3

Regression and Variable Selection

You see, I always keep my sums.
—**Ian Rankin**, ***Strip Jack***.—

Roadmap

Linear regression is one of the most widely used tools in statistics for analyzing the (linear) influence of some variables or some factors on others and thus to uncover explanatory and predictive patterns. This chapter details the Bayesian analysis of the linear (or regression) model both in terms of prior specification (Zellner's G-prior) and in terms of variable selection, the next chapter appearing as a sequel for nonlinear dependence structures. The reader should be warned that, given that these models are the only conditional models where explicit computation can be conducted, this chapter contains a fair amount of matrix calculus. The photograph at the top of this page is a picture of processionary caterpillars, in connection (for once!) with the benchmark dataset used in this chapter.

J.-M. Marin and C.P. Robert, *Bayesian Essentials with R*, Springer Texts in Statistics, DOI 10.1007/978-1-4614-8687-9_3,

3.1 Linear Models

A large proportion of statistical analyses deal with the representation of dependences among several observed quantities. For instance, which social factors influence unemployment duration and the probability of finding a new job? Which economic indicators are best related to recession occurrences? Which physiological levels are most strongly correlated with aneurysm strokes? From a statistical point of view, the ultimate goal of these analyses is thus to find a proper representation of the conditional distribution, $f(y|\theta, \mathbf{x})$, of an observable variable y given a vector of observables $\mathbf{x}$, based on a sample of $\mathbf{x}$ and y. While the overall estimation of the conditional density f is usually beyond our ability, the estimation of θ and possibly of restricted features of f is possible within the Bayesian framework, as shown in this chapter.

The variable of primary interest, y, is called the *response* or the *outcome* variable; we assume here that this variable is continuous, but we will completely relax this assumption in the next chapter. The variables $\mathbf{x} = (x_1, \ldots, x_p)$ are called *explanatory variables* and may be discrete, continuous, or both. One sometimes picks a single variable x_j to be of primary interest. We then call it the *treatment* variable, labeling the other components of x as *control* variables, meaning that we want to address the (linear) influence of x_j on y once the linear influence of all the other variables has been taken into account (as in medical studies). The distribution of y given $\mathbf{x}$ is typically studied in the context of a set of *units* or experimental *subjects*, $i = 1, \ldots, n$, such as patients in a hospital ward, on which both y_i and $x_{i1}, \ldots, x_{ip}$ are measured. 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 **caterpillar** dataset exploited in this chapter was extracted from a 1973 study on pine processionary[1]caterpillars: it assesses the influence of some forest settlement characteristics on the development of caterpillar colonies. This dataset was first published and studied in Tomassone et al. (1993). The response variable is the logarithmic transform of the average number of nests of caterpillars per tree (as the one in the picture at the beginning of this chapter) in an area of $500\,\mathrm{m}^2$ (which corresponds to the last column in **caterpillar**). There are $p = 8$ potential explanatory variables defined on $n = 33$ areas, as follows

[1]These caterpillars derive their name from their habit of moving over the ground in incredibly long head-to-tail monk-like processions when leaving their nest to create a new colony.

x_1 is the altitude (in meters),
x_2 is the slope (in degrees),
x_3 is the number of pine trees in the area,
x_4 is the height (in meters) of the tree sampled at the center of the area,
x_5 is the orientation of the area (from 1 if southbound to 2 otherwise),
x_6 is the height (in meters) of the dominant tree,
x_7 is the number of vegetation strata,
x_8 is the mix settlement index (from 1 if not mixed to 2 if mixed).

The goal of the regression analysis is to decide which explanatory variables have a strong influence on the number of nests and how these influences overlap with one another. As shown by Fig. 3.1, some of these variables clearly have a restricting influence on the number of nests, as for instance with x_5, x_7 and x_8. We use the following R code to produce Fig. 3.1 (the way we created the objects `y` and `X` will be described later).

```
> par(mfrow=c(2,4),mar=c(4.2,2,2,1.2))
> for (j in 1:8) plot(X[,j],y,xlab=vnames[j],pch=19,
+ col="sienna4",xaxt="n",yaxt="n")
```

While many models and thus many dependence structures can be proposed for dependent datasets like **caterpillar**, in this chapter we only focus on the Gaussian linear regression model, namely the case when $\mathbb{E}[y|x,\theta]$ is linear in x and the noise is normal.

The *ordinary normal linear regression* model is such that, using a matrix representation,

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

where $\mathcal{N}_n$ denotes the normal distribution in dimension n, and thus the 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 chapter are all conditional on the regressors, we omit the conditioning on $\mathbf{X}$ to simplify the notations.

For **caterpillar**, where $n = 33$ and $p = 8$, we thus assume that the expected lognumber y_i of caterpillar nests per tree over an area is modeled as a linear combination of an intercept and eight predictor variables ($i = 1, \ldots, n$),

$$\mathbb{E}[y_i|\alpha,\boldsymbol{\beta},\sigma^2] = \alpha + \sum_{j=1}^{8} \beta_j x_{ij} ,$$

