The Classical Linear Regression Model

Andrea Carriero

January 2018

Bayesian estimation of CLRM

- We will now go more into the details of the estimation.
- The likelihood function is

$$p(Y|\beta,\sigma^2) = (2\pi)^{-\frac{T}{2}} \left|\sigma^2 I_T\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(Y-X\beta)'\left(\sigma^2 I_T\right)^{-1}(Y-X\beta)\right]$$

- Depending on the form of prior we will have different models/posteriors. The most used priors are:
  1. Fix the error variance with an estimate (Theil's mixed estimator)
  2. The Natural Conjugate N-IG prior
  3. The Independent N-IG prior
  4. The diffuse (Jeffrey's) prior
  5. The Normal-diffuse (Zellner's) prior

Theil mixed estimation - OLS with uncertain restrictions

- Consider the model:

$$Y = X\beta + \varepsilon; \qquad \varepsilon \sim N(0, \hat{\sigma}^2 I_T),$$

where we are assuming $\hat{\sigma}^2$ is a frequentist estimate of the (non-random) error variance σ^2

- Consider the set of uncertain (fuzzy) restrictions:

$$\beta \sim N(\beta_0, \Sigma_0)$$

- This can be written:

$$\overset{Y^*}{\begin{bmatrix} Y \\ \beta_0 \end{bmatrix}} = \overset{X^*}{\begin{bmatrix} X \\ I \end{bmatrix}} \beta + \overset{\varepsilon^*}{\begin{bmatrix} \varepsilon \\ u \end{bmatrix}}; \quad Var(\varepsilon^*) = \overset{\Sigma^*}{\begin{bmatrix} \hat{\sigma}^2 I_T & 0 \\ 0 & \Sigma_0 \end{bmatrix}}$$

where $-u \equiv (\beta - \beta_0) \sim N(0, \Sigma_0) \Rightarrow \beta_0 = \beta + u$

- This system can be estimated with GLS (Theil and Goldberger 1960). The GLS estimator $\bar{b}$ is:

$$\bar{b} = \left(X^{*\prime}\Sigma^{*-1}X^*\right)^{-1}\left(X^{*\prime}\Sigma^{*-1}Y^*\right)$$

Theil mixed estimation - GLS

- The GLS estimator $\bar{b}$ is:

$$\begin{aligned}
\bar{b} &= \left(X^{*\prime}\Sigma^{*-1}X^{*}\right)^{-1}\left(X^{*\prime}\Sigma^{*-1}Y^{*}\right) \\
&= \left(\begin{bmatrix} X' & I \end{bmatrix}\begin{bmatrix} (\hat{\sigma}^2)^{-1}I_T & 0 \\ 0 & \Sigma_0^{-1} \end{bmatrix}\begin{bmatrix} X \\ I \end{bmatrix}\right)^{-1} \\
&\quad \times\left(\begin{bmatrix} X' & I \end{bmatrix}\begin{bmatrix} (\hat{\sigma}^2)^{-1}I_T & 0 \\ 0 & \Sigma_0^{-1} \end{bmatrix}\begin{bmatrix} Y \\ \beta_0 \end{bmatrix}\right) \\
&= (X'(\hat{\sigma}^2)^{-1}X+\Sigma_0^{-1})^{-1}(X'(\hat{\sigma}^2)^{-1}Y+\Sigma_0^{-1}\beta_0)
\end{aligned}$$

which is a mix of the data and the restrictions.

Theil mixed estimation - Bayesian interpretation

- This model is -basically- the one we have seen in the very beginning, except there we assumed knowledge of σ^2 while here we use an estimate $\hat{\sigma}^2$
- The prior for β is:

$$\beta \sim N(\beta_0, \Sigma_0)$$

- The parameter σ^2 is assumed fixed (i.e. not random) and estimated ($\hat{\sigma}^2$) in a preliminary step
- The posterior is:

$$\beta|y \sim N(\beta_1, \Sigma_1)$$

with

$$\Sigma_1 = \left(\Sigma_0^{-1} + \frac{1}{\hat{\sigma}^2}X'X\right)^{-1}, \ \beta_1 = \Sigma_1\left(\Sigma_0^{-1}\beta_0 + \frac{1}{\hat{\sigma}^2}X'Y\right)$$

Theil mixed estimation - posterior derivation

Prior:

$$p(\beta) \propto \exp\left[-0.5\left(\beta-\beta_0\right)'\Sigma_0^{-1}\left(\beta-\beta_0\right)\right]$$

Likelihood:

$$p(Y|\beta) \propto \exp[-0.5\left(Y-X\beta\right)'\left(Y-X\beta\right)/\hat{\sigma}^2]$$

Posterior kernel:

$$\begin{aligned} p(\beta|Y) &\propto \exp[-0.5\{\left(\beta-\beta_0\right)'\Sigma_0^{-1}\left(\beta-\beta_0\right)+\left(Y-X\beta\right)'\left(Y-X\beta\right)/\hat{\sigma}^2\}] \\ &\propto \exp[-0.5\left(\beta-\beta_1\right)'\Sigma_1^{-1}\left(\beta-\beta_1\right)] \end{aligned}$$

where:

$$\Sigma_1 = \left(\Sigma_0^{-1}+\frac{1}{\hat{\sigma}^2}X'X\right)^{-1},\ \beta_1 = \Sigma_1\left(\Sigma_0^{-1}\beta_0+\frac{1}{\hat{\sigma}^2}X'Y\right)$$

Prior selection

- Can we estimate the optimal degree of "uncertainty" of the restrictions?
- Yes. For simplicity, let us say:

$$\beta \sim N(\beta_0 = 0, \Sigma_0 = \lambda^{-1} I_k)$$

that is, all of the k restrictions are mutually independent

- Consider the simplest possible case in which the parameter λ can take just 2 values: λ_1 and λ_2
- Basically, these correspond to two different models:

$$\begin{aligned} M_1 &: \quad Y = X\beta + \varepsilon; \qquad \varepsilon \sim N(0, \hat{\sigma}^2 I_T), \ \beta \sim N(0, \lambda_1^{-1} I_k) \\ M_2 &: \quad Y = X\beta + \varepsilon; \qquad \varepsilon \sim N(0, \hat{\sigma}^2 I_T), \ \beta \sim N(0, \lambda_2^{-1} I_k) \end{aligned}$$

How can I choose among the two models?

