Contents

10

Bayesian treatment effects without joint distribution of outcomes

This chapter continues discussion of Bayesian strategies for identifying treatment effects but now without the joint distribution of outcomes (Chib [2007]). Bayesian data augmentation again replaces the classical control or replacement function.

10.1 Treatment effects and counterfactuals

Suppose we observe treatment or no treatment and the associated outcome, $Y = DY_1 + (1 - D) Y_0$, where

$$\begin{array}{rcl} Y_1 & = & \beta_1 + V_1 \\ Y_0 & = & \beta_0 + V_0 \end{array}$$

and a representative sample is

Y	D	Y_1	Y_0	V_1	V_0
15	1	15	10	0	$-3\frac{1}{3}$
10	0	15	10	0	$-3\frac{1}{3}$
20	0	20	20	5	$6\frac{2}{3}$
20	0	20	20	5	$6\frac{2}{3}$
10	1	10	10	-5	$-3\frac{1}{3}$
10	0	10	10	-5	$-3\frac{1}{3}$

Further, we have the following binary instrument at our disposal z where the representative values $Z = \left[\begin{array}{cc} \iota & z \end{array} \right]$ are

ι	z
1	1
1	0
1	1
1	0
1	1
1	0

and we perceive latent utility, D^*, to be related to choice via the instruments.

$$D^* = Z\theta + V_D$$

and observed choice is

$$D = \left\{ \begin{array}{cc} 1 & D^* > 0 \\ 0 & \text{otherwise} \end{array} \right.$$

This is the exact setup we discussed earlier in IV example 4 of chapter 3.

10.2 Posterior distribution

Define the augmented data along with the observed outcome $j = 0, 1$ as

$$Y_{ji}^* = \left[\begin{array}{c} Y_{ji} \\ D_i^* \end{array} \right]$$

Also, let

$$H_i = \left[\begin{array}{ccc} X_i \left(1 - D_i\right) & X_i D_i & 0 \\ 0 & 0 & Z_i \end{array} \right]$$

and

$$\beta = \left[\begin{array}{c} \beta_0 \\ \beta_1 \\ \theta \end{array} \right]$$

where X is a matrix of outcome regressors, in the current example it is simply ι, a vector of ones, as there are no outcome covariates. Hence, a compact model is

$$Y_{ji}^* = H_i \beta + \varepsilon_{ji}$$

where $\varepsilon_{ji} = \left[\begin{array}{c} V_{ji} \\ V_{Di} \end{array} \right]$ and $\Sigma_{ji} = Var\left[\varepsilon_{ji}\right] = \frac{1}{\lambda_i}\Omega_j = \frac{1}{\lambda_i}\left[\begin{array}{cc} \eta_j^2 & \omega_j \\ \omega_j & 1 \end{array} \right]$ for $j = 0, 1$.

10.2.1 Likelihood function

As usual the posterior distribution is proportional to the likelihood function times the prior distribution. The likelihood function is

$$Y_{ji}^{*} \sim N\left(H_i\beta, \Sigma_{ji}\right)$$

Or,

$$\ell\left(\beta, \Sigma_{ji} \mid Y_{ji}^{*}, D_i, X_i, Z_i\right) \propto \left|\Sigma_{ji}\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(Y_{ji}^{*} - H_i\beta\right)^{T} \Sigma_{ji}^{-1}\left(Y_{ji}^{*} - H_i\beta\right)\right]$$

10.2.2 Prior distribution

Frequently, relatively diffuse priors are chosen such that the data dominates the posterior distribution. Chib's prior distribution for β is $p(\beta) \sim N\left(\beta_0, V_\beta\right)$ where $\beta_0 = 0, V_\beta = 20I$. Let $\sigma_j^2 = \eta_j^2 - \omega_j^2 > 0$, then η_j^2 is defined in terms of σ_j^2 and ω_j^2. To simplify posterior draws for Σ_{ji}, treat λ_i and σ_j^2 independently and ω_j conditional on σ_j^2. Accordingly, the prior for λ_i is $p(\lambda_i) \sim \mathit{Gamma}\left(\frac{\nu}{2}, \frac{2}{\nu}\right)$ where $\nu = 15$, the prior for η_j^2 is inverse gamma, $p\left(\sigma_j^2\right) \sim \mathit{Inverse{-}gamma}\left(\frac{\nu_{j0}}{2}, \frac{2}{d_{j0}}\right)$ where $\nu_{j0} = 4.22$ and $d_{j0} = 2.22$, and the prior for ω_j is conditional normal, $p\left(\omega_j \mid \eta_j^2\right) \sim N\left(m_{j0}, \sigma_j^2 M_{j0}\right)$ where $m_{j0} = 0$ and $M_{j0} = 10$. Hence, the joint conjugate prior is normal-inverse gamma-normal-gamma.

$$\begin{aligned}
p\left(\beta, \Sigma_{ji}\right) &= p(\beta)\, p\left(\sigma_j^2\right) p\left(\omega_j \mid \sigma_j^2\right) p\left(\lambda_i\right) \\
&\propto \left|V_\beta\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\beta - \beta_0\right)^{T} V_\beta^{-1}\left(\beta - \beta_0\right)\right] \\
&\quad \times \left(\sigma_j^2\right)^{-\frac{\nu_{j0}}{2}-1} \exp\left[-\frac{d_{j0}}{2\sigma_j^2}\right] \\
&\quad \times \left(\sigma_j^2 M_{j0}\right)^{-\frac{1}{2}} \exp\left[-\frac{\left(\omega_j - m_{j0}\right)^2}{2\sigma_j^2 M_{j0}}\right] \\
&\quad \times \left(\lambda_i\right)^{\frac{\nu}{2}-1} \exp\left[-\frac{\nu\lambda_i}{2}\right]
\end{aligned}$$

10.2.3 Posterior distribution

Now, the posterior distribution (or posterior kernel) is

$$p\left(\beta, \Sigma_{ji}, D_i^{*} \mid Y_i, D_i, X_i, Z_i\right) \propto p\left(\beta, \Sigma_{ji}\right) \ell\left(\beta, \Sigma_{ji} \mid Y_{ji}^{*}, D_i, X_i, Z_i\right)$$

Unlike the strategy discussed in the previous chapter, the posterior distribution does not involve the counterfactual outcomes.

10.3 Gibbs sampler for treatment effects

As is frequently the case, it's much easier to simulate from the recognizable conditional posterior distributions via a Gibbs sampler than simulate from the unrecognizable joint posterior distribution. We follow Chib [2007] and generate conditional posterior draws in three blocks.

10.3.1 Full conditional posterior distributions

Let Γ_{-x} denote all parameters other than x. The *McMC* algorithm cycles through three blocks as follows.

First block

Sample $\psi_j = \left(\sigma_j^2, \omega_j\right)$, $j = (0,1)$, conditioned on $\left(Y_j^*, D, X, Z, \beta, \lambda\right)$ by composition. That is, draw marginalized σ_j^2 on integrating out ω_j.

$$p\left(\sigma_j^2 \mid \Gamma_{-\psi_j}, Y_j^*, D, X, Z\right) \propto \int \frac{p\left(\beta, \Sigma_j, D^* \mid Y, D, X, Z\right)}{p\left(\beta\right) p\left(\lambda\right) p\left(Y_j^* \mid D, X, Z, \beta, \lambda\right)} d\omega_j$$

$$\begin{aligned}
&\propto \frac{p\left(\beta\right) p\left(\lambda\right) \int p\left(\sigma_j^2\right) p\left(\omega_j \mid \sigma_j^2\right) \ell\left(\beta, \Sigma_j \mid Y_j^*, D, X, Z\right) d\omega_j}{p\left(\beta\right) p\left(\lambda\right) p\left(Y_j^* \mid D, X, Z, \beta, \lambda\right)} \\
&\propto \frac{p\left(\sigma_j^2\right) p\left(Y_j^* \mid D, X, Z, \beta, \lambda, \sigma_j^2\right)}{p\left(Y_j^* \mid D, X, Z, \beta, \lambda\right)} \\
&\propto \frac{p\left(\sigma_j^2, Y_j^* \mid D, X, Z, \beta, \lambda\right)}{p\left(Y_j^* \mid D, X, Z, \beta, \lambda\right)} \\
&\propto p\left(\sigma_j^2 \mid Y_j^*, D, X, Z, \beta, \lambda\right)
\end{aligned}$$

