# The case of linear regression models

**Jean-Michel Marin**

Université de Montpellier
Institut Montpelliérain Alexander Grothendieck

## Plan

## The model

The dataset is then made up of the reunion of the vector of outcomes

$$\mathbf{y} = (y_1, \ldots, y_n)$$

and the $n \times p$ matrix of explanatory variables

$$\mathbf{X} = [\mathbf{x}_1 \quad \ldots \quad \mathbf{x}_p] = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ x_{31} & x_{32} & \ldots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} .$$

The ordinary Gaussian linear regression model is such that:

$$\mathbf{y}|\alpha, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}_n\left(\alpha \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}, \sigma^2\, \mathbf{I}_n\right)\ ,$$

$y_i$'s are independent normal random variables with

$$\mathbb{E}[y_i|\alpha, \boldsymbol{\beta}, \sigma^2] = \alpha + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\,, \quad \mathbb{V}[y_i|\alpha, \boldsymbol{\beta}, \sigma^2] = \sigma^2\,.$$

Given that the models studied in this section are all conditional on the regressors, we omit the conditioning on $\mathbf{X}$ to simplify the notations.

We assume that $\text{rank}\,[\mathbf{1}_n \quad \mathbf{X}] = p + 1$.

$$\ell(\alpha, \boldsymbol{\beta}, \sigma^2|\mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}\right)\right\}.$$

**Natural conjugate prior family**

$$(\alpha, \boldsymbol{\beta})|\sigma^2 \sim \mathcal{N}_{p+1}((\tilde{\alpha}, \tilde{\boldsymbol{\beta}}), \sigma^2 M^{-1}),$$

conditional on $\sigma^2$ and

$$\sigma^2 \sim \mathcal{IG}(a, b).$$

**Even in the presence of genuine information on the parameters, the hyperparameters $M$, $a$ and $b$ are very difficult to specify and the posterior distributions.**

Ridge regression: $(\tilde{\alpha}, \tilde{\boldsymbol{\beta}}) = 0_{p+1}$ and $M = I_n$

**Zellner's $G$-prior**

$$\boldsymbol{\beta}|\alpha, \sigma^2 \sim \mathcal{N}_p \left( \tilde{\boldsymbol{\beta}}, g\sigma^2 (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \right) ,$$

and a noninformative prior distribution is imposed on the pair $(\alpha, \sigma^2)$,

$$\pi \left( \alpha, \sigma^2 \right) \propto \sigma^{-2} \,.$$

The factor $g$ can be interpreted as being inversely proportional to the amount of information available in the prior relative to the sample.

For instance, setting $g = n$ gives the prior the same weight as one observation of the sample.

We will use this as our default value.

When $p > 0$,

$$\alpha|\sigma^2, \mathbf{y} \sim \mathcal{N}_1\left(\bar{\mathbf{y}}, \sigma^2/n\right) ,$$

$$\boldsymbol{\beta}|\mathbf{y}, \sigma^2 \sim \mathcal{N}_p\left(\frac{g}{g+1}\left(\hat{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}/g\right), \frac{\sigma^2 g}{g+1}\left\{\mathbf{X}^{\mathrm{T}}\mathbf{X}\right\}^{-1}\right) ,$$

where $\hat{\boldsymbol{\beta}} = \left\{\mathbf{X}^{\mathrm{T}}\mathbf{X}\right\}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ is the maximum likelihood and least squares estimator of $\boldsymbol{\beta}$.

The posterior independence between $\alpha$ and $\boldsymbol{\beta}$ is due to the fact that $\mathbf{X}$ is centered and that $\alpha$ and $\boldsymbol{\beta}$ are a priori independent.

$$\sigma^2|\mathbf{y} \sim \mathcal{IG}\left[(n-1)/2, s^2 + (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})/(g+1)\right]$$

where $s^2 = (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n - \mathbf{X}\hat{\boldsymbol{\beta}})$

When $p = 0$,

$$\alpha|\mathbf{y}, \sigma^2 \sim \mathcal{N}\left(\bar{\mathbf{y}}, \sigma^2/n\right) ,$$

$$\sigma^2|\mathbf{y} \sim \mathcal{IG}\left[(n-1)/2, (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n)^{\mathrm{T}}(\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n)/2\right] .$$

