Normal Linear Regression Model with Conjugate Priors

R scripts:
`mod6s2a, mod6s2b, mod6s2c, mod6s2d`

"Conjugate" refers to the property of a prior to generate, when combined with the likelihood function, a posterior that has the same density as the prior itself. Conjugate priors are mainly chosen for computational convenience, as they lead to exact analytical expressions for the posterior and the marginal likelihood. In practice, they are very restrictive, and in few real applications will we be able to formalize our prior understanding of the world in conjugate terms. However, this particular model is very useful for pedagogical purposes, and is thus a good starting point.

Importantly, it illustrates that Bayesian modeling is not inherently synonymous with "simulation techniques". The latter is simply a computational tool that broadens the choice and thus applicability of Bayesian models.

Model Specification

Step 1:
Start by writing down the regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim n\left(\mathbf{0}, \sigma^2 \mathbf{I}\right) \tag{1}$$

For this model with conjugate priors it is more convenient to work with the "precision" $h = \sigma^{-2}$ instead of the variance for the error term. Thus, our model parameters about which we wish to learn are $\boldsymbol{\theta} = \left[\boldsymbol{\beta}' \quad h\right]'$.

Step 2:
Next, write down the likelihood function (LHF) for this model. It's the same you would use for Maximum Likelihood Estimation (MLE), except perhaps for the switch from σ^2 to *h*. Nothing Bayesian yet.

$$p\left(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{X}\right) = \left(2\pi\right)^{-n/2} h^{n/2} \exp\left(-\tfrac{h}{2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)\right) \tag{2}$$

For this model we need to use a lot of "tricks" to get convenient and intuitive expressions for interim and final results. First, we will express the LHF in terms of the OLS estimator **b**, and the summed squared errors (*SSE*). Note the following equivalence:

$$(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}+\mathbf{X}\mathbf{b}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}+\mathbf{X}\mathbf{b}-\mathbf{X}\mathbf{b}) =$$

$$\left(\mathbf{y}-\mathbf{X}\mathbf{b}+\left(\mathbf{X}(\mathbf{b}-\boldsymbol{\beta})\right)\right)'\left(\mathbf{y}-\mathbf{X}\mathbf{b}+\left(\mathbf{X}(\mathbf{b}-\boldsymbol{\beta})\right)\right) =$$

$$(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b}) + 2\left(\mathbf{X}(\mathbf{b}-\boldsymbol{\beta})\right)'(\mathbf{y}-\mathbf{X}\mathbf{b}) + (\mathbf{b}-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\mathbf{b}-\boldsymbol{\beta}) =$$

$$(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b}) + (\mathbf{b}-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\mathbf{b}-\boldsymbol{\beta}) = SSE + (\mathbf{b}-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\mathbf{b}-\boldsymbol{\beta}) \tag{3}$$

since

$$\left(\mathbf{X}(\mathbf{b}-\boldsymbol{\beta})\right)'(\mathbf{y}-\mathbf{X}\mathbf{b}) = (\mathbf{b}-\boldsymbol{\beta})'\mathbf{X}'(\mathbf{y}-\mathbf{X}\mathbf{b}) = (\mathbf{b}-\boldsymbol{\beta})'\left(\mathbf{X}'\mathbf{y}-\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right) = 0$$

Note that only in the last line do we use the explicit form of the OLS estimator. Thus, we can express the LHF as

$$p(\mathbf{y}\,|\,\boldsymbol{\theta},\mathbf{X}) = (2\pi)^{-n/2}\,h^{n/2}\exp\left(-\tfrac{h}{2}\left(SSE + (\mathbf{b}-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\mathbf{b}-\boldsymbol{\beta})\right)\right) \tag{4}$$

Step 3:

Next, we need to define the prior distributions for our parameters. Here, we choose "conjugate priors" that, when combined with the likelihood, yield a posterior from the exact same conjugate family. A trademark of conjugate priors is that the prior for one set of parameters depends on the prior for another set. In other words, the parameters of interest do NOT have independent priors. Here, the prior of $\boldsymbol{\beta}$, conditional on *h*, is normal, and the prior of *h* is gamma.