Working through the details, we have

$$\begin{aligned}
& \frac{p\left(\beta\right)p\left(\lambda\right)\int p\left(\sigma_j^2\right)p\left(\omega_j \mid \sigma_j^2\right)\ell\left(\beta,\Sigma_j \mid Y_j^*,D,X,Z\right)d\omega_j}{p\left(\beta\right)p\left(\lambda\right)p\left(Y_j^* \mid D,X,Z,\beta,\lambda\right)} \\
\propto \quad & \frac{\int p\left(\sigma_j^2\right)p\left(\omega_j \mid \sigma_j^2\right)\ell\left(\beta,\Sigma_j \mid Y_j^*,D,X,Z\right)d\omega_j}{p\left(Y_j^* \mid D,X,Z,\beta,\lambda\right)} \\
\propto \quad & \int p\left(\sigma_j^2\right)p\left(\omega_j \mid \sigma_j^2\right)\ell\left(\beta,\Sigma_j \mid Y_j^*,D,X,Z\right)d\omega_j \\
\propto \quad & \int \left(\sigma_j^2\right)^{-\frac{\nu_{j0}}{2}-1}\exp\left[-\frac{d_{j0}}{2\sigma_j^2}\right]\left(\sigma_j^2 M_{j0}\right)^{-\frac{1}{2}}\exp\left[-\frac{\left(\omega_j-m_{j0}\right)^2}{2\sigma_j^2 M_{j0}}\right] \\
& \times\prod_{i=1}^{n_j}\lambda_i\left(\sigma_j^2\right)^{-\frac{n_j}{2}} \\
& \times\exp\left[-\frac{1}{2,\sigma_j^2}\sum_{i=1}^{n_j}\lambda_i\begin{bmatrix} V_{ji} & V_{Di}\end{bmatrix}\begin{bmatrix} 1 & -\omega_j \\ -\omega_j & \sigma_j^2+\omega^2\end{bmatrix}\begin{bmatrix} V_{ji} \\ V_{Di}\end{bmatrix}\right]d\omega_j
\end{aligned}$$

On making some judicious substitutions into the likelihood function, we have

$$\begin{aligned}
\propto \quad & \int \left(\sigma_j^2\right)^{-\frac{\nu_{j0}}{2}-1}\exp\left[-\frac{d_{j0}}{2\sigma_j^2}\right]\left(\sigma_j^2 M_{j0}\right)^{-\frac{1}{2}}\exp\left[-\frac{\left(\omega_j-m_{j0}\right)^2}{2\sigma_j^2 M_{j0}}\right] \\
& \times\prod_{i=1}^{n_j}\lambda_i\left(\sigma_j^2\right)^{-\frac{n_j}{2}} \\
& \times\exp\left[-\frac{1}{2\sigma_j^2}\sum_{i=1}^{n_j}\lambda_i\begin{bmatrix} V_{ji} & V_{Di}\end{bmatrix}\begin{bmatrix} 1 & -\omega_j \\ -\omega_j & \sigma_j^2+\omega^2\end{bmatrix}\begin{bmatrix} V_{ji} \\ V_{Di}\end{bmatrix}\right]d\omega_j
\end{aligned}$$

Collecting terms gives

$$\begin{aligned}
\propto \quad & \prod_{i=1}^{n_j}\lambda_i M_{j0}^{-\frac{1}{2}} \\
& \times\left(\sigma_j^2\right)^{-\frac{\nu_{j0}+3+n_j}{2}}\exp\left[-\frac{d_{j0}}{2\sigma_j^2}\right]\int\exp\left[-\frac{1}{2\sigma_j^2}\frac{\left(\omega_j-m_{j0}\right)^2}{M_{j0}}\right] \\
& \times\exp\left[-\frac{1}{2\sigma_j^2}\sum_{i=1}^{n_j}\lambda_i\begin{bmatrix} V_{ji} & V_{Di}\end{bmatrix}\begin{bmatrix} 1 & -\omega_j \\ -\omega_j & \sigma_j^2+\omega^2\end{bmatrix}\begin{bmatrix} V_{ji} \\ V_{Di}\end{bmatrix}\right]d\omega_j
\end{aligned}$$

Integrating out ω_j and dropping some normalizing constants leaves

$$\propto\left(\sigma_j^2\right)^{-\frac{\nu_{j0}+3+n_j}{2}}\exp\left[-\frac{d_{j0}+d_j+\sum_{i=1}^{n_j}\lambda_i u_i^2}{2\sigma_j^2}\right]$$

where n_j is the sample size for subsample j,

$$d_j = \left(V_j - m_{j0}V_{Dj}\right)^T \left(\Lambda_j^{-1} + u_j M_{j0} u_j^T\right)^{-1} \left(V_j - m_{j0}V_{Dj}\right),$$

Λ_j is a diagonal matrix with λ_i along the diagonal, $V_j = Y_j - X_j\beta_j$, and $V_{Dj} = D_j^* - Z_j\theta$. Since we're conditioning on λ_i and $V_{Di} = D_i^* - Z_i\theta$, $\sum_{i=1}^{n_j}\lambda_i V_{Di}^2$ is a normalizing constant and can be dropped from the expression for the kernel. Hence, the kernel is recognizable as inverse gamma.

$$\begin{aligned}\sigma_j^2 \quad &\mid \quad \Gamma_{-\psi_j}, Y_j^*, D, X, Z \\ &\sim \quad \textit{inverse gamma}\left(\sigma_j^2 \mid \Gamma_{-\psi_j}, Y_j^*, D, X, Z; \frac{\nu_{j0}+n_j}{2}, \frac{2}{d_{j0}+d_j}\right)\end{aligned}$$

Now, draw ω_j conditional on σ_j^2.

$$\begin{aligned}p\left(\omega_j \mid \Gamma_{-\omega_j}, Y_j^*, D, X, Z\right) &\propto \frac{p\left(\beta, \Sigma_j \mid Y_j^*, D, X, Z\right)}{p(\beta)\, p\left(\sigma_j^2\right) p(\lambda)\, p\left(Y_j^* \mid D, X, Z, \beta, \lambda, \sigma_j^2\right)} \\ &\propto \frac{p(\beta)\, p\left(\omega_j \mid \sigma_j^2\right) p\left(\sigma_j^2\right) p(\lambda)\, \ell\left(\beta, \Sigma_j \mid Y_j^*, D, X, Z\right)}{p(\beta)\, p\left(\sigma_j^2\right) p(\lambda)\, p\left(Y_j^* \mid D, X, Z, \beta, \lambda, \sigma_j^2\right)} \\ &\propto \frac{p\left(\omega_j \mid \sigma_j^2\right) p\left(Y_j^* \mid, D, X, Z, \beta, \omega_j, \lambda, \sigma_j^2\right)}{p\left(Y_j^* \mid D, X, Z, \beta, \lambda, \sigma_j^2\right)} \\ &\propto \frac{p\left(\omega_j, Y_j^* \mid, D, X, Z, \beta, \lambda, \sigma_j^2\right)}{p\left(Y_j^* \mid D, X, Z, \beta, \lambda, \sigma_j^2\right)} \\ &\propto p\left(\omega_j \mid Y_j^*, D, X, Z, \beta, \lambda, \sigma_j^2\right)\end{aligned}$$

Working through the details, we have

$$\begin{aligned}&\frac{p(\beta)\, p(\lambda)\, p\left(\sigma_j^2\right) p\left(\omega_j \mid \sigma_j^2\right) \ell\left(\beta, \Sigma_j \mid Y_j^*, D, X, Z\right)}{p(\beta)\, p(\lambda)\, p\left(\sigma_j^2\right) p\left(Y_j^* \mid D, X, Z, \beta, \lambda, \sigma_j^2\right)} \\ \propto\quad &\frac{p\left(\omega_j \mid \sigma_j^2\right) \ell\left(\beta, \Sigma_j \mid Y_j^*, D, X, Z\right)}{p\left(Y_j^* \mid D, X, Z, \beta, \lambda, \sigma_j^2\right)} \\ \propto\quad &\int p\left(\sigma_j^2\right) p\left(\omega_j \mid \sigma_j^2\right) \ell\left(\beta, \Sigma_j \mid Y_j^*, D, X, Z\right) d\omega_j \\ \propto\quad &\exp\left[-\frac{\left(\omega_j - m_{j0}\right)^2}{2\sigma_j^2 M_{j0}}\right] \\ &\times \exp\left[-\frac{1}{2, \sigma_j^2}\sum_{i=1}^{n_j}\lambda_i \begin{bmatrix} V_{ji} & V_{Di}\end{bmatrix}\begin{bmatrix} 1 & -\omega_j \\ -\omega_j & \sigma_j^2 + \omega^2\end{bmatrix}\begin{bmatrix} V_{ji} \\ V_{Di}\end{bmatrix}\right]\end{aligned}$$