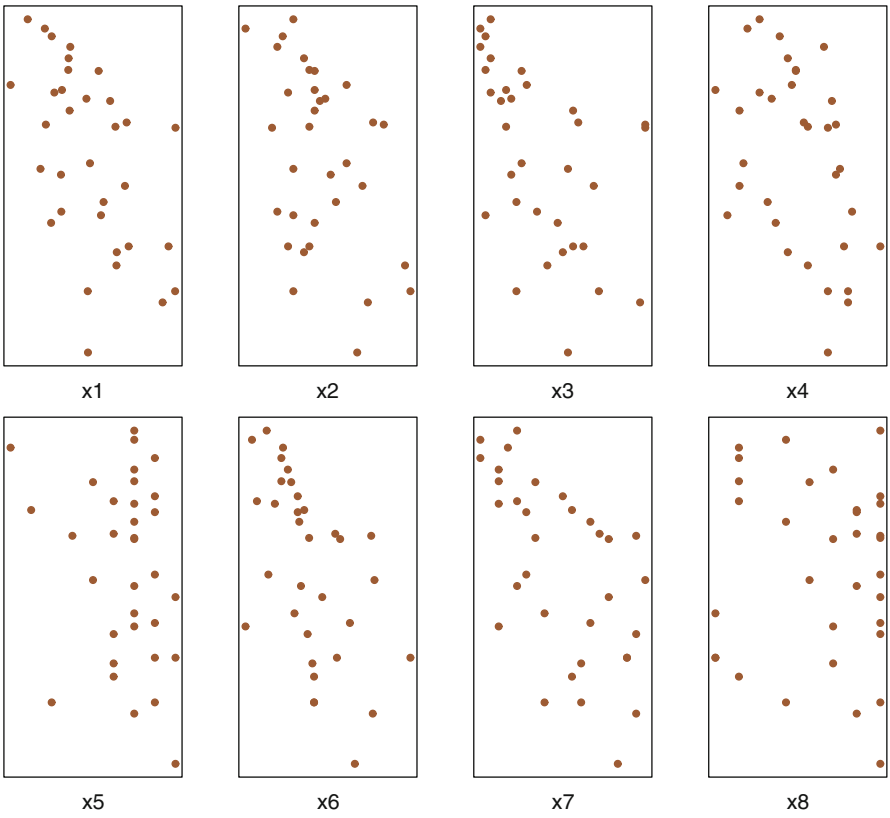

Fig. 3.1. Dataset **caterpillar**: Plot of the pairs $(\mathbf{x}_j, \mathbf{y})$ $(1 \leq j \leq 8)$

while the variation around this expectation is supposed to be normally distributed. Note that it is also customary to assume that the y_i's are conditionally independent.

The **caterpillar** dataset is called by the command `data(caterpillar)` and is made of the following rows:

```
1200 22 1 4 1.1 5.9 1.4 1.4 2.37
1342 28 8 4.4 1.5 6.4 1.7 1.7 1.47
....
1229 21 11 5.8 1.8 10 2.3 2 0.21
1310 36 17 5.2 1.9 10.3 2.6 2 0.03
```

The first eight columns correspond to the explanatory variables and the last column is the response variable, i.e. the lognumber of caterpillar nests. The following R code is an example for starting with this **caterpillar** dataset:

```
> y=log(caterpillar$y)
> X=as.matrix(caterpillar[,1:8])
```

There is a difference between using finite-valued regressors like x_7 in caterpillar and using *categorical* variables (or *factors*), which also take a finite number of values but whose range has no numerical meaning. For instance, if x denotes the socio-professional category of an employee, this variable may range from 1 to 9 for a rough grid of socio-professional activities, or it may range from 1 to 89 on a finer grid, and the numerical values are not comparable. It thus makes little sense to involve x directly in the regression, and the usual approach is to replace the single regressor x (taking values in $\{1, \ldots, m\}$, say) with m indicator (or *dummy*) variables $x_1 = \mathbb{I}_1(x)$, ..., $x_m = \mathbb{I}_m(x)$. In essence, a different constant (or *intercept*) β_j is used in the regression for each class of categorical variable: it is invoked in the linear regression under the form

$$\ldots + \beta_1 \mathbb{I}_1(x) + \ldots + \beta_m \mathbb{I}_m(x) + \ldots .$$

Note that there is an identifiability issue related with this model since the sum of the indicators is always equal to one. In a Bayesian perspective, identifiability can be achieved via the prior distribution. However, we can also impose an identifiability constraint on the parameters, for instance by omitting one class (such as $\beta_1 = 0$). We pursue this direction further in Sects. 4.5.1 and 6.2.

3.2 Classical Least Squares Estimator

Before fully launching into the description of the Bayesian approach to the linear model, we recall the basics of the classical processing of this model (in particular, to relate the Bayesian perspective to the results provided by standard software such as R lm output). For instance, the parameter $\boldsymbol{\beta}$ can obviously be estimated via maximum likelihood estimation. In order to avoid non-identifiability and uniqueness problems, we assume that $[\mathbf{1}_n \quad \mathbf{X}]$ is of full rank, that is, rank $[\mathbf{1}_n \quad \mathbf{X}] = p+1$. This also means that there is no redundant structure among the explanatory variables.[2] We suppose in addition that $p + 1 < n$ in order to obtain well-defined estimates for all parameters. Notice that, since the inferential process is conditioned on the design matrix $\mathbf{X}$, we choose to standardize the data, namely to center and to scale the columns of $\mathbf{X}$ so that the estimated values of $\boldsymbol{\beta}$ are truly comparable. For this purpose, we use the R function `scale`:

```
> X=scale(X)
```

[2] Hence, the exclusion of one of the classes for categorical variables.

The likelihood $\ell(\alpha, \boldsymbol{\beta}, \sigma^2|\mathbf{y})$ of the *standard normal linear model* is provided by the following matrix representation:

$$\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}\right)\right\}. \tag{3.1}$$

The maximum likelihood estimators of α and $\boldsymbol{\beta}$ are then the solution of the (least squares) minimization problem

$$\begin{aligned}&\min_{\alpha,\boldsymbol{\beta}} \left(\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}\right)\\ &\quad = \min_{\alpha,\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \alpha - \beta_1 x_{i1} - \ldots - \beta_p x_{ip}\right)^2,\end{aligned}$$

If we denote by $\bar{\mathbf{y}} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i$ the empirical mean of the y_i's and recall that, $\mathbf{1}_n^{\mathrm{T}}\mathbf{X} = \mathbf{0}_n^{\mathrm{T}}$ because of the standardization step, we have a Pythagorean decomposition of the above norm as

$$\begin{aligned}&\left(\mathbf{y}-\alpha\mathbf{1}_n-\mathbf{X}\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\mathbf{y}-\alpha\mathbf{1}_n-\mathbf{X}\boldsymbol{\beta}\right)\\ &= \left(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n-\mathbf{X}\boldsymbol{\beta}+(\bar{\mathbf{y}}-\alpha)\mathbf{1}_n\right)^{\mathrm{T}}\left(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n-\mathbf{X}\boldsymbol{\beta}+(\bar{\mathbf{y}}-\alpha)\mathbf{1}_n\right)\\ &= \left(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n-\mathbf{X}\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n-\mathbf{X}\boldsymbol{\beta}\right)+2(\bar{\mathbf{y}}-\alpha)\mathbf{1}_n^{\mathrm{T}}\left(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n-\mathbf{X}\boldsymbol{\beta}\right)+n(\bar{\mathbf{y}}-\alpha)^2\\ &= \left(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n-\mathbf{X}\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n-\mathbf{X}\boldsymbol{\beta}\right)+n(\bar{\mathbf{y}}-\alpha)^2.\end{aligned}$$