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

$$\boldsymbol{\beta}\,|\,h \sim n\left(\boldsymbol{\mu}_0, h^{-1}\mathbf{V}_0\right), \qquad h \sim g(\tau_0, \nu_0)$$

$$p(\boldsymbol{\beta}\,|\,h) = (2\pi)^{-k/2}\left|h^{-1}\mathbf{V}_0\right|^{-1/2}\exp\left(-\tfrac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)'\left(h^{-1}\mathbf{V}_0\right)^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\right) \tag{5}$$

$$p(h) = \left(\frac{2\tau_0}{\nu_0}\right)^{-\nu_0/2}\Gamma\left(\tfrac{\nu_0}{2}\right)^{-1}h^{\left(\frac{\nu_0-2}{2}\right)}\exp\left(-\frac{h\nu_0}{2\tau_0}\right) \qquad \text{with} \quad \Gamma(x) = \int_0^\infty t^{x-1}\exp(-t)\,dt$$

where $\boldsymbol{\mu}_0$ is the prior mean for $\boldsymbol{\beta}$, $\mathbf{V}_0$ is the unscaled prior variance-covariance matrix for $\boldsymbol{\beta}$, and τ_0 and ν_0 are the mean and degrees of freedom (DoF) for the gamma distribution. (Note: In later examples and models we might express the gamma density in form of its shape and scale instead of mean and DoF. Here we're sticking with the version chosen in KPT, ch. 10). Note:

$$\left|h^{-1}\mathbf{V}_0\right| = h^{-k}\left|\mathbf{V}_0\right| \qquad \text{(e.g. Greene, A-48)}$$

$$\left(h^{-1}\mathbf{V}_0\right)^{-1} = h(\mathbf{V}_0)^{-1} \tag{6}$$

R script `mod6s2a` illustrates the synergies between these conjugate priors. The upshot is that with conjugate priors, the prior settings for one parameter will affect the prior density of another. A bit of trial and error (and density plotting) will be needed to arrive at the desired prior densities for all parameters.

Step 4:
Write down the complete posterior distribution.

$$
\begin{aligned}
&p(\boldsymbol{\beta},h\,|\,\mathbf{y},\mathbf{X}) = \frac{p(\boldsymbol{\beta},h)\,p(\mathbf{y}\,|\,\boldsymbol{\beta},h,\mathbf{X})}{p(\mathbf{y}\,|\,\mathbf{X})} \propto p(\boldsymbol{\beta},h)\,p(\mathbf{y}\,|\,\boldsymbol{\beta},h,\mathbf{X}) = p(\boldsymbol{\beta}\,|\,h)\,p(h)\,p(\mathbf{y}\,|\,\boldsymbol{\beta},h,\mathbf{X}) = \\
&(2\pi)^{-k/2}\,h^{k/2}\,|\mathbf{V_0}|^{-1/2}\exp\left(-\tfrac{h}{2}(\boldsymbol{\beta}-\boldsymbol{\mu_0})'\,\mathbf{V_0^{-1}}\,(\boldsymbol{\beta}-\boldsymbol{\mu_0})\right)* \\
&\left(\frac{2\tau_0}{\nu_0}\right)^{-\nu_0/2}\Gamma\left(\tfrac{\nu_0}{2}\right)^{-1}\,h^{\left(\frac{\nu_0-2}{2}\right)}\exp\left(-\frac{h\nu_0}{2\tau_0}\right)* \\
&(2\pi)^{-n/2}\,h^{n/2}\exp\left(-\tfrac{h}{2}\left(SSE+(\mathbf{b}-\boldsymbol{\beta})'\,\mathbf{X}'\mathbf{X}(\mathbf{b}-\boldsymbol{\beta})\right)\right)
\end{aligned}
\tag{7}
$$

Note the switch from equality ("=") to proportionality ($\propto$) in the first line. This simplifies our posterior computation since we don't need the marginal likelihood $p(\mathbf{y}|\mathbf{X})$ to find the posterior densities for our parameters.