Collecting terms we have

$$\propto \exp\Bigg[-\frac{1}{2\sigma_j^2}\Bigg(\frac{\left(\omega_j - m_{j0}\right)^2}{2\sigma_j^2 M_{j0}} + \sum_{i=1}^{n_j}\lambda_i \left\{\begin{array}{c} V_{ji}^2 - \omega_j V_{Di} V_{ji} \\ + \left(\sigma_j^2 + \omega_j^2\right) V_{Di}^2 - \omega_j V_{ji} V_{Di}\end{array}\right\}\Bigg)\Bigg)\Bigg]$$

Hence,

$$\omega_j \mid \Gamma_{-\omega_j}, Y_j^*, D, X, Z \sim N\left(\omega_j \mid \Gamma_{-\omega_j}, Y_j^*, D, X, Z; b_j, \sigma_j^2 B_j\right)$$

where $b_j = B_j\left(M_{j0}^{-1} m_{j0} + V_{Dj}^T \Lambda V_j\right)$ and $B_j = \left(M_{j0}^{-1} + V_{Dj}^T \Lambda V_{Dj}\right)^{-1}$.[1]

This strategy warrants comment as b_j and B_j are based on truncated data (for subsample j) and the truncated covariance U_{Dj} and variance U_{DD} differ from their full support counterparts, W_{Dj} and W_{DD}, while ω_j is the full support covariance between V_D and V_j. Even though the distribution for V_D is truncated and the covariance and variance of truncated random variables differs from their counterparts for the full support population, the ratio of the truncated covariance $U_{DD} W_{Dj} W_{DD}^{-1}$ to truncated variance U_{DD} equals the ratio of the population covariance V_{Dj} to population variance V_{DD} (see Johnson and Kotz [1972, pp. 204-205].

$$\frac{U_{DD} W_{Dj} W_{DD}^{-1}}{U_{DD}} = \frac{W_{Dj}}{W_{DD}}$$

Further, since W_{DD} is normalized to unity we have

$$\frac{W_{Dj}}{W_{DD}} = \frac{W_{Dj}}{1} = \omega_j$$

Therefore, this important indicator of the extent of endogeneity, ω_j, is identified by our procedure.

Second block

With prior distribution $p(\beta) \sim N(\beta_0, V_\beta)$, the posterior distribution for the parameters conditional on $Y^*, D, X, Z, \psi_0, \psi_1, \Lambda$ is

$$\begin{aligned} p(\beta \mid \Gamma_{-\beta}, Y^*, D, X, Z) &\propto \frac{p(\beta, \Sigma_j, D^* \mid Y, D, X, Z)}{p(\Sigma_j)\, p(Y^* \mid D, X, Z, \Sigma_j)} \\ &\propto \frac{p(\beta)\, \ell(\beta, \Sigma_j \mid Y^*, D, X, Z)}{p(Y^* \mid D, X, Z, \Sigma_j)} \\ &\propto \frac{p(\beta, Y^* \mid D, X, Z, \Sigma_j)}{p(Y^* \mid D, X, Z, \Sigma_j)} \\ &\propto p(\beta \mid Y^*, D, X, Z, \Sigma_j) \end{aligned}$$

[1] The difference between expanding the expression in brackets above and the normal kernel indicated here,

$$\frac{1}{2\sigma_j^2}\left\{\begin{array}{c} V_j^T \Lambda V_j + \sigma_j^2 V_D^T \Lambda V_D \\ +B M_{j0}^{-1}\left[-2 m_{j0} V_j^T \Lambda V_D + m_{j0}^2 V_D^T \Lambda V_D - M_{j0}\left(V_j^T \Lambda V_D\right)^2\right] \end{array}\right\},$$

is entirely constants absorbed through normalization.

Working through the details yields a standard Bayesian (GLS or SUR) regression result. One can think of the prior as providing one subsample result and the data another subsample result such that the posterior is a precision-weighted average of the two subsamples.

$$p(\beta)\,\ell(\beta,\Sigma_j \mid Y^*,D,X,Z) \propto p(\beta,Y^* \mid D,X,Z,\Sigma_j)$$

$$\begin{aligned}\propto \quad & \exp\left[-\frac{1}{2}(\beta-\beta_0)^T V_\beta^{-1}(\beta-\beta_0)\right]\\ & \times\exp\left[-\frac{1}{2}(Y^*-H_0\beta)^T \Lambda_0\Omega_0^{-1}(Y^*-H_0\beta)\right]\\ & \times\exp\left[-\frac{1}{2}(Y^*-H_1\beta)^T \Lambda_1\Omega_1^{-1}(Y^*-H_1\beta)\right]\end{aligned}$$

Hence, we can draw β from

$$\beta \mid \Gamma_{-\beta},Y^*,D,X,Z \sim N\left(\beta \mid Y^*,D,X,Z,\Sigma_j;\widehat{\beta},\Sigma_\beta\right)$$

where

$$\widehat{\beta} = \Sigma_\beta\left[V_\beta^{-1}\beta_0 + A_0 + A_1\right]$$

$$A_j = \sum_{i\in n_j}\lambda_i H_{ji}^T\Omega_j^{-1}Y_i^*$$

$$\Sigma_\beta = \left[V_\beta^{-1} + \sum_{i\in n_0}\lambda_i H_{0i}^T\Omega_0^{-1}H_{0i} + \sum_{i\in n_1}\lambda_i H_{1i}^T\Omega_1^{-1}H_{1i}\right]^{-1}$$

Again, any difference in the two expressions does not involve β and is absorbed via normalization.

Third block

Sample (D_{ji}^*,λ_i) conditioned on $Y_{ji},D_i,X_i,Z_i,\beta_j,\psi_j$ by composition. That is, marginalize Y_{ji}^* by integrating out λ_i. Then, draw λ_i conditional on D_i^*. $Y_{ji}^* = (Y_{ji},D_i^*)$ marginalized is bivariate Student t(ν) and the conditional distribution of $(D_i^* \mid Y_{ji})$ is univariate, noncentral scaled Student t$(\mu_{ji},h_{ji}\phi_j^2;\nu+1)$ where

$$\mu_{ji} = Z_i\theta + \frac{\omega_j}{\eta_j^2}\left(Y_{ji}-X_i\beta_j\right)$$

$$h_{ji} = \frac{\nu}{\nu+1}\left[1+\frac{\left(Y_{ji}-X_i\beta_j\right)^2}{\nu\left(\sigma_j^2+\omega_j^2\right)}\right]$$

and

$$\phi_j^2 = 1-\frac{\omega_j^2}{\eta_j^2}$$

In other words, the conditional posterior for D_i^* is

$$p\left(D_i^* \mid \Gamma_{-D^*}, Y_{ji}, D_i, X_i, Z_i\right) \propto \int \frac{p\left(\beta_j, \Sigma_j, D_i^* \mid Y_{ji}, D_i, X_i, Z_i\right)}{p\left(\psi_j\right) p\left(\beta_j\right) p\left(Y_{ji} \mid D_i, X_i, Z_i, \beta_j, \psi_j\right)} d\lambda_i$$

$$\begin{aligned}
&\propto \int \frac{p\left(\psi_j\right) p\left(\beta_j\right) p\left(\lambda_i\right) \ell\left(\beta_j, \Sigma_j \mid Y_{ji}^*, D_i, X_i, Z_i\right)}{p\left(\psi_j\right) p\left(\beta_j\right) p\left(Y_{ji} \mid D_i, X_i, Z_i, \beta_j, \psi_j\right)} d\lambda_i \\
&\propto \frac{p\left(Y_{ji}, D_i^* \mid D_i, X_i, Z_i, \beta_j, \psi_j\right)}{p\left(Y_{ji} \mid D_i, X_i, Z_i, \beta_j, \psi_j\right)} \\
&\propto p\left(D_i^* \mid Y_{ji}, D_i, X_i, Z_i, \beta_j, \psi_j\right)
\end{aligned}$$

We work through the details carefully as mistakes are easily made in this case (at least for me).