Indeed, $\mathbf{1}_n^{\mathrm{T}}\left(\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}\right) = (n\bar{\mathbf{y}} - n\bar{\mathbf{y}}) = 0$. Therefore, the likelihood $\ell(\alpha, \boldsymbol{\beta}, \sigma^2|\mathbf{y})$ is given by

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

We get from the above decomposition that

$$\hat{\alpha} = \bar{\mathbf{y}}, \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}(\mathbf{y} - \bar{y}).$$

In geometrical terms, $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ is the orthogonal projection of $\mathbf{y}$ on the linear subspace spanned by the columns of $[\mathbf{1}_n \quad \mathbf{X}]$. It is quite simple to check that $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ is an unbiased estimator of (α, β). In fact, the Gauss–Markov theorem (see, e.g., Christensen, 2002) states that $(\hat{\alpha}, \hat{\beta})$ is the *best* linear unbiased estimator of (α, β). This means that, for all $a \in \mathbb{R}^{p+1}$, and with the abuse of notation that, here, $(\hat{\alpha}, \hat{\beta})$ represents a column vector,

$$\mathbb{V}(a^{\mathsf{T}}(\hat{\alpha}, \hat{\beta})|\alpha, \boldsymbol{\beta}, \sigma^2) \leq \mathbb{V}(a^{\mathsf{T}}(\tilde{\alpha}, \tilde{\beta})|\alpha, \boldsymbol{\beta}, \sigma^2)$$

for any unbiased linear estimator $(\tilde{\alpha}, \tilde{\beta})$ of (α, β). (Note that the property of unbiasedness is not particularly appealing when considered on its own.)

An unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\,(\mathbf{y} - \hat{\alpha}\mathbf{1}_n - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathsf{T}}(\mathbf{y} - \hat{\alpha}\mathbf{1}_n - \mathbf{X}\hat{\boldsymbol{\beta}}) = \frac{s^2}{n-p-1},$$

and $\hat{\sigma}^2(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$ approximates the covariance matrix of $\hat{\boldsymbol{\beta}}$. Note that the MLE of σ^2 is not $\hat{\sigma}^2$ but $\tilde{\sigma}^2 = s^2/n$.

The standard *t-statistics* are defined as ($j = 1, \ldots, p$)

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2\omega_{jj}}} \sim \mathscr{T}(n-p-1, 0, 1)\,,$$

where ω_{jj} denotes the (j,j)-th element of the matrix $(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$. These t-statistic are used in classical tests, for instance for testing $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$, the former being accepted at level γ if

$$|\hat{\beta}_j|/\hat{\sigma}\sqrt{\omega_{jj}} < F^{-1}_{n-p-1}(1-\gamma/2)$$

the $(1-\gamma/2)$th quantile of the Student's t $\mathscr{T}(n-p-1, 0, 1)$ distribution (with location parameter 0 and scale parameter 1). The frequentist argument in using this bound (see Casella and Berger, 2001) is that the so-called *p-value* is smaller than γ,

$$p_j = P_{H_0}(|T_j| > |t_j|) < \gamma.$$

Note that these statistics T_j can also be used when constructing marginal frequentist confidence intervals on the β_j's like

$$\left\{\beta_j;\ \left|\beta_j - \hat{\beta}_j\right| \leq \hat{\sigma}\sqrt{\omega_{jj}}\,F^{-1}_{n-p-1}(1-\gamma/2)\right\} = \left\{\beta_j;\ |T_j| \leq \hat{\sigma}\sqrt{\omega_{jj}}\,F^{-1}_{n-p-1}(1-\gamma/2)\right\}.$$

From a Bayesian perspective, we far from advocate the use of p-values in Bayesian settings or elsewhere since they suffer many defects (exposed for instance in Robert, 2007, Chap. 5), one being that they are often wrongly interpreted as probabilities of the null hypotheses.

For **caterpillar**, the unbiased estimate of σ^2 is equal to 0.7781 and the maximum likelihood estimates of α and of the components β_j produced by the R command

```
> summary(lm(y~X))
```

are given in Fig. 3.2, along with the least squares estimates of their respective standard deviations and p-values. According to the classical paradigm, the coefficients β_1, β_2 and β_7 are the only ones considered to be *significant*.

We stress here that conditioning on $\mathbf{X}$ is valid only when $\mathbf{X}$ is *exogenous*, that is, only when we can write the joint distribution of $(\mathbf{y}, \mathbf{X})$ as

$$f(\mathbf{y}, \mathbf{X}|\alpha, \boldsymbol{\beta}, \sigma^2, \delta) = f(\mathbf{y}|\alpha, \boldsymbol{\beta}, \sigma^2, \mathbf{X}) f(\mathbf{X}|\delta)\,,$$

where $(\alpha, \boldsymbol{\beta}, \sigma^2)$ and δ are fixed parameters. We can thus ignore $f(\mathbf{X}|\delta)$ if the parameter δ is only a nuisance parameter since this part is independent[3] of $(\alpha, \boldsymbol{\beta}, \sigma^2)$. The practical advantage of using a regression model as above is that it is much easier to specify a realistic conditional distribution of one variable given p others rather than a joint distribution on all $p + 1$ variables. Note that if $\mathbf{X}$ is not *exogenous*, for instance when $\mathbf{X}$ involves past values of $\mathbf{y}$ (see Chap. 7), the joint distribution must be used instead.

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.4710 -0.4474 -0.1769  0.6121  1.5602

lm(formula = y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4710 -0.4474 -0.1769  0.6121  1.5602