At this point it is customary (for ANY Bayesian model) to further ignore (= drop) all terms that are multiplicatively unrelated to our parameters of interest. Thus:

$$
\begin{aligned}
&p(\boldsymbol{\beta}\,|\,h)\,p(h)\,p(\mathbf{y}\,|\,\boldsymbol{\beta},h,\mathbf{X}) \propto \\
&h^{k/2}\exp\left(-\tfrac{h}{2}(\boldsymbol{\beta}-\boldsymbol{\mu_0})'\,\mathbf{V_0^{-1}}\,(\boldsymbol{\beta}-\boldsymbol{\mu_0})\right)*h^{\left(\frac{\nu_0-2}{2}\right)}\exp\left(-\frac{h\nu_0}{2\tau_0}\right)* \\
&h^{n/2}\exp\left(-\tfrac{h}{2}\left(SSE+(\mathbf{b}-\boldsymbol{\beta})'\,\mathbf{X}'\mathbf{X}(\mathbf{b}-\boldsymbol{\beta})\right)\right).
\end{aligned}
\tag{8}
$$

Collecting terms the posterior kernel further simplifies to

$$
\begin{aligned}
&p(\boldsymbol{\beta}\,|\,h)\,p(h)\,p(\mathbf{y}\,|\,\boldsymbol{\beta},h,\mathbf{X}) \propto \\
&h^{\frac{k+\nu_0-2+n}{2}}\exp\left(-\tfrac{h}{2}\left(\nu_0\tau_0^{-1}+SSE+(\boldsymbol{\beta}-\boldsymbol{\mu_0})'\,\mathbf{V_0^{-1}}\,(\boldsymbol{\beta}-\boldsymbol{\mu_0})+(\mathbf{b}-\boldsymbol{\beta})'\,\mathbf{X}'\mathbf{X}(\mathbf{b}-\boldsymbol{\beta})\right)\right).
\end{aligned}
\tag{9}
$$

Here comes another trick:

Let $\mathbf{V_1} = \left(\mathbf{V_0^{-1}} + \mathbf{X'X}\right)^{-1}$ and $\boldsymbol{\mu_1} = \mathbf{V_1}\left(\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \mathbf{X'Xb}\right) = \mathbf{V_1}\left(\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \mathbf{X'y}\right)$

Then:

$$\begin{aligned}
&(\boldsymbol{\beta} - \boldsymbol{\mu_0})' \mathbf{V_0^{-1}} (\boldsymbol{\beta} - \boldsymbol{\mu_0}) + (\mathbf{b} - \boldsymbol{\beta})' \mathbf{X'X}(\mathbf{b} - \boldsymbol{\beta}) = \\
&\boldsymbol{\beta'}\mathbf{V_0^{-1}}\boldsymbol{\beta} - \boldsymbol{\mu_0'}\mathbf{V_0^{-1}}\boldsymbol{\beta} - \boldsymbol{\beta'}\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \boldsymbol{\mu_0'}\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \boldsymbol{\beta'}\mathbf{X'X}\boldsymbol{\beta} - \boldsymbol{\beta'}\mathbf{X'Xb} - \mathbf{b'X'X}\boldsymbol{\beta} + \mathbf{b'X'Xb} = \\
&\boldsymbol{\beta'}\left(\mathbf{V_0^{-1}} + \mathbf{X'X}\right)\boldsymbol{\beta} - \boldsymbol{\beta'}\left(\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \mathbf{X'Xb}\right) - \left(\boldsymbol{\mu_0'}\mathbf{V_0^{-1}} + \mathbf{b'X'X}\right)\boldsymbol{\beta} + \boldsymbol{\mu_0'}\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \mathbf{b'X'Xb} = \\
&\boldsymbol{\beta'}\mathbf{V_1^{-1}}\boldsymbol{\beta} - \boldsymbol{\beta'}\mathbf{V_1^{-1}}\boldsymbol{\mu_1} - \boldsymbol{\mu_1'}\mathbf{V_1^{-1}}\boldsymbol{\beta} + \boldsymbol{\mu_0'}\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \mathbf{b'X'Xb}, \quad \text{since}
\end{aligned} \tag{10}$$

$$\mathbf{V_1^{-1}}\boldsymbol{\mu_1} = \mathbf{V_1^{-1}}\mathbf{V_1}\left(\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \mathbf{X'Xb}\right) = \left(\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \mathbf{X'Xb}\right) \quad \text{and}$$

$$\left(\boldsymbol{\mu_0'}\mathbf{V_0^{-1}} + \mathbf{b'X'X}\right) = \left(\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \mathbf{X'Xb}\right)' = \left(\mathbf{V_1^{-1}}\boldsymbol{\mu_1}\right)' = \boldsymbol{\mu_1'}\mathbf{V_1^{-1}}.$$