$$\begin{aligned}
&\frac{p\left(Y_{ji}, D_i^* \mid D_i, X_i, Z_i, \beta_j, \psi_j\right)}{p\left(Y_{ji} \mid D_i, X_i, Z_i, \beta_j, \psi_j\right)} \\
\propto\ &\int_0^\infty \frac{p\left(Y_{ji}, D_i^*, \lambda_i \mid D_i, X_i, Z_i, \beta_j, \psi_j\right)}{\int_{-\infty}^\infty \left[\int_0^\infty p\left(Y_{ji}, \lambda_i, D_i^* \mid D_i, X_i, Z_i, \beta_j, \psi_j\right) d\lambda_i\right] dD_i^*} d\lambda_i \\
\propto\ &\int \frac{\lambda_i^{\frac{\nu}{2}} \exp\left[-\frac{\lambda_i}{2}\left\{\nu + \frac{\left(Y_{ji}-X_i\beta_j\right)^2 - 2\omega\left(Y_{ji}-X_i\beta_j\right)\left(D_i^*-Z_i\theta\right) + \left(\sigma^2+\omega^2\right)\left(D_i^*-Z_i\theta\right)^2}{\sigma_j^2}\right\}\right]}{\int \lambda_i^{\frac{\nu-1}{2}} \exp\left[-\frac{\lambda_i}{2}\left\{\nu + \frac{\left(Y_{ji}-X_i\beta_j\right)^2}{\sigma_j^2+\omega_j^2}\right\}\right] d\lambda_i} d\lambda_i \\
\propto\ &\int \frac{\lambda_i^{\frac{\nu}{2}} \exp\left[-\frac{\lambda_i}{2}\left\{\nu + \frac{\left(Y_{ji}-X_i\beta_j\right)^2 - 2\omega_j\left(Y_{ji}-X_i\beta_j\right)\left(D_i^*-Z_i\theta\right) + \left(\sigma_j^2+\omega_j^2\right)\left(D_i^*-Z_i\theta\right)^2}{\sigma_j^2}\right\}\right]}{\left(\nu + \frac{\left(Y_{ji}-X_i\beta_j\right)^2}{\sigma_j^2+\omega_j^2}\right)^{-\frac{\nu+1}{2}}} d\lambda_i \\
\propto\ &\int \lambda_i^{\frac{\nu}{2}} \exp\left[-\frac{\lambda_i}{2}\left\{\begin{array}{c} \nu + \frac{\left(Y_{ji}-X_i\beta_j\right)^2 - 2\omega_j\left(Y_{ji}-X_i\beta_j\right)\left(D_i^*-Z_i\theta\right)}{\sigma_j^2} \\ + \frac{+\left(\sigma_j^2+\omega_j^2\right)\left(D_i^*-Z_i\theta\right)^2}{\sigma_j^2} \end{array}\right\}\right] d\lambda_i \\
\propto\ &\left[\frac{\sigma_j^2\left(\nu + \left(D_i^* - Z_i\theta\right)^2\right) + \left(Y_{ji} - X_i\beta_j - \omega_j\left(D_i^* - Z_i\theta\right)\right)^2}{\sigma_j^2}\right]^{-\frac{v+2}{2}}
\end{aligned}$$

This looks like the kernel for a univariate Student t with $\nu + 1$ degrees of freedom. Some manipulation bears this out and identifies the centrality and scale parameters. Work with the term in brackets

$$
\begin{aligned}
&\frac{\sigma_j^2\left(\nu+\left(D_i^*-Z_i\theta\right)^2\right)+\left(Y_{ji}-X_i\beta_j-\omega_j\left(D_i^*-Z_i\theta\right)\right)^2}{\sigma_j^2} \\
= \quad & \nu+\frac{\frac{\sigma_j^2}{\sigma_j^2+\omega_j^2}\left(D_i^*-Z_i\theta\right)^2+\frac{\left(Y_{ji}-X_i\beta_j-\omega_j\left(D_i^*-Z_i\theta\right)\right)^2}{\sigma_j^2+\omega_j^2}}{\frac{\sigma_j^2}{\sigma_j^2+\omega_j^2}} \\
= \quad & \nu+\frac{\frac{\sigma_j^2\left(D_i^*-Z_i\theta\right)^2+\left(Y_{ji}-X_i\beta_j\right)^2-2\omega_j\left(D_i^*-Z_i\theta\right)\left(Y_{ji}-X_i\beta_j\right)+\omega_j^2\left(D_i^*-Z_i\theta\right)}{\sigma_j^2+\omega_j^2}}{\frac{\sigma_j^2}{\sigma_j^2+\omega_j^2}} \\
= \quad & \nu+\frac{\left(D_i^*-Z_i\theta\right)^2+\frac{-2\omega_j\left(D_i^*-Z_i\theta\right)\left(Y_{ji}-X_i\beta_j\right)+\left(Y_{ji}-X_i\beta_j\right)^2}{\sigma_j^2+\omega_j^2}}{\frac{\sigma_j^2}{\sigma_j^2+\omega_j^2}} \\
= \quad & \nu+\frac{\left(D_i^*-Z_i\theta-\frac{\omega_j}{\sigma_j^2+\omega_j^2}\left(Y_{ji}-X_i\beta_j\right)\right)^2+\frac{\sigma_j^2\left(Y_{ji}-X_i\beta_j\right)^2}{\left(\sigma_j^2+\omega_j^2\right)^2}}{\frac{\sigma^2}{\sigma^2+\omega^2}} \\
= \quad & \nu+\frac{\left(Y_{ji}-X_i\beta_j\right)^2}{\left(\sigma_j^2+\omega_j^2\right)}+\frac{\left(D_i^*-Z_i\theta-\frac{\omega_j}{\sigma_j^2+\omega_j^2}\left(Y_{ji}-X_i\beta_j\right)\right)^2}{\frac{\sigma_j^2}{\sigma_j^2+\omega_j^2}} \\
= \quad & \frac{\nu\left(\sigma_j^2+\omega_j^2\right)+\left(Y_{ji}-X_i\beta_j\right)^2}{\left(\sigma_j^2+\omega_j^2\right)}+\frac{\left(D_i^*-Z_i\theta-\frac{\omega_j}{\sigma_j^2+\omega_j^2}\left(Y_{ji}-X_i\beta_j\right)\right)^2}{\frac{\sigma_j^2}{\sigma_j^2+\omega_j^2}} \\
= \quad & h_j\left(\nu+1\right)+\frac{\left(D_i^*-Z_i\theta-\frac{\omega_j}{\sigma_j^2+\omega_j^2}\left(Y_{ji}-X_i\beta_j\right)\right)^2}{\frac{\sigma^2}{\sigma_j^2+\omega_j^2}}
\end{aligned}
$$

recall

$$
h_{ji}=\frac{\nu}{\nu+1}\left[1+\frac{\left(Y_{ji}-X_i\beta_j\right)^2}{\nu\left(\sigma_j^2+\omega_j^2\right)}\right]
$$

which can be rewritten as

$$
\frac{1}{\nu+1}\left[\frac{\nu\left(\sigma_j^2+\omega_j^2\right)+\left(Y_{ji}-X_i\beta_j\right)^2}{\left(\sigma_j^2+\omega_j^2\right)}\right]
$$

and the term in brackets is identical to our expression above prior to substituting in $h_j\left(\nu+1\right)$. Placing this expression back into the kernel expression, we have

$$\left[h_{ji}\left(\nu+1\right)+\frac{\left(D_{i}^{*}-Z_{i}\theta-\frac{\omega_{j}}{\sigma_{j}^{2}+\omega_{j}^{2}}\left(Y_{ji}-X_{i}\beta_{j}\right)\right)^{2}}{\frac{\sigma_{j}^{2}}{\sigma_{j}^{2}+\omega_{j}^{2}}}\right]^{-\frac{v+2}{2}}$$

$$\propto\quad\left[\left(\nu+1\right)+\frac{\left(D_{i}^{*}-Z_{i}\theta-\frac{\omega}{\sigma^{2}+\omega^{2}}\left(Y_{ji}-X_{i}\beta_{j}\right)\right)^{2}}{h_{ji}\frac{\sigma^{2}}{\sigma^{2}+\omega^{2}}}\right]^{-\frac{v+2}{2}}$$

Hence,

$$D_{i}^{*}\mid\Gamma_{-D^{*}},Y_{ji},D_{i},X_{i},Z_{i}\sim t\left(D_{i}^{*}\mid Y_{ji},D_{i},X_{i},Z_{i},\beta_{j},\psi_{j};\mu_{ji},h_{ji}\phi_{j}^{2},\nu+1\right)$$

Simulation involves truncated Student t draws with D_{i}^{*} stemming from $t\left(\mu_{0i},h_{0i}\phi_{0}^{2},\nu+1\right)I_{(-\infty,0)}$ if $D_{i}=0$ and from $t\left(\mu_{1i},h_{1i}\phi_{1}^{2},\nu+1\right)I_{(0,\infty)}$ if $D_{i}=1$ $(i\leq n)$.

