Gibbs Sampling in Endogenous Variables Models

Econ 690

Purdue University

Outline

Motivation

- In this lecture we take up issues related to posterior simulation in models with endogeneity concerns.
- By "endogeneity," we are referring to the case where a covariate (usually the object of interest) is potentially correlated with the error term.
- In this case, OLS estimates (in the case of a linear model) will be biased and inconsistent in general.
- Moreover, interest often centers on providing a "causal" interpretation associated with the slope coefficient, but if such correlation is present, this interpretation cannot be given.
- A simple example motivates.

- The following public service announcement frequently used to air on California TV:
- A young boy walked down the street and approaches an older, disheveled-looking man, who was holding out a cup, presumably asking the boy to drop change into this cup.
- The boy looks at the man, and then you hear the following voice-over: "High school dropouts make 42 percent less, on average, than high school graduates"
- What happens next is that the man takes some change out of the cup and decides to hand it to the boy (presumably he is a dropout, and so the implication is that if you drop out of high school, you will be in bad financial shape).

- To get this result, a simple log wage regression was run, with a high school dropout indicator on the right-hand side. The coefficient from the OLS regression was -.42, and thus the interpretation.
- But, should we believe this as a "causal" statement?
- Doesn't it seem likely that the dropout variable will be correlated with other omitted characteristics in the wage equation?
- In this lecture, we discuss models that are appropriate for these types of situations, and will give better estimates of the "causal" impact of interest.

- Before going into details about models and Gibbs, we first review some important identification issues.
- Specifically, we show that an instrument (a variable which is conditionally correlated with the endogenous variable, but can be excluded from the outcome equation) is necessary in the purely linear model for identification purposes *even under our assumption of normality!*

Identification Issues

Consider, for simplicity, a continuous "endogenous" variable x and a continuous outcome y.

We abstract from the presence of other covariates and consider the model:

$$\begin{aligned} x &= \epsilon_1 \\ y &= \beta x + \epsilon_2 \end{aligned}$$

where

$$\left[\begin{array}{c} \epsilon_1 \\ \epsilon_2 \end{array} \right] \Big| x \sim N \left[\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \left(\begin{array}{cc} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{array} \right) \right].$$

Note, in this case, that the triangularity of our model guarantees that the Jacobian of the transformation from ϵ to $[y \ x]$ is unity.

With a unit Jacobian, the joint distribution of a representative (y, x) pair in the sample is

$$p(x, y|\beta, \Sigma) = \phi\left[\begin{pmatrix} x \\ y \end{pmatrix}; \begin{pmatrix} 0 \\ \beta x \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right].$$

Note that, when evaluating this likelihood,

$$|\Sigma| = \sigma_1^2\sigma_2^2 - \sigma_{12}^2 \equiv a^2$$

and

$$\Sigma^{-1} = \frac{1}{a^2}\begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix}.$$

Therefore,

$$p(x, y|\beta, \sigma) \propto \frac{1}{a} \exp\left(-\frac{1}{2a^2}[x \;\; (y-\beta x)] \begin{bmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{bmatrix} \begin{bmatrix} x \\ y-\beta x \end{bmatrix}\right)$$

Which is

$$\frac{1}{a} \exp\left(-\frac{1}{2a^2}[\sigma_2^2 x - \sigma_{12}(y-\beta x) \;\; : \;\; -\sigma_{12}x + \sigma_1^2(y-\beta x)] \begin{bmatrix} x \\ y-\beta x \end{bmatrix}\right)$$

or

$$\frac{1}{a} \exp\left(-\frac{1}{2a^2}\left[\sigma_2^2 x^2 - 2\sigma_{12}x(y-\beta x) + \sigma_1^2(y-\beta x)^2\right]\right).$$

Is it obvious that there is an identification problem here? (That is, we will get the same likelihood at different configurations of the parameters σ_1^2, σ_2^2, σ_{12} and β?)

Let's keep going and see ...