Marginal data density

- We can compute $p(Y|M_1)$ and $p(Y|M_2)$. These are the data densities under the two alternative models
- Recall that, in general:

$$1 = \int p(\beta|Y)d\beta = \int \frac{p(Y|\beta) \times p(\beta)}{p(Y)} d\beta = \frac{1}{p(Y)} \int \underbrace{p(Y|\beta) \times p(\beta)}_{posterior\ \ kernel} d\beta$$

$$\longrightarrow \quad p(Y) = \int p(Y|\beta) \times p(\beta) d\beta$$

- We just have to apply this to the two models, choose the model with the highest $p(Y|M_j)$

$$p(Y|M_j) = \int p(Y|\beta, M_j) \times p(\beta|M_j) d\beta, \quad j = 1, 2$$

- The value $p(Y|M_j)$ is the marginal data density for model M_j (marginal likelihood)
- In general, computation of the MDD is not easy, as it requires integration of $\int p(Y|\beta, M_i) \times p(\beta|M_i)$

Marginal data density

- However Theil estimation is one case in which computation of the MDD is easy.
- Indeed, in this case we can use the Bayes formula:

$$p(\beta|Y) = \frac{p(Y|\beta) \times p(\beta)}{p(Y)} \rightarrow p(Y) = \frac{p(Y|\beta) \times p(\beta)}{p(\beta|Y)}$$

  and note that the quantity on the RHS is known.
- This is a consequence of conjugacy: since both the numerator $p(Y|\beta) \times p(\beta)$ and the denominator $p(\beta|Y)$ share the same kernel (Gaussian) by construction, they will simplify and all is left is the ratio of the integrating constants!
- So far we have only used the posterior kernel, we now need the properly normalized posterior.

Theil mixed estimation - marginal data density

Let us see how this works:

$$p_{M_j}(Y) = \frac{\overbrace{\left(\begin{array}{c} (2\pi)^{-\frac{T}{2}} \left|\hat{\sigma}^2 I_T\right|^{-\frac{1}{2}} \times \\ \exp\left[\begin{array}{c} -\frac{1}{2}(Y - X\beta)' \\ (\sigma^2 I_T)^{-1}(Y - X\beta) \end{array} \right] \end{array} \right)}^{p_{M_j}(Y|\beta)} \times \overbrace{\left(\begin{array}{c} (2\pi)^{-\frac{k}{2}} |\Sigma_0|^{-\frac{1}{2}} \times \\ \exp\left[\begin{array}{c} -\frac{1}{2}(\beta - \beta_0)' \\ \Sigma_0^{-1}(\beta - \beta_0) \end{array} \right] \end{array} \right)}^{p_{M_j}(\beta)}}{\underbrace{\left((2\pi)^{-\frac{k}{2}} |\Sigma_1|^{-\frac{1}{2}} \times \exp\left[-\frac{1}{2}\left((\beta - \beta_1)' \Sigma_1^{-1} (\beta - \beta_1) \right) \right] \right)}_{p_{M_j}(\beta|Y)}}$$

Theil mixed estimation - marginal data density

$$p_{M_j}(Y) = \frac{\overbrace{\left((2\pi)^{-\frac{T}{2}} \left|\hat{\sigma}^2 I_T\right|^{-\frac{1}{2}} \times (2\pi)^{-\frac{k}{2}} |\Sigma_0|^{-\frac{1}{2}} \times \exp\left[-\frac{1}{2}\left(\begin{array}{c} (\beta-\beta_1)'\Sigma_1^{-1}(\beta-\beta_1) \\ \boxed{-\beta_1'\Sigma_1^{-1}\beta_1+\beta_0'\Sigma_0^{-1}\beta_0+Y'(\hat{\sigma}^2)^{-1}Y} \end{array}\right)\right]\right)}^{p_{M_j}(Y|\beta)\times p_{M_j}(\beta)}}{\underbrace{\left((2\pi)^{-\frac{k}{2}}|\Sigma_1|^{-\frac{1}{2}} \times \exp\left[-\frac{1}{2}\left((\beta-\beta_1)'\Sigma_1^{-1}(\beta-\beta_1)\right)\right]\right)}_{p_{M_j}(\beta|Y)}}$$

$$= (2\pi)^{-\frac{T}{2}} \left|\hat{\sigma}^2 I_T\right|^{-\frac{1}{2}} |\Sigma_0|^{-\frac{1}{2}} |\Sigma_1|^{\frac{1}{2}} \times \exp[-\frac{1}{2}(\beta_1'\Sigma_1^{-1}\beta_1 + \beta_0'\Sigma_0^{-1}\beta_0 + Y'(\hat{\sigma}^2)^{-1}Y)]$$