Simplifying further:

$$\begin{aligned}
&\boldsymbol{\beta'}\mathbf{V_1^{-1}}\boldsymbol{\beta} - \boldsymbol{\beta'}\mathbf{V_1^{-1}}\boldsymbol{\mu_1} - \boldsymbol{\mu_1'}\mathbf{V_1^{-1}}\boldsymbol{\beta} + \boldsymbol{\mu_0'}\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \mathbf{b'X'Xb} = \\
&\boldsymbol{\beta'}\mathbf{V_1^{-1}}\boldsymbol{\beta} - \boldsymbol{\beta'}\mathbf{V_1^{-1}}\boldsymbol{\mu_1} - \boldsymbol{\mu_1'}\mathbf{V_1^{-1}}\boldsymbol{\beta} + \boldsymbol{\mu_1'}\mathbf{V_1^{-1}}\boldsymbol{\mu_1} - \boldsymbol{\mu_1'}\mathbf{V_1^{-1}}\boldsymbol{\mu_1} + \boldsymbol{\mu_0'}\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \mathbf{b'X'Xb} = \\
&(\boldsymbol{\beta} - \boldsymbol{\mu_1})' \mathbf{V_1^{-1}} (\boldsymbol{\beta} - \boldsymbol{\mu_1}) - \boldsymbol{\mu_1'}\mathbf{V_1^{-1}}\boldsymbol{\mu_1} + \boldsymbol{\mu_0'}\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \mathbf{b'X'Xb}.
\end{aligned} \tag{11}$$

The last three terms can be further "massaged" to yield:

$$\begin{aligned}
&\boldsymbol{\mu_0'}\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \mathbf{b'X'Xb} - \boldsymbol{\mu_1'}\mathbf{V_1^{-1}}\boldsymbol{\mu_1} = \\
&\boldsymbol{\mu_0'}\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \mathbf{b'X'Xb} - \left(\left(\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \mathbf{X'Xb}\right)' \mathbf{V_1}\left(\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \mathbf{X'Xb}\right)\right) = \\
&\boldsymbol{\mu_0'}\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \mathbf{b'X'Xb} - \left(\boldsymbol{\mu_0'}\mathbf{V_0^{-1}}\mathbf{V_1}\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \mathbf{b'X'XV_1}\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \boldsymbol{\mu_0'}\mathbf{V_0^{-1}}\mathbf{V_1}\mathbf{X'Xb} + \mathbf{b'X'XV_1}\mathbf{X'Xb}\right) = \\
&\boldsymbol{\mu_0'}\mathbf{V_0^{-1}}\boldsymbol{\mu_0} - \boldsymbol{\mu_0'}\mathbf{V_0^{-1}}\mathbf{V_1}\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \boldsymbol{\mu_0'}\mathbf{X'XV_1}\mathbf{V_0^{-1}}\boldsymbol{\mu_0} - \boldsymbol{\mu_0'}\mathbf{X'XV_1}\mathbf{V_0^{-1}}\boldsymbol{\mu_0} + \\
&\mathbf{b'X'Xb} - \mathbf{b'X'XV_1}\mathbf{X'Xb} + \mathbf{b'X'XV_1}\mathbf{V_0^{-1}}\mathbf{b} - \mathbf{b'X'XV_1}\mathbf{V_0^{-1}}\mathbf{b} - \\
&\mathbf{b'X'XV_1}\mathbf{V_0^{-1}}\boldsymbol{\mu_0} - \boldsymbol{\mu_0'}\mathbf{V_0^{-1}}\mathbf{V_1}\mathbf{X'Xb} = \\
&(\mathbf{b} - \boldsymbol{\mu_0})' \mathbf{X'XV_1}\mathbf{V_0^{-1}}(\mathbf{b} - \boldsymbol{\mu_0}) \qquad \text{since}
\end{aligned} \tag{12}$$

$$\mathbf{b'X'Xb} - \mathbf{b'X'XV_1}\mathbf{X'Xb} - \mathbf{b'X'XV_1}\mathbf{V_0^{-1}}\mathbf{b} = \mathbf{b'}\left(\mathbf{X'X}\left(\mathbf{I} - \mathbf{V_1}\left(\mathbf{X'X} + \mathbf{V_0^{-1}}\right)\right)\right)\mathbf{b} = 0$$

$$\boldsymbol{\mu_0'}\mathbf{V_0^{-1}}\boldsymbol{\mu_0} - \boldsymbol{\mu_0'}\mathbf{V_0^{-1}}\mathbf{V_1}\mathbf{V_0^{-1}}\boldsymbol{\mu_0} - \boldsymbol{\mu_0'}\mathbf{X'XV_1}\mathbf{V_0^{-1}}\boldsymbol{\mu_0} = \boldsymbol{\mu_0'}\left(\mathbf{V_0^{-1}}\left(\mathbf{I} - \mathbf{V_1}\left(\mathbf{V_0^{-1}} + \mathbf{X'X}\right)\right)\right)\boldsymbol{\mu_0} = 0$$