Consider the term in the inside of the exponential kernel, and let's expand it:

$$\begin{aligned}
\sigma_2^2 x^2 - 2\sigma_{12} x(y-\beta x) + \sigma_1^2 (y-\beta x)^2 &= \sigma_1^2 [(y-\beta x)^2 - 2\frac{\sigma_{12}}{\sigma_1^2} x(y-\beta x) + \frac{\sigma_2^2}{\sigma_1^2} x^2] \\
&= \sigma_1^2 \left([(y-\beta x) - \frac{\sigma_{12}}{\sigma_1^2} x]^2 + \frac{\sigma_2^2}{\sigma_1^2} x^2 - \frac{\sigma_{12}^2}{\sigma_1^4} x^2 \right) \\
&= \sigma_1^2 \left([(y-\beta x) - \frac{\sigma_{12}}{\sigma_1^2} x]^2 \right) + \frac{x^2}{\sigma_1^2} a^2
\end{aligned}$$

Subbing this back into our expression for the likelihood gives

$$p(y, x|\beta, \Sigma) \propto \frac{1}{a} \exp\left(-\frac{\sigma_1^2}{2a^2}[y - x(\beta + \frac{\sigma_{12}}{\sigma_1^2})]^2\right) \exp\left(-\frac{x^2}{2\sigma_1^2}\right).$$

So, the likelihood is completely determined by three parameters:

$$a, \quad \sigma_1^2, \quad \text{and} \quad \psi = (\beta + \sigma_{12}/\sigma_1^2).$$

However, for the estimation of our model, we need four parameters: σ_1^2, σ_2^2, σ_{12} and β.

It is clear, then, that different combinations of these four values can yield the same likelihood function, which depends on only three of these objects. Thus, this model, as written is *not fully identified*!

Also note the close connection of this result with breaking the joint distribution of y and x into $p(x)$ and $p(y|x)$.

Specifically,

$$y|x \sim N\left[\beta x + \frac{\sigma_{12}}{\sigma_1^2}x, \sigma_1^{-2}a^2\right].$$

and

$$x \sim N(0, \sigma_1^2)$$

From here, it is obvious that we can identify σ_1^2 from the x marginal, a^2 from the variance of the conditional, and $\beta + \sigma_{12}/\sigma_1^2$ from the conditional, but that is all - β is not separately identified.

What would you expect to see if you fit a Gibbs sampler to this model?

Identification via IV's

Instruments, however, are one potential vehicle for identification.

There are others, of course, including assuming conditional independence or through cross-equation restrictions. These assumptions, however, are typically not convincing.

To see the value of IV's, consider the model:

$$\begin{aligned} x &= z\delta + \epsilon \\ y &= x\beta + w\theta + u, \end{aligned}$$

with a similar joint distributional assumption on ϵ and u.

In this case, the x marginal is:

$$x \sim N(z\delta, \sigma_\epsilon^2),$$

and thus δ and σ_ϵ^2 are separately identifiable.

In addition,

$$y|x \sim N[x\beta + w\theta + (\sigma_{\epsilon u}/\sigma_\epsilon^2)(x - z\delta), \sigma_u^2(1 - \rho_{\epsilon u}^2)].$$

or

$$y|x \sim N[(\beta + \sigma_{\epsilon u}/\sigma_\epsilon^2)x + w\theta - (\sigma_{\epsilon u}/\sigma_\epsilon^2)z\delta, \sigma_u^2(1 - \rho_{\epsilon u}^2)].$$

Now, consider the case where $w = z$ (i.e., there are no instruments). Then, we have

$$x \sim N(z\delta, \sigma_\epsilon^2)$$

and

$$y|x \sim N[(\beta + \sigma_{\epsilon u}/\sigma_\epsilon^2)x + w[\theta - (\sigma_{\epsilon u}/\sigma_\epsilon^2)\delta], \sigma_u^2(1 - \rho_{\epsilon u}^2)].$$

So, we have 6 parameters: β, δ, θ, σ_ϵ^2, σ_u^2 and $\sigma_{\epsilon u}$, but only 5 "equations" we can use to estimate them!

However, if there is at least one element in z that is not in w, then:

$$y|x \sim N[(\beta + \sigma_{\epsilon u}/\sigma_\epsilon^2)x + w\theta - (\sigma_{\epsilon u}/\sigma_\epsilon^2)z\delta, \sigma_u^2(1 - \rho_{\epsilon u}^2)].$$

and all the parameters are identified (how, intuitively, does this happen?)

A Standard Binary Treatment Model

Consider the following model with a continuous outcome y and a discrete treatment variable D:

-

where

$$\begin{pmatrix} \epsilon_i \\ u_i \end{pmatrix} \overset{iid}{\sim} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\epsilon^2 & \sigma_{u\epsilon} \\ \sigma_{u\epsilon} & 1 \end{pmatrix}\right] \equiv N(0, \Sigma).$$

The second equation of the system is a latent variable equation which generates D, i.e., $D = I(D^* > 0)$, like the probit model previously discussed.
We seek to describe a Gibbs sampling algorithm for fitting this two equation system.