Theil mixed estimation - marginal data density

Note that $Y'Y = \hat{Y}'_{LS}\hat{Y}_{LS} + \hat{\varepsilon}'_{LS}\hat{\varepsilon}_{LS} = \hat{\beta}'_{LS}X'X\hat{\beta}_{LS} + \hat{\varepsilon}'_{LS}\hat{\varepsilon}_{LS}$, therefore:

$$\begin{aligned} p_{M_j}(Y) &= (2\pi)^{-\frac{T}{2}} \left|\hat{\sigma}^2 I_T\right|^{-\frac{1}{2}} |\Sigma_0|^{-\frac{1}{2}} |\Sigma_1|^{\frac{1}{2}} \\ &\quad \times \exp[-\frac{1}{2}(\beta'_1\Sigma_1^{-1}\beta_1 + \beta'_0\Sigma_0^{-1}\beta_0 + \hat{\beta}'_{LS}X'X\hat{\beta}_{LS}/\hat{\sigma}^2 + \hat{\varepsilon}'_{LS}\hat{\varepsilon}_{LS}/\hat{\sigma}^2)], \end{aligned}$$

and by re-grouping we can see:

$$\begin{aligned} p_{M_j}(Y) &= (2\pi\hat{\sigma}^2)^{-\frac{T}{2}} \exp[-\frac{1}{2}\hat{\varepsilon}'_{LS}\hat{\varepsilon}_{LS}/\hat{\sigma}^2] \times \exp[-\frac{1}{2}\hat{\beta}'_{LS}(\hat{\sigma}^2(X'X)^{-1})^{-1}\hat{\beta}_{LS}] \\ &\quad \times |\Sigma_0|^{-\frac{1}{2}} \times \exp[-\frac{1}{2}\beta'_0\Sigma_0^{-1}\beta_0] \\ &\quad \times |\Sigma_1|^{\frac{1}{2}} \times \exp[-\frac{1}{2}\beta'_1\Sigma_1^{-1}\beta_1], \end{aligned} \tag{1}$$

where the first line of (1) comes from the likelihood and represents the LS fit of the model. Therefore the first line is going to be the same across all models (since the models we considered **here** only differ in the prior while they share the same likelihood).

Theil mixed estimation - marginal data density

The second line of (1) is a sum of squares of prior moments, which are chosen by the econometrician. Finally, in this model the joint/marginal posterior moments Σ_1 and β_1 are known in closed form:

$$\Sigma_1 = \left(\Sigma_0^{-1} + \frac{1}{\hat{\sigma}^2} X'X\right)^{-1}, \; \beta_1 = \Sigma_1 \left(\Sigma_0^{-1}\beta_0 + \frac{1}{\hat{\sigma}^2} X'Y\right),$$

which means that the term

$$\beta_1' \Sigma_1^{-1} \beta_1 = \left(\Sigma_0^{-1}\beta_0 + \frac{1}{\hat{\sigma}^2} X'Y\right)' \left(\Sigma_0^{-1} + \frac{1}{\hat{\sigma}^2} X'X\right)^{-1} \left(\Sigma_0^{-1}\beta_0 + \frac{1}{\hat{\sigma}^2} X'Y\right)$$

depends only on the prior moments β_0, Σ_0, and the data.
It follows that the MDD for this model is readily available.

Theil mixed estimation - example

Going back to our example, considering the two models:

$$\begin{aligned} M_1 &: \quad Y = X\beta + \varepsilon; \qquad \varepsilon \sim N(0, \hat{\sigma}^2 I_T),\ \beta \sim N(0, \lambda_1^{-1} I_k) \\ M_2 &: \quad Y = X\beta + \varepsilon; \qquad \varepsilon \sim N(0, \hat{\sigma}^2 I_T),\ \beta \sim N(0, \lambda_2^{-1} I_k) \end{aligned}$$

where $\Sigma_0^{-1}(M_1) = \lambda_1 I, \Sigma_0^{-1}(M_2) = \lambda_2 I$. We can simply compute $p(Y|M_1)$ and $p(Y|M_2)$ and compute the ratio $\frac{p(Y|M_1)}{p(Y|M_2)}$ which in this case is:

$$\frac{\lambda_1^{k/2} |\Sigma_1(\lambda_1)|^{\frac{1}{2}} \times \exp[-\frac{1}{2}(\beta_0' \lambda_1 \beta_0 - \beta_1'(\lambda_1)\Sigma_1^{-1}(\lambda_1)\beta_1(\lambda_1))]}{\lambda_2^{k/2} |\Sigma_1(\lambda_2)|^{\frac{1}{2}} \times \exp[-\frac{1}{2}(\beta_0' \lambda_2 \beta_0 - \beta_1'(\lambda_2)\Sigma_1^{-1}(\lambda_2)\beta_1(\lambda_2))]}$$

where

$$\begin{aligned} \Sigma_1(\lambda_1) &= (\lambda_1 I + X'X/\hat{\sigma}^2)^{-1},\ \Sigma_1(\lambda_2) = (\lambda_2 I + X'X/\hat{\sigma}^2)^{-1} \\ \beta_1(\lambda_1) &= \Sigma_1(\lambda_1 \beta_0 + X'Y/\hat{\sigma}^2),\ \beta_1(\lambda_2) = \Sigma_2(\lambda_2 \beta_0 + X'Y/\hat{\sigma}^2) \end{aligned}$$

The Natural Conjugate N-IG prior

Now let us consider the task of estimating σ^2. In the mixed estimation, we have that:

$$\Sigma_1 = \left(\Sigma_0^{-1} + \frac{1}{\hat{\sigma}^2}X'X\right)^{-1}, \ \beta_1 = \Sigma_1\left(\Sigma_0^{-1}\beta_0 + \frac{1}{\hat{\sigma}^2}X'Y\right)$$

rewrite as:

$$\Sigma_1 = \left(\frac{1}{\hat{\sigma}^2}\Sigma_0^{*-1} + \frac{1}{\hat{\sigma}^2}X'X\right)^{-1}, \ \beta_1 = \Sigma_1\left(\frac{1}{\hat{\sigma}^2}\Sigma_0^{*-1}\beta_0 + \frac{1}{\hat{\sigma}^2}X'Y\right)$$

with $\Sigma_0 = \hat{\sigma}^2\Sigma_0^*$ and $\Sigma_1 = \hat{\sigma}^2\Sigma_1^*$ and:

$$\Sigma_1^* = \left(\Sigma_0^{*-1} + X'X\right)^{-1}, \ \beta_1 = \left(\Sigma_0^{*-1} + X'X\right)^{-1}\left(\Sigma_0^{*-1}\beta_0 + X'Y\right)$$

The Natural Conjugate N-IG prior

- The prior for β is:

$$\beta|\sigma^2 \sim N(\beta_0, \sigma^2\Sigma_0^*)$$

  therefore we have $\Sigma_0 = \sigma^2\Sigma_0^*$.

