Selection, Treatment, and Switching Models

KPT, Ch. 14

Matlab scripts: `mod9_select_data, mod9_select, mod9_select_door2door, mod9_treat_data, mod9_treat, mod9_treat_naive, mod9_treatJobTrain, mod9_treatJobTrainNaive, mod9_switch_data, mod9_switch`

Matlab functions: `gs_select, gs_treat, gs_switch, iw_sig11`

We will encounter two new estimation challenges for these models: A restricted covariance matrix, and drawing latent data from multivariate densities with inequality restrictions.

Sample Selection Model

Basic features

The basic sample selection model (often also referred to as "Heckman selection model" or "Type II Tobit") is structured as follows:

$$
\begin{array}{llll}
y_{1i}^{*} = \mathbf{x}_{1i}'\boldsymbol{\beta}_1 + \varepsilon_{1i} & & y_{2i}^{*} = \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + \varepsilon_{2i} & \\
\begin{bmatrix}\varepsilon_1 \\ \varepsilon_2\end{bmatrix} \sim n\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \boldsymbol{\Sigma}\right) & & \boldsymbol{\Sigma} = \begin{bmatrix}1 & \sigma_{12} \\ \sigma_{12} & \sigma_{22}\end{bmatrix} & \\
y_{1i} = 1 & \text{if} & y_{1i}^{*} > 0, & = 0 \text{ otherwise} \\
y_{2i} = y_{2i}^{*} & \text{if} & y_{1i}^{*} > 0, & \text{unobserved otherwise}
\end{array}
\tag{1}
$$

Since the outcome for the first equation (often called "participation equation" or "hurdle equation") is only observed in binary form (as in the Probit), the variance for the first error term is not identified and arbitrarily set to one. This generates a constrained covariance matrix for $\boldsymbol{\varepsilon}_i = \begin{bmatrix}\varepsilon_{1i} & \varepsilon_{2i}\end{bmatrix}'$, which renders a generic *IW* density (as used e.g. for the SUR model) unsuitable for obtaining draws of $\boldsymbol{\Sigma}$. We will deal with this issue below.

Conceptually, the model structure in (1) arises because there are unobservables in the hurdle and outcome equation that are correlated. A classic example would be a woman's decision to join the labor force, which may be a function of unobserved innate drive, ability, or "spunk". This quality will also affect the outcome of interest, e.g. hourly wage. If this link between the two equations is ignored, results flowing solely from the outcome equation will suffer from sample selection bias.

For example, let's assume our interest is in the marginal effects of regressors on (potential) female wage for the entire population of female workers (including women who may not currently be in the labor market). Thus, we are primarily interested in $\boldsymbol{\beta}_2$. Ignoring the selection problem, we would naively use

the second equation only, run a basic normal regression model, and estimate $\boldsymbol{\beta}_2$. Under sample selection, i.e. if $\sigma_{12} \neq 0$, these estimates will be biased. Intuitively, this bias arises because under sample selection women who work share an unobserved quality that increases the likelihood of passing the first hurdle. The sub-sample that's left for the second equation is thus systematically different for the general population, and will be affected differently by regressors than those who don't. Econometrically, the bias can be best illustrated when comparing conditional regression means or expectations.

For the naïve model we have

$$E\left(y_{2i}^{*}\right) = \mathbf{x}_{\mathbf{2i}}'\boldsymbol{\beta}_{\mathbf{2}} = \omega_{2i} \tag{2}$$

In reality, however, the second equation is based on a conditional expectation, i.e

$$\begin{aligned} &E\left[\, y_{2i}^{*} \mid y_{1i}^{*} > 0 \,\right] = E\left[\, \omega_{2i} + \varepsilon_{2i} \mid y_{1i}^{*} > 0 \,\right] = \omega_{2i} + E\left[\, \varepsilon_{2i} \mid y_{1i}^{*} > 0 \,\right] = \\ &\omega_{2i} + E\left[\varepsilon_{2i} \mid \varepsilon_{1i} > -\omega_{1i}\right] = \omega_{2i} + \sigma_{12}\frac{\phi\left(\omega_{1i}\right)}{\Phi\left(\omega_{1i}\right)} \qquad \text{where } \omega_{1i} = \mathbf{x}_{\mathbf{1i}}'\boldsymbol{\beta}_{\mathbf{1}} \end{aligned} \tag{3}$$

The fraction in the last term (Inverse Mills ratio, IMR) is always positive. Thus, the relative magnitude of the conditional vs. the unconditional mean depends on the sign of the covariance. If $\sigma_{12} > 0$ the conditional mean exceeds the unconditional mean. Intuition for the labor market example: In that case an increased propensity to join the labor force is positively correlated with an increased wage. Since we condition the distribution of y_{2i}^{*} exclusively on positive outcomes for the hurdle, we would naturally expect the conditional mean to be higher. The opposite holds if $\sigma_{12} < 0$. In either case, the estimates for $\boldsymbol{\beta}_{\mathbf{2}}$ will adjust for the added term, and thus be different from the estimates generated by the naïve model. Specifically, if $\sigma_{12} > 0$ we would expect the naive estimate for $\boldsymbol{\beta}_{\mathbf{2}}$ to be biased upwards to make up for the missing positive term, and downwards if $\sigma_{12} < 0$.

Thus, a key focus in the estimation of the sample selection model will lie on the covariance σ_{12}. A natural "restricted" candidate model would be one with $\sigma_{12} = 0$, in which case the two equations are independent. For our labor market example this would imply that women that work are no different (based on unobservables in equation 1) than women who don't , and we can interpret any inference flowing from equation 2 to hold for the general population.

Another important feature of the sample selection model is the "identification restriction", which stipulates that the first equation must include at least one explanatory variable that is not also present in the second equation. This can often be tricky to satisfy as there is substantial overlap between $\mathbf{x}_{\mathbf{1i}}$ and $\mathbf{x}_{\mathbf{2i}}$ in many applications.

Estimation

As for the Probit and Tobit models we assume that explanatory data are observed for all individuals, regardless of the outcome for the first hurdle. Our estimation strategy will have features of the SUR model (since we have multiple linked equations per individual) and of the Probit and Tobit models (since we will work with latent dependent variables).

Denoting the number of regressors in the first and second equation as k_1 and k_2, respectively, we can write the pair of latent observations for a given individual as:

$$\mathbf{y_i^*} = \mathbf{X_i}\boldsymbol{\beta} + \boldsymbol{\varepsilon_i} \quad \text{with} \quad \boldsymbol{\varepsilon_i} \sim n(\mathbf{0}, \boldsymbol{\Sigma}),$$
$$\mathbf{X_i} = \begin{bmatrix} \mathbf{x'_{1i}} & \mathbf{0} \\ \mathbf{0} & \mathbf{x'_{2i}} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{bmatrix} \qquad \boldsymbol{\varepsilon_i} = \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \tag{4}$$

In analogy to the full model over all individuals can be written as

$$\mathbf{y^*} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$
$$\mathbf{y^*} = \underset{(nx2)x1}{\begin{bmatrix} \mathbf{y_1^*} \\ \mathbf{y_1^*} \\ \vdots \\ \mathbf{y_n^*} \end{bmatrix}} \qquad \mathbf{X} = \underset{(nx2)xk}{\begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \\ \vdots \\ \mathbf{X_n} \end{bmatrix}} \qquad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon_1} \\ \boldsymbol{\varepsilon_2} \\ \vdots \\ \boldsymbol{\varepsilon_n} \end{bmatrix} \sim n(\mathbf{0}, \boldsymbol{\Omega}) \qquad \text{with} \tag{5}$$
$$\boldsymbol{\Omega} = \underset{(nx2)x(nx2)}{\begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \cdots & \mathbf{0} \\ \mathbf{0} & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma} \end{bmatrix}}$$

To construct the likelihood function for this model, we need to review some basic results for joint densities with inequality restrictions for one or more equations. Generically, the truncated joint density for two random variables *y* and *z* can be written as

$$f(y < a, z) = \int_{-\infty}^{a} f(y, z)\,dy = \int_{-\infty}^{a} f(y \mid z)\, f(z)\,dy = f(z) \int_{-\infty}^{a} f(y \mid z)\,dy \tag{6}$$

Also, if *y* and *z* are bivariate normal, the conditional densities are normal as well with the following moments:

$$\begin{aligned} &z \mid y \sim n\left(\mu_{z.y}, \sigma^2_{z.y}\right) \qquad \mu_{z.y} = \mu_z + \frac{\sigma_{yz}}{\sigma_y^2}\left(y - \mu_y\right) \qquad \sigma^2_{z.y} = \sigma^2_z - \frac{\sigma^2_{yz}}{\sigma^2_y} \\ &y \mid z \sim n\left(\mu_{y.z}, \sigma^2_{y.z}\right) \qquad \mu_{y.z} = \mu_y + \frac{\sigma_{yz}}{\sigma_z^2}\left(z - \mu_z\right) \qquad \sigma^2_{y.z} = \sigma^2_y - \frac{\sigma^2_{yz}}{\sigma^2_z} \end{aligned} \tag{7}$$

For our sample likelihood there are *two possible observable outcomes,* $pr\left(y_{1i}=1,y_{2i}\right)$ and $pr\left(y_{1i}=0\right)$. The second one can be immediately expressed as

$$pr\left(y_{1i}=0\right)=pr\left(y_{1i}^{*}\le 0\right)=pr\left(\varepsilon_{1i}\le -\mathbf{x_{1i}'\beta_1}\right)=\Phi\left(-\mathbf{x_{1i}'\beta_1}\right) \tag{8}$$