The posterior for λ_{i} conditional on $Y_{ji}^{*},D_{i},X_{i},Z_{i},\beta_{j},\psi_{0},\psi_{1}$ is

$$p\left(\lambda_{i}\mid\Gamma_{-\lambda_{i}},Y_{ji}^{*},D_{i},X_{i},Z_{i}\right)\propto\frac{p\left(\beta_{j},\Sigma_{j},D_{i}^{*}\mid Y_{ji},D_{i},X_{i},Z_{i}\right)}{p\left(\psi_{j}\right)p\left(\beta_{j}\right)p\left(Y_{ji}^{*}\mid D_{i},X_{i},Z_{i},\beta_{j},\psi_{j}\right)}$$

$$\begin{aligned}
&\propto\quad\frac{p\left(\psi_{j}\right)p\left(\beta_{j}\right)p\left(\lambda_{i}\right)\ell\left(\beta_{j},\Sigma_{j}\mid Y_{ji}^{*},D_{i},X_{i},Z_{i}\right)}{p\left(\psi_{j}\right)p\left(\beta_{j}\right)p\left(Y_{ji}^{*}\mid D_{i},X_{i},Z_{i},\beta_{j},\psi_{j}\right)}\\
&\propto\quad\frac{p\left(\lambda_{i},Y_{ji}^{*}\mid D_{i},X_{i},Z_{i},\beta_{j},\psi_{j}\right)}{p\left(Y_{ji}^{*}\mid D_{i},X_{i},Z_{i},\beta_{j},\psi_{j}\right)}\\
&\propto\quad p\left(\lambda_{i}\mid Y_{ji}^{*},D_{i},X_{i},Z_{i},\beta_{j},\psi_{j}\right)\\
&\propto\quad\lambda_{i}^{\frac{\nu}{2}}\exp\left[-\frac{\lambda_{i}\nu}{2}-\frac{\lambda_{i}G_{i}}{2}\right]
\end{aligned}$$

Hence,

$$\lambda_{i}\mid\Gamma_{-\lambda_{i}},Y_{i}^{*},D_{i},X_{i},Z_{i}\sim gamma\left(\lambda_{i}\mid Y_{ji}^{*},D_{i},X_{i},Z_{i},\beta_{j},\psi_{j};\frac{\nu+2}{2},\frac{2}{\nu+G_{ji}}\right)$$

where $G_{ji}=\left(Y_{ji}^{*}-H_{i}\beta_{j}\right)^{T}\Omega_{j}^{-1}\left(Y_{ji}^{*}-H_{i}\beta_{j}\right)$.

The McMC algorithm continues by returning to the first block and repeating the cycle conditional on the last draw until we have a representative, convergent sample of *McMC* simulated draws. As usual, starting values for the Gibbs sampler are varied to test convergence of the posterior distributions (adequate coverage of the sample space). Stationary convergence plots and quickly dampening autocorrelation plots support the notion of representative posterior draws.

10.4 Predictive average treatment effects

To keep things interesting, we're going to define a different set of average treatment effects. That is, we're going to focus on predictive or out-of-sample treatment effects. The idea involves a straightforward generalization of Bayesian updating to include $Y_{j,n+1}$. In principle, we draw $Y_{j,n+1}$, $j = 0, 1$ from the marginal posterior by marginalizing the unknowns

$$\begin{aligned} p\left(Y_{j,n+1} \mid Y, H, D\right) &= \int p\left(Y_{j,n+1} \mid H_{n+1}, \lambda_{n+1}, \Lambda, \beta_j, \psi_j, Y, H, D\right) \\ &\quad \times p\left(H_{n+1}, \lambda_{n+1}, \Lambda, \beta_j, \psi_j \mid Y, H, D\right) d\lambda_{n+1} d\Lambda \\ &\quad d\beta_j d\psi_j dH_{n+1} \end{aligned}$$

where

$$\begin{aligned} & p\left(Y_{j,n+1} \mid H_{n+1}, \lambda_{n+1}, \Lambda, \beta_j, \psi_j, Y, H, D\right) \\ = \; & p\left(Y_{j,n+1} \mid H_{n+1}, \lambda_{n+1}, \Lambda, \beta_j, \psi_j\right) \sim N\left(X_{n+1}\beta_j, \frac{\eta_j^2}{\lambda_{n+1}}\right) \end{aligned}$$

While analytical integration is daunting, it is straightforward to sample from the Markov chain if we approximate the distribution for H_{n+1} by its empirical distribution.

10.4.1 Predicted ATE

For the gth ($g \leq M$) iteration of the *McMC* algorithm, when the chain is defined by $\left(\Lambda^{(g)}, \beta_j^{(g)}, \psi_j^{(g)}\right)$, we draw $Y_{j,n+1}^{(g)}$ from the predictive distribution by adding the following steps to our previously defined *McMC* algorithm. To accommodate inherently unobservable heterogeneity, the following algorithm differs from Chib [2007].

- Draw $X_{n+1}^{(g)}$ by assigning probability $\frac{1}{n}$ to each row of regressors.
- Draw $\lambda_{n+1}^{(g)}$ from gamma$\left(\frac{\nu}{2}, \frac{2}{\nu}\right)$.
- Draw $V_{D,n+1}^{(g)}$ from $N\left(0, \frac{1}{\lambda_{n+1}^{(g)}}\right)$.
- Draw $D_{n+1}^{*(g)} = Z_{n+1}^{(g)}\theta^{(g)} + V_{D,n+1}^{(g)}$.
- Draw $Y_{j,n+1}^{(g)} \sim N\left(X_{n+1}^{(g)}\beta_j^{(g)} + \omega_j^{(g)}V_{D,n+1}^{(g)}, \frac{\sigma_j^{2(g)}}{\lambda_{n+1}^{(g)}}\right)$ for $j = 0, 1$.

This produces the desired sample for the predictive distribution

$$\left\{Y_{j,n+1}^{(1)}, \ldots, Y_{j,n+1}^{(M)}\right\}$$

A sample of predictive conditional treatment effects is then defined by

$$\begin{aligned} & \left\{\Delta_{n+1}^{(1)} \mid X_{n+1}^{(1)}, \ldots, \Delta_{n+1}^{(M)} \mid X_{n+1}^{(M)}\right\} \\ \equiv & \left\{\left(Y_{1,n+1}^{(1)} - Y_{0,n+1}^{(1)} \mid X_{n+1}^{(1)}\right), \ldots, Y\left(_{1,n+1}^{(M)} - Y_{0,n+1}^{(M)} \mid X_{n+1}^{(M)}\right)\right\} \end{aligned}$$

By iterated expectations, the predictive (unconditional) average treatment effect is

$$\begin{aligned} & E_{X_{n+1}}\left[E\left[\Delta_{n+1} \mid X_{n+1}, Y, H, D\right]\right] \\ = & \int Y_{1,n+1} p\left(Y_1 \mid X_{n+1}, Y, H, D\right) \\ & \times p\left(X_{n+1} \mid Y, H, D\right) dY_1 dX_{n+1} \\ & - \int Y_{0,n+1} p\left(Y_0 \mid X_{n+1}, Y, H, D\right) \\ & \times p\left(X_{n+1} \mid Y, H, D\right) dY_0 dX_{n+1} \\ = & \int Y_{1,n+1} p\left(Y_1, X_{n+1} \mid Y, H, D\right) dY_1 dX_{n+1} \\ & - \int Y_{0,n+1} p\left(Y_0, X_{n+1} \mid Y, H, D\right) dY_0 dX_{n+1} \end{aligned}$$

Then estimated ATE is

$$estATE = \frac{1}{M} \sum_{g=1}^{M} \left(Y_{1,n+1}^{(g)} - Y_{0,n+1}^{(g)}\right)$$

This latter quantity — our estimate of the predictive average treatment effect — is simply the average over our post-convergence difference in predictive draws.

10.4.2 Predicted LATE

Local average treatment effects are defined for the subsample of compliers. For binary instrument z, compliers are identified by $D_1 - D_0 = 1$ where D_j refers to the selection variable when the instrument z has value j. We add the following steps to the algorithm to identify compliers for our predictive sample. Let $Z_{j,n+1}^{(g)} = Z_{n+1}^{(g)}$ where j refers to the value of the binary instrument $z_{n+1}^{(g)}$.

- Set $z_{n+1}^{(g)} = 0$, then $D_{0,n+1}^{(g)} = I\left(Z_{0,n+1}^{(g)} \theta^{(g)} + V_{D,n+1}^{(g)} > 0\right)$.
- Set $z_{n+1}^{(g)} = 1$, then $D_{1,n+1}^{(g)} = I\left(Z_{1,n+1}^{(g)} \theta^{(g)} + V_{D,n+1}^{(g)} > 0\right)$.
- Define $C_{n+1}^{(g)} = I\left(D_{1,n+1}^{(g)} - D_{0,n+1}^{(g)} = 1\right)$.