- The prior for σ^2 is:

$$\sigma^2 \sim \Gamma^{-1}\left(\frac{\nu_0}{2}, \frac{s_0^2}{2}\right) \Leftrightarrow \frac{1}{\sigma^2} = h \sim \Gamma\left(\frac{\nu_0}{2}, \frac{s_0^2}{2}\right)$$

- The posterior is:

$$\begin{aligned} \sigma^2|y &\sim \Gamma^{-1}\left(\frac{\nu_1}{2}, \frac{s_1^2}{2}\right) \\ \beta|\sigma^2, y &\sim N(\beta_1, \sigma^2\Sigma_1^*) \end{aligned}$$

  where $\Sigma_1 = \sigma^2\Sigma_1^*$

The Natural Conjugate N-IG prior: posterior moments

- The moments of the posterior are:

$$
\begin{aligned}
\nu_1 &= \nu_0 + T \\
s_1^2 &= s_0^2 + Q \\
\sigma^2 \Sigma_1^* &= \sigma^2 \left(\Sigma_0^{*-1} + X'X \right)^{-1} \longrightarrow \Sigma_1^* = \left(\Sigma_0^{*-1} + X'X \right)^{-1} \\
\beta_1 &= \left(\Sigma_0^{*-1} + X'X \right)^{-1} \left(\Sigma_0^{*-1} \beta_0 + X'X\hat{\beta} \right)
\end{aligned}
$$

with:

$$
\begin{aligned}
Q &= s + s_\beta \\
s &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) \\
s_\beta &= (\beta_1 - \hat{\beta})'X'X(\beta_1 - \hat{\beta}) + (\beta_1 - \beta_0)'\Sigma_0^{*-1}(\beta_1 - \beta_0)
\end{aligned}
$$

which -note- only depends on the prior and posterior moments, and the data (but not on the actual draws of β)

Posterior derivation

It is useful to re-write it as follows:

$$
\begin{aligned}
p(Y|\beta,\sigma^2) &= (2\pi)^{-\frac{T}{2}}\left|\sigma^2 I_T\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(Y-X\beta)'\left(\sigma^2 I_T\right)^{-1}(Y-X\beta)\right] \\
&= (2\pi\sigma^2)^{-\frac{T}{2}}\exp\left[\begin{array}{c} -\frac{1}{2\sigma^2}(Y-X\hat{\beta})'(Y-X\hat{\beta}) \\ -\frac{1}{2\sigma^2}(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\end{array}\right]
\end{aligned}
$$

where we have used $(Y-X\beta+X\hat{\beta}-X\hat{\beta})$ and completed the squares, exploiting the orthogonality $\hat{\varepsilon}'X(\beta-\hat{\beta})=0$. Then by breaking $\sigma^{2-\frac{T}{2}}$ into $(\sigma^2)^{-\frac{k}{2}}$ and $(\sigma^2)^{-\frac{T-k-2}{2}-1}$ and defining $s=(Y-X\hat{\beta})'(Y-X\hat{\beta})$ we have:

$$
\begin{aligned}
p(Y|\beta,\sigma^2) &\propto (\sigma^2)^{-\frac{T-k-2}{2}-1}\exp\left[-\frac{s}{2\sigma^2}\right] \\
&\quad (\sigma^2)^{-\frac{k}{2}}\exp\left[-\frac{1}{2\sigma^2}(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\right]
\end{aligned}
$$

which is the kernel of a Normal-Gamma form, with $\beta|\sigma^2$ a normal and σ^2 a Gamma with $T-k-2$ shape and scale s.

The Natural Conjugate N-IG prior: joint and marginals

Prior:

$$\begin{aligned} p(\beta,\sigma^2) &= p(\beta|\sigma^2)p(\sigma^2) \\ &= \left(2\pi\sigma^2\right)^{-\frac{v_0}{2}-1}\exp\left[-\frac{s_0}{2\sigma^2}\right] \\ &\quad (\sigma^2)^{-\frac{k}{2}}\exp\left[-\frac{1}{2}(\beta-\beta_0)'\frac{1}{\sigma^2}\Sigma_0^{*-1}(\beta-\beta_0)\right] \end{aligned}$$

where - note - $\frac{1}{\sigma^2}\Sigma_0^{*-1}=\Sigma_0^{-1}$

Posterior:

$$\begin{aligned} p(\beta,\sigma^2|Y) &\propto (\sigma^2)^{-\frac{T-k-2}{2}-1}\exp\left[-\frac{s}{2\sigma^2}\right]\left(\sigma^2\right)^{-\frac{v_0}{2}-1}\exp\left[-\frac{s_0}{2\sigma^2}\right] \\ &\quad (\sigma^2)^{-\frac{k}{2}}\exp\left[-\frac{1}{2\sigma^2}(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\right] \\ &\quad (\sigma^2)^{-\frac{k}{2}}\exp\left[-\frac{1}{2\sigma^2}(\beta-\beta_0)'\Sigma_0^{*-1}(\beta-\beta_0)\right] \end{aligned}$$

Note that the last two terms can be put together because $-\frac{1}{2\sigma^2}$ appears in the prior!