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.81328    0.15356  -5.296 1.97e-05 ***
Xx1         -0.52722    0.21186  -2.489   0.0202 *
Xx2         -0.39286    0.16974  -2.315   0.0295 *
Xx3          0.65133    0.38670   1.684   0.1051
Xx4         -0.29048    0.31551  -0.921   0.3664
Xx5         -0.21645    0.16865  -1.283   0.2116
Xx6          0.29361    0.53562   0.548   0.5886
Xx7         -1.09027    0.47020  -2.319   0.0292 *
Xx8         -0.02312    0.17225  -0.134   0.8944
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 0.8821 on 24 degrees of freedom
Multiple R-squared: 0.6234,Adjusted R-squared: 0.4979
```

Fig. 3.2. Dataset **caterpillar**: R output providing the least squares estimates of the regression coefficients along with their standard significance analysis

[3] From a Bayesian point of view, note that we would also need to impose prior independence between $(\alpha, \boldsymbol{\beta}, \sigma^2)$ and δ to achieve this separation.

3.3 The Jeffreys Prior Analysis

Considering only the case of a complete lack of prior information on the parameters of the linear model, we first describe a noninformative solution based on the Jeffreys prior. It is rather easy to show that the Jeffreys prior in this case is

$$\pi^J(\alpha, \boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}\,,$$

which is equivalent to a flat prior on $(\alpha, \boldsymbol{\beta}, \log \sigma^2)$. We recall that

$$\begin{aligned}
\ell(\alpha, \boldsymbol{\beta}, \sigma^2|\mathbf{y}) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}\right)\right\} \times \\
&\quad \exp\left\{-\frac{n}{2\sigma^2}(\bar{\mathbf{y}} - \alpha)^2\right\} \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{y} - \hat{\alpha}\mathbf{1}_n - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^{\mathrm{T}}\left(\mathbf{y} - \hat{\alpha}\mathbf{1}_n - \mathbf{X}\hat{\boldsymbol{\beta}}\right)\right\} \times \\
&\quad \exp\left\{-\frac{n}{2\sigma^2}(\hat{\alpha} - \alpha)^2 - \frac{1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right\}.
\end{aligned}$$

The corresponding posterior distribution is therefore

$$\begin{aligned}
\pi^J(\alpha, \boldsymbol{\beta}, \sigma^2|\mathbf{y}) &\propto \left(\sigma^{-2}\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \hat{\alpha}\mathbf{1}_n - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathsf{T}}(\mathbf{y} - \hat{\alpha}\mathbf{1}_n - \mathbf{X}\hat{\boldsymbol{\beta}})\right\} \times \\
&\quad \sigma^{-2} \exp\left\{-\frac{n}{2\sigma^2}(\hat{\alpha} - \alpha)^2 - \frac{1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right\} \\
&\propto \left(\sigma^{-2}\right)^{-p/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right\} \times \\
&\quad \left(\sigma^{-2}\right)^{-1/2} \exp\left\{-\frac{n}{2\sigma^2}(\hat{\alpha} - \alpha)^2\right\} \\
&\quad \left(\sigma^{-2}\right)^{-(n-p-1)/2-1} \exp\left\{-\frac{1}{2\sigma^2}s^2\right\}.
\end{aligned}$$

From this expression, we deduce the following (conditional and marginal) posterior distributions

$$\begin{aligned}
\alpha|\sigma^2, \mathbf{y} &\sim \mathcal{N}\left(\hat{\alpha}, \sigma^2/n\right), \\
\boldsymbol{\beta}|\sigma^2, \mathbf{y} &\sim \mathcal{N}_p\left(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\right), \\
\sigma^2|\mathbf{y} &\sim \mathcal{IG}((n-p-1)/2, s^2/2)\,.
\end{aligned}$$

As in every analysis involving an improper prior, one needs to check that the corresponding posterior distribution is proper. In this case, $\pi(\alpha, \boldsymbol{\beta}, \sigma^2|\mathbf{y})$ is proper when both $n > p + 1$ and $\text{rank}\,[\mathbf{1}_n \quad \mathbf{X}] = p + 1$. The former constraint requires that there be at least as many data points as there are parameters in the model, and, as already explained above, the latter is obviously necessary for identifiability reasons.

The corresponding Bayesian estimates of α, $\boldsymbol{\beta}$ and σ^2 are thus given by

$$\mathbb{E}^\pi[\alpha|\mathbf{y}] = \hat{\alpha}\,, \quad \mathbb{E}^\pi[\boldsymbol{\beta}|\mathbf{y}] = \hat{\boldsymbol{\beta}} \quad \text{and} \quad \mathbb{E}^\pi[\sigma^2|\mathbf{y}] = \frac{s^2}{n-p-3}\,,$$

respectively. Unsurprisingly, the Jeffreys prior estimate of α is the empirical mean. Further, the posterior expectation of $\boldsymbol{\beta}$ is the maximum likelihood estimate. Note also that the Jeffreys prior estimate of σ^2 is larger (and thus more pessimistic) than both the maximum likelihood estimate s^2/n and the classical unbiased estimate $s^2/(n-p-1)$.

The marginal posterior distribution of β_j associated with the above joint distribution is

$$\mathscr{T}(n-p-1, \hat{\beta}_j, \omega_{jj}s^2/(n-p-1))\,,$$

(recall that $\omega_{jj} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}_{(j,j)}$). Hence, the similarity with a frequentist analysis of this model is very strong since the classical $(1-\gamma)$ confidence interval and the Bayesian HPD interval on β_j coincide, even though they have different interpretations. They are both equal to

$$\left\{\beta_j; |\beta_j - \hat{\beta}_j| \leq F^{-1}_{n-p-1}(1-\gamma/2)\sqrt{\omega_{jj}s^2/(n-p-1)}\right\}\,.$$

For **caterpillar**, the Bayes estimate of σ^2 is equal to 0.8489. Figure 3.3 provides the corresponding (marginal) 95 % HPD intervals for each component of β. (It is obtained by the `plotCI` function, part of the `gplots` package.) Note that while some of these credible intervals include the value $\beta_j = 0$ (represented by the dashed line), they do not necessarily validate acceptance of the null hypothesis $H_0 : \beta_j = 0$, which must be tested through a Bayes factor, as described below. This distinction is a major difference from the classical approach, where confidence intervals are dual sets of acceptance regions.

3.4 Zellner's *G*-Prior Analysis

From this section onwards,[4] we concentrate on a different noninformative approach which was proposed by Arnold Zellner[5] to handle linear regression from a Bayesian perspective. This approach is a middle-ground perspective where some prior information may be available on $\boldsymbol{\beta}$ and it is called *Zellner's G-prior*, the "*G*" being the symbol used by Zellner in the prior variance.

[4] In order to keep this coverage of *G*-priors simple and self-contained, we made several choices in the presentation that the most mature readers will possibly find arbitrary, but this cannot be avoided if we want to keep the chapter at a reasonable length.

[5] Arnold Zellner was a famous Bayesian econometrician, who wrote two reference books on Bayesian econometrics (Zellner, 1971, 1984)