A sample of predictive conditional treatment effects for the subsample of compliers is then defined by

$$\begin{aligned}
&\left\{\left(\Delta^{(1)}_{n+1} \mid X^{(1)}_{n+1}, C^{(g)}_{n+1} = 1\right), \ldots, \left(\Delta^{(M_c)}_{n+1} \mid X^{(M_c)}_{n+1}, C^{(g)}_{n+1} = 1\right)\right\} \\
\equiv &\left\{\begin{array}{c} \left(Y^{(1)}_{1,n+1} - Y^{(1)}_{0,n+1} \mid X^{(1)}_{n+1}, C^{(g)}_{n+1} = 1\right), \ldots, \\ Y\left(^{(M_c)}_{1,n+1} - Y^{(M_c)}_{0,n+1} \mid X^{(M_c)}_{n+1}, C^{(g)}_{n+1} = 1\right) \end{array}\right\}
\end{aligned}$$

where $M_c = \sum_{g=1}^{M} C^{(g)}_{n+1}$. By iterated expectations, the predictive (unconditional) local average treatment effect is

$$\begin{aligned}
& E_{X_{n+1}}\left[E\left[\Delta_{n+1} \mid X_{n+1}, Y, H, D, C^{(g)}_{n+1} = 1\right]\right] \\
= & \int Y_{1,n+1} p\left(Y_1 \mid X_{n+1}, Y, H, D, C^{(g)}_{n+1} = 1\right) \\
& \times p\left(X_{n+1} \mid Y, H, D, C^{(g)}_{n+1} = 1\right) dY_1 dX_{n+1} \\
& - \int Y_{0,n+1} p\left(Y_0 \mid X_{n+1}, Y, H, D, C^{(g)}_{n+1} = 1\right) \\
& \times p\left(X_{n+1} \mid Y, H, D, C^{(g)}_{n+1} = 1\right) dY_0 dX_{n+1} \\
= & \int Y_{1,n+1} p\left(Y_1, X_{n+1} \mid Y, H, D, C^{(g)}_{n+1} = 1\right) dY_1 dX_{n+1} \\
& - \int Y_{0,n+1} p\left(Y_0, X_{n+1} \mid Y, H, D, C^{(g)}_{n+1} = 1\right) dY_0 dX_{n+1}
\end{aligned}$$

Then, estimated $LATE$ is

$$estLATE = \frac{1}{M_c} \sum_{g=1}^{M_c} \left(Y^{(g)}_{1,n+1} - Y^{(g)}_{0,n+1} \mid C^{(g)}_{n+1} = 1\right)$$

10.4.3 Predicted ATT

The draws are in place for predicted treatment effects on treated and untreated except for $D^{(g)}_{n+1}$.

- Draw j from Bernoulli$\left(\frac{\sum_{i=1}^{n} D_i}{n}\right)$ and let $D^{(g)}_{n+1} = D^{(g)}_{j,n+1}$ where $D^{(g)}_{j,n+1}$ was defined in the previous step for $j = 0, 1$.

For $D^{(g)}_{n+1} = 1$ the predicted treatment effect on the treated is defined by

$$\begin{aligned}
&\left\{\left(\Delta^{(1)}_{n+1} \mid X^{(1)}_{n+1}, D^{(g)}_{n+1} = 1\right), \ldots, \left(\Delta^{(M_D)}_{n+1} \mid X^{(M_D)}_{n+1}, D^{(g)}_{n+1} = 1\right)\right\} \\
\equiv &\left\{\begin{array}{c} \left(Y^{(1)}_{1,n+1} - Y^{(1)}_{0,n+1} \mid X^{(1)}_{n+1}, D^{(g)}_{n+1} = 1\right), \ldots, \\ Y\left(^{(M_D)}_{1,n+1} - Y^{(M_D)}_{0,n+1} \mid X^{(M_D)}_{n+1}, D^{(g)}_{n+1} = 1\right) \end{array}\right\}
\end{aligned}$$

where $M_D = \sum_{g=1}^{M} D_{n+1}^{(g)}$. By iterated expectations, the predictive (unconditional) average treatment effect on the treated is

$$\begin{aligned}
& E_{X_{n+1}}\left[E\left[\Delta_{n+1} \mid X_{n+1}, Y, H, D, D_{n+1}^{(g)}=1\right]\right] \\
= & \int Y_{1,n+1} p\left(Y_1 \mid X_{n+1}, Y, H, D, D_{n+1}^{(g)}=1\right) \\
& \times p\left(X_{n+1} \mid Y, H, D, D_{n+1}^{(g)}=1\right) dY_1 dX_{n+1} \\
& -\int Y_{0,n+1} p\left(Y_0 \mid X_{n+1}, Y, H, D, D_{n+1}^{(g)}=1\right) \\
& \times p\left(X_{n+1} \mid Y, H, D, D_{n+1}^{(g)}=1\right) dY_0 dX_{n+1} \\
= & \int Y_{1,n+1} p\left(Y_1, X_{n+1} \mid Y, H, D, D_{n+1}^{(g)}=1\right) dY_1 dX_{n+1} \\
& -\int Y_{0,n+1} p\left(Y_0, X_{n+1} \mid Y, H, D, D_{n+1}^{(g)}=1\right) dY_0 dX_{n+1}
\end{aligned}$$

Then, estimated ATT is

$$estATT = \frac{\sum_{g=1}^{M}\left(Y_{1,n+1}^{(g)} - Y_{0,n+1}^{(g)} \mid D_{n+1}^{(g)}=1\right)}{\sum_{g=1}^{M} D_{n+1}^{(g)}}$$

10.4.4 Predicted ATUT

Analogously, for $D_{n+1}^{(g)} = 0$ the predicted treatment effect on the untreated is defined by

$$\begin{aligned}
& \left\{\left(\Delta_{n+1}^{(1)} \mid X_{n+1}^{(1)}, D_{n+1}^{(g)}=0\right), \ldots, \left(\Delta_{n+1}^{\left(M_{D_0}\right)} \mid X_{n+1}^{\left(M_{D_0}\right)}, D_{n+1}^{(g)}=0\right)\right\} \\
\equiv & \left\{\begin{array}{c}
\left(Y_{1,n+1}^{(1)} - Y_{0,n+1}^{(1)} \mid X_{n+1}^{(1)}, D_{n+1}^{(g)}=0\right), \ldots, \\
Y\left({}_{1,n+1}^{\left(M_{D_0}\right)} - Y_{0,n+1}^{\left(M_{D_0}\right)} \mid X_{n+1}^{\left(M_{D_0}\right)}, D_{n+1}^{(g)}=0\right)
\end{array}\right\}
\end{aligned}$$

where $M_{D_0} = \sum_{g=1}^{M} 1 - D_{n+1}^{(g)} = M - M_D$. By iterated expectations, the predictive (unconditional) average treatment effect on the treated is

$$\begin{aligned}
& E_{X_{n+1}} \left[E \left[\Delta_{n+1} \mid X_{n+1}, Y, H, D, D_{n+1}^{(g)} = 0 \right] \right] \\
= & \int Y_{1,n+1} p \left(Y_1 \mid X_{n+1}, Y, H, D, D_{n+1}^{(g)} = 0 \right) \\
& \times p \left(X_{n+1} \mid Y, H, D, D_{n+1}^{(g)} = 1 \right) dY_1 dX_{n+1} \\
& - \int Y_{0,n+1} p \left(Y_0 \mid X_{n+1}, Y, H, D, D_{n+1}^{(g)} = 0 \right) \\
& \times p \left(X_{n+1} \mid Y, H, D, D_{n+1}^{(g)} = 1 \right) dY_0 dX_{n+1} \\
= & \int Y_{1,n+1} p \left(Y_1, X_{n+1} \mid Y, H, D, D_{n+1}^{(g)} = 0 \right) dY_1 dX_{n+1} \\
& - \int Y_{0,n+1} p \left(Y_0, X_{n+1} \mid Y, H, D, D_{n+1}^{(g)} = 0 \right) dY_0 dX_{n+1}
\end{aligned}$$

Then, estimated $ATUT$ is

$$estATUT = \frac{\sum_{g=1}^{M} \left(Y_{1,n+1}^{(g)} - Y_{0,n+1}^{(g)} \mid D_{n+1}^{(g)} = 0 \right)}{\sum_{g=1}^{M} 1 - D_{n+1}^{(g)}}$$

10.5 Return to the treatment effect example

Initially, we employ Bayesian data augmentation via a Gibbs sampler on the treatment effect problem outlined above. Recall this example was employed in the projections notes to illustrate where the inverse-Mills ratios control functions strategy based on the full complement of instruments[2] was exceptionally effective.