The Natural Conjugate N-IG prior: joint and marginals

Posterior:

$$p(\beta,\sigma^2|Y) \propto (\sigma^2)^{-\frac{T+v_0}{2}-1}\exp\left[-\frac{s_0+s}{2\sigma^2}\right] (\sigma^2)^{-\frac{k}{2}}\exp\left[\begin{array}{c} -\frac{1}{2\sigma^2}(\beta-\hat{\beta})'X'X(\beta-\hat{\beta}) \\ -\frac{1}{2\sigma^2}(\beta-\beta_0)'\Sigma_0^{*-1}(\beta-\beta_0)\end{array}\right]$$

we want to separate the part depending from β. Complete the squares and regroup:

$$\begin{aligned}(\beta-\hat{\beta})'X'X(\beta-\hat{\beta}) &= \beta'X'X\beta-\beta'X'X\hat{\beta}-\hat{\beta}'X'X\beta+\hat{\beta}'X'X\hat{\beta} \\ &+ \\ (\beta-\beta_0)'\Sigma_0^{*-1}(\beta-\beta_0) &= \beta'\Sigma_0^{*-1}\beta-\beta'\Sigma_0^{*-1}\beta_0-\beta_0'\Sigma_0^{*-1}\beta+\beta_0'\Sigma_0^{*-1}\beta_0\end{aligned}$$

$$\begin{aligned}= \beta'\underbrace{(X'X+\Sigma_0^{*-1})}_{\Sigma_1^*}\beta-\beta'(\underbrace{X'X\hat{\beta}-\Sigma_0^{*-1}\beta_0}_{\Sigma_1^*\beta_1})-\underbrace{(\hat{\beta}'X'X+\beta_0'\Sigma_0^{*-1})}_{\beta_1'\Sigma_1^*}\beta+ \\ +\hat{\beta}'X'X\hat{\beta}+\beta_0'\Sigma_0^{*-1}\beta_0\end{aligned}$$

The Natural Conjugate N-IG prior: joint and marginals

Recall from the same step as previous lecture that $\beta'\Sigma_1^*\beta - \beta'\Sigma_1^*\beta_1 - \beta_1\Sigma_1^*\beta$ can be written as $(\beta-\beta_1)'\Sigma_1^{*-1}(\beta-\beta_1) - \beta_1'\Sigma_1^{*-1}\beta_1$. Then we have

$$\begin{aligned} & (\beta-\hat{\beta})'X'X(\beta-\hat{\beta}) + (\beta-\beta_0)'\Sigma_0^{*-1}(\beta-\beta_0) \\ = \quad & (\beta-\beta_1)'\Sigma_1^{*-1}(\beta-\beta_1) - \beta_1'\Sigma_1^{*-1}\beta_1 + \hat{\beta}'X'X\hat{\beta} + \beta_0'\Sigma_0^{*-1}\beta \end{aligned}$$

and:

$$\begin{aligned} p(\beta,\sigma^2|Y) \quad \propto \quad & (\sigma^2)^{-\frac{T+v_0}{2}-1}\exp\left[-\frac{s_0+s+s_\beta}{2\sigma^2}\right] \\ & (\sigma^2)^{-\frac{k}{2}}\exp\left[-\frac{1}{2\sigma^2}(\beta-\beta_1)'\Sigma_1^{*-1}(\beta-\beta_1)\right] \end{aligned}$$

with

$$\begin{aligned} s_\beta \quad = \quad & -\beta_1'\Sigma_1^{*-1}\beta_1 + \hat{\beta}'X'X\hat{\beta} + \beta_0'\Sigma_0^{*-1}\beta_0 \\ = \quad & (\beta_1-\hat{\beta})'X'X(\beta_1-\hat{\beta}) + (\beta_1-\beta_0)'\Sigma_0^{*-1}(\beta_1-\beta_0) \end{aligned}$$

Note that $Q = s + s_\beta = Y'Y - \beta_1'\Sigma_1^{*-1}\beta_1 + \beta_0'\Sigma_0^{*-1}\beta_0$

The Natural Conjugate N-IG prior: joint and marginals

- We can integrate out the parameter σ^2 and get the marginal $\beta|y$:

$$\begin{aligned} p(\beta|Y) &= \int (\beta|\sigma^2, Y)p(\sigma^2|Y)d\sigma^2 \\ &= c^{-1}\left[s_1 + (\beta-\beta_1)'\Sigma_1^{*-1}(\beta-\beta_1)\right]^{-\frac{\nu_1+k}{2}} \end{aligned}$$

  with $c = \frac{\pi^{k/2}\Gamma\left(\frac{\nu_1}{2}\right)}{\Gamma\left(\frac{\nu_1+k}{2}\right)}|\Sigma_1^{*-1}|^{-1/2}s_1^{-\nu_1/2}$ (Dickey 1967 parameterization).

- The distribution resulting from the integration is known, it is a multivariate t:

$$\beta|Y \sim t(\beta_1, \Sigma_1^*, s_1^2, \nu_1)$$

- To draw from this distribution we can simply use MC simulation, drawing from $p(\sigma^2|Y)$ and then $(\beta|\sigma^2, Y)$:

$$p(\beta, \sigma^2)|Y = (\beta|\sigma^2, Y)p(\sigma^2|Y)$$

  which gives the joint (and the marginals).

The Natural Conjugate N-IG prior: marginal likelihood

- The same applies to the prior (and likelihood):

$$p(\beta) = \int (\beta|\sigma^2)p(\sigma^2)d\sigma^2 \sim t(\beta_0, \Sigma_0, s_0^2, \nu_0)$$

- The prior, the posterior, and the likelihood, are of the same form (conjugacy):

$$\begin{aligned} \beta|\sigma^2 &\sim N;\ \sigma^2 \sim \Gamma^{-1} \longrightarrow \beta \sim t \\ \beta|\sigma^2, Y &\sim N;\ \sigma^2|Y \sim \Gamma^{-1} \longrightarrow \beta|Y \sim t \\ Y|\beta, \sigma^2 &\sim N;\ \longrightarrow Y|\beta \sim t \end{aligned}$$