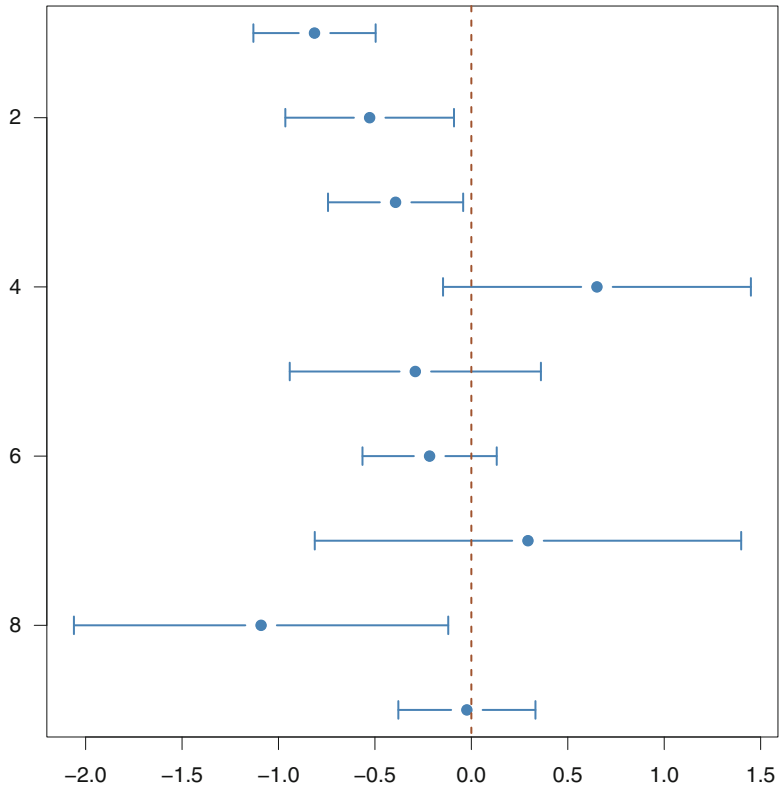

Fig. 3.3. Dataset **caterpillar**: Range of the credible 95 % HPD intervals for α (*top row*) and each component of β when using the Jeffreys prior

3.4.1 A Semi-noninformative Solution

When considering the likelihood (3.1) its shape is both Gaussian and Inverse Gamma, indeed, $\boldsymbol{\beta}$ given σ^2 appears in a Gaussian-like expression, while σ^2 involves an Inverse Gamma expression. This structure leads to a natural conjugate prior family, of the form

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

conditional on σ^2, where M is a $(p+1, p+1)$ positive definite symmetric matrix, and for σ^2,

$$\sigma^2 \sim \mathcal{IG}(a, b), \qquad a, b > 0\,.$$

(The conjugacy can be easily checked by the reader.) Even in the presence of genuine information on the parameters, the hyperparameters M, a and b are very difficult to specify. Moreover, the posterior distributions, notably the posterior variances are sensitive to the specification of these hyper-parameters.

Therefore, given that a natural conjugate prior for the linear regression model has severe limitations, a more elaborate strategy is called for. The idea at the core of Zellner's G-prior modeling is to allow the experimenter to introduce (possibly weak) information about the location parameter of the regression but to bypass the most difficult aspects of the prior specification, namely the derivation of the prior correlation structure. This structure is fixed in Zellner's proposal since the prior corresponds to

$$\boldsymbol{\beta}|\alpha,\sigma^2 \sim \mathcal{N}_p\left(\tilde{\boldsymbol{\beta}}, g\sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right), \tag{3.2}$$

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

$$\pi\left(\alpha,\sigma^2\right) \propto \sigma^{-2}. \tag{3.3}$$

Zellner's G-prior is thus decomposed as a (conditional) Gaussian prior for $\boldsymbol{\beta}$ and an improper (Jeffreys) prior for (α, σ^2). This modelling somehow appears as a data-dependent prior through its dependence on $\mathbf{X}$, but this is not a genuine issue[6] since the *whole* model is conditional on $\mathbf{X}$. The experimenter thus restricts prior determination to the choices of $\tilde{\boldsymbol{\beta}}$ and of the constant g. As we will see once the posterior distribution is constructed, 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.

> Genuine data-dependent priors are not acceptable in a Bayesian analysis because they use the data *twice* and fail to enjoy the basic convergence properties of the Bayes estimators. (See Carlin and Louis, 1996, for a comparative study of the corresponding so-called *empirical Bayes* estimators.)

Note that, in the initial proposition of Zellner (1984), the parameter α is not modelled by a flat prior distribution. It was instead considered to be a component of the vector $\boldsymbol{\beta}$. (This was also the approach adopted in Marin and Robert 2007.) However, endowing α with a flat prior ensures the location-scale invariance of the analysis, which means that changes in location or scale on $\mathbf{y}$ (like a switch from Celsius to Fahrenheit degrees for temperatures) do not impact on the resulting inference.

We are now engaging into some algebra that will expose the properties of the G-posterior. First, we assume $p > 0$, meaning that there is at least one explanatory variable in the model. We define the matrix $\mathbf{P} = \mathbf{X}\left\{\mathbf{X}^{\mathrm{T}}\mathbf{X}\right\}^{-1}\mathbf{X}^{\mathrm{T}}$. The prior $\pi\left(\alpha,\boldsymbol{\beta},\sigma^2\right)$ can then be decomposed as

[6] This choice is more problematic when conditioning on $\mathbf{X}$ is no longer possible, as for instance when $\mathbf{X}$ contains lagged dependent variables (Chap. 7) or endogenous variables.

$$\pi\left(\alpha,\boldsymbol{\beta},\sigma^2\right) \propto (\sigma^2)^{-p/2}\exp\left[-\frac{1}{2g\sigma^2}\left\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}-2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{P}\mathbf{X}\tilde{\boldsymbol{\beta}}\right\}\right]\times$$
$$\sigma^{-2}\exp\left(-\frac{1}{2g\sigma^2}\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{P}\mathbf{X}\tilde{\boldsymbol{\beta}}\right),$$

since $\mathbf{X}^{\mathrm{T}}\mathbf{P}\mathbf{X}=\mathbf{X}^{\mathrm{T}}\mathbf{X}$. Therefore,

$$\begin{aligned}\pi\left(\alpha,\boldsymbol{\beta},\sigma^2|\mathbf{y}\right) \propto{}& (\sigma^2)^{-n/2-p/2-1}\\ &\exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n-\mathbf{X}\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n-\mathbf{X}\boldsymbol{\beta}\right)\right\}\times\\ &\exp\left\{-\frac{n}{2\sigma^2}(\bar{\mathbf{y}}-\alpha)^2\right\}\times\exp\left\{-\frac{1}{2g\sigma^2}\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{P}\mathbf{X}\tilde{\boldsymbol{\beta}}\right\}\times\\ &\exp\left\{-\frac{1}{2g\sigma^2}\left[\boldsymbol{\beta}\mathrm{T}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}-2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{P}\mathbf{X}\tilde{\boldsymbol{\beta}}\right]\right\}.\end{aligned}$$

Since $\mathbf{1}_n^{\mathrm{T}}\mathbf{X}=\mathbf{0}_p$, we deduce that

$$\begin{aligned}\pi\left(\alpha,\boldsymbol{\beta},\sigma^2|\mathbf{y}\right) \propto{}& (\sigma^2)^{-n/2-p/2-1}\exp\left\{-\frac{1}{2\sigma^2}\left[\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}-2\mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}\right]\right\}\times\\ &\exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n\right)^{\mathrm{T}}\left(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n\right)\right\}\times\\ &\exp\left\{-\frac{n}{2\sigma^2}(\bar{\mathbf{y}}-\alpha)^2\right\}\times\exp\left\{-\frac{1}{2g\sigma^2}\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{P}\mathbf{X}\tilde{\boldsymbol{\beta}}\right\}\times\\ &\exp\left\{-\frac{1}{2g\sigma^2}\left[\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}-2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{P}\mathbf{X}\tilde{\boldsymbol{\beta}}\right]\right\}.\end{aligned}$$

Since $\mathbf{P}\mathbf{X}=\mathbf{X}$, we deduce that, conditionally on $\mathbf{y}$, $\mathbf{X}$ and σ^2, the parameters α and $\boldsymbol{\beta}$ are independent and such that

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

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

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 α and $\boldsymbol{\beta}$ is due to the fact that $\mathbf{X}$ is centered and that α and $\boldsymbol{\beta}$ are a priori independent.