The representative sample is

Y	D	Y_1	Y_0	Z_1	Z_2	Z_3	Z_4
15	1	15	9	5	4	3	1
14	1	14	10	−6	−5	−4	−2
13	1	13	11	0	0	0	1
13	0	11	13	0	0	1	0
14	0	10	14	0	1	0	0
15	0	9	15	1	0	0	0

[2] Typically, we're fortunate to identify any instruments. In the example, the instruments form a basis for the nullspace to the outcomes, Y_1 and Y_0. In this (linear or Gaussian) sense, we've exhausted the potential set of instruments.

which is repeated 200 times to create a sample of $n = 1,200$ observations. The Gibbs sampler employs $15,000$ draws from the conditional posteriors. The first $5,000$ draws are discarded as burn-in, then sample statistics are based on the remaining $10,000$ draws. First, estimated coefficients for outcome and selection equations are tabulated. This is followed by variance-covariance estimates in a second table.

statistic	β_1	β_0	θ_1	θ_2	θ_3	θ_4
mean	13.80	13.78	0.380	−0.147	−0.845	0.842
median	13.80	13.78	0.381	−0.147	−0.845	0.842
standard dev	0.036	0.043	0.033	0.038	0.042	0.042
quantiles:						
0.01	13.72	13.68	0.301	−0.238	−0.942	0.747
0.05	13.74	13.71	0.325	−0.195	−0.914	0.773
0.10	13.76	13.73	0.337	−0.195	−0.899	0.788
0.25	13.75	13.78	0.358	−0.172	−0.874	0.813
0.75	13.83	13.81	0.403	−0.122	−0.817	0.870
0.90	13.85	13.84	0.422	−0.098	−0.791	0.895
0.95	13.86	13.85	0.435	−0.084	−0.776	0.910
0.99	13.89	13.88	0.457	−0.059	−0.747	0.937

Sample statistics for the model coefficients of the data augmented Gibbs sampler applied to the sparse data example

statistic	λ_0	λ_1	ω_0	ω_1	η_0^2	η_1^2
mean	1.051	1.042	−0.754	0.710	1.026	0.878
median	1.008	1.000	−0.753	0.710	1.021	0.876
standard dev			0.069	0.044	0.090	0.056
quantiles:						
0.01	0.387	0.384	−0.919	0.603	0.840	0.754
0.05	0.528	0.523	−0.868	0.637	0.888	0.789
0.10	0.616	0.610	−0.844	0.653	0.914	0.807
0.25	0.785	0.778	−0.801	0.680	0.963	0.839
0.75	1.269	1.260	−0.706	0.740	1.084	0.914
0.90	1.541	1.530	−0.667	0.766	1.145	0.950
0.95	1.720	1.709	−0.642	0.781	1.183	0.971
0.99	2.091	2.079	−0.600	0.812	1.258	1.015

Sample statistics for the covariance parameters of the data augmented Gibbs sampler applied to the sparse data example

The results demonstrate selection bias as the means are biased upward from 12. This does not bode well for effective estimation of marginal or average treatment effects. Estimated average treatment effects are tabulated below; recall DGP parameters are $ATE = 0, ATT = 4, ATUT = -4$, amd $LATE = -2$.

statistic	$estATE$	$estATT$	$estATUT$	$estLATE$
mean	0.035	0.558	-0.523	0.035
median	0.019	0.000	0.000	0.019
standard dev	1.848	1.205	1.175	1.848
quantiles:				
0.01	-4.366	-1.720	-4.366	-4.366
0.05	-2.944	-0.705	-2.935	-2.944
0.10	-2.254	-0.128	-2.211	-2.254
0.25	-1.164	0.000	-0.980	-1.164
0.75	1.221	1.046	0.000	1.221
0.90	2.331	2.281	0.179	2.331
0.95	3.083	3.074	0.754	3.083
0.99	4.487	4.487	1.815	4.487

Sample statistics for estimated average treatment effects of the data augmented Gibbs sampler applied to the sparse data example

The average treatment effects on the treated and untreated suggest heterogeneity but are grossly understated compared to the DGP averages of 4 and -4. Next, we revisit the sparse data problem and attempt to consider what is left out of our model specification.

10.6 Instrumental variable restrictions

To fully exploit the instruments with the limited sample space our data covers we'll need to add data augmentation of counterfactuals to our algorithm. Counterfactuals are drawn conditional on everything else, $Y^*, H, D, \beta, \Sigma_j$, $j = 0, 1$.

$$Y^{miss} = DY_0^{miss} + (1 - D)\, Y_1^{miss}$$

where

$$Y_{ji}^{miss} \sim N\left(X_{ji}\beta_j + \omega_j V_{Di}, \frac{\sigma_j^2}{\lambda_i} \right)$$

and $V_{Di} = D_i^* - Z_i\theta$. That is, we're still not employing the full distribution of outcomes; implicitly V_0 and V_1 are treated as independent. Then, we can add the instrumental variable restrictions discussed in chapter 9 for this sparse data case. That is, counterfactual outcomes are drawn such that they are independent of the instruments. $DY + (1 - D)\, Y^{draw}$ and $DY^{draw} + (1 - D)\, Y$ are independent of Z. This is our IV data augmented Gibbs sampler treatment effect analysis.

To implement this we add the following steps along with the counterfactual draws above to the previous Gibbs sampler. Minimize the distance of Y^{draw} from Y^{miss} such that $Y_1^* = DY + (1-D)\,Y^{draw}$ and $Y_0^* = DY^{draw} + (1-D)\,Y$ are orthogonal to the instruments, Z.

$$\begin{array}{ll} \min\limits_{Y^{draw}} & \left(Y^{draw} - Y^{miss}\right)^T \left(Y^{draw} - Y^{miss}\right) \\ s.t. & Z^T \left[\begin{array}{cc} DY + (1-D)\,Y^{draw} & DY^{draw} + (1-D)\,Y \end{array}\right] = 0 \end{array}$$

where the constraint is $p \times 2$ zeroes and p is the number of columns in Z (the number of instruments). Hence, the IV $McMC$ outcome draws are

$$Y_1^* = DY + (1-D)\,Y^{draw}$$

and

$$Y_0^* = DY^{draw} + (1-D)\,Y$$

Rather than repeat results for the IV data augmented Gibbs sampler applied to the sparse data example presented in the previous chapter (results are similar), we turn to a prototypical (nonsparse data) example to evaluate the efficacy of our Bayesian treatment effects without the joint outcome distribution strategy.

10.7 Prototypical example

Treatment effects for the following DGP are analyzed via a Bayesian data augmentation without full outcome distributions strategy. The DGP is

$$\begin{array}{rcl} EU & = & \gamma_0 + \gamma_1 x + \gamma_2 z + V_D \\ & = & -1 + 1x + 1z + V_D \\ Y_1 & = & \alpha_1 + \beta_1 x + V_1 \\ & = & 2 + 10x + V_1 \\ Y_0 & = & \alpha_0 + \beta_0 x + V_0 \\ & = & 1 + 2x + V_0 \end{array}$$

where x is uniform$(0,1)$ and z is binary (Bernoulli(0.5)) and variance-covariance for (V_D, V_1, V_0) is

$$\left[\begin{array}{ccc} 1 & 0.7 & -0.7 \\ 0.7 & 1 & -0.1 \\ -0.7 & -0.1 & 1 \end{array}\right]$$

The Gibbs sampler employs $15,000$ draws from the conditional posteriors.[3] The first $5,000$ draws are discarded as burn-in, then sample statistics

[3] The results presented in this chapter employ two-stage estimation. That is, a Gibbs sampler is employed to estimate the selection equation. Then, the coefficient means

are based on the remaining 10,000 draws. First, estimated coefficients for outcome and selection equations are tabulated. This is followed by variance-covariance estimates in a second table

statistic	α_0	β_0	α_1	β_1	γ_0	γ_1	γ_2
mean	1.184	2.094	2.232	9.750	-0.863	0.806	0.985
median	1.184	2.094	2.233	9.750	-0.864	0.806	0.986
standard dev	0.075	0.137	0.088	0.132	0.067	0.106	0.051
quantiles:							
0.01	1.005	1.779	2.023	9.446	-1.017	0.559	0.867
0.05	1.060	1.868	2.085	9.535	-0.973	0.633	0.902
0.10	1.087	1.920	2.119	9.582	-0.950	0.671	0.919
0.25	1.134	2.001	2.175	9.661	-0.908	0.735	0.950
0.75	1.234	2.187	2.292	9.839	-0.818	0.878	1.021
0.90	1.277	2.272	2.344	9.918	-0.777	0.944	1.051
0.95	1.304	2.320	2.376	9.967	-0.753	0.984	1.068
0.99	1.356	2.413	2.433	10.06	-0.706	1.052	1.106