- An important advantage is that then we can apply Bayes formula to obtain:

$$\underset{data\ density}{p(Y)} = \frac{\underset{likelihood}{p(Y|\beta)} \times \underset{prior}{p(\beta)}}{\underset{posterior}{p(\beta|Y)}}$$

  which is also a multivariate t (known).
- Drawback: have to specify the prior on β as a function of σ^2

The Natural Conjugate N-IG prior: marginal likelihood

The marginal likelihood is multivariate t:

$$Y = X\beta + \varepsilon$$

with $\varepsilon \sim N(0, \sigma^2 I)$. Since $\beta|\sigma^2 \sim N(\beta_0, \sigma^2\Sigma_0^*)$ then $X\beta|\sigma^2 \sim N(X\beta_0, \sigma^2 X\Sigma_0^* X')$. It follows that

$$Y|\sigma^2 \sim N(X\beta_0, \sigma^2(X\Sigma_0^* X' + I))$$

because ε and β are independent when conditioning on σ^2. This is a normal, and σ^2 an inverse gamma, so integrating this out gives a t:

$$Y \sim t(X\beta_0, (X\Sigma_0^* X' + I), s_0, v_0)$$

which has pdf:

$$p(Y) = \frac{\left[s_0 + (Y - X\beta_0)'(X\Sigma_0^* X' + I)^{-1}(Y - X\beta_0)\right]^{-\frac{v_0+T}{2}}}{\frac{\pi^{T/2}\Gamma\left(\frac{v_0}{2}\right)}{\Gamma\left(\frac{v_0+T}{2}\right)}|(X\Sigma_0^* X' + I)|^{-1/2} s_0^{-v_0/2}}$$

The Natural Conjugate N-IG prior: marginal likelihood

The expression further simplifies if one notes that $v_{0+T} = v_1$, and

$$s_0 + (Y - X\beta_0)'(X\Sigma_0^* X' + I)^{-1}(Y - X\beta_0) = s_0 + s + s_\beta = s_1,$$

and

$$\left|X\Sigma_0^* X' + I\right| = |\Sigma_0^*|\,|\Sigma_1^*|^{-1}$$

giving:

$$p(Y) = \pi^{-T/2}\frac{\Gamma\left(\frac{v_1}{2}\right)}{\Gamma\left(\frac{v_0}{2}\right)}\frac{|\Sigma_1^*|^{-1/2}}{\left|\Sigma_0^*\right|^{-1/2}}\frac{[s_1]^{-v_1/2}}{[s_0]^{-v_0/2}}$$

note this is very easy to compute just on the basis of prior moments and data.

The diffuse (Jeffrey's) prior

- Ensures uninformativeness, regardless of transformations of the model (invariance)
- It is specified as follows:

$$p(\beta, \sigma^2) \propto 1/\sigma^2$$

- And it gives:

$$\sigma^2|y \sim \Gamma^{-1}\left(\frac{\nu_1}{2}, \frac{s_1^2}{2}\right);\ \beta|\sigma^2, y \sim N(\beta_1, \sigma^2\Sigma_1)$$

with

$$\begin{aligned} \nu_1 &= \nu_0 + T;\ \ s_1^2 = s_0^2 + (Y - X\hat{\beta})'(Y - X\hat{\beta}) \\ \sigma^2\Sigma_1 &= \sigma^2 (X'X)^{-1};\ \beta_1 = (X'X)^{-1}(X'X\hat{\beta}) \end{aligned}$$

- Note this is the "limit" of the conjugate N-IW prior when the prior precision tends to 0.
- Marginal $\beta|y$ is a multivariate t. This is the equivalent of classical estimation of the CLRM.
- Marginal likelihood is 0.

The Independent N-IG prior

- The prior for β and σ^2 is:

$$\beta \sim N(\beta_0, \Sigma_0);\ \sigma^2 \sim \Gamma^{-1}\left(\frac{\nu_0}{2}, \frac{s_0^2}{2}\right)$$

  where -note- Σ_0 no longer depends on σ^2
- The posterior is:

$$\begin{aligned} p(\beta, \sigma^2 | Y) \quad \propto \quad & (\sigma^2)^{-\frac{T+\nu_0}{2}-1} \exp\left[-\frac{s_0 + s}{2\sigma^2}\right] \\ & (\sigma^2)^{-\frac{k}{2}} \exp\left[\begin{array}{c} -\frac{1}{2\sigma^2}(\beta-\hat{\beta})'X'X(\beta-\hat{\beta}) \\ -\frac{1}{2}(\beta-\beta_0)'\Sigma_0^{-1}(\beta-\beta_0) \end{array}\right] \end{aligned}$$

- This still contains the kernels of a Normal and a Gamma, but due to the lack of a common scale factor $\frac{1}{\sigma^2}$ in the prior and the likelihood we cannot proceed as we did for the conjugate case.
- Since the kernel of $p(\beta)$ does not have a $\frac{1}{\sigma^2}$ the term $-\frac{1}{2\sigma^2}(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})$ cannot be eliminated from the kernel for σ^2

The Independent N-IG prior

- We can still write:

$$p(\beta|\sigma^2, Y) \propto (\sigma^2)^{-\frac{k}{2}} \exp\left[-\frac{1}{2\sigma^2}(\beta-\beta_1)'\Sigma_1^{-1}(\beta-\beta_1)\right]$$

- Recognizing the kernel of $p(\beta|\sigma^2)$ gives:

$$\beta|\sigma^2, Y \sim N(\beta_1, \Sigma_1);$$

with

$$\Sigma_1 = \left(\Sigma_0^{-1} + \frac{1}{\sigma^2}X'X\right)^{-1};\ \beta_1 = \Sigma_1\left(\Sigma_0^{-1}\beta_0 + \frac{1}{\sigma^2}X'X\hat{\beta}\right)$$

- Therefore we can still derive $\beta|\sigma^2$, but note that now also its mean β_1 depends on σ^2.