The first one takes a bit of extra work to bring it into a manageable form. We have

$$\begin{aligned} &pr\left(y_{1i}=1,y_{2i}\right)=f\left(\varepsilon_{1i}>-\mathbf{x_{1i}'\beta_1},\varepsilon_{2i}\right)= \\ &\int_{-\mathbf{x_{1i}'\beta_1}}^{\infty} f\left(\varepsilon_{1i},\varepsilon_{2i}\right)d\varepsilon_{1i}=\int_{-\mathbf{x_{1i}'\beta_1}}^{\infty} f\left(\varepsilon_{1i}\mid\varepsilon_{2i}\right)f\left(\varepsilon_{2i}\right)d\varepsilon_{1i}= \\ &f\left(\varepsilon_{2i}\right)\int_{-\mathbf{x_{1i}'\beta_1}}^{\infty} f\left(\varepsilon_{1i}\mid\varepsilon_{2i}\right)d\varepsilon_{1i}=\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\tfrac{1}{2\sigma_2^2}\left(y_{2i}-\mathbf{x_{2i}'\beta_2}\right)^2\right)\int_{-\mathbf{x_{1i}'\beta_1}}^{\infty} f\left(\varepsilon_{1i}\mid\varepsilon_{2i}\right)d\varepsilon_{1i} \end{aligned} \tag{9}$$

where we use $\varepsilon_{2i}=y_{2i}-\mathbf{x_{2i}'\beta_2}$ in the last expression. To solve $\int_{-\mathbf{x_{1i}'\beta_1}}^{\infty} f\left(\varepsilon_{1i}\mid\varepsilon_{2i}\right)d\varepsilon_{1i}$, we note that

$$\begin{aligned} &\int_{-\mathbf{x_{1i}'\beta_1}}^{\infty} f\left(\varepsilon_{1i}\mid\varepsilon_{2i}\right)d\varepsilon_{1i}=pr\left(\left(\varepsilon_{1i}\mid\varepsilon_{2i}\right)>-\mathbf{x_{1i}'\beta_1}\right)=pr\left(\frac{\left(\varepsilon_{1i}\mid\varepsilon_{2i}\right)-\mu_{1.2}}{\sigma_{1.2}}<\frac{\mathbf{x_{1i}'\beta_1}+\mu_{1.2}}{\sigma_{1.2}}\right)= \\ &\Phi\left(\frac{\mathbf{x_{1i}'\beta_1}+\mu_{1.2}}{\sigma_{1.2}}\right) \qquad \text{where} \\ &\mu_{1.2}=0+\frac{\sigma_{12}}{\sigma_2^2}\left(\varepsilon_{2i}-0\right)=\frac{\sigma_{12}}{\sigma_2^2}\left(y_{2i}-\mathbf{x_{2i}'\beta_2}\right) \qquad \sigma_{1.2}=\sqrt{1-\frac{\sigma_{12}^2}{\sigma_2^2}}=\sqrt{1-\rho^2} \end{aligned} \tag{10}$$

We can now write the sample likelihood concisely as

$$p\left(\mathbf{y}\mid\boldsymbol{\beta},\boldsymbol{\Sigma}\right)=\prod_{i=1}^{n}\left(\Phi\left(-\mathbf{x_{1i}'\beta_1}\right)\right)^{1-y_{1i}}*\left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\tfrac{1}{2\sigma_2^2}\left(y_{2i}-\mathbf{x_{2i}'\beta_2}\right)^2\right)\Phi\left(\frac{\mathbf{x_{1i}'\beta_1}+\frac{\sigma_{12}}{\sigma_2^2}\left(y_{2i}-\mathbf{x_{2i}'\beta_2}\right)}{\sqrt{1-\rho^2}}\right)\right)^{y_{1i}} \tag{11}$$

The next step would be to assign priors to $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. However, this is where we run into the problem of a constrained covariance matrix. Several approaches have been suggested to handle this issue. For this simple problem with a single constrained variance, the *algorithm proposed by Nobile (2000)* provides fast and accurate results. Without going into the details for this algorithm, we can simply note that it allows you to draw from the Inverse Wishart as usual. The variance constraint is "automatically" taken care of. Thus, if you want to specify an IW prior with degrees of freedom ν_0 and scale matrix $\mathbf{S_0}$, with the (1,1) element constrained to a constant (here "1") you can use the Matlab function `iw_sig11(v0, inv(S0),1)`.

Thus, we can write our priors as

$$p(\boldsymbol{\beta},\boldsymbol{\Sigma}) = p(\boldsymbol{\beta})\,p(\boldsymbol{\Sigma}) \quad \text{where}$$

$$p(\boldsymbol{\beta}) = (2\pi)^{-k/2}\left|\mathbf{V_0}\right|^{-1/2}\exp\left(-\tfrac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu_0})'\,\mathbf{V_0^{-1}}(\boldsymbol{\beta}-\boldsymbol{\mu_0})\right) \tag{12}$$

$$p(\boldsymbol{\Sigma}) = \left(2^{\nu_0}\pi^{1/2}\prod_{i=1}^{2}\Gamma\left(\frac{\nu_0+1-i}{2}\right)\right)^{-1} * \left|\mathbf{S_0}\right|^{\nu_0/2}\left|\boldsymbol{\Sigma}\right|^{-(\nu_0+3)/2}\exp\left(-\frac{1}{2}tr\left(\mathbf{S_0}\cdot\boldsymbol{\Sigma}^{-1}\right)\right)I(\Sigma_{11}=1)$$

Setting $M = 2$ in the *IW* density. Combining the priors with the likelihood, and dropping all terms that are multiplicatively unrelated to our parameters of interest yields the posterior kernel

$$p(\boldsymbol{\beta},\boldsymbol{\Sigma}\,|\,\mathbf{y},\mathbf{X}) \propto$$

$$\left|\boldsymbol{\Sigma}\right|^{-(\nu_0+3)/2}\exp\left(-\frac{1}{2}tr\left(\mathbf{S_0}\cdot\boldsymbol{\Sigma}^{-1}\right)\right)\exp\left(-\tfrac{1}{2}\left((\boldsymbol{\beta}-\boldsymbol{\mu_0})'\,\mathbf{V_0^{-1}}(\boldsymbol{\beta}-\boldsymbol{\mu_0})\right)\right) * \tag{13}$$

$$\prod_{i=1}^{n}\left(\Phi\left(-\mathbf{x_{1i}'}\boldsymbol{\beta_1}\right)\right)^{1-y_{1i}} * \left(\Phi\left(\frac{\mathbf{x_{1i}'}\boldsymbol{\beta_1}+\frac{\sigma_{12}}{\sigma_2^2}\left(y_{2i}-\mathbf{x_{2i}'}\boldsymbol{\beta_2}\right)}{\sqrt{1-\rho^2}}\right)\right)^{y_{1i}}$$

As you can imagine by now, this would not yield "well-behaved" conditional posterior densities for either $\boldsymbol{\beta}$ or $\boldsymbol{\Sigma}$, since both are represented in the "unruly" likelihood function. We will turn again to the vector of latent outcomes, $\mathbf{y}^*$, for help. Conceptually, the joint augmented posterior kernel can be expressed as

$$p\left(\boldsymbol{\beta},\boldsymbol{\Sigma},\mathbf{y}^*\,|\,\mathbf{y}\right) \propto p(\boldsymbol{\beta})\,p(\boldsymbol{\Sigma})\,p\left(\mathbf{y}^*\,|\,\boldsymbol{\beta},\boldsymbol{\Sigma}\right)p\left(\mathbf{y}\,|\,\mathbf{y}^*,\boldsymbol{\beta},\boldsymbol{\Sigma}\right) \tag{14}$$

As for the Probit, Tobit, and OP, we need to take a closer look at the terms $p\left(\mathbf{y}^*\,|\,\boldsymbol{\beta},\boldsymbol{\Sigma}\right)$ and $p\left(\mathbf{y}\,|\,\mathbf{y}^*,\boldsymbol{\beta},\boldsymbol{\Sigma}\right)$ to assure that the augmentation will indeed make our life easier. The conditional prior $p\left(\mathbf{y}^*\,|\,\boldsymbol{\beta},\boldsymbol{\Sigma}\right)$ describes essentially the likelihood of a basic SUR model, i.e.

$$p\left(\mathbf{y}^*\,|\,\boldsymbol{\beta},\boldsymbol{\Sigma},\mathbf{X}\right) = \prod_{i=1}^{n}(2\pi)^{-1}\left|\boldsymbol{\Sigma}\right|^{-1/2}\exp\left(-\tfrac{1}{2}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)\right) =$$

$$(2\pi)^{-n}\left|\boldsymbol{\Sigma}\right|^{-n/2}\exp\left(-\tfrac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)\right) \tag{15}$$

The conditional likelihood function (the last term in (14)) can be written as

$$p\left(\mathbf{y}\,|\,\boldsymbol{\beta},\boldsymbol{\Sigma},\mathbf{y}^*\right) = p\left(\mathbf{y}\,|\,\mathbf{y}^*\right) = \prod_{i=1}^{n}\left(I\left(y_{i1}=0\right)I\left(-\infty < y_{i1}^* \le 0\right) + I\left(y_{i1}=1, y_{i2}=y_{i2}^*\right)I\left(y_{i1}^*>0, y_{i2}^*\right)\right) \tag{16}$$

As for the Probit and Tobit, our data augmentation step de-links the original likelihood from our main parameters of interest, which facilitates their drawing in the GS.

The augmented posterior kernel can now be written in its explicit form as

$$
\begin{aligned}
&p\left(\boldsymbol{\beta},\boldsymbol{\Sigma},\mathbf{y}^{*}\mid\mathbf{y},\mathbf{X}\right)\propto\\
&\left|\boldsymbol{\Sigma}\right|^{-(v_0+3+n)/2}\exp\left(-\frac{1}{2}tr\left(\mathbf{S_0}\cdot\boldsymbol{\Sigma}^{-1}\right)\right)I\left(\Sigma_{11}=1\right)*\exp\left(-\tfrac{1}{2}\left(\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathbf{0}}\right)'\mathbf{V_0^{-1}}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathbf{0}}\right)\right)\right)*\\
&\exp\left(-\tfrac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y_i^{*}}-\mathbf{X_i}\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i^{*}}-\mathbf{X_i}\boldsymbol{\beta}\right)\right)*\\
&\prod_{i=1}^{n}\left(I\left(y_{i1}=0\right)I\left(-\infty<y_{i1}^{*}\le 0\right)+I\left(y_{i1}=1,y_{i2}=y_{i2}^{*}\right)I\left(y_{i1}^{*}>0,y_{i2}^{*}\right)\right)
\end{aligned}
\tag{17}
$$

The conditional posterior kernel for $\boldsymbol{\beta}$ now takes the following form:

$$
p\left(\boldsymbol{\beta}\mid\boldsymbol{\Sigma},\mathbf{y}^{*},\mathbf{X}\right)\propto\exp\left(-\tfrac{1}{2}\left(\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathbf{0}}\right)'\mathbf{V_0^{-1}}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathbf{0}}\right)+\sum_{i=1}^{n}\left(\mathbf{y_i^{*}}-\mathbf{X_i}\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i^{*}}-\mathbf{X_i}\boldsymbol{\beta}\right)\right)\right)
\tag{18}
$$