Sample statistics for the model coefficients of the data augmented Gibbs sampler applied to the prototypical example

statistic	λ_0	λ_1	ω_0	ω_1	η_0^2	η_1^2
mean	0.996	0.993	-0.488	0.599	0.723	0.824
median	0.951	0.948	-0.488	0.599	0.720	0.821
standard dev			0.044	0.050	0.051	0.065
quantiles:						
0.01	0.346	0.346	-0.593	0.486	0.615	0.688
0.05	0.481	0.479	-0.562	0.519	0.643	0.724
0.10	0.566	0.564	-0.545	0.536	0.659	0.744
0.25	0.732	0.729	-0.518	0.565	0.686	0.778
0.75	1.213	1.209	-0.458	0.633	0.756	0.867
0.90	1.483	1.479	-0.431	0.663	0.790	0.909
0.95	1.663	1.658	-0.415	0.681	0.811	0.935
0.99	2.036	2.033	-0.387	0.717	0.852	0.990

Sample statistics for the covariance parameters of the data augmented Gibbs sampler applied to the prototypical example

from this first stage are employed in the Gibbs sampler for estimating outcome model parameters, covariances, and treatment effects. Simultaneous estimation of selection and outcome parameters (not reported), along the lines suggested by Chib [2007], produces similar results (especially for the average treatmet effects). Results differ primarily in that selection equation parameters are dampened (around 0.5 in absolute value) and variances are dampened (approximately 0.5) for simultaneous estimation.

These parameter estimates are quite close to their DGP counterparts which is encouraging for identifying treatment effects.

Estimated average treatment effects are tabulated below; sample average treatment effects are $ATE = 4.962, ATT = 6.272, ATUT = 3.620$, amd $LATE = 4.804$.

statistic	$estATE$	$estATT$	$estATUT$	$estLATE$
mean	4.866	5.963	3.684	4.840
median	4.855	6.075	3.593	4.843
standard dev	2.757	2.513	2.510	2.262
quantiles:				
0.01	-1.145	0.680	-1.724	0.261
0.05	0.415	1.798	-0.273	1.261
0.10	1.294	2.570	0.451	1.808
0.25	2.798	4.009	1.804	3.041
0.75	6.942	7.828	5.587	4.843
0.90	8.473	9.131	7.034	6.674
0.95	9.245	9.958	7.789	7.815
0.99	10.859	11.27	9.017	8.389

Sample statistics for estimated average treatment effects of the data augmented Gibbs sampler applied to the prototypical example

These results correspond reasonably well with the *DGP*. This Bayesian strategy without full outcome distributions seems to have potential for identifying outcome heterogeneity.

10.7.1 High correlation example

One concern with this identification strategy might be an extreme correlation *DGP*. That is, we identify treatment effects without (bounding unidentified parameters for) the full joint distribution for outcomes. Implicitly, at least for treatment effects on treated and untreated we're behaving as if unobservable outcome with and without treatment (V_1, V_0) are independent. Suppose we have the same setup as the previous prototypical example except the correlation between unobserved expected utility, V_D, and unobserved outcome with treatment, V_1, is $\rho_{D1} = 0.9$, and with unobserved outcome without treatment, V_0, is $\rho_{D0} = -0.9$. Then, the bounds on the unidentifiable parameter, ρ_{10}, are $(-1.0, -0.62)$; in other words, $\rho_{10} = 0$ is not included within bounds which preserve positive definiteness of variance-covariance matrix, Σ. A natural question then is how effective is the treatment effect identification strategy without joint outcome distributions. Next, we report the results of a simulation experiment addressing this out of bounds correlation issue.

Suppose the *DGP* is

$$\begin{aligned} EU &= \gamma_0 + \gamma_1 x + \gamma_2 z + V_D \\ &= -1 + 1x + 1z + V_D \\ Y_1 &= \alpha_1 + \beta_1 x + V_1 \\ &= 2 + 10x + V_1 \\ Y_0 &= \alpha_0 + \beta_0 x + V_0 \\ &= 1 + 2x + V_0 \end{aligned}$$

where x is uniform$(0,1)$ and z is binary (Bernoulli(0.5)) and variance-covariance for (V_D, V_1, V_0) is

$$\begin{bmatrix} 1 & 0.9 & -0.9 \\ 0.9 & 1 & -0.8 \\ -0.9 & -0.8 & 1 \end{bmatrix}$$

The Gibbs sampler employs $15,000$ draws from the conditional posteriors. The first $5,000$ draws are discarded as burn-in, then sample statistics are based on the remaining $10,000$ draws. First, estimated coefficients for outcome and selection equations are tabulated. This is followed by variance-covariance estimates in a second table

statistic	α_0	β_0	α_1	β_1	γ_0	γ_1	γ_2
mean	1.201	2.272	2.258	9.658	−1.048	1.103	1.057
median	1.202	2.271	2.582	9.658	−1.049	1.104	1.057
standard dev	0.062	0.113	0.075	0.116	0.068	0.112	0.043
quantiles:							
0.01	1.051	2.009	2.408	9.387	−1.209	0.839	0.957
0.05	1.098	2.087	2.456	9.468	−1.161	0.920	0.986
0.10	1.121	2.127	2.484	9.507	−1.136	0.960	1.002
0.25	1.160	2.195	2.532	9.580	−1.094	1.029	1.029
0.75	1.243	2.347	2.634	9.734	−1.003	1.180	1.086
0.90	1.282	2.416	2.679	9.807	−0.960	1.249	1.111
0.95	1.303	2.457	2.705	9.851	−0.934	1.290	1.127
0.99	1.348	2.537	2.751	9.933	−0.893	1.366	1.116

Sample statistics for the model coefficients of the data augmented Gibbs sampler applied to the $|\rho_{10}| >> 0$ example

statistic	λ_0	λ_1	ω_0	ω_1	η_0^2	η_1^2
mean	0.969	0.960	-0.584	0.597	0.569	0.591
median	0.926	0.915	-0.584	0.598	0.567	0.590
standard dev			0.035	0.034	0.037	0.037
quantiles:						
0.01	0.312	0.302	-0.665	0.517	0.487	0.510
0.05	0.447	0.436	-0.640	0.541	0.510	0.532
0.10	0.534	0.523	-0.629	0.553	0.523	0.544
0.25	0.702	0.692	-0.607	0.574	0.543	0.565
0.75	1.189	1.179	-0.561	0.620	0.593	0.615
0.90	1.461	1.452	-0.539	0.640	0.618	0.638
0.95	1.641	1.632	-0.526	0.653	0.631	0.652
0.99	2.016	2.006	-0.501	0.675	0.659	0.680

Sample statistics for the covariance parameters of the data augmented Gibbs sampler applied to the $|\rho_{10}| >> 0$ example

These parameter estimates are quite close to their DGP counterparts which is encouraging for identifying treatment effects. Outcome error variances are an exception as they are dampened considerably (from 1 to ~ 0.6); on the other hand, correlations with unobserved expected utility are dampened relatively little (from 0.9 in absolute value to ~ 0.8).

Estimated average treatment effects are tabulated below; sample average treatment effects are $ATE = 5.144, ATT = 6.815, ATUT = 3.370$, amd $LATE = 4.922$.

statistic	$estATE$	$estATT$	$estATUT$	$estLATE$
mean	5.070	6.344	3.744	5.044
median	5.073	6.411	3.691	5.081
standard dev	2.579	2.232	2.223	1.893
quantiles:				
0.01	-0.467	1.644	-1.028	1.093
0.05	0.901	2.649	0.269	2.005
0.10	1.719	3.302	0.903	2.484
0.25	3.164	4.701	2.070	3.591
0.75	6.984	7.999	5.406	6.502
0.90	8.443	9.190	6.787	7.560
0.95	9.228	9.833	7.446	8.031
0.99	10.585	11.16	8.377	8.862

Sample statistics for estimated average treatment effects of the data augmented Gibbs sampler applied to the $|\rho_{10}| >> 0$ example

These results correspond remarkably well with the DGP and with unreported joint outcome distribution (bounding of the unidentified parameter as discussed in the previous chapter) identification strategy results in spite of our bounding concerns.