The Independent N-IG prior

- We can still write:

$$p\left(\sigma^2|Y,\beta\right) \propto (\sigma^2)^{-\frac{T+\nu_0}{2}-1}\exp\left[-\frac{s_0+s+(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})}{2\sigma^2}\right]$$
$$= (\sigma^2)^{-\frac{T+\nu_0}{2}-1}\exp\left[-\frac{s_0+(Y-X\beta)'(Y-X\beta)}{2\sigma^2}\right]$$

  where $s = Y'Y - \hat{\beta}'X'X\hat{\beta}$.
- This is a conditional posterior $\sigma|\beta$, with s_1 depending on β:

$$\sigma^2|\beta, Y \sim \Gamma^{-1}\left(\frac{\nu_1}{2}, \frac{s_1^2}{2}\right)$$

  with

$$\begin{aligned}\nu_1 &= \nu_0 + T\\ s_1^2 &= s_0^2 + (y-X\beta)'(y-X\beta)\end{aligned}$$

- Note this is the same we derived in lecture 1.

The Independent N-IG prior

- Therefore, the (conditional) posteriors are:

$$\beta|\sigma^2, Y \sim N(\beta_1, \Sigma_1); \ \sigma^2|\beta, Y \sim \Gamma^{-1}\left(\frac{\nu_1}{2}, \frac{s_1^2}{2}\right)$$

with moments:

$$\begin{aligned}
\nu_1 &= \nu_0 + T \\
s_1^2 &= s_0^2 + (y - X\beta)'(y - X\beta) \\
\Sigma_1 &= \left(\Sigma_0^{-1} + \frac{1}{\sigma^2}X'X\right)^{-1} \\
\beta_1 &= \left(\Sigma_0^{-1} + \frac{1}{\sigma^2}X'X\right)^{-1}\left(\Sigma_0^{-1}\beta_0 + \frac{1}{\sigma^2}X'X\hat{\beta}\right)
\end{aligned}$$

The Independent N-IG prior: Gibbs sampling

- This prior is only conditionally conjugate.
- We can only obtain closed form solutions for $\beta|\sigma^2, Y$ and $\sigma^2|\beta, Y$
- Simple MC is not an option to obtain the joint $\beta, \sigma^2 | Y$. Other methods are needed $\longrightarrow$ Gibbs Sampling (MCMC)
- Integrating out σ^2 analytically is not an option, since to draw it one needs to know β. Computing the marginal likelihood in closed form is not feasible.
- However, the model is more flexible than the one with conjugate prior for it does not require proportionality between the prior on β and the error variance.

The Normal-diffuse prior (Zellner 1971)

- The uninformativeness is only imposed on the error variance.
- It is specified as follows:

$$\beta \sim N(\beta_0, \Sigma_0);\ p(\sigma^2) \propto 1/\sigma^2$$

- The (conditional) posteriors are:

$$\beta|\sigma^2, Y \sim N\left(\beta_1, \Sigma_1\right);\ \sigma^2|\beta, Y \sim \Gamma^{-1}\left(\frac{\nu_1}{2}, \frac{s_1^2}{2}\right)$$

with moments:

$$\begin{aligned} \nu_1 &= T;\ s_1^2 = (y - X\beta)'(y - X\beta); \\ \Sigma_1 &= \left(\Sigma_0^{-1} + \frac{1}{\sigma^2}X'X\right)^{-1};\ \beta_1 = \Sigma_1\left(\Sigma_0^{-1}\beta_0 + \frac{1}{\sigma^2}X'X\hat{\beta}\right) \end{aligned}$$

that is, the "limit" of the independent N-IW prior, when $\nu_0 \to 0,\ s_0^2 \to 0$

- Estimation via Gibbs sampling

Gibbs sampling

1. Set starting values for $x_1 ....x_k$

$$x_1^{j=0}, ..., x_k^{j=0}$$

2. Sample $x_1^{j=1}$ from $f\left(x_1^1 | x_2^0, x_3^0 ..., x_k^0\right)$,then sample $x_2^{j=1}$ from $f\left(x_2^1 | x_1^1, x_3^0, ..., x_k^0\right)$,
... then sample $x_i^{j=1}$ from

$$f\left(x_i^1 | x_1^1, x_2^1, ..., x_{i-1}^1, x_{i+1}^0, ..., x_k^0\right)$$

.... finally, sample $x_k^{j=1}$ from $f\left(x_k^1 | x_1^1, x_2^1, ..., x_{k-1}^1\right)$.

3. This completes iteration $j = 1$. Set $j = 2$ and repeat until $j = J$:

$$f\left(x_i^j | x_1^j, x_2^j, ..., x_{i-1}^j, x_{i+1}^{j-1}, ..., x_k^{j-1}\right)$$

File **example_gibbs.m**

Gibbs sampling

- Gibbs sampling is a special case of more general MCMC sampling
- As $J \longrightarrow \infty$ the joint and marginal distributions of simulated $\{x_1^j, ...x_K^j\}_{j=1}^m$ converge at an exponential rate to the joint and marginal distributions of $x_1 ....x_k$
- For simple models (e.g. linear regressions, also multivariate), this happens *really fast*
  - How to evaluate if convergence happened?
- By construction the Gibbs sampler produces draws than are autocorrelated
  - Some burn-in required
  - What is the efficiency/mixing?