This is equivalent to the conditional posterior for the basic SUR model, and we can immediately derive the conditional posterior moments as:

$$
\boldsymbol{\beta}\mid\boldsymbol{\Sigma},\mathbf{y}^{*},\mathbf{X}\sim n\left(\boldsymbol{\mu}_{\mathbf{1}},\mathbf{V}_{\mathbf{1}}\right)\quad\text{with}\quad\mathbf{V}_{\mathbf{1}}=\left(\mathbf{V_0^{-1}}+\sum_{i=1}^{n}\mathbf{X}_{\mathbf{i}}'\boldsymbol{\Sigma}^{-1}\mathbf{X_i}\right)^{-1}\text{and}\quad\boldsymbol{\mu}_{\mathbf{1}}=\mathbf{V}_{\mathbf{1}}\left(\mathbf{V_0^{-1}}\boldsymbol{\mu}_{\mathbf{0}}+\sum_{i=1}^{n}\mathbf{X}_{\mathbf{i}}'\boldsymbol{\Sigma}^{-1}\mathbf{y_i^{*}}\right)
\tag{19}
$$

For the conditional posterior of $\boldsymbol{\Sigma}$ we have

$$
\begin{aligned}
&p\left(\boldsymbol{\Sigma}\mid\boldsymbol{\beta},\mathbf{y}^{*},\mathbf{X}\right)\propto\\
&\left|\boldsymbol{\Sigma}\right|^{-(v_0+3+n)/2}\exp\left(-\tfrac{1}{2}\left(tr\left(\mathbf{S_0}\cdot\boldsymbol{\Sigma}^{-1}\right)+\sum_{i=1}^{n}\left(\mathbf{y_i^{*}}-\mathbf{X_i}\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i^{*}}-\mathbf{X_i}\boldsymbol{\beta}\right)\right)\right)I\left(\Sigma_{11}=1\right)
\end{aligned}
\tag{20}
$$

This is again analogous to the SUR model (except for the variance restriction), leading to

$$
\begin{aligned}
&\boldsymbol{\Sigma}\mid\boldsymbol{\beta},\mathbf{y}^{*},\mathbf{X}\sim IW\left(v_1,\mathbf{S_1}\right)I\left(\Sigma_{11}=1\right)\qquad\text{with}\\
&v_1=v_0+n\qquad\mathbf{S_1}=\mathbf{S_0}+\sum_{i=1}^{n}\left(\mathbf{y_i^{*}}-\mathbf{X_i}\boldsymbol{\beta}\right)\left(\mathbf{y_i^{*}}-\mathbf{X_i}\boldsymbol{\beta}\right)'
\end{aligned}
\tag{21}
$$

The conditional posterior kernel for latent data vector $\mathbf{y}^{*}$ is given by

$$
\begin{aligned}
&p\left(\mathbf{y}^{*}\mid\boldsymbol{\beta},\boldsymbol{\Sigma},\mathbf{y},\mathbf{X}\right)\propto\\
&\exp\left(-\tfrac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y_i^{*}}-\mathbf{X_i}\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i^{*}}-\mathbf{X_i}\boldsymbol{\beta}\right)\right)*\\
&\prod_{i=1}^{n}\left(I\left(y_{i1}=0\right)I\left(-\infty<y_{i1}^{*}\le 0\right)+I\left(y_{i1}=1,y_{i2}=y_{i2}^{*}\right)I\left(y_{i1}^{*}>0,y_{i2}^{*}\right)\right)
\end{aligned}
\tag{22}
$$

At the individual level this implies

$$
\begin{aligned}
&p\left(\mathbf{y_i^*} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{y}, \mathbf{X_i}\right) \propto \\
&\exp\left(-\tfrac{1}{2}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)' \boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)\right)* \\
&\left(I\left(y_{i1}=0\right)I\left(-\infty < y_{i1}^* \le 0\right)+I\left(y_{i1}=1, y_{i2}=y_{i2}^*\right)I\left(y_{i1}^*>0, y_{i2}^*\right)\right)
\end{aligned}
\tag{23}
$$

For the $y_{i1}=0$ case we have

$$
\begin{aligned}
&p\left(\mathbf{y_i^*} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, y_{i1}=0, \mathbf{X_i}\right) \propto \\
&\exp\left(-\tfrac{1}{2}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)' \boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)\right)* I\left(-\infty < y_{i1}^* \le 0\right)
\end{aligned}
\tag{24}
$$

This implies a draw of $\mathbf{y_i^*}=\begin{bmatrix} y_{1i}^* & y_{2i}^* \end{bmatrix}'$ from a joint normal density, with y_{i1}^* truncated to the negative domain, i.e. from

$$
p\left(y_{1i}^*<0, y_{2i}^* \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right)=\int_{-\infty}^{0} f\left(y_{1i}^*, y_{2i}^* \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right) dy_{1i}^* \tag{25}
$$

Where *f (.)* denotes the bivariate normal density. This can be computationally implemented as follows:

1. Using $p\left(y_{1i}^*<0, y_{2i}^* \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right)=p\left(y_{2i}^* \mid y_{1i}^*<0, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right) p\left(y_{1i}^*<0 \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right)$, we can first draw y_{1i}^* from its marginal normal density with mean $\mathbf{x_{1i}'\boldsymbol{\beta}_1}$ and variance 1, truncated from above at zero.

2. To obtain draws from $p\left(y_{2i}^* \mid y_{1i}^*<0, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right)$ first note that $y_{2i}^* \mid y_{1i}^* \sim n\left(\mathbf{x_{2i}'\boldsymbol{\beta}_2}+\sigma_{12}\left(y_{1i}^*-\mathbf{x_{1i}'\boldsymbol{\beta}_1}\right), \sigma_{22}\left(1-\rho^2\right)\right)$ where y_{1i}^* are the draws we just obtained in the previous step. Then draw y_{2i}^* from this *untruncated* distribution (no truncation is needed here, since, conditional on $y_{1i}^*<0$, y_{2i}^* can be anything!)

For the $y_{i1}=1$ case we have

$$
\begin{aligned}
&p\left(\mathbf{y_i^*} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, y_{i1}=1, \mathbf{X_i}\right) \propto \\
&\exp\left(-\tfrac{1}{2}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)' \boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)\right)* I\left(y_{i1}^*>0, y_{i2}^*\right)
\end{aligned}
\tag{26}
$$

This implies a draw of $\mathbf{y_i^*} = \begin{bmatrix} y_{1i}^* & y_{2i}^* \end{bmatrix}'$ from a joint normal density, with y_{i1}^* truncated to the positive domain, i.e. from

$$p\left(y_{1i}^* > 0, y_{2i}^* \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right) = \int_0^{\infty} f\left(y_{1i}^*, y_{2i}^* \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right) dy_{1i}^* \tag{27}$$

However, since in this case y_{2i} is observed, we have $p\left(y_{1i}^* > 0, y_{2i}^* \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right) = p\left(y_{1i}^* > 0 \mid y_{2i}^*, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right) p\left(y_{2i}^* \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right) = p\left(y_{1i}^* > 0 \mid y_{2i}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right)$. Thus, we simply draw y_{1i}^* from its conditional normal density with mean $\mathbf{x_{1i}'\beta_1} + \frac{\sigma_{12}}{\sigma_2^2}\left(y_{2i}^* - \mathbf{x_{2i}'\beta_2}\right)$ and variance $1-\rho^2$, *truncated from below at zero*.

Predictive Constructs of Interest

The first predictive construct of interest for the basic selection model is usually the probability of mastering the first hurdle, i.e

$$pr\left(y_1^* > 0 \mid \mathbf{x_{1p}}, \boldsymbol{\beta_1}\right) = \Phi\left(\mathbf{x_{1p}'\beta_1}\right) = \Phi\left(\omega_{1p}\right) \tag{28}$$

Note that even though σ_{12} does not figure in this expression, the resulting PPD for a naïve probit model (ignoring the correlation across equations) will still be misleading since the posterior density of $\boldsymbol{\beta_1}$ will be biased. However, the degree of inferential error for this first predictive outcome of interest is likely smaller than the error for the second outcome, which is a direct function of σ_{12}. It is the expected value for the second equation outcome, conditional on mastering the first equation hurdle, i.e.

$$E\left[y_{2p} \mid y_{1p}^* > 0, \mathbf{x_{1p}}, \mathbf{x_{2p}}, \boldsymbol{\beta}, \boldsymbol{\Sigma}\right] = \omega_{2p} + \sigma_{12}\frac{\phi\left(\omega_{1p}\right)}{\Phi\left(\omega_{1p}\right)} \qquad \text{where } \omega_{1p} = \mathbf{x_{1p}'\beta_1}, \quad \omega_{2p} = \mathbf{x_{2p}'\beta_2} \tag{29}$$

As mentioned above, a simple regression model (ignoring both the the link with the first equation AND the left side truncation in the outcome variable) would estimate this as

$$E\left[y_{2p} \mid \mathbf{x_{2p}}, \boldsymbol{\beta_2}\right] = \omega_{2p} \tag{30}$$

A Tobit model would produce (see earlier lecture notes)

$$E\left[y_{2p} \mid y_{2p}^* > 0, \mathbf{x_{2p}}, \boldsymbol{\beta_2}\right] = \omega_{2p} + \sqrt{\sigma_{22}}\frac{\phi\left(\omega_{2p}\right)}{\Phi\left(\omega_{2p}\right)} \tag{31}$$

i.e. it would correctly censor y_2 to the left, but attribute the censoring rule to the same regressors as those that affect the conditional mean of the outcome variable.

The third construct that can be of interest in some applications is the expected value of the *latent* variable for the second equation, i.e.

$$E\left[y_{2p}^{*} \mid \mathbf{x_{2p}}, \boldsymbol{\beta}_{\mathbf{2}} \right] = \omega_{2p} \tag{32}$$

Naturally, a naïve regression model or a Tobit would produce the same expression, but would suffer from biased posterior results for $\boldsymbol{\beta}_{\mathbf{2}}$.

