The case of linear models

Jean-Michel Marin

University of Montpellier
Faculty of Sciences

HAX918X / 2023-2024

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 model

The ordinary Gaussian linear regression model is such that

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

The y_i are independent Gaussian random variables with

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

$$\mathbb{V}[y_i|\alpha, \boldsymbol{\beta}, \sigma^2] = \sigma^2$$

We omit the conditioning on $\mathbf{X}$ to simplify the notations

The model

We assume that rank $[\mathbf{1}_n \quad \mathbf{X}] = p + 1$
identifiability contraint

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

Natural conjugate prior family

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

$$\sigma^2 \sim \mathscr{I}\mathscr{G}(a, b)$$

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

Zellner's G-prior

X is centered

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

and a noninformative prior distribution is imposed on the pair (α, σ^2)

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

Zellner's G-prior

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

Zellner's G-prior

When $p > 0$

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

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

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

The posterior independence between α and β is due to the fact that X is centered and that α and β are a priori independent

Zellner's G-prior

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

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

When $p = 0$,

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

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

Zellner's G-prior

We can derive from the previous derivations that

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

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

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

Zellner's G-prior

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

When $g = 1$, the prior information has the same weight as this amount: the Bayesian estimate of β 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

Zellner's G-prior

If $p \neq 0$, the evidence of the linear regression model is

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

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

$$\kappa = s^2 + (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^{\mathsf{T}} \mathbf{X}^{\mathsf{T}} \mathbf{X} (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \big/ (g+1)$$

Zellner's G-prior

If $p = 0$, a similar expression emerges, the evidence of the null model is

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

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

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 α, which means that each regression model includes an intercept

This assumption allows us to take the *same* improper prior on (α, σ^2) for all of those models

Model choice

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)

$$\mathbb{P}^{\pi}(\mathfrak{M} = 1|\mathbf{y}) \propto (g_1+1)^{-p_1/2}$$

$$\left[s_1^2 + (\tilde{\boldsymbol{\beta}}^1 - \hat{\boldsymbol{\beta}}^1)^{\mathsf{T}}(\mathbf{X}^1)^{\mathsf{T}}\mathbf{X}^1(\tilde{\boldsymbol{\beta}}^1 - \hat{\boldsymbol{\beta}}^1)/(g_1+1)\right]^{-(n-1)/2}$$

$$\mathbb{P}^{\pi}(\mathfrak{M} = 2|\mathbf{y}) \propto (g_2+1)^{-p_2/2}$$

$$\left[s_2^2 + (\tilde{\boldsymbol{\beta}}^2 - \hat{\boldsymbol{\beta}}^2)^{\mathsf{T}}(\mathbf{X}^2)^{\mathsf{T}}\mathbf{X}^2(\tilde{\boldsymbol{\beta}}^2 - \hat{\boldsymbol{\beta}}^2)/(g_2+1)\right]^{-(n-1)/2}$$

Prediction

The prediction of $m \geqslant 1$ future observations from units for which the explanatory variables $\tilde{\mathbf{X}}$ have been observed or set is also based on the posterior distribution

The m-vector $\tilde{\mathbf{y}}$ have a Gaussian distribution with expectation $\alpha \mathbf{1}_m + \tilde{\mathbf{X}}\beta$ and variance $\sigma^2 \mathbf{I}_m$

Prediction

Conditional on σ^2, the vector $\tilde{\mathbf{y}}$ of future observations has a Gaussian distribution

We can derive its expectation by averaging over α and β

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

This representation is quite intuitive, being the product of the matrix of explanatory variables $\tilde{\mathbf{X}}$ by the Bayesian estimator of β

Prediction

Similarly, we can compute

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

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

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

Prediction

Conditionally on σ^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 $\sigma^2 \frac{g}{g+1} \tilde{\mathbf{X}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\tilde{\mathbf{X}}^\mathsf{T}$, which corresponds to the uncertainty about β

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

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

Prediction

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

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