Almost done – one final simplification:

$$
\begin{aligned}
&\left(\mathbf{b}-\boldsymbol{\mu}_{\mathbf{0}}\right)' \mathbf{X}'\mathbf{X}\mathbf{V}_{\mathbf{1}}\mathbf{V}_{\mathbf{0}}^{\mathbf{-1}}\left(\mathbf{b}-\boldsymbol{\mu}_{\mathbf{0}}\right) = \\
&\left(\mathbf{b}-\boldsymbol{\mu}_{\mathbf{0}}\right)' \mathbf{X}'\mathbf{X}\left(\mathbf{V}_{\mathbf{0}}^{\mathbf{-1}}+\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{V}_{\mathbf{0}}^{\mathbf{-1}}\left(\mathbf{b}-\boldsymbol{\mu}_{\mathbf{0}}\right) = \\
&\left(\mathbf{b}-\boldsymbol{\mu}_{\mathbf{0}}\right)' \mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{V}_{\mathbf{0}}\left(\mathbf{V}_{\mathbf{0}}^{\mathbf{-1}}+\mathbf{X}'\mathbf{X}\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)^{-1}\mathbf{V}_{\mathbf{0}}^{\mathbf{-1}}\mathbf{V}_{\mathbf{0}}\left(\mathbf{b}-\boldsymbol{\mu}_{\mathbf{0}}\right) = \\
&\left(\mathbf{b}-\boldsymbol{\mu}_{\mathbf{0}}\right)' \left(\mathbf{V}_{\mathbf{0}}+\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)^{-1}\left(\mathbf{b}-\boldsymbol{\mu}_{\mathbf{0}}\right).
\end{aligned}
\tag{13}
$$

After all this tedious work we can now write the posterior kernel as:

$$
\begin{aligned}
&p\left(\boldsymbol{\beta},h\,|\,\mathbf{y},\mathbf{X}\right)\propto \\
&h^{\frac{k+v_0-2+n}{2}}\exp\left(-\tfrac{h}{2}\left(v_0\tau_0^{-1}+SSE+\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathbf{1}}\right)'\mathbf{V}_{\mathbf{1}}^{\mathbf{-1}}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathbf{1}}\right)+\left(\mathbf{b}-\boldsymbol{\mu}_{\mathbf{0}}\right)'\left(\mathbf{V}_{\mathbf{0}}+\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)^{-1}\left(\mathbf{b}-\boldsymbol{\mu}_{\mathbf{0}}\right)\right)\right)= \\
&h^{\frac{k+v_0-2+n}{2}}\exp\left(-\tfrac{h}{2}\left(v_0\tau_0^{-1}+SSE+\left(\mathbf{b}-\boldsymbol{\mu}_{\mathbf{0}}\right)'\left(\mathbf{V}_{\mathbf{0}}+\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)^{-1}\left(\mathbf{b}-\boldsymbol{\mu}_{\mathbf{0}}\right)\right)\right)*\exp\left(-\tfrac{h}{2}\left(\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathbf{1}}\right)'\mathbf{V}_{\mathbf{1}}^{\mathbf{-1}}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathbf{1}}\right)\right)\right).
\end{aligned}
\tag{14}
$$

Step 5:
It is customary in Bayesian posterior analysis to break the posterior kernel into conditional components that have a well-understood density. Here, we want to first look at the posterior for $\boldsymbol{\beta}$, conditional on *h*. This means we can ignore all terms in (14) that are multiplicatively unrelated to $\boldsymbol{\beta}$. We get:

$$
p\left(\boldsymbol{\beta}\,|\,h,\mathbf{y},\mathbf{X}\right)\propto\exp\left(-\tfrac{h}{2}\left(\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathbf{1}}\right)'\mathbf{V}_{\mathbf{1}}^{\mathbf{-1}}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathbf{1}}\right)\right)\right)=\left(-\tfrac{1}{2}\left(\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathbf{1}}\right)'\left(h^{-1}\mathbf{V}_{1}\right)^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathbf{1}}\right)\right)\right).
\tag{15}
$$

This is *exactly* the kernel of a multivariate normal distribution. Thus we have established that

$$
\boldsymbol{\beta}\,|\,h,\mathbf{y},\mathbf{X}\sim n\left(\boldsymbol{\mu}_{\mathbf{1}},h^{-1}\mathbf{V}_{\mathbf{1}}\right)
\tag{16}
$$

While this is not our end-product, it is a useful finding which we will also exploit in future models.