The fourth construct of occasional interest is the marginal effect of a given regressor on the conditional expectation. There are three separate expressions, depending if the regressor appears only in the first, only in the second, or both equations. The corresponding marginal effects (evaluated at the sample mean $\bar{\mathbf{x}}$) are as follows:

$$\begin{aligned}
&\frac{\partial E\left[y_2 \mid y_1^{*} > 0, \bar{\mathbf{x}}, \boldsymbol{\beta}, \boldsymbol{\Sigma} \right]}{\partial\left(x_{1,j} \right)} = -\beta_{1,j}\sigma_{12}\frac{\phi\left(\bar{\omega}_1\right)}{\Phi\left(\bar{\omega}_1\right)}\left(\frac{\phi\left(\bar{\omega}_1\right)}{\Phi\left(\bar{\omega}_1\right)} + \bar{\omega}_1 \right) \\
&\frac{\partial E\left[y_2 \mid y_1^{*} > 0, \bar{\mathbf{x}}, \boldsymbol{\beta}, \boldsymbol{\Sigma} \right]}{\partial\left(x_{2,j} \right)} = \beta_{2,j} \\
&\frac{\partial E\left[y_2 \mid y_1^{*} > 0, \bar{\mathbf{x}}, \boldsymbol{\beta}, \boldsymbol{\Sigma} \right]}{\partial\left(x_{1,j}, x_{2j} \right)} = \beta_{2,j} - \beta_{1,j}\sigma_{12}\frac{\phi\left(\bar{\omega}_1\right)}{\Phi\left(\bar{\omega}_1\right)}\left(\frac{\phi\left(\bar{\omega}_1\right)}{\Phi\left(\bar{\omega}_1\right)} + \bar{\omega}_1 \right) \qquad \text{with} \qquad \bar{\omega}_1 = \bar{\mathbf{x}}_{\mathbf{1}}'\boldsymbol{\beta}_{\mathbf{1}}
\end{aligned} \tag{33}$$

Naturally, a Tobit model can only produce $\dfrac{\partial E\left[y_p \mid y_p^{*} > 0, \bar{\mathbf{x}}, \boldsymbol{\beta}, \sigma^2 \right]}{\partial x_j} = \beta_j \Phi\left(\frac{\bar{\mathbf{x}}'\boldsymbol{\beta}}{\sigma} \right)$.

For 0/1 regressors, use

$$\begin{aligned}
&E\left[y_2 \mid y_1^{*} > 0, \bar{\mathbf{x}}_{-j}, x_j = 1, \boldsymbol{\beta}, \boldsymbol{\Sigma} \right] - E\left[y_2 \mid y_1^{*} > 0, \bar{\mathbf{x}}_{-j}, x_j = 0, \boldsymbol{\beta}, \boldsymbol{\Sigma} \right] = \\
&\left(\bar{\omega}_2 \mid x_{2,j} = 1 \right) + \sigma_{12}\frac{\phi\left(\bar{\omega}_1 \mid x_{1,j} = 1 \right)}{\Phi\left(\omega_1 \mid x_{1,j} = 1 \right)} - \left(\bar{\omega}_2 \mid x_{2,j} = 0 \right) + \sigma_{12}\frac{\phi\left(\bar{\omega}_1 \mid x_{1,j} = 0 \right)}{\Phi\left(\omega_1 \mid x_{1,j} = 0 \right)}
\end{aligned} \tag{34}$$

In contrast, for the Tobit we used

$$E\left(y_p \mid y_p^* > 0, \overline{\mathbf{x}}_{-j}, x_j = 1\right) - E\left(y_p \mid y_p^* > 0, \overline{\mathbf{x}}_{-j}, x_j = 0\right) =$$
$$\left(\overline{\mathbf{x}}'\boldsymbol{\beta} + \sigma \frac{\phi\left(\frac{\overline{\mathbf{x}}'\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{\overline{\mathbf{x}}'\boldsymbol{\beta}}{\sigma}\right)} \mid x_j = 1\right) - \left(\overline{\mathbf{x}}'\boldsymbol{\beta} + \sigma \frac{\phi\left(\frac{\overline{\mathbf{x}}'\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{\overline{\mathbf{x}}'\boldsymbol{\beta}}{\sigma}\right)} \mid x_j = 0\right) \tag{35}$$

Matlab implementation:
Scripts `mod9_select_data,` and `mod9_selec` implement this model using simulated data. An empirical application to door-t0-door fundraising is given in script `mod9_select_door2door.`

Treatment – Effect Model

The treatment-effect (TE) model has a similar structure as the Sample-Selection (SS) model. It focuses on the causal impact of a binary "treatment" on a continuous outcome. So to start, our main equation of interest can be written as

$$y_i = \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + \gamma T_i + \varepsilon_{2i} \tag{36}$$

where $T_i = 1$ if individual *i* received the treatment, and 0 otherwise. If T_i was truly exogenous, we could stop right here and estimate the model as a generic normal regression. An exogenous treatment would be any outside intervention that is imposed on a given individual, rather than her choosing it. For example, a letter sent to randomly chosen households reminding them of designated watering days would be such a treatment. Or randomly choosing some wild horses to be trained before releasing them for adoption would be another.

However, in many cases the treatment is chosen by the individual herself. Examples are: Decision to enter a job training program, decision to start / quit smoking, decision to engage in criminal behavior, choice of schooling, etc. In these situations there could be the problem of *unobserved confounding*: If there exists an unobserved effect or characteristic that drives an individual towards a treatment decision, and that is correlated with unobservables in the outcome equation, our measure of the causal treatment effect (i.e. our estimate of γ) would be biased. A classic example would be that an individual with more "spunk" is more likely to enter a training program, but would have likely performed above average without the program as well (as measured by the dependent variable in the outcome equation). If mostly "talented" individuals take the training, and mostly "untalented" don't, the causal impact of the treatment would be overestimated.

Econometrically, we can model this as

$$
\begin{aligned}
&T_i^* = \mathbf{x}_{1i}'\boldsymbol{\beta}_1 + \varepsilon_{1i} \\
&y_i = \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + \gamma T_i + \varepsilon_{2i} \\
&\begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{bmatrix} \sim n\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} \right) \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \\
&T_i = 1 \qquad \text{if} \qquad T_i^* > 0, \qquad = 0 \text{ otherwise}
\end{aligned} \tag{37}
$$

Note that we observe y_i regardless of that individual's treatment decision. Thus, in contrast to the SS model, data attrition via a "first hurdle" is not an issue in the TE model. We will also assume that all regressors are observed, regardless of the treatment decision. At the individual level we can write the model as

$$
\begin{aligned}
&\mathbf{y}_\mathbf{i}^* = \mathbf{X}_\mathbf{i}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_\mathbf{i} \qquad \text{with} \qquad \boldsymbol{\varepsilon}_\mathbf{i} \sim n(\mathbf{0}, \boldsymbol{\Sigma}), \\
&\mathbf{y}_\mathbf{i}^* = \begin{bmatrix} T_i^* \\ y_i \end{bmatrix} \quad \mathbf{X}_\mathbf{i} = \begin{bmatrix} \mathbf{x}_{1\mathbf{i}}' & \mathbf{0} & 0 \\ \mathbf{0} & \mathbf{x}_{2\mathbf{i}}' & T_i \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \gamma \end{bmatrix} \qquad \boldsymbol{\varepsilon}_\mathbf{i} = \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{bmatrix}
\end{aligned} \tag{38}
$$

For the full sample of n individuals the model can be written as

$$
\begin{aligned}
&\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\
&\mathbf{y}^* = \underset{(nx2)x1}{\begin{bmatrix} \mathbf{y}_1^* \\ \mathbf{y}_2^* \\ \vdots \\ \mathbf{y}_\mathbf{n}^* \end{bmatrix}} \qquad \mathbf{X} = \underset{(nx2)xk}{\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_\mathbf{n} \end{bmatrix}} \qquad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \gamma \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_\mathbf{n} \end{bmatrix} \sim n(\mathbf{0}, \boldsymbol{\Omega}) \qquad \text{with} \\
&\boldsymbol{\Omega} = \underset{(nx2)x(nx2)}{\begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \cdots & \mathbf{0} \\ \mathbf{0} & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma} \end{bmatrix}}
\end{aligned} \tag{39}
$$

For our sample likelihood there are again *two possible observable outcomes*, $pr(T_i = 1, y_i)$ and $pr(T_i = 0, y_i)$. Building on our insights from the SS model, we can express these outcomes as

$$
\begin{aligned}
&pr(T_i = 1, y_i) = f\left(\varepsilon_{1i} > -\mathbf{x}_{1\mathbf{i}}'\boldsymbol{\beta}_\mathbf{1}, \varepsilon_{2i}\right) = \\
&f(\varepsilon_{2i}) \int_{-\mathbf{x}_{1\mathbf{i}}'\boldsymbol{\beta}_\mathbf{1}}^{\infty} f(\varepsilon_{1i} \mid \varepsilon_{2i})\, d\varepsilon_{1i} = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\tfrac{1}{2\sigma_2^2}\left(y_i - \mathbf{x}_{2\mathbf{i}}'\boldsymbol{\beta}_\mathbf{2} - \gamma\right)^2 \right) \Phi\left(\frac{\mathbf{x}_{1\mathbf{i}}'\boldsymbol{\beta}_\mathbf{1} + \frac{\sigma_{12}}{\sigma_2^2}\left(y_i - \mathbf{x}_{2\mathbf{i}}'\boldsymbol{\beta}_\mathbf{2} - \gamma\right)}{\sqrt{1-\rho^2}} \right) \\
&\rho = \frac{\sigma_{12}}{\sqrt{\sigma_2^2}}
\end{aligned} \tag{40}
$$

and

$$pr\left(T_i=0,y_i\right)=f\left(\varepsilon_{1i}<-\mathbf{x}_{\mathbf{1i}}'\boldsymbol{\beta}_{\mathbf{1}},\varepsilon_{2i}\right)=$$
$$f\left(\varepsilon_{2i}\right)\int_{-\infty}^{-\mathbf{x}_{\mathbf{1i}}'\boldsymbol{\beta}_{\mathbf{1}}}f\left(\varepsilon_{1i}\mid\varepsilon_{2i}\right)d\varepsilon_{1i}=\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\tfrac{1}{2\sigma_2^2}\left(y_i-\mathbf{x}_{\mathbf{2i}}'\boldsymbol{\beta}_{\mathbf{2}}\right)^2\right)\Phi\left(\frac{-\mathbf{x}_{\mathbf{1i}}'\boldsymbol{\beta}_{\mathbf{1}}-\frac{\sigma_{12}}{\sigma_2^2}\left(y_i-\mathbf{x}_{\mathbf{2i}}'\boldsymbol{\beta}_{\mathbf{2}}\right)}{\sqrt{1-\rho^2}}\right) \tag{41}$$