Moreover, the posterior distribution of σ^2 is given by

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

where $I\mathscr{G}(a,b)$ is an inverse Gamma distribution with mean $b/(a-1)$ and where $s^2=(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n-\mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y}-\bar{\mathbf{y}}\mathbf{1}_n-\mathbf{X}\hat{\boldsymbol{\beta}})$ corresponds to the (classical) residual sum of squares.

The previous derivation assumes that $p > 0$. In the special case $p = 0$, which will later be used as a null model in hypothesis testing, similar arguments lead to

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

(There is no β when $p = 0$, as this corresponds to the constant mean model.)

Recalling the double expectation formulas

$$\mathbb{E}\left[\mathbb{E}\left[X|Y\right]\right] = \mathbb{E}\left[X\right] \quad \text{and} \quad \mathbb{V}(X) = \mathbb{V}[\mathbb{E}(X|Y)] + \mathbb{E}[\mathbb{V}(X|Y)]$$

for $\mathbb{V}(X|Y) = \mathbb{E}[(X - \mathbb{E}(X|Y))^2\big|Y]$, we can derive from the previous derivations that

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

and that

$$\mathbb{V}^{\pi}(\alpha|\mathbf{y}) = \mathbb{V}(\bar{\mathbf{y}}|\mathbf{y}) + \mathbb{E}\left[\frac{\sigma^2}{n}\middle|\mathbf{y}\right] = \kappa/n(n-3) ,$$

where

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

With a bit of extra algebra, we can recover the whole distribution of α from

$$\pi(\alpha, \sigma^2|\mathbf{y}) \propto (\sigma^{-2})^{(n-1)/2+1+1/2} \exp\left\{-\frac{n}{2\sigma^2}(\alpha - \bar{\mathbf{y}})^2\right\} \exp\left\{-\frac{\kappa}{2\sigma^2}\kappa\right\} ,$$

namely

$$\pi(\alpha|\mathbf{y}) \propto \left[1 + \frac{n(\alpha - \bar{\mathbf{y}})^2}{\kappa}\right]^{-n/2} .$$

This means that the marginal posterior distribution of α—the distribution of α given only $\mathbf{y}$ and $\mathbf{X}$—is a Student's t distribution with $n-1$ degrees of freedom, a location parameter equal to $\bar{\mathbf{y}}$ and a scale parameter equal to $\kappa/n(n-1)$.

If we now turn to the parameter $\boldsymbol{\beta}$, by the same double expectation formula, we derive that

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

This result gives its meaning to the above point relating g with the amount of information contained in the dataset. For instance, when $g = 1$, the prior information has the same weight as this amount. In this case, 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. For instance, when g goes to infinity, the posterior mean converges to $\hat{\boldsymbol{\beta}}$.

Based on similar derivations, we can compute the posterior variance of $\boldsymbol{\beta}$. Indeed,

$$\begin{aligned}\mathbb{V}^{\pi}(\boldsymbol{\beta}|\mathbf{y}) &= \mathbb{V}\left[\frac{g}{g+1}(\hat{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}/g)|\mathbf{y}\right] + \mathbb{E}\left[\frac{g\sigma^2}{g+1}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right] \\ &= \frac{\kappa g}{(g+1)(n-3)}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,.\end{aligned}$$

Once more, it is possible to integrate out σ^2 in

$$\begin{aligned}\pi(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) \propto (\sigma^2)^{-p/2} \exp\left(-\frac{g+1}{2g\sigma^2}\big\{\boldsymbol{\beta} - \mathbb{E}^{\pi}\left[\boldsymbol{\beta}|\mathbf{y}\right]\big\}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\big\{\boldsymbol{\beta} - \mathbb{E}^{\pi}\left[\boldsymbol{\beta}|\mathbf{y}\right]\big\}\right) \\ \times (\sigma^2)^{-(n-1)/2-1} \exp\left(-\frac{1}{2\sigma^2}\kappa\right)\,,\end{aligned}$$

leading to

$$\pi(\boldsymbol{\beta}|\mathbf{y}) \propto \left[1 + \frac{g+1}{g\kappa}\big\{\boldsymbol{\beta} - \mathbb{E}^{\pi}\left[\boldsymbol{\beta}|\mathbf{y}\right]\big\}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\big\{\boldsymbol{\beta} - \mathbb{E}^{\pi}\left[\boldsymbol{\beta}|\mathbf{y}\right]\big\}\right]\,.$$

Therefore, the marginal posterior distribution of $\boldsymbol{\beta}$ is also a multivariate Student's t distribution with $n-1$ degrees of freedom, location parameter equal to $\dfrac{g}{g+1}(\hat{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}/g)$ and scale parameter equal to $\dfrac{g\kappa}{(g+1)(n-1)}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$.

The standard Bayes estimator of σ^2 for this model is the posterior expectation

$$\mathbb{E}^{\pi}\left[\sigma^2|\mathbf{y}\right] = \frac{\kappa}{n-3} = \frac{s^2 + (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})/(g+1)}{n-3}\,.$$

In the special case $p = 0$, by using similar arguments, we get

$$\mathbb{E}^{\pi}\left[\sigma^2|\mathbf{y}\right] = \frac{(\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n)^{\mathrm{T}}(\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n)}{n-3} = \frac{s^2}{n-3}\,,$$

which is the same expectation as with the Jeffreys prior.

HPD regions on subvectors of the parameter $\boldsymbol{\beta}$ can be derived in a straightforward manner from this marginal posterior distribution of $\boldsymbol{\beta}$. For a single parameter, we have for instance

$$\beta_j|\mathbf{y} \sim \mathscr{T}\left(n-1, \frac{g}{g+1}\left(\frac{\tilde{\beta}_j}{g}+\hat{\beta}_j\right), \frac{g\kappa}{(n-1)(g+1)}\omega_{jj}\right),$$

where ω_{jj} is the (j,j)-th element of the matrix $(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$. If we set

$$\zeta = (\tilde{\boldsymbol{\beta}} + g\hat{\boldsymbol{\beta}})/(g+1)$$

the transform

$$\beta_j - \zeta_j \Big/ \sqrt{\frac{g\kappa}{(n-1)(g+1)}\omega_{jj}}$$

is (marginally) distributed as a standard t distribution with $n-1$ degrees of freedom. A $(1-\gamma)$ HPD interval on β_j has therefore

$$\zeta_j \pm \sqrt{\frac{g\kappa}{(n-1)(g+1)}\omega_{jj}}F_{n-1}^{-1}(1-\gamma/2)$$

as bounds, where F_{n-1}^{-1} denotes the quantile function of the $\mathscr{T}(n-1,0,1)$ distribution.

3.4.2 The BayesReg R Function

We have created in `bayess` an R function called `BayesReg` to implement Zellner's G-prior analysis within R. The purpose is dual: first, this R function shows how easily automated this approach can be. Second, it also illustrates how it is possible to get exactly the same type of output as the standard R function `summary(lm(y~X))`.

The following R code is extracted from this function `BayesReg` and used to calculate the Bayes estimates. As an aside, notice that we use the function `stop` in order to end the calculations if the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is not invertible.