Step 6:
To derive the marginal posterior density for *h*, we can use a "reversed" version of Bayes' Rule:

$$
p\left(h\,|\,\mathbf{y},\mathbf{X}\right)=\frac{p\left(\boldsymbol{\beta},h\,|\,\mathbf{y},\mathbf{X}\right)}{p\left(\boldsymbol{\beta}\,|\,h,\mathbf{y},\mathbf{X}\right)}\propto\frac{p\left(\boldsymbol{\beta}\,|\,h\right)p\left(h\right)p\left(\mathbf{y}\,|\,\boldsymbol{\beta},h,\mathbf{X}\right)}{p\left(\boldsymbol{\beta}\,|\,h,\mathbf{y},\mathbf{X}\right)}
\tag{17}
$$

Thus, we should get the kernel for $p\left(h\,|\,\mathbf{y},\mathbf{X}\right)$ by dividing the joint posterior kernel in (14) by the full conditional density of $\boldsymbol{\beta}$. Dropping terms that are multiplicatively unrelated to *h*, this yields:

$$
\begin{aligned}
&p(h\,|\,\mathbf{y},\mathbf{X}) \propto h^{\frac{k+v_0-2+n}{2}} \exp\left(-\tfrac{h}{2}\left(v_0\tau_0^{-1} + SSE + (\mathbf{b}-\boldsymbol{\mu}_\mathbf{0})'\left(\mathbf{V_0} + (\mathbf{X}'\mathbf{X})^{-1}\right)^{-1}(\mathbf{b}-\boldsymbol{\mu}_\mathbf{0})\right)\right) * \\
&\exp\left(-\tfrac{h}{2}\left((\boldsymbol{\beta}-\boldsymbol{\mu}_\mathbf{1})'\,\mathbf{V_1^{-1}}(\boldsymbol{\beta}-\boldsymbol{\mu}_\mathbf{1})\right)\right) * \\
&\left((2\pi)^{-k/2} h^{k/2} |\mathbf{V_1}|^{-1/2} \exp\left(-\tfrac{h}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_1)'\,\mathbf{V_1^{-1}}(\boldsymbol{\beta}-\boldsymbol{\mu}_1)\right)\right)^{-1} \propto \\
&h^{\frac{v_0-2+n}{2}} \exp\left(-\tfrac{h}{2}\left(v_0\tau_0^{-1} + SSE + (\mathbf{b}-\boldsymbol{\mu}_\mathbf{0})'\left(\mathbf{V_0} + (\mathbf{X}'\mathbf{X})^{-1}\right)^{-1}(\mathbf{b}-\boldsymbol{\mu}_\mathbf{0})\right)\right) = \\
&h^{\frac{v_1-2}{2}} \exp\left(-\tfrac{h}{2}\left(v_0\tau_0^{-1} + v_1\tau_1^{-1}\right)\right) \qquad \text{where} \\
&v_1 = v_0 + n \qquad \text{and} \qquad \tau_1^{-1} = \frac{1}{v_1}\left(v_0\tau_0^{-1} + SSE + (\mathbf{b}-\boldsymbol{\mu}_\mathbf{0})'\left(\mathbf{V_0} + (\mathbf{X}'\mathbf{X})^{-1}\right)^{-1}(\mathbf{b}-\boldsymbol{\mu}_\mathbf{0})\right)
\end{aligned}
\tag{18}
$$

Comparing the final expression to the prior of h in (5), we can see that we have, again, the kernel of a gamma distribution. Thus we can state that

$$h\,|\,\mathbf{y},\mathbf{X} \sim g(\tau_1, v_1) \tag{19}$$

The marginal posterior for $\boldsymbol{\beta}$ can now be derived by marginalizing the joint posterior over *h*, i.e

$$p(\boldsymbol{\beta}\,|\,\mathbf{y},\mathbf{X}) = \int p(\boldsymbol{\beta}, h\,|\,\mathbf{y},\mathbf{X})\,dh = \int p(\boldsymbol{\beta}\,|\,h,\mathbf{y},\mathbf{X})\; p(h\,|\,\mathbf{y},\mathbf{X})\,dh \tag{20}$$