We can now write the sample likelihood concisely as

$$p\left(\mathbf{y}\mid\boldsymbol{\beta},\boldsymbol{\Sigma}\right)=\prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\tfrac{1}{2\sigma_2^2}\left(y_{2i}-\mathbf{x}_{\mathbf{2i}}'\boldsymbol{\beta}_{\mathbf{2}}-\gamma\right)^2\right)\Phi\left(\frac{\mathbf{x}_{\mathbf{1i}}'\boldsymbol{\beta}_{\mathbf{1}}+\frac{\sigma_{12}}{\sigma_2^2}\left(y_{2i}-\mathbf{x}_{\mathbf{2i}}'\boldsymbol{\beta}_{\mathbf{2}}-\gamma\right)}{\sqrt{1-\rho^2}}\right)\right)^{T_i}*$$
$$\left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\tfrac{1}{2\sigma_2^2}\left(y_{2i}-\mathbf{x}_{\mathbf{2i}}'\boldsymbol{\beta}_{\mathbf{2}}\right)^2\right)\Phi\left(\frac{-\mathbf{x}_{\mathbf{1i}}'\boldsymbol{\beta}_{\mathbf{1}}-\frac{\sigma_{12}}{\sigma_2^2}\left(y_{2i}-\mathbf{x}_{\mathbf{2i}}'\boldsymbol{\beta}_{\mathbf{2}}\right)}{\sqrt{1-\rho^2}}\right)\right)^{1-T_i} \tag{42}$$

Using the same priors as for the SS model the augmented posterior kernel takes the following form:

$$p\left(\boldsymbol{\beta},\boldsymbol{\Sigma},\mathbf{y}^*\mid\mathbf{y},\mathbf{X}\right)\propto$$
$$\left|\boldsymbol{\Sigma}\right|^{-\left(\nu_0+3+n\right)/2}\exp\left(-\frac{1}{2}tr\left(\mathbf{S_0}\cdot\boldsymbol{\Sigma}^{-1}\right)\right)I\left(\Sigma_{11}=1\right)*\exp\left(-\tfrac{1}{2}\left(\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathbf{0}}\right)'\mathbf{V_0^{-1}}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathbf{0}}\right)\right)\right)*$$
$$\exp\left(-\tfrac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)\right)* \tag{43}$$
$$\prod_{i=1}^{n}\left(I\left(T_i=0,y_i\right)I\left(T_i^*<0,y_i\right)+I\left(T_i=1,y_i\right)I\left(T_i^*>0,y_i\right)\right)$$

The conditional posteriors for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ take the exact same form as for the SS model. The conditional posterior for the latent data is given as

$$p\left(\mathbf{y}^*\mid\boldsymbol{\beta},\boldsymbol{\Sigma},\mathbf{y},\mathbf{X}\right)\propto$$
$$\exp\left(-\tfrac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)\right)* \tag{44}$$
$$\prod_{i=1}^{n}\left(I\left(T_i=0,y_i\right)I\left(T_i^*<0,y_i\right)+I\left(T_i=1,y_i\right)I\left(T_i^*>0,y_i\right)\right)$$

At the individual level this implies

$$p\left(\mathbf{y_i^*} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{y}, \mathbf{X_i}\right) \propto$$
$$\exp\left(-\tfrac{1}{2}\left(\mathbf{y_i^*} - \mathbf{X_i}\boldsymbol{\beta}\right)' \boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i^*} - \mathbf{X_i}\boldsymbol{\beta}\right)\right) * \tag{45}$$
$$I\left(T_i = 0, y_i\right) I\left(T_i^* < 0, y_i\right) + I\left(T_i = 1, y_i\right) I\left(T_i^* > 0, y_i\right)$$

For the $T_i = 0$ case we have

$$p\left(\mathbf{y_i^*} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, T_i = 0, \mathbf{X_i}\right) \propto$$
$$\exp\left(-\tfrac{1}{2}\left(\mathbf{y_i^*} - \mathbf{X_i}\boldsymbol{\beta}\right)' \boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i^*} - \mathbf{X_i}\boldsymbol{\beta}\right)\right) * I\left(T_i^* < 0, y_i\right) \tag{46}$$

This implies a draw of $\mathbf{y_i^*} = \begin{bmatrix} T_i^* & y_i \end{bmatrix}'$ from a joint normal density, with T_i^* truncated to the negative domain, i.e. from

$$p\left(T_i^* < 0, y_i \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right) = \int_{-\infty}^{0} f\left(T_i^*, y_i \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right) d\,T_i^* \tag{47}$$

Where *f (.)* denotes the bivariate normal density. This can be computationally implemented as follows:

Using $p\left(T_i^* < 0, y_i \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right) = p\left(T_i^* < 0 \mid y_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right) p\left(y_i \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right) = p\left(T_i^* < 0 \mid y_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right)$, since y_i is observed. Thus, we simply draw T_i^* from its conditional normal density with mean $\mathbf{x'_{1i}}\boldsymbol{\beta_1} + \frac{\sigma_{12}}{\sigma_2^2}\left(y_i - \mathbf{x'_{2i}}\boldsymbol{\beta_2}\right)$ and variance $1 - \rho^2$, *truncated from above at zero.*

Similarly, for the $T_i = 1$ case we have

$$p\left(\mathbf{y_i^*} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, T_i = 1, \mathbf{X_i}\right) \propto$$
$$\exp\left(-\tfrac{1}{2}\left(\mathbf{y_i^*} - \mathbf{X_i}\boldsymbol{\beta}\right)' \boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i^*} - \mathbf{X_i}\boldsymbol{\beta}\right)\right) * I\left(T_i^* > 0, y_i\right) \tag{48}$$

This implies a draw of $\mathbf{y_i^*} = \begin{bmatrix} T_i^* & y_i \end{bmatrix}'$ from a joint normal density, with T_i^* truncated to the positive domain, i.e. from

$$p\left(T_i^* > 0, y_i \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right) = \int_{0}^{\infty} f\left(T_i^*, y_i \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right) d\,T_i^* \tag{49}$$

This can be computationally implemented by simply drawing T_i^* from its conditional normal density with mean $\mathbf{x'_{1i}}\boldsymbol{\beta_1} + \frac{\sigma_{12}}{\sigma_2^2}\left(y_i - \mathbf{x'_{2i}}\boldsymbol{\beta_2} - \gamma\right)$ and variance $1 - \rho^2$, *truncated from below at zero.*

Matlab implementation:
See Matlab scripts / functions `mod9_treat_data` , `mod9_treat`, and `gs_treat` for the implementation of this model using simulated data. An empirical example using a popular job training data set is given in scripts `mod9_treatJobTrain` and `mod9_treatJobTrainNaive`.

Switching Regression Model

This model is closely related to the previous two. It is also known as the "Roy Model" (after Roy, 1951) and "type 5 Tobit" (as listed in Amemiya, 1985). It is a three equation model, where the first equation determines an endogenous treatment or selection outcome, and the remaining two equations describe the main outcome of interest for the treated and untreated sample, respectively. Thus, we now allow the outcome equation to differ (in terms of explanatory variables and / or coefficients) depending on the treatment decision.

Econometrically, this model is interesting since it shows how in Bayesian analysis we can sometimes learn about a parameter that is not directly identified.

The model can be written as

$$
\begin{aligned}
y_{1i}^{*} &= \mathbf{x_{1i}'}\boldsymbol{\beta_1} + \varepsilon_{1i} \\
y_{2i}^{*} &= \mathbf{x_{2i}'}\boldsymbol{\beta_2} + \varepsilon_{2i} \\
y_{3i}^{*} &= \mathbf{x_{3i}'}\boldsymbol{\beta_3} + \varepsilon_{3i} \\
\begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \end{bmatrix} &\sim n\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} \right) \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{bmatrix} \\
y_{1i} &= 1 \quad \text{if} \quad y_{1i}^{*} > 0, \quad = 0 \text{ otherwise} \\
y_{2i} &= y_{2i}^{*} \quad \text{if} \quad y_{1i} = 0 \quad = \text{unobserved otherwise} \\
y_{3i} &= y_{3i}^{*} \quad \text{if} \quad y_{1i} = 1 \quad = \text{unobserved otherwise}
\end{aligned}
\tag{50}
$$

The new element is that different outcome equations apply to different individuals. As for the two preceding models, the first equation must contain at least one regressor that is not included in the other two equations for identification purposes. As before, we assume that explanatory variables are observed for all individuals and equations, regardless of the first hurdle result. Although not necessary, most existing applications set $\mathbf{x_{2i}} = \mathbf{x_{3i}}$.

Learning about σ_{23}

Since *either* equ. 2 *or* equ. 3 apply to a given individual (i.e. a given individual cannot be treated AND untreated at the same time), the covariance between equations 2 and 3 (σ_{23}) does not enter the sample

likelihood, i.e. it is not identified. Yet, we will be able to derive its posterior density, based on information through priors (weak, if priors are vague) and – due to the error correlation –information via remaining model parameters.

The details about this indirect learning are given in Lie et al (2004). I will synthesize their argument:

First note that for $\boldsymbol{\Sigma}$ to be a well-behaved, positive definite variance matrix, we need

$$\left|\boldsymbol{\Sigma}\right| = \sigma_{22}\sigma_{33}\left(\left(1-\frac{\sigma_{12}^2}{\sigma_{22}}\right)\left(1-\frac{\sigma_{13}^2}{\sigma_{33}}\right)-\left(\frac{\sigma_{23}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{33}}}-\frac{\sigma_{12}\sigma_{13}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{33}}}\right)^2\right)>0 \qquad \rightarrow$$

$$\left(1-\frac{\sigma_{12}^2}{\sigma_{22}}\right)\left(1-\frac{\sigma_{13}^2}{\sigma_{33}}\right)>\left(\frac{\sigma_{23}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{33}}}-\frac{\sigma_{12}\sigma_{13}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{33}}}\right)^2 \qquad \rightarrow \qquad (51)$$

$$\sigma_{12}\sigma_{13}-\sqrt{\sigma_{22}}\sqrt{\sigma_{33}}\left(\left(1-\frac{\sigma_{12}^2}{\sigma_{22}}\right)\left(1-\frac{\sigma_{13}^2}{\sigma_{33}}\right)\right)^{1/2}<\sigma_{23}<\sigma_{12}\sigma_{13}+\sqrt{\sigma_{22}}\sqrt{\sigma_{33}}\left(\left(1-\frac{\sigma_{12}^2}{\sigma_{22}}\right)\left(1-\frac{\sigma_{13}^2}{\sigma_{33}}\right)\right)^{1/2}$$

