\right) \propto \prod_{i=1}^{N} \Pr\left(y_i|y_i^*\right) \exp\left(-0.5\left(y_i^* - X_i\beta\right)^2\right)$$

Each of the terms in this product represents the kernel of an independent truncated normal distribution for y_i^*. When $y_i = 1$, the term is a normal distribution that is left truncated at $y_i^* = 0$. When $y_i = 0$, the term is a normal distribution that is right truncated at $y_i^* = 0$. As $p(Y^*|\beta, Y, X)$ is proportional to this product, this probability distribution is then given by:

$$Y^*|\beta, Y, X \sim \prod_{i=1}^{N} \left(N_{0^+}(X_i\beta, 1)\right)^{y_i} \left(N_{0^-}(X_i\beta, 1)\right)^{1-y_i}$$

where N_{0^+} and N_{0^-} indicate left and right truncation at 0 respectively. We can sample this truncated distribution via rejection sampling:

1. Draw y_i^* from $N(x_i\beta, 1)$.
2. If $y_i = 1$ and $y_i^* \geq 0$ or $y_i = 0$ and $y_i^* < 0$ then accept the draw of y_i^*. Otherwise, go to 1 and redraw.

   Repeat 1 and 2 for $i = 1, 2, \cdots, N$

- The Gibbs sampler for the probit model with normal prior for β is then implemented as follows.

  1. Generate N independent random variates $y_i^{*[g+1]}$, $i = 1, \cdots, N$ from $N_{0^+}(X_i\beta, 1)$ if $y_i = 1$ and from $N_{0^-}(X_i\beta, 1)$ if $y_i = 0$.
  2. Generate a random value of $\beta^{[g+1]}$ from $N\left(\overline{\mu}, \overline{V}^{-1}\right)$
  3. Increment g and go to 1.

  The sampler is initialized with an arbitrary initial value for β, denoted $\beta^{[0]}$:

- Once convergence has occurred, the sampler defined above will provide draws of $\beta^{[g]}$ and $y^{*[g]}$ from $p(\beta, y^*|Y)$, $g = 1, 2, \cdots, G$. Of course, the $\beta^{[g]}$ are also draws from $p(\beta|Y)$.

- With probit models, we are often interested not in the parameter β, but in the marginal effects:

$$\frac{\partial \Pr(y_i = 1|x_i, \beta)}{\partial x_{ik}} = \frac{\partial \Phi(x_i\beta)}{\partial x_{ik}} = \beta_k \phi(x_i\beta)$$

where $\phi\left(\cdot\right)$ is the standard normal probability distribution function. As this is a deterministic function of β, we can simulate a draw from the posterior distribution of this marginal effect as:

$$\beta_k^{[g]}\phi\left(x_i\beta^{[g]}\right)$$

6 Exercise: Forecasting U.S. Recessions with a Probit Model

In the zipped collection of files titled "Probit GS.zip" you will find a monthly dataset called "Recession_Forecast_Data.xlsx." This holds data on a variable called NBER, which is a $\{0,1\}$ variable measuring U.S. recession (1) and expansion (0) dates. It also holds four variables that we will use to predict the NBER variable, which include the Federal Funds Rate, the S&P500 stock price index, the 10 year Treasury Bond yield, and the 3 month Treasury Bill yield.

From these four variables we will define three predictor variables:

1. FFR: The level of the Federal Funds Rate
2. $SP500\_Return$: The three month growth rate of the S&P 500 stock price index
3. $Term\_Spread$: The difference between the 10 year Treasury Bond and the 3 month Treasury Bill yields

We will be interested in using these predictor variables to forecast the NBER variable 12 months ahead. The probit model will then be:

$$NBER_t = \begin{cases} 1 \ \text{ if } \ y_t^* \geq 0 \\ 0 \ \text{ if } \ y_t^* < 0 \end{cases}$$

where:

$$y^*_{t+12} = \beta_1 + \beta_2 FFR_t + \beta_3 SP500\_Return_t + \beta_4 Term\_Spread_t + \epsilon_{t+12}$$

$$\epsilon_{t+12} \sim N(0,1)$$

The Matlab files "Probit_Recession_Forecast.m," "gen_ystar" and "gen_beta" are set up to sample from the joint posterior of β and the latent variable using the Gibbs Sampler. Work with these files to make sure you understand how they work. Suggested activities include:

1. Estimate this model over the time period from January 1960 to February 2014. Plot estimates of the posterior density function for β_2, β_3 and β_4. Also plot the posterior mean of $\Pr(NBER_t = 1|Y)$, where t covers the in-sample estimation period.

2. Now plot an estimate of the posterior predictive distribution for the "out-of-sample" prediction $\Pr(NBER_{T+12} = 1|Y_T)$, where T is the end of the sample period, February 2014. So you will be producing a forecast of $NBER_t$, where $t =$ February 2015. Also report the median of this posterior predictive distribution.

3. Modify the program to mimic a forecaster who is using this model during 2006 and 2007 to forecast recessions 12 months ahead. In particular, estimate the model over rolling samples starting with January 1960 - January 2007, and ending with January 1960 - December 2007. Report information from the posterior predictive distribution for each of these samples - you decide what and how you will report this information. Did this forecasting model forecast the "Great Recession?"