We can derive from the previous derivations that

$$\mathbb{E}^\pi\left[\alpha|\mathbf{y}\right] = \mathbb{E}^\pi\left[\mathbb{E}^\pi\left(\alpha|\sigma^2,\mathbf{y}\right)|\mathbf{y}\right] = \mathbb{E}^\pi\left[\bar{\mathbf{y}}|\mathbf{y}\right] = \bar{\mathbf{y}}$$

$$\begin{aligned}\mathbb{E}^\pi\left[\boldsymbol{\beta}|\mathbf{y}\right] &= \mathbb{E}^\pi\left[\mathbb{E}^\pi\left(\boldsymbol{\beta}|\sigma^2,\mathbf{y}\right)\Big|\mathbf{y}\right] \\ &= \mathbb{E}^\pi\left[\frac{g}{g+1}(\hat{\boldsymbol{\beta}}+\tilde{\boldsymbol{\beta}}/g)\Big|\mathbf{y}\right] \\ &= \frac{g}{g+1}(\hat{\boldsymbol{\beta}}+\tilde{\boldsymbol{\beta}}/g)\,.\end{aligned}$$

This result gives its meaning to the above point relating $g$ with the amount of information contained in the dataset.

$$\mathbb{E}^{\pi}\left[\boldsymbol{\beta}|\mathbf{y}\right] = \frac{g}{g+1}(\hat{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}/g)$$

When $g = 1$, the prior information has the same weight as this amount: the Bayesian estimate of $\boldsymbol{\beta}$ is the average between the least square estimator and the prior expectation.

**The larger $g$ is, the weaker the prior information and the closer the Bayesian estimator is to the least squares estimator.**

When considering the marginal likelihood at the core of the Bayes factors, we have, if $p \neq 0$,

$$f(\mathbf{y}) = \int \left( \int \int f(\mathbf{y}|\alpha, \boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta}|\alpha, \sigma^2)\pi(\sigma^2, \alpha)\mathrm{d}\alpha\mathrm{d}\boldsymbol{\beta} \right) \mathrm{d}\sigma^2 \,,$$

$$f(\mathbf{y}) = \frac{\delta\Gamma((n-1)/2)}{\pi^{(n-1)/2} n^{1/2}} (g+1)^{-p/2} \kappa^{-(n-1)/2} \,.$$

$$\begin{aligned}\kappa &= (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n)^{\mathrm{T}}(\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n) + \frac{1}{g+1} \left\{ -g\mathbf{y}^{\mathrm{T}}\mathbf{P}\mathbf{y} + \tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{P}\mathbf{X}\tilde{\boldsymbol{\beta}} - 2\mathbf{y}^{\mathrm{T}}\mathbf{P}\mathbf{X}\tilde{\boldsymbol{\beta}} \right\} \\ &= s^2 + (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})/(g+1) \,.\end{aligned}$$

If $p = 0$, a similar expression emerges:

$$f(\mathbf{y}) = \int \left( \int f(\mathbf{y}|\alpha, \sigma^2)\pi(\alpha, \sigma^2)\mathrm{d}\alpha \right) \mathrm{d}\sigma^2 \,,$$

$$f(\mathbf{y}) = \frac{\delta\Gamma((n-1)/2)}{\pi^{(n-1)/2} n^{1/2}} \left[ (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n)^{\mathrm{T}} (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n) \right]^{-(n-1)/2}$$

as the evidence associated with this “null” model.

## Model choice

The computation of model's posterior probabilities is plagued by the inability to include generic improper prior distributions.

In order to bypass this difficulty, we will assume that all the linear models under comparison do include the parameter $\alpha$, which means that each regression model includes an intercept.

This assumption allows us to take the *same* improper prior on $(\alpha, \sigma^2)$ for all of those models.

When we compare two sets of regressors, we have to handle two regression matrices, $\mathbf{X}^1$ and $\mathbf{X}^2$, with respective dimensions $(n, p_1)$ and $(n, p_2)$, extracted from the original matrix $\mathbf{X}$ by removing some columns.

$$\mathbb{P}^\pi(\mathfrak{M} = 1|\mathbf{y}) \propto (g_1 + 1)^{-p_1/2}$$
$$\left[s_1^2 + (\tilde{\boldsymbol{\beta}}^1 - \hat{\boldsymbol{\beta}}^1)^{\mathrm{T}}(\mathbf{X}^1)^{\mathrm{T}}\mathbf{X}^1(\tilde{\boldsymbol{\beta}}^1 - \hat{\boldsymbol{\beta}}^1)\big/(g_1 + 1)\right]^{-(n-1)/2}$$

$$\mathbb{P}^\pi(\mathfrak{M} = 2|\mathbf{y}) \propto (g_2 + 1)^{-p_2/2}$$
$$\left[s_2^2 + (\tilde{\boldsymbol{\beta}}^2 - \hat{\boldsymbol{\beta}}^2)^{\mathrm{T}}(\mathbf{X}^2)^{\mathrm{T}}\mathbf{X}^2(\tilde{\boldsymbol{\beta}}^2 - \hat{\boldsymbol{\beta}}^2)\big/(g_2 + 1)\right]^{-(n-1)/2}$$

**Prediction**

The prediction of $m \geq 1$ future observations from units for which the explanatory variables $\tilde{\mathbf{X}}$—but not the outcome variable $\tilde{\mathbf{y}}$—have been observed or set is also based on the posterior distribution.