Thus, conditional on the other elements of $\boldsymbol{\Sigma}$ the unidentified covariance is bound between

$$\underline{\sigma}_{23} = \sigma_{12}\sigma_{13}-\sqrt{\sigma_{22}}\sqrt{\sigma_{33}}\left(\left(1-\frac{\sigma_{12}^2}{\sigma_{22}}\right)\left(1-\frac{\sigma_{13}^2}{\sigma_{33}}\right)\right)^{1/2} \qquad \text{and}$$

$$\bar{\sigma}_{23} = \sigma_{12}\sigma_{13}+\sqrt{\sigma_{22}}\sqrt{\sigma_{33}}\left(\left(1-\frac{\sigma_{12}^2}{\sigma_{22}}\right)\left(1-\frac{\sigma_{13}^2}{\sigma_{33}}\right)\right)^{1/2} \qquad (52)$$

As described in Li et al. the other model parameters and the data add no further contribution to learning about σ_{23} once we condition its posterior on the other elements of $\boldsymbol{\Sigma}$, i.e.

$$p\left(\sigma_{23}\,|\,\boldsymbol{\Sigma}_{\text{-}23},\boldsymbol{\theta},\mathbf{y},\mathbf{X}\right) = p\left(\sigma_{23}\,|\,\boldsymbol{\Sigma}_{\text{-}23}\right) \qquad (53)$$

where $\boldsymbol{\Sigma}_{\text{-}23}$ captures all other terms in $\boldsymbol{\Sigma}$ and $\boldsymbol{\theta}$ collects all other model parameters. Now define $R\left(\sigma_{23}\right)$ the joint parameter space of $\boldsymbol{\Sigma}_{\text{-}23}$, given σ_{23}, that assures that $\boldsymbol{\Sigma}$ remains positive definite. We can then write the marginal posterior of σ_{23} as:

$$
\begin{aligned}
&p(\sigma_{23}\,|\,\mathbf{y},\mathbf{X}) = \int_{R(\boldsymbol{\Sigma}_{-23})}\int_{\boldsymbol{\theta}} p(\sigma_{23},\boldsymbol{\Sigma}_{-23},\boldsymbol{\theta}\,|\,\mathbf{y},\mathbf{X})\,d\boldsymbol{\theta}\,d\,\boldsymbol{\Sigma}_{-23} = \\
&\int_{R(\boldsymbol{\Sigma}_{-23})}\int_{\boldsymbol{\theta}} p(\sigma_{23}\,|\,\boldsymbol{\Sigma}_{-23},\boldsymbol{\theta},\mathbf{y},\mathbf{X})\;p(\boldsymbol{\Sigma}_{-23},\boldsymbol{\theta}\,|\,\mathbf{y},\mathbf{X})\,d\boldsymbol{\theta}\,d\,\boldsymbol{\Sigma}_{-23} = \\
&\int_{R(\boldsymbol{\Sigma}_{-23})}\int_{\boldsymbol{\theta}} p(\sigma_{23}\,|\,\boldsymbol{\Sigma}_{-23})\;p(\boldsymbol{\Sigma}_{-23},\boldsymbol{\theta}\,|\,\mathbf{y},\mathbf{X})\,d\boldsymbol{\theta}\,d\,\boldsymbol{\Sigma}_{-23} = \\
&\int_{R(\boldsymbol{\Sigma}_{-23})} p(\sigma_{23}\,|\,\boldsymbol{\Sigma}_{-23})\,p(\boldsymbol{\Sigma}_{-23}\,|\,\mathbf{y},\mathbf{X})\left(\int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}\,|\,\boldsymbol{\Sigma}_{-23},\mathbf{y},\mathbf{X})\,d\boldsymbol{\theta}\right)\,d\,\boldsymbol{\Sigma}_{-23} = \\
&\int_{R(\boldsymbol{\Sigma}_{-23})} p(\sigma_{23}\,|\,\boldsymbol{\Sigma}_{-23})\,p(\boldsymbol{\Sigma}_{-23}\,|\,\mathbf{y},\mathbf{X})\;d\,\boldsymbol{\Sigma}_{-23}
\end{aligned}
\tag{54}
$$

where we use result (53) in the third line. The final result shows that the marginal posterior of σ_{23} can be expressed as a function solely of the remaining elements in $\boldsymbol{\Sigma}$. Furthermore, if the joint posterior of $\boldsymbol{\Sigma}_{-23}$ is highly informative or 'tight ' around some point $\hat{\boldsymbol{\Sigma}}_{-23}$, then

$$
\begin{aligned}
&p(\sigma_{23}\,|\,\mathbf{y},\mathbf{X}) \approx p\left(\sigma_{23}\,|\,\hat{\boldsymbol{\Sigma}}_{-23}\right) \qquad or \qquad pr\left(\underline{\hat{\sigma}}_{23} \le \sigma_{23} \le \bar{\hat{\sigma}}_{23}\,|\,\mathbf{y},\mathbf{X}\right) \approx 1 \qquad \text{where} \\
&\underline{\hat{\sigma}}_{23} = \underline{\sigma}_{23}\,|\,\boldsymbol{\Sigma}_{-23} = \hat{\boldsymbol{\Sigma}}_{-23}, \quad \bar{\hat{\sigma}}_{23} = \bar{\sigma}_{23}\,|\,\boldsymbol{\Sigma}_{-23} = \hat{\boldsymbol{\Sigma}}_{-23}
\end{aligned}
\tag{55}
$$

are posterior estimates of the unidentified covariances given in (52).

Thus, as stated in Li et al, information conveyed from the data on the identified elements of $\boldsymbol{\Sigma}$"spills over" and revises our belief about the conditional support of σ_{23}. In other words, the marginal prior and posterior for σ_{23} will generally differ, suggesting that learning has taken place. This can be easily verified by plotting both prior and posterior after running the Gibbs Sampler, as we will show below. Poirier (1998) provides other examples of learning about non-identified parameters.

Estimation

Before we work out the details for our Gibbs Sampler we need to review the formulas for bi-conditional normal densities, i.e. normal densities for a random variable, conditioned on *two* other normal variates. These densities will be needed to draw latent data in the Gibbs Sampler: Assume random vector $\mathbf{x}$ is multinormal with mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$. Partition $\mathbf{x}$ arbitrarily into sub-vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, and $\boldsymbol{\Sigma}$ conformably into $\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}'_{12} \\ \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$. As shown e.g. in Greene, Theorem B.7, p. 1014, the conditional densities for each sub-vector are then given as:

$$
\begin{aligned}
&\mathbf{x_1} \mid \mathbf{x_2} \sim n\left(\boldsymbol{\mu}_{\mathbf{1.2}}, \boldsymbol{\Sigma}_{\mathbf{1.2}}\right) && \mathbf{x_2} \mid \mathbf{x_1} \sim n\left(\boldsymbol{\mu}_{\mathbf{2.1}}, \boldsymbol{\Sigma}_{\mathbf{2.1}}\right) \\
&\boldsymbol{\mu}_{\mathbf{1.2}} = \boldsymbol{\mu}_{\mathbf{1}} + \boldsymbol{\Sigma}_{\mathbf{12}}' \boldsymbol{\Sigma}_{\mathbf{22}}^{\mathbf{-1}} \left(\mathbf{x_2} - \boldsymbol{\mu}_{\mathbf{2}}\right) && \boldsymbol{\mu}_{\mathbf{2.1}} = \boldsymbol{\mu}_{\mathbf{2}} + \boldsymbol{\Sigma}_{\mathbf{12}}' \boldsymbol{\Sigma}_{\mathbf{22}}^{\mathbf{-1}} \left(\mathbf{x_1} - \boldsymbol{\mu}_{\mathbf{1}}\right) \\
&\boldsymbol{\Sigma}_{\mathbf{1.2}} = \boldsymbol{\Sigma}_{\mathbf{11}} - \boldsymbol{\Sigma}_{\mathbf{12}}' \boldsymbol{\Sigma}_{\mathbf{22}}^{\mathbf{-1}} \boldsymbol{\Sigma}_{\mathbf{12}} && \boldsymbol{\Sigma}_{\mathbf{2.1}} = \boldsymbol{\Sigma}_{\mathbf{22}} - \boldsymbol{\Sigma}_{\mathbf{12}} \boldsymbol{\Sigma}_{\mathbf{22}}^{\mathbf{-1}} \boldsymbol{\Sigma}_{\mathbf{12}}'
\end{aligned}
\tag{56}
$$

For our GS we will require the conditional densities for $y_2^* \mid \begin{bmatrix} y_1^* \\ y_3^* \end{bmatrix} = y_2^* \mid \mathbf{y}_{\mathbf{13}}^*$, and $y_3^* \mid \begin{bmatrix} y_1^* \\ y_2^* \end{bmatrix} = y_3^* \mid \mathbf{y}_{\mathbf{12}}^*$. For completeness, I also added the density for $y_1^* \mid \begin{bmatrix} y_2^* \\ y_3^* \end{bmatrix} = y_1^* \mid \mathbf{y}_{\mathbf{23}}^*$.

Let

$$
\begin{aligned}
&E\left(y_1^*\right) = \mathbf{x_1'}\boldsymbol{\beta}_{\mathbf{1}} = \omega_1, && E\left(y_2^*\right) = \mathbf{x_2'}\boldsymbol{\beta}_{\mathbf{2}} = \omega_2, && E\left(y_3^*\right) = \mathbf{x_3'}\boldsymbol{\beta}_{\mathbf{3}} = \omega_3, \\
&E\left(\mathbf{y}_{\mathbf{12}}^*\right) = \begin{bmatrix} \omega_1 \\ \omega_2 \end{bmatrix} = \boldsymbol{\omega}_{\mathbf{12}}, && E\left(\mathbf{y}_{\mathbf{13}}^*\right) = \begin{bmatrix} \omega_1 \\ \omega_3 \end{bmatrix} = \boldsymbol{\omega}_{\mathbf{13}}, && E\left(\mathbf{y}_{\mathbf{23}}^*\right) = \begin{bmatrix} \omega_2 \\ \omega_3 \end{bmatrix} = \boldsymbol{\omega}_{\mathbf{23}}, \\
&\boldsymbol{\Sigma}_{\mathbf{1r}} = \begin{bmatrix} \sigma_{12} \\ \sigma_{13} \end{bmatrix}, && \boldsymbol{\Sigma}_{\mathbf{2r}} = \begin{bmatrix} \sigma_{12} \\ \sigma_{23} \end{bmatrix}, && \boldsymbol{\Sigma}_{\mathbf{3r}} = \begin{bmatrix} \sigma_{13} \\ \sigma_{23} \end{bmatrix} \\
&\boldsymbol{\Sigma}_{\mathbf{12}} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} && \boldsymbol{\Sigma}_{\mathbf{13}} = \begin{bmatrix} \sigma_{11} & \sigma_{13} \\ \sigma_{13} & \sigma_{33} \end{bmatrix} && \boldsymbol{\Sigma}_{\mathbf{23}} = \begin{bmatrix} \sigma_{22} & \sigma_{23} \\ \sigma_{23} & \sigma_{33} \end{bmatrix}
\end{aligned}
\tag{57}
$$