All functions in the integrand are fully known, so this is straightforward, but tedious. We jump directly to the final result. It turns out that the marginal posterior of $\boldsymbol{\beta}$ follows a *t*- distribution with mean $\boldsymbol{\mu}_\mathbf{1}$, variance $\tau_1^{-1}\mathbf{V}_1$, and DoF v_1, i.e.

$$\boldsymbol{\beta}\,|\,\mathbf{y},\mathbf{X} \sim t\left(\boldsymbol{\mu}_\mathbf{1}, \tau_1^{-1}\mathbf{V}_1, v_1\right). \tag{21}$$

Step 7:

The final construct of interest is the marginal likelihood p(**y**). It is useful for model comparison and hypothesis testing. For the conjugate normal model it has an analytical solution derived via

$$p(\mathbf{y}) = \int p(\mathbf{y}\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta})\,d\boldsymbol{\theta} \qquad\qquad \boldsymbol{\theta} = \begin{bmatrix}\boldsymbol{\beta}' & h\end{bmatrix}' \tag{22}$$

We will again skip the details of this derivation. The final result is

$$p(\mathbf{y}) = \frac{\Gamma(v_1/2)\left(v_0\tau_0^{-1}\right)^{v_0/2}}{\Gamma(v_0/2)\pi^{n/2}}\left(\frac{|\mathbf{V_1}|}{|\mathbf{V_0}|}\right)^{1/2}\left(v_1\tau_1^{-1}\right)^{-v_1/2} \tag{23}$$

R Implementation

R script `mod6s2b` follows these lecture notes exactly. Since with conjugate priors we know the exact form of marginal posteriors for all parameters, there is no "posterior simulation", such as a Gibbs Sampler, required for this model.

We are usually interested in the posterior mean and standard deviation for our parameters. From the lecture notes we know that

$$\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X} \sim t\left(\boldsymbol{\mu}_{\mathbf{1}}, \tau_1^{-1}\mathbf{V}_1, v_1\right) \qquad \text{and} \qquad h \mid \mathbf{y}, \mathbf{X} \sim g\left(\tau_1, v_1\right)$$

Thus, focusing on a specific element of $\boldsymbol{\beta}$, say β_j, we immediately know $E\left(\beta_j\right) = \mu_{1,j}$ and $std\left(\beta_j\right) = \sqrt{\tau_1^{-1}\mathbf{V}_{1,jj}}$, where $\mu_{1,j}$ is the jth element of $\boldsymbol{\mu}_{\mathbf{1}}$, and $\mathbf{V}_{\mathbf{1},jj}$ is the (j,j) element of $\mathbf{V}_{\mathbf{1}}$ (i.e. the j^{th} diagonal element). Similarly, we know $E\left(h\right) = \tau_1$ and, by the properties of the gamma density we can directly compute $std\left(h\right) = \sqrt{\frac{2\tau_1^2}{v_1}}$. These posterior moments are sent to the log file.

Script `mod6s2b` also takes draws from the prior and posterior distributions for all parameters and compares the resulting densities in a set of nonparametric plots. It is clear that the posterior densities are much tighter than the priors in all cases. Thus, the sample data have indeed contributed to "learning" about the parameters.

Model Comparison

We compare models with different sets of regressors using their (logged) marginal likelihood and Bayes Factors. Specifically, we estimate a known-to-be- false model that drops the last regressor in **X** in script `mod6s2c`.

Scripts `mod6s2d` loads the marginal likelihood from both models and computes the logged Bayes Factor of model 1 ("main") over model 2 ("constrained").

Kass and Raftery (1995) propose the following decision rules for one model over another based on logged Bayes Factors:

Log BF	Strength of evidence in favor of model with higher marg. LH
0 - 6.9	"positive"
6.9 – 11.5	"strong"
>11.5	"decisive"

In this case, the first model receives "decisive" support from the sample data. In other words, it is much more probable for the first model to have generated the sample data than for the second model.

References:

Kass, R. E. and A. E. Raftery, 1995, Bayes Factors. Journal of the American Statistical Association, 90, 773-795.