Logically enough, were $\alpha$, $\boldsymbol{\beta}$ and $\sigma^2$ known quantities, the $m$-vector $\tilde{\mathbf{y}}$ would then have a Gaussian distribution with mean $\alpha\mathbf{1}_m + \tilde{\mathbf{X}}\boldsymbol{\beta}$ and variance $\sigma^2\mathbf{I}_m$.

Conditional on $\sigma^2$, the vector $\tilde{\mathbf{y}}$ of future observations has a Gaussian distribution and we can derive its expectation–used as our Bayesian estimator–by averaging over $\alpha$ and $\boldsymbol{\beta}$,

$$\begin{aligned}
\mathbb{E}^\pi[\tilde{\mathbf{y}}|\sigma^2, \mathbf{y}] &= \mathbb{E}^\pi[\mathbb{E}^\pi(\tilde{\mathbf{y}}|\alpha, \boldsymbol{\beta}, \sigma^2, \mathbf{y})|\sigma^2, \mathbf{y}] \\
&= \mathbb{E}^\pi[\alpha \mathbf{1}_m + \tilde{\mathbf{X}}\boldsymbol{\beta}|\sigma^2, \mathbf{y}] \\
&= \hat{\alpha}\mathbf{1}_m + \tilde{\mathbf{X}}\frac{\tilde{\boldsymbol{\beta}} + g\hat{\boldsymbol{\beta}}}{g+1},
\end{aligned}$$

which is independent from $\sigma^2$. This representation is quite intuitive, being the product of the matrix of explanatory variables $\tilde{\mathbf{X}}$ by the Bayesian estimator of $\boldsymbol{\beta}$.

Similarly, we can compute

$$\begin{aligned}\mathbb{V}^\pi(\tilde{\mathbf{y}}|\sigma^2,\mathbf{y}) &= \mathbb{E}^\pi[\mathbb{V}^\pi(\tilde{\mathbf{y}}|\alpha,\boldsymbol{\beta},\sigma^2,\mathbf{y})|\sigma^2,\mathbf{y}] \\ &\quad +\mathbb{V}^\pi(\mathbb{E}^\pi(\tilde{\mathbf{y}}|\alpha,\boldsymbol{\beta},\sigma^2,\mathbf{y})|\sigma^2,\mathbf{y}) \\ &= \mathbb{E}^\pi[\sigma^2 I_m|\sigma^2,\mathbf{y}] + \mathbb{V}^\pi(\alpha\mathbf{1}_m + \tilde{X}\boldsymbol{\beta}|\sigma^2,\mathbf{y}) \\ &= \sigma^2\left(I_m + \frac{g}{g+1}\tilde{\mathbf{X}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}\right).\end{aligned}$$

Due to this factorisation, and the fact that the conditional expectation does not depend on $\sigma^2$, we thus obtain

$$\mathbb{V}^\pi(\tilde{\mathbf{y}}|\mathbf{y}) = \hat{\sigma}^2\left(I_m + \frac{g}{g+1}\tilde{\mathbf{X}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}\right).$$

Conditionally on $\sigma^2$, the posterior predictive variance has two terms, the first term being $\sigma^2 I_m$, which corresponds to the sampling variation, and the second one being $\sigma^2 \frac{g}{g+1} \tilde{\mathbf{X}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}$, which corresponds to the uncertainty about $\boldsymbol{\beta}$.

HPD credible regions and tests can then be conducted based on this conditional predictive distribution

$$\tilde{\mathbf{y}}|\mathbf{y}, \sigma^2 \sim \mathcal{N}\left(\mathbb{E}^\pi[\tilde{\mathbf{y}}], \mathbb{V}^\pi(\tilde{\mathbf{y}}|\mathbf{y}, \sigma^2)\right) \ .$$

Integrating $\sigma^2$ out to produce the marginal distribution of $\tilde{\mathbf{y}}$ leads to a multivariate Student's $t$ distribution

$$\begin{aligned}\tilde{\mathbf{y}}|\mathbf{y} \quad \sim \quad & \mathcal{T}_m \bigg( n, \hat{\alpha}\mathbf{1}_m + g\tilde{\boldsymbol{\beta}}/(g+1), \\ & \frac{s^2 + \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}}}{n} \left\{ \mathbf{I}_m + \tilde{\mathbf{X}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\tilde{\mathbf{X}}^{\mathrm{T}} \right\} \bigg) .\end{aligned}$$