We can then derive the required conditional densities as

$$
\begin{aligned}
&y_1^* \mid \mathbf{y}_{\mathbf{23}}^* \sim n\left(\omega_{1.23}, \sigma_{1.23}\right) && \text{with} \\
&\omega_{1.23} = \omega_1 + \boldsymbol{\Sigma}_{\mathbf{1r}}' \boldsymbol{\Sigma}_{\mathbf{23}}^{\mathbf{-1}} \left(\mathbf{y}_{\mathbf{23}}^* - \boldsymbol{\omega}_{\mathbf{23}}\right) && \sigma_{1.23} = 1 - \boldsymbol{\Sigma}_{\mathbf{1r}}' \boldsymbol{\Sigma}_{\mathbf{23}}^{\mathbf{-1}} \boldsymbol{\Sigma}_{\mathbf{1r}} \\
&y_2^* \mid \mathbf{y}_{\mathbf{13}}^* \sim n\left(\omega_{2.13}, \sigma_{2.13}\right) && \text{with} \\
&\omega_{2.13} = \omega_2 + \boldsymbol{\Sigma}_{\mathbf{2r}}' \boldsymbol{\Sigma}_{\mathbf{13}}^{\mathbf{-1}} \left(\mathbf{y}_{\mathbf{13}}^* - \boldsymbol{\omega}_{\mathbf{13}}\right) && \sigma_{2.13} = \sigma_{22} - \boldsymbol{\Sigma}_{\mathbf{2r}}' \boldsymbol{\Sigma}_{\mathbf{13}}^{\mathbf{-1}} \boldsymbol{\Sigma}_{\mathbf{2r}} \\
&y_3^* \mid \mathbf{y}_{\mathbf{12}}^* \sim n\left(\omega_{3.12}, \sigma_{3.12}\right) && \text{with} \\
&\omega_{3.12} = \omega_3 + \boldsymbol{\Sigma}_{\mathbf{3r}}' \boldsymbol{\Sigma}_{\mathbf{12}}^{\mathbf{-1}} \left(\mathbf{y}_{\mathbf{12}}^* - \boldsymbol{\omega}_{\mathbf{12}}\right) && \sigma_{3.12} = \sigma_{33} - \boldsymbol{\Sigma}_{\mathbf{3r}}' \boldsymbol{\Sigma}_{\mathbf{12}}^{\mathbf{-1}} \boldsymbol{\Sigma}_{\mathbf{3r}}
\end{aligned}
\tag{58}
$$

For a given individual, we can write the system as for the two preceding models, plus an additional equation:

$$
\begin{aligned}
&\mathbf{y_i^*} = \mathbf{X_i}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{\mathbf{i}} \quad \text{with} \quad \boldsymbol{\varepsilon}_{\mathbf{i}} \sim n\left(\mathbf{0}, \boldsymbol{\Sigma}\right), \\
&\mathbf{y_i^*} = \begin{bmatrix} y_{1i}^* \\ y_{2i}^* \\ y_{3i}^* \end{bmatrix} \quad \mathbf{X_i} = \begin{bmatrix} \mathbf{x_{1i}'} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{x_{2i}'} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{x_{3i}'} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_{\mathbf{1}} \\ \boldsymbol{\beta}_{\mathbf{2}} \\ \boldsymbol{\beta}_{\mathbf{2}} \end{bmatrix} \qquad \boldsymbol{\varepsilon}_{\mathbf{i}} = \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \end{bmatrix}
\end{aligned}
\tag{59}
$$

Also, building on the two previous model, it is straightforward to derive the likelihood function as

$$p(\mathbf{y}\,|\,\boldsymbol{\beta},\boldsymbol{\Sigma}) = \prod_{i=1}^{n}\left(f\left(\varepsilon_{1i} < -\mathbf{x}_{1i}'\boldsymbol{\beta}_1, \varepsilon_{2i}\right)\right)^{1-y_{1i}}\left(f\left(\varepsilon_{1i} > -\mathbf{x}_{1i}'\boldsymbol{\beta}_1, \varepsilon_{3i}\right)\right)^{y_{1i}} =$$

$$\prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi\sigma_{22}}}\exp\left(-\tfrac{1}{2\sigma_{22}}\left(y_{2i}-\mathbf{x}_{2i}'\boldsymbol{\beta}_2\right)^2\right)\Phi\left(\frac{-\mathbf{x}_{1i}'\boldsymbol{\beta}_1 - \frac{\sigma_{12}}{\sigma_{22}}\left(y_{2i}-\mathbf{x}_{2i}'\boldsymbol{\beta}_2\right)}{\sqrt{1-\rho_{12}^2}}\right)\right)^{1-y_{1i}} *$$

$$\left(\frac{1}{\sqrt{2\pi\sigma_{33}}}\exp\left(-\tfrac{1}{2\sigma_{33}}\left(y_{3i}-\mathbf{x}_{3i}'\boldsymbol{\beta}_3\right)^2\right)\Phi\left(\frac{\mathbf{x}_{1i}'\boldsymbol{\beta}_1 + \frac{\sigma_{12}}{\sigma_{33}}\left(y_{3i}-\mathbf{x}_{3i}'\boldsymbol{\beta}_3\right)}{\sqrt{1-\rho_{13}^2}}\right)\right)^{y_{1i}} \quad \text{with} \tag{60}$$

$$\rho_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_{22}}} \qquad \rho_{13} = \frac{\sigma_{13}}{\sqrt{\sigma_{33}}}$$

We will augment the model with latent data for all three equations. The unobserved latent data for equations two and three is also often referred to as "missing data".

Using the same priors as for the SS and TE models the augmented posterior kernel takes the following form:

$$p\left(\boldsymbol{\beta},\boldsymbol{\Sigma},\mathbf{y}^*\,|\,\mathbf{y},\mathbf{X}\right) \propto$$

$$|\boldsymbol{\Sigma}|^{-(\nu_0+4+n)/2}\exp\left(-\frac{1}{2}tr\left(\mathbf{S_0}\cdot\boldsymbol{\Sigma}^{-1}\right)\right)I\left(\Sigma_{11}=1\right)*\exp\left(-\tfrac{1}{2}\left(\left(\boldsymbol{\beta}-\boldsymbol{\mu}_0\right)'\mathbf{V_0^{-1}}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_0\right)\right)\right)*$$

$$\exp\left(-\tfrac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)\right)* \tag{61}$$

$$\prod_{i=1}^{n}\left(I\left(y_{1i}=0,y_{2i}=y_{2i}^*\right)I\left(y_{1i}^*<0,y_{2i}^*\right)+I\left(y_{1i}=1,y_{3i}=y_{3i}^*\right)I\left(y_{1i}^*>0,y_{3i}^*\right)\right)$$

The conditional posteriors for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ take the exact same form as for the SS model (except for the "4" instead of "3" in $|\boldsymbol{\Sigma}|^{-(\nu_0+4+n)/2}$). The conditional posterior for the latent data is given as

$$p\left(\mathbf{y}^*\,|\,\boldsymbol{\beta},\boldsymbol{\Sigma},\mathbf{y},\mathbf{X}\right) \propto$$

$$\exp\left(-\tfrac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)\right)* \tag{62}$$

$$\prod_{i=1}^{n}\left(I\left(y_{1i}=0,y_{2i}=y_{2i}^*\right)I\left(y_{1i}^*<0,y_{2i}^*\right)+I\left(y_{1i}=1,y_{3i}=y_{3i}^*\right)I\left(y_{1i}^*>0,y_{3i}^*\right)\right)$$

At the individual level this implies

$$
\begin{aligned}
&p\left(\mathbf{y_i^*} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{y}, \mathbf{X_i}\right) \propto \\
&\exp\left(-\tfrac{1}{2}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)' \boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)\right)* \\
&I\left(y_{1i}=0, y_{2i}=y_{2i}^*\right) I\left(y_{1i}^*<0, y_{2i}^*\right)+I\left(y_{1i}=1, y_{3i}=y_{3i}^*\right) I\left(y_{1i}^*>0, y_{3i}^*\right)
\end{aligned} \tag{63}
$$

For the $y_{i1}=0$ case we have

$$
\begin{aligned}
&p\left(\mathbf{y_i^*} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathrm{y}_{1\mathrm{i}}=0, \mathbf{X_i}\right) \propto \\
&\exp\left(-\tfrac{1}{2}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)' \boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)\right)* I\left(y_{1i}^*<0, y_{2i}^*\right)
\end{aligned} \tag{64}
$$

Thus, we need draws of $\mathbf{y_i^*}=\left[y_{1i}^* \quad y_{2i}^* \quad y_{3i}^*\right]'$ conditional on: (i) $y_{1i}^*<0$ and (ii) y_{2i}^* is observed.

In other words, we need to draw $\left[y_{1i}^* \quad y_{3i}^*\right]' \mid y_{2i}^*$ with y_{i1}^* truncated to the negative domain, i.e. from

$$
p\left(y_{1i}^*<0, y_{3i}^* \mid y_{2i}^*, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right)=\int_{-\infty}^{0} f\left(y_{1i}^*, y_{3i}^* \mid y_{2i}^*, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right) dy_{1i}^* \tag{65}
$$

Where *f (.)* denotes the bivariate normal density. This can be computationally implemented as follows:

1. Using $p\left(y_{1i}^*<0, y_{3i}^* \mid y_{2i}^*, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right)=p\left(y_{3i}^* \mid y_{1i}^*<0, y_{2i}^*, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right) p\left(y_{1i}^*<0 \mid y_{2i}^*, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right)$, we can first draw y_{1i}^* from its conditioned-on y_{2i}^* normal density with mean $\mathbf{x_{1i}'\beta_1}+\frac{\sigma_{12}}{\sigma_{22}}\left(\mathbf{y_{2i}}-\mathbf{x_{2i}'\beta_2}\right)$ and variance $1-\rho_{12}^2$, truncated from above at zero.