Gibbs sampling - convergence and mixing

- **Convergence**
  - Time series plots
  - Tests of equality across independent chains e.g. Geweke's (1992) convergence diagnostic test for equal means
  - Potential Scale Reduction Factors (PSRF), Gelman and Rubin (1992)
- **Mixing**
  - Time series plots
  - Autocorrelograms
  - Inefficiency factors (IF) give an idea of how far we are from i.i.d. sampling
- Check out the R / MATLAB packages COnvergence DiAgnostics

Convergence - PSRF

- Gelman and Rubin (1992) and Brooks, S.P. and Gelman, A. (1998)
- Let $\{\theta_{mj}\}_{j=1}^{J}$ be the $m-$th simulated chain, $m = 1, \ldots, M$. Let $\hat{\theta}_m$ and $\hat{\sigma}_m^2$ be the sample posterior mean and variance of the $m-$th chain, and let the overall sample posterior mean be $\hat{\theta} = \Sigma_{m+1}^{M} \hat{\theta}_m / M$.
- There are two ways to estimate the variance of the stationary distribution σ^2:
  - The mean of the empirical variance within each chain:

$$W = \frac{1}{M} \Sigma_{m+1}^{M} \hat{\sigma}_m^2$$

  - The empirical variance from all chains combined:

$$V = \frac{J-1}{J} W + \frac{M+1}{MJ} B,$$

where $B = \frac{J}{M-1} \Sigma_{m+1}^{M} (\hat{\theta}_m - \hat{\theta})^2$ is the empirical between-chain variance

Convergence - PSRF

- If the chains have converged, then both W and V are unbiased. Otherwise the first method will underestimate the variance, since the individual chains have not had time to range all over the stationary distribution, and the second method will overestimate the variance, since the starting points were chosen to be overdispersed.
- The convergence diagnostic is:

$$PSRF = \sqrt{\frac{V}{W}}$$

- Brooks and Gelman (1997) have suggested, if $PSRF < 1.2$ for all model parameters, one can be fairly confident that convergence has been reached.
- More reassuring (and common) is to apply the more stringent condition $PSRF < 1.1$

Mixing - inefficiency factors

- The inefficiency factor (IF) $1 + 2\sum_{k=1}^{\infty} \rho_k$, where ρ_k is the k-th order autocorrelation. This is the inverse of the relative numerical efficiency measure of Geweke (1992). Usually estimated as the spectral density at frequency zero with Newey-West kernel (with a 4% bandwidth).
- i.i.d. sampling (e.g. MC sampling) features IF=1, here IF < 20 are considered good.
- Note that mixing can *always* be improved artificially by a practice called thinning. Thinning (or skip sampling) is only advisable if you have space constraints, since it always implies loss of information

GLRM: autocorrelated errors

- Gibbs sampling is powerful. For example, we can easily extend the model we are considering:

$$y_t = \beta x_t + \varepsilon_t \tag{2}$$
$$\varepsilon_t = \phi \varepsilon_{t-1} + u_t, \; u_t \sim iidN(0, \sigma_u^2) \tag{3}$$

  In this model there are 3 (groups of) parameters: β, ϕ, σ_u^2. Also, we have $\sigma_\varepsilon^2 = Var(\varepsilon_t) = \sigma_u^2/(1-\phi^2)$
- Consider the Cochrane-Orcutt transformation:

$$P = \begin{bmatrix} -\phi & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & & & -\phi \quad 1 \end{bmatrix}$$

$$Py = PX\beta + P\varepsilon \tag{4}$$

  where $[P\varepsilon]_t = -\phi\varepsilon_{t-1} + \varepsilon_t$.
- The model in (4) is a Generalized LRM, with error variance $Var(P\varepsilon) = PVar(\varepsilon)P' = \sigma_\varepsilon^2 PP' = \frac{\sigma_u^2}{1-\phi^2} PP' = \Omega(\phi, \sigma_u^2)$.

GLRM: autocorrelated errors

- We have:

$$P(\phi)y = P(\phi)X\beta + P(\phi)\varepsilon \tag{5}$$

which has likelihood:

$$p(Y|\beta,\sigma^2,\phi) = (2\pi)^{-\frac{T}{2}} |\Omega|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(Y-X\beta)'(\Omega)^{-1}(Y-X\beta)\right]$$

where recall $\Omega = \Omega(\phi,\sigma_u^2)$

- One can specify:

$$\beta \sim N(\beta_0,\Sigma_0);\ \phi \sim N(\phi_0,\Sigma_{\phi_0});\ \sigma_u^2 \sim \Gamma^{-1}\left(\frac{\nu_0}{2},\frac{s_0^2}{2}\right)$$

- Under knowledge of σ_u^2 and ϕ, this gives the following posterior for β

$$\beta|\phi,\sigma^2,y \sim N(\beta_1,\Sigma_1)$$

$$\Sigma_1 = \left(\Sigma_0^{-1} + X'\Omega(\phi,\sigma_u^2)^{-1}X\right)^{-1},\ \beta_1 = \Sigma_1\left(\Sigma_0^{-1}\beta_0 + X'\Omega(\phi,\sigma_u^2)^{-1}Y\right)$$

which is simply the average of a GLS estimator and the prior.

GLRM: autocorrelated errors

- Then, under knowledge of β, it is easy to use the model in (2) to derive $\varepsilon|\beta, y$ and this can be used as an observable in (3). This gives

$$\varepsilon_1 = \varepsilon_0\phi + u, \;\; u_t \sim N(0, \sigma_u^2 I_{T-1}) \tag{6}$$

  which is a standard linear regression model with AR coefficient ϕ and error variance σ_u^2.
- Given the prior specified in (3), the posteriors will be:

$$\begin{aligned}
\phi|\beta, \sigma_u^2, y &\sim N(\phi_1, \Sigma_{\phi_1}); \\
\Sigma_{\phi_1} &= \left(\Sigma_{\phi_0}^{-1} + \frac{1}{\sigma_u^2}\varepsilon_0'\varepsilon_0\right)^{-1}, \; \beta_1 = \Sigma_{\phi_1}\left(\Sigma_{\phi_0}^{-1}\phi_0 + \frac{1}{\sigma_u^2}\varepsilon_0'\varepsilon_1\right) \\
\sigma_u^2|\beta, \phi, y &\sim \Gamma^{-1}\left(\frac{\nu_0 + T}{2}, \frac{s_0^2 + (\varepsilon_1 - \varepsilon_0\phi)'(\varepsilon_1 - \varepsilon_0\phi)}{2}\right)
\end{aligned}$$