First, note that we can stack the model into the form

$$\tilde{y}_i = X_i\beta + \tilde{u}_i,$$

where

$$\tilde{y}_i = \begin{bmatrix} y_i \\ D_i^* \end{bmatrix}, \quad X_i = \begin{bmatrix} 1 & D_i & 0 \\ 0 & 0 & z_i \end{bmatrix} \quad \beta = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \theta \end{bmatrix}, \quad \tilde{u}_i = \begin{bmatrix} \epsilon_i \\ u_i \end{bmatrix}.$$

It this form, posterior simulation seems identical to the SUR model. However, there is one small complication (what is it?) Our covariance matrix Σ is not unrestricted, and in fact, the (2,2) element must be unity for identification purposes.

This precludes the use of a standard Wishart prior and typical conjugate analysis.

To facilitate drawing the parameters of the covariance matrix Σ, it is desirable to work with the population expectation of ϵ given u:

-

where $v_i \sim N(0, \sigma_v^2)$, and $\sigma_v^2 \equiv \sigma_\epsilon^2 - \sigma_{u\epsilon}^2$.

So, we can work with an equivalent version of the model:

-

where u and v are independently distributed. In this parameterization,Σ takes the form:

-

We complete the model by choosing priors of the following forms:

$$\begin{aligned} \beta &\sim N(\mu_\beta, V_\beta) \\ \sigma_{u\epsilon} &\sim N(\mu_0, V_0) \\ \sigma_v^2 &\sim IG(a, b), \end{aligned}$$

Finally, note that Σ is positive definite for $\sigma_v^2 > 0$, which is enforced through our prior.

We work with an augmented posterior distribution of the form

$$p(\beta, D^*, \sigma_v^2, \sigma_{u\epsilon}|y, D).$$

As for the complete conditional for β, its derivation follows identically to the SUR model:

$$\beta|D^*, \Sigma, y, D \sim N(D_\beta d_\beta, D_\beta)$$

where

$$D_\beta = (\sum_i X_i'\Sigma^{-1}X_i + V_\beta^{-1})^{-1}, \quad d_\beta = \sum_i X_i'\Sigma^{-1}\tilde{y}_i + V_\beta^{-1}\mu_\beta.$$

(Note that Σ is known given σ_v^2 and $\sigma_{u\epsilon}$).

As for the posterior conditional for each D_i^*, we must first break our likelihood contributions into a conditional for $D_i^*|y_i$ and a marginal for y_i. Thus,

As for the parameters of the covariance matrix, let us go back to our earlier version of the model:

$$\begin{aligned} y_i &= \alpha_0 + \alpha_1 D_i + \sigma_{u\epsilon} u_i + v_i \\ D_i^* &= z_i\theta + u_i, \end{aligned}$$

Note that, conditioned on θ and D^*, the errors u are "known" and thus we can treat u as a typical regressor in the first equation when sampling from the posterior conditional for $\sigma_{u\epsilon}$:

-

where

-

Finally,

-

The Gibbs sampler proceeds by cycling through all of these conditionals, and it is easy to simulate draws from each of these.

Generalized Tobit Models

- There have been many generalizations of the Tobit model, and these generalizations are closely linked to the model just presented.
- In these generalized Tobit models, there is often the problem of incidental truncation: the value of an outcome variable y is only observed in magnitude depending on the sign of some other variable, say z (and not y itself).
- In this question, we take up Bayesian estimation of the most general Type 5 tobit model (using Amemiya's enumeration system), often referred to as a model of potential outcomes.
- Inference in the remaining generalized models follows similarly, and if you understand these two models, you should be able to work your way through any of these.

The model we consider is given as follows:

-

In the above, y_1 and y_0 are continuous *outcome* variables, with y_1 denoting the treated outcome and y_0 denoting the untreated outcome.

The variable D^* is, again, a latent variable which generates an *observed binary treatment decision* D according to:

-

One interesting feature of this model is that only one outcome is observed for each observation in the sample. That is, if we let y denote the observed vector of outcomes in the data, we can write

-

In other words, if an individual takes the treatment ($D_i = 1$), then we only observe his / her treated outcome y_{1i}, and conversely, if $D_i = 0$, we only observe y_{0i}.