2. We then draw $y_3^* \mid \mathbf{y_{12}^*}$ from its bi-conditional normal density, with moments given in (58).

For the $y_{1i}=1$ case we have

$$
\begin{aligned}
&p\left(\mathbf{y_i^*} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, y_{1i}=1, \mathbf{X_i}\right) \propto \\
&\exp\left(-\tfrac{1}{2}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)' \boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i^*}-\mathbf{X_i}\boldsymbol{\beta}\right)\right)* I\left(y_{1i}^*>0, y_{3i}^*\right)
\end{aligned} \tag{66}
$$

Thus, we need draws of $\mathbf{y_i^*}=\left[y_{1i}^* \quad y_{2i}^* \quad y_{3i}^*\right]'$ conditional on: (i) $y_{1i}^*>0$ and (ii) y_{3i}^* is observed.

In other words, we need to draw $\left[y_{1i}^* \quad y_{2i}^*\right]' \mid y_{3i}^*$ with y_{i1}^* truncated to the positive domain, i.e. from

$$p\left(y_{1i}^{*}>0, y_{2i}^{*} \mid y_{3i}^{*}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right)=\int_{0}^{\infty} f\left(y_{1i}^{*}, y_{2i}^{*} \mid y_{3i}^{*}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right) dy_{1i}^{*} \tag{67}$$

Where *f (.)* denotes the bivariate normal density. This can be computationally implemented as follows:

1. Using $p\left(y_{1i}^{*}>0, y_{2i}^{*} \mid y_{3i}^{*}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right)=p\left(y_{2i}^{*} \mid y_{1i}^{*}>0, y_{3i}^{*}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right) p\left(y_{1i}^{*}>0 \mid y_{3i}^{*}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X_i}\right)$, we can first draw y_{1i}^{*} from its conditioned-on y_{3i}^{*} normal density with mean $\mathbf{x_{1i}'\beta_1}-\frac{\sigma_{13}}{\sigma_{33}}\left(\mathbf{y_{3i}}-\mathbf{x_{3i}'\beta_3}\right)$ and variance $1-\rho_{13}^{2}$, truncated from below at zero.

2. We then draw $y_{2}^{*} \mid \mathbf{y_{13}^{*}}$ from its bi-conditional normal density, with moments given in (58).

Matlab examples: `mod9_switch_data, mod9_switch`

Post-Estimation

Two of the most important outcomes of interest in the program evaluation literature are the Average Treatment Effect (ATE) and the Treatment Effect on the Treated (TT). The first refers to the expected treatment effect for a randomly drawn person from the underlying population of interest. The second refers to the expected treatment effect on those who actually receive the treatment.

Let $\mathbf{x_p}$ denote some common predictive settings for regressors in equations 2 and 3, and let $\Delta_p=y_{3p}^{*}-y_{2p}^{*}=\mathbf{x_p'}\left(\boldsymbol{\beta_3}-\boldsymbol{\beta_2}\right)+\left(\varepsilon_3-\varepsilon_2\right)$. Then in a classical setting the Average Treatment Effect would be computed as

$$ATE_p=E\left(\Delta_p\right)=E\left(\mathbf{x_p'}\left(\boldsymbol{\beta_3}-\boldsymbol{\beta_2}\right)+\left(\varepsilon_3-\varepsilon_2\right)\right)=\mathbf{x_p'}\left(\boldsymbol{\beta_3}-\boldsymbol{\beta_2}\right) \tag{68}$$

To compute the (classical) TT, we need to introduce additional notation. Assume that regressors for the treatment equation are set to $\mathbf{z_p}$, s.t. $y_{1p}^{*}=\mathbf{z_p'\beta_1}+\varepsilon_1$. Also let $\Delta_\varepsilon=\varepsilon_3-\varepsilon_2$ with $E\left(\Delta_\varepsilon\right)=0, V\left(\Delta_\varepsilon\right)=\sigma_{2-3}^{2}=\sigma_{33}+\sigma_{22}-2\sigma_{23}$. Also note that $\operatorname{cov}\left(\varepsilon_1, \Delta_\varepsilon\right)=\sigma_{1,3-2}=\operatorname{cov}\left(\varepsilon_1,\left(\varepsilon_3-\varepsilon_2\right)\right)=\operatorname{cov}\left(\varepsilon_1, \varepsilon_3\right)-\operatorname{cov}\left(\varepsilon_1, \varepsilon_2\right)=\sigma_{13}-\sigma_{12}$.

Then

$$TT_p=E\left(\Delta_p \mid y_{1p}^{*}>0\right)=\mathbf{x_p'}\left(\boldsymbol{\beta_3}-\boldsymbol{\beta_2}\right)+E\left(\Delta_\varepsilon \mid \varepsilon_1<\mathbf{z_p'\beta_1}\right)=$$
$$\mathbf{x_p'}\left(\boldsymbol{\beta_3}-\boldsymbol{\beta_2}\right)+\sigma_{1,2-3}\frac{\phi\left(\mathbf{z_p'\beta_1}\right)}{\Phi\left(\mathbf{z_p'\beta_1}\right)} \tag{69}$$

Note that, not surprisingly, neither the expression for ATE nor TT include σ_{23}. The Bayesian approach has the advantage that learning about σ_{23} is possible during posterior analysis. This allows for the construction of *PPDs* for ATE and TT, which are much more informative than the simple expectations given in (68) and (69).

The PPD for ATE is formally given by

$$\begin{aligned} &p\left(\Delta_p \mid \mathbf{x_p}, \mathbf{X}, \mathbf{y}\right) = \int_{\mathbf{\Gamma}} p\left(\Delta_p \mid \mathbf{x_p}, \mathbf{\Gamma}\right) p\left(\mathbf{\Gamma} \mid \mathbf{y}, \mathbf{X}\right) d\,\mathbf{\Gamma} \quad \text{with} \\ &\Delta_p \mid \mathbf{x_p}, \mathbf{\Gamma} \sim n\left(\mathbf{x_p}'\left(\boldsymbol{\beta_3} - \boldsymbol{\beta_2}\right), \sigma_{22} + \sigma_{33} - 2\sigma_{23}\right) \end{aligned} \tag{70}$$

where $\mathbf{\Gamma}$ comprises all model parameters. The posterior mean of $\Delta_p \mid \mathbf{x_p}, \mathbf{X}, \mathbf{y}$ will be the analog of the classical ATE measure. However, to go beyond the simple expectation of this distribution and to simulate the entire distribution, knowledge of σ_{23} is required, as is evident from the formula.

Similarly, for the PPD of TT we have

$$p\left(\Delta_p \mid \mathbf{x_p}, \mathbf{z_p}, y_{1p}^* > 0, \mathbf{X}, \mathbf{y}\right) = \int_{\mathbf{\Gamma}} p\left(\Delta_p \mid \mathbf{x_p}, \mathbf{z_p}, y_{1p}^* > 0, \mathbf{\Gamma}\right) p\left(\mathbf{\Gamma} \mid \mathbf{y}, \mathbf{X}\right) d\,\mathbf{\Gamma} \qquad \text{with}$$

$$p\left(\Delta_p \mid \mathbf{x_p}, \mathbf{z_p}, y_{1p}^* > 0, \mathbf{\Gamma}\right) = \frac{p\left(\Delta_p, y_{1p}^* > 0 \mid \mathbf{x_p}, \mathbf{z_p}, \mathbf{\Gamma}\right)}{p\left(y_{1p}^* > 0 \mid \mathbf{x_p}, \mathbf{z_p}, \mathbf{\Gamma}\right)} =$$

$$\Phi\left(\mathbf{z_p}'\boldsymbol{\beta_1}\right)^{-1} \int_{\mathbf{z_p}'\boldsymbol{\beta_1}}^{\infty} p\left(\Delta_p, \varepsilon_1 \mid \mathbf{x_p}, \mathbf{z_p}, \mathbf{\Gamma}\right) d\varepsilon_1 =$$

$$\Phi\left(\mathbf{z_p}'\boldsymbol{\beta_1}\right)^{-1} \int_{\mathbf{z_p}'\boldsymbol{\beta_1}}^{\infty} f\left(\varepsilon_1 \mid \Delta_\varepsilon, \mathbf{x_p}, \mathbf{z_p}, \mathbf{\Gamma}\right) p\left(\Delta_p \mid \mathbf{x_p}, \mathbf{z_p}, \mathbf{\Gamma}\right) d\varepsilon_1 =$$

$$\frac{p\left(\Delta_p \mid \mathbf{x_p}, \mathbf{z_p}, \mathbf{\Gamma}\right)}{\Phi\left(\mathbf{z_p}'\boldsymbol{\beta_1}\right)} \int_{\mathbf{z_p}'\boldsymbol{\beta_1}}^{\infty} f\left(\varepsilon_1 \mid \Delta_\varepsilon, \mathbf{x_p}, \mathbf{z_p}, \mathbf{\Gamma}\right) d\varepsilon_1 =$$

$$\frac{p\left(\Delta_p \mid \mathbf{x_p}, \mathbf{z_p}, \mathbf{\Gamma}\right)}{\Phi\left(\mathbf{z_p}'\boldsymbol{\beta_1}\right)} \Phi\left(\frac{\mathbf{z_p}'\boldsymbol{\beta_1} - \frac{\sigma_{1,3-2}}{\sigma_{3-2}^2}\left(\Delta_p - \mathbf{x_p}'\left(\boldsymbol{\beta_3} - \boldsymbol{\beta_2}\right)\right)}{\sqrt{1 - \frac{\sigma_{1,3-2}^2}{\sigma_{3-2}^2}}}\right) \qquad \text{with}$$

$$p\left(\Delta_p \mid \mathbf{x_p}, \mathbf{z_p}, \mathbf{\Gamma}\right) = n\left(\mathbf{x_p}'\left(\boldsymbol{\beta_3} - \boldsymbol{\beta_2}\right), \sigma_{22} + \sigma_{33} - 2\sigma_{23}\right) \tag{71}$$

To draw $p\left(\Delta_p \mid \mathbf{x_p}, \mathbf{z_p}, y_{1p}^* > 0, \mathbf{\Gamma}\right)$, we first draw y_{1p}^* from its truncated marginal density, then we draw Δ_p from its conditional normal with mean $\mathbf{x_p}'\left(\boldsymbol{\beta_3} - \boldsymbol{\beta_2}\right) + \sigma_{1,3-2}\left(y_{1p} - \mathbf{z_p}'\boldsymbol{\beta_1}\right)$ and variance $\sigma_{3-2}^2 - \sigma_{1,3-2}^2$

References

Li, M., D. J. Poirier and J. L. Tobias. 2004. Do dropouts suffer from dropping out? Estimation and prediction of outcome gains in a generalized selection model. *Journal of Applied Econometrics* **19:** 203-225.

Poirier, D. J. 1998. Revising beliefs in non-identified models. *Econometric Theory* **14:** 483-509.