We make the following joint Normality assumption:

$$U_i \overset{iid}{\sim} N(0, \Sigma),$$

where

$$U_i = [U_{Di}\ U_{1i}\ U_{0i}]' \quad \text{and} \quad \Sigma \equiv \begin{bmatrix} 1 & \rho_{1D}\sigma_1 & \rho_{0D}\sigma_0 \\ \rho_{1D}\sigma_1 & \sigma_1^2 & \rho_{10}\sigma_1\sigma_0 \\ \rho_{0D}\sigma_0 & \rho_{10}\sigma_1\sigma_0 & \sigma_0^2 \end{bmatrix}.$$

Finally, the following priors are employed:

$$\beta \equiv \begin{bmatrix} \theta \\ \beta_1 \\ \beta_0 \end{bmatrix} \sim N(\mu_\beta, V_\beta)$$

and

$$\Sigma^{-1} \sim W([\rho R]^{-1}, \rho) I(\sigma^2_{D^*} = 1)$$

where the indicator function is added to the standard Wishart prior so that the variance parameter in the latent data D^* equation is normalized to unity.

We seek to show how the Gibbs sampler can be used to fit this model.

First, it is useful to describe how MLE-based frequentist inference might proceed.

When $D_i = 1$, we observe y_{1i} and the event that $D_i = 1$. Conversely, when $D_i = 0$, we observe y_{0i} and the event that $D_i = 0$. We thus obtain the likelihood function:

$$\begin{aligned} L(\beta, \Sigma) &= \prod_{\{i:D_i=1\}} p(y_{1i}, D_i^* > 0) \prod_{\{i:D_i=0\}} p(y_{0i}, D_i^* \leq 0) \\ &= \prod_{\{i:D_i=1\}} \int_0^\infty p(y_{1i}, D_i^*) dD_i^* \prod_{\{i:D_i=0\}} \int_{-\infty}^0 p(y_{0i}, D_i^*) dD_i^* \\ &= \prod_{\{i:D_i=1\}} \int_0^\infty p(D_i^* | y_{1i}) p(y_{1i}) dD_i^* \prod_{\{i:D_i=0\}} \int_{-\infty}^0 p(D_i^* | y_{0i}) p(y_{0i}) dD_i^*. \end{aligned}$$

The conditional and marginal densities in the above expression can be determined from the assumed joint Normality of the error vector. Performing the required calculations we obtain

$$
\begin{aligned}
L(\beta, \Sigma) &= \prod_{\{i:D_i=1\}} \Phi\left(\frac{\tilde{U}_{Di} + \rho_{1D}\tilde{U}_{1i}}{(1-\rho_{1D}^2)^{-1/2}}\right)\frac{1}{\sigma_1}\phi(\tilde{U}_{1i}) \\
&\quad \prod_{\{i:D_i=0\}} \left[1 - \Phi\left(\frac{\tilde{U}_{Di} + \rho_{0D}\tilde{U}_{0i}}{(1-\rho_{0D}^2)^{-1/2}}\right)\right]\frac{1}{\sigma_0}\phi(\tilde{U}_{0i}),
\end{aligned}
$$

where

$$
\begin{aligned}
\tilde{U}_{Di} &= z_i\theta \\
\tilde{U}_{1i} &= (y_{1i} - x_i\beta_1)/\sigma_1 \\
\tilde{U}_{0i} &= (y_{0i} - x_i\beta_0)/\sigma_0,
\end{aligned}
$$

and $\phi(\cdot)$ denotes the standard Normal density.

To implement the Gibbs sampler, we will work with the augmented joint posterior $p(y^{Miss}, D^*, \beta, \Sigma^{-1}|y, D)$.

As for the first of these,

$$y_i^{Miss}|\Gamma_{-y_i^{Miss}}, y, D \overset{ind}{\sim} N((1-D_i)\mu_{1i} + (D_i)\mu_{0i}, (1-D_i)\omega_{1i} + (D_i)\omega_{0i})$$

where

$$\begin{aligned}
\mu_{1i} &= x_i\beta_1 + (D_i^* - z_i\theta)\left[\frac{\sigma_0^2\sigma_{1D} - \sigma_{10}\sigma_{0D}}{\sigma_0^2 - \sigma_{0D}^2}\right] + (y_i - x_i\beta_0)\left[\frac{\sigma_{10} - \sigma_{0D}\sigma_{1D}}{\sigma_0^2 - \sigma_{0D}^2}\right] \\
\mu_{0i} &= x_i\beta_0 + (D_i^* - z_i\theta)\left[\frac{\sigma_1^2\sigma_{0D} - \sigma_{10}\sigma_{1D}}{\sigma_1^2 - \sigma_{1D}^2}\right] + (y_i - x_i\beta_1)\left[\frac{\sigma_{10} - \sigma_{0D}\sigma_{1D}}{\sigma_1^2 - \sigma_{1D}^2}\right] \\
\omega_{1i} &= \sigma_1^2 - \frac{\sigma_{1D}^2\sigma_0^2 - 2\sigma_{10}\sigma_{0D}\sigma_{1D} + \sigma_{10}^2}{\sigma_0^2 - \sigma_{0D}^2} \\
\omega_{0i} &= \sigma_0^2 - \frac{\sigma_{0D}^2\sigma_1^2 - 2\sigma_{10}\sigma_{0D}\sigma_{1D} + \sigma_{10}^2}{\sigma_1^2 - \sigma_{1D}^2},
\end{aligned}$$

Note that this construction automatically samples from the posterior conditional for y_1 if $D = 0$, and samples from the posterior conditional for y_0 if $D = 1$.

As for the latent data D_i^*, it is also drawn from its conditional Normal, though it is truncated by the observed value of D_i:

$$D_i^* | \Gamma_{-D_i^*}, y, D \overset{ind}{\sim} \begin{cases} TN_{(0,\infty)}(\mu_{Di}, \omega_{Di}) & \text{if } D_i = 1 \\ TN_{(-\infty,0]}(\mu_{Di}, \omega_{Di}) & \text{if } D_i = 0 \end{cases}, \quad i = 1, 2, \cdots, n,$$

where

$$\begin{aligned} \mu_{Di} &= z_i\theta + \left(D_i y_i + (1 - D_i) y_i^{Miss} - x_i\beta_1\right) \left[\frac{\sigma_0^2\sigma_{1D} - \sigma_{10}\sigma_{0D}}{\sigma_1^2\sigma_0^2 - \sigma_{10}^2}\right] + \\ &\quad \left(D_i y_i^{Miss} + (1 - D_i) y_i - x_i\beta_0\right) \left[\frac{\sigma_1^2\sigma_{0D} - \sigma_{10}\sigma_{1D}}{\sigma_1^2\sigma_0^2 - \sigma_{10}^2}\right], \end{aligned}$$

$$\omega_{Di} = 1 - \frac{\sigma_{1D}^2\sigma_0^2 - 2\sigma_{10}\sigma_{0D}\sigma_{1D} + \sigma_1^2\sigma_{0D}^2}{\sigma_1^2\sigma_0^2 - \sigma_{10}^2},$$

Given these drawn quantities, we then compute the *complete data* vector

$$r_i^* = \begin{bmatrix} D_i^* \\ D_i y_i + (1 - D_i) y_i^{Miss} \\ D_i y_i^{Miss} + (1 - D_i) y_i \end{bmatrix}.$$

and use this in the simulation steps associated with the remaining posterior conditionals.

As for the regression parameters β, it, again, follows similarly to the SUR model results:

$$\beta|\Gamma_{-\beta}, y, D \sim N(\mu_\beta, \omega_\beta),$$

where

$$\begin{aligned} \mu_\beta &= [W'(\Sigma^{-1} \otimes I_n)W + V_\beta^{-1}]^{-1}[W'(\Sigma^{-1} \otimes I_n)\overline{y} + V_\beta^{-1}\mu_\beta] \\ \omega_\beta &= [W'(\Sigma^{-1} \otimes I_n)W + \underline{V}_\beta^{-1}]^{-1}, \end{aligned}$$

and

$$W_{3n\times k} \equiv \begin{bmatrix} Z & 0 & 0 \\ 0 & X & 0 \\ 0 & 0 & X \end{bmatrix}, \quad \text{and} \quad \overline{y}_{3n\times 1} \equiv \begin{bmatrix} D^* \\ Dy + (1-D)y^{Miss} \\ Dy^{Miss} + (1-D)y \end{bmatrix}.$$

Finally, for the inverse covariance matrix Σ, note that when our indicator function in our prior is one, the derivations are identical to those of the standard Wishart posterior. Thus, we obtain:

$$\Sigma^{-1}|\Gamma_{-\Sigma}, y, D \sim W\left(\left[\sum_{i=1}^{n}(r_i^* - \tilde{W}_i\beta)(r_i^* - \tilde{W}_i\beta)' + \rho R\right]^{-1}, n+\rho\right) I(\sigma_{D^*}^2 = 1),$$

where

$$\tilde{W}_i \equiv \begin{bmatrix} z_i & 0 & 0 \\ 0 & x_i & 0 \\ 0 & 0 & x_i \end{bmatrix}.$$

This can not be sampled by drawing from a Wishart. However, Nobile (2000 *Journal of Econometrics*) provides an algorithm for drawing from a Wishart, given a restriction on a diagonal element. Such an algorithm can be employed to generate draws from this restricted Wishart conditional.

Finally, note that there are other interesting features of this potential outcomes model, including the issue of the non-identified cross-regime correlation parameter ρ_{10}.