```
if (det(t(X)%*%X)<=1e-7)
stop("Design matrix has too low a rank!",call.=FALSE)
```

We also stress the use of `scale` below to standardize the explanatory variables.

```
X=as.matrix(X)
n=length(y)
p=dim(X)[2]
X=scale(X)
U=solve(t(X)%*%X)%*%t(X)
# MLE
alphaml=mean(y)
betaml=U%*%y
s2=t(y-alphaml-X%*%betaml)%*%(y-alphaml-X%*%betaml)
kappa=as.numeric(s2+t(betatilde-betaml)%*%t(X)%*%X%*%
   (betatilde-betaml)/(g+1))
```

```
malphabayes=alphaml
mbetabayes=g/(g+1)*(betaml+betatilde/g)
msigma2bayes=kappa/(n-3)
valphabayes=kappa/(n*(n-3))
vbetabayes=diag(kappa*g/((g+1)*(n-3))*solve(t(X)%*%X))
vsigma2bayes=2*kappa^2/((n-3)*(n-4))
postmean=c(malphabayes,mbetabayes)
postsd=sqrt(c(valphabayes,vbetabayes))
# evidence of the model
intlike=(g+1)^(-p/2)*kappa^(-(n-1)/2)
```

We will see further aspects of `BayesReg` in the following sections.

3.4.3 Bayes Factors and Model Comparison

One important inferential issue pertaining to linear models is to test whether or not a specific explanatory variable is truly explanatory or, in other words, to decide which explanatory variables should be kept within the model. This leads to tests on the nullity of some elements of the parameter $\boldsymbol{\beta}$. Following the general testing methodology presented in Chap. 2, these tests can be conducted using Bayes factors. In the case of linear models and under Zellner's G-priors, those Bayes factors are actually available in closed form.

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

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

with

$$\begin{aligned} f(\mathbf{y}|\alpha, \boldsymbol{\beta}, \sigma^2) \pi(\boldsymbol{\beta}|\alpha, \sigma^2) = {} & \frac{\left|\mathbf{X}^{\mathrm{T}}\mathbf{X}\right|^{1/2}}{(2\pi\sigma^2)^{(n+p)/2} g^{p/2}} \exp\left\{ -\frac{n}{2\sigma^2}(\alpha - \bar{\mathbf{y}})^2 \right\} \times \\ & \exp\left\{ -\frac{1}{2\sigma^2}(\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}) \right\} \times \\ & \exp\left\{ -\frac{1}{2g\sigma^2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) \right\} , \end{aligned}$$

and $\pi(\alpha, \sigma^2) = \delta\sigma^{-2}$ (where δ is an arbitrary constant). Thus

$$\begin{aligned} f(\mathbf{y}) &= \delta n^{-1/2}(g+1)^{-p/2}(2\pi)^{-(n-1)/2} \int (\sigma^2)^{-(n-1)/2-1} \exp\left(-\frac{1}{2\sigma^2}\kappa \right) \mathrm{d}\sigma^2 \\ &= \frac{\delta\Gamma((n-1)/2)}{\pi^{(n-1)/2} n^{1/2}} (g+1)^{-p/2} \left[s^2 + (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) / (g+1) \right]^{-(n-1)/2} , \\ &= \frac{\delta\Gamma((n-1)/2)}{\pi^{(n-1)/2} n^{1/2}} (g+1)^{-p/2} \kappa^{-(n-1)/2} . \end{aligned} \tag{3.4}$$

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

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

with

$$\begin{aligned} f(\mathbf{y}|\alpha, \sigma^2)\pi(\alpha, \sigma^2) &= \frac{\delta(\sigma^2)^{-1}}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2}(\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n)^{\mathsf{T}}(\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n) \right\} \\ &= \frac{\delta(\sigma^2)^{-n/2-1}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n)^{\mathsf{T}}(\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n) \right\} \times \\ &\quad \exp\left\{ -\frac{n}{2\sigma^2}(\alpha - \bar{\mathbf{y}})^2 \right\} . \end{aligned}$$

The integration in both α and σ^2 can then be conducted in closed form and we obtain

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

as the evidence associated with this "null" model. The evidence corresponds to `intlike0` in the `BayesReg` code.

As pointed out in Chap. 2, the computation of Bayes factors 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 (and hence the *same* arbitrary constant δ) on (α, σ^2) for all of those models. Otherwise, the Bayes factors simply cannot be correctly defined.

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. From a Bayesian perspective, using Zellner's G-prior modelling in both cases, we are thus comparing model $\mathfrak{M}_1$

$$\begin{aligned} \mathbf{y}|\alpha, \boldsymbol{\beta}^1, \sigma^2 &\sim \mathcal{N}_n\left(\alpha\mathbf{1}_n + \mathbf{X}^1\boldsymbol{\beta}^1, \sigma^2\,\mathbf{I}_n\right) , \\ \boldsymbol{\beta}^1|\alpha, \sigma^2 &\sim \mathcal{N}_{p_1}\left(\tilde{\boldsymbol{\beta}}^1, g_1\sigma^2((\mathbf{X}^1)^{\mathrm{T}}\mathbf{X}^1)^{-1}\right), \quad p_1 \neq 0 \\ \pi\left(\alpha, \sigma^2\right) &\propto \sigma^{-2} , \end{aligned}$$

with model $\mathfrak{M}_2$:

$$\begin{aligned} \mathbf{y}|\alpha, \boldsymbol{\beta}^2, \sigma^2 &\sim \mathcal{N}_n\left(\alpha\mathbf{1}_n + \mathbf{X}^2\boldsymbol{\beta}^2, \sigma^2\,\mathbf{I}_n\right) , \\ \boldsymbol{\beta}^2|\alpha, \sigma^2 &\sim \mathcal{N}_{p_2}\left(\tilde{\boldsymbol{\beta}}^2, g_2\sigma^2((\mathbf{X}^2)^{\mathrm{T}}\mathbf{X}^2)^{-1}\right), \quad p_2 \neq 0 \\ \pi\left(\alpha, \sigma^2\right) &\propto \sigma^{-2} . \end{aligned}$$

Using the above derivations, the Bayes factor between model $\mathfrak{M}_1$ and model $\mathfrak{M}_2$ is then given by

$$B_{12}(\mathbf{y}) = \frac{(g_1+1)^{-p_1/2}\left[s_1^2 + (\tilde{\boldsymbol{\beta}}^1 - \hat{\boldsymbol{\beta}}^1)^{\mathrm{T}}(\mathbf{X}^1)^{\mathrm{T}}\mathbf{X}^1(\tilde{\boldsymbol{\beta}}^1 - \hat{\boldsymbol{\beta}}^1)/(g_1+1)\right]^{-(n-1)/2}}{(g_2+1)^{-p_2/2}\left[s_2^2 + (\tilde{\boldsymbol{\beta}}^2 - \hat{\boldsymbol{\beta}}^2)^{\mathrm{T}}(\mathbf{X}^2)^{\mathrm{T}}\mathbf{X}^2(\tilde{\boldsymbol{\beta}}^2 - \hat{\boldsymbol{\beta}}^2)/(g_2+1)\right]^{-(n-1)/2}}\,.$$

For **caterpillar**, if we have to test the null hypothesis $H_0 : \beta_8 = \beta_9 = 0$, using $\tilde{\beta}^1 = 0_8$, $\tilde{\beta}^2 = 0_6$, and an arbitrary[7] $g_1 = g_2 = 100$, in Zellner's G-priors, we obtain $B_{12}^{\pi} = 0.0165$ when model $\mathfrak{M}_2$ corresponds to H_0. Using Jeffreys' scale of evidence (provided in Chap. 2), this implies that $\log_{12}(B_{12}^{\pi}) = -1.78$, hence that the posterior distribution appears to strongly favor H_0.

More generally, using $\tilde{\boldsymbol{\beta}} = 0_8$ and $g = 100$, we can produce a Bayesian regression output, programmed in R, which mimics a standard software regression output like `lm`: besides the estimation of the β_j's via their posterior expectation, we include the Bayes factors B_{12}^j, in the log scale $\log_{10}\left(B_{12}^j\right)$, corresponding to testing the null hypotheses $H_0 : \beta_j = 0$. (The stars are related to Jeffreys' scale of evidence.)

The R code corresponding to this "standard" output is also part of the R function `BayesReg`:

```
bayesfactor=rep(0,p)
p0=p-1 # remove one variate
X0=X[,-j]
U0=solve(t(X0)%*%X0)%*%t(X0)
betatilde0=U0%*%X%*%betatilde
betaml0=U0%*%y
s20=t(y-alphaml-X0%*%betaml0)%*%(y-alphaml-X0%*%betaml0)
kappa0=as.numeric(s20+t(betatilde0-betaml0)%*%t(X0)%*%
X0%*%(betatilde0-betaml0)/(g+1))
intlike0=(g+1)^(-p0/2)*kappa0^(-(n-1)/2)
bayesfactor[j]=intlike/intlike0
```

where `intlike` is the marginal likelihood for the full model. (The way this computation is repeated and used to mimic the output of the `lm` function can be found by reading the function `BayesReg`.)

For the **caterpillar** dataset, $\tilde{\boldsymbol{\beta}} = 0_8$ and $g = n = 33$, the G-prior estimate of σ^2 is equal to 0.653, while the posterior means and standard variations of the β_j's are given below. We can immediately spot that the (most) significant explanatory variables are the same ones as those selected by `lm`, x_1, x_2, and x_7. Note, however, that this output does not rigorously validate the selection of the submodel with the covariates x_1, x_2, and x_7, as it does not produce the Bayes factor associated with this (sub)model and the full model.

[7] Arbitrary means here that this choice is no more justified than any other. We will see later that $g_j = n$ is the recommended or default value for non-informative settings.

```
> res1=BayesReg(y,X)

          PostMean PostStError Log10bf EvidAgaH0
Intercept  -0.8133      0.1407
x1         -0.5039      0.1883  0.7224      (**)
x2         -0.3755      0.1508  0.5392      (**)
x3          0.6225      0.3436 -0.0443
x4         -0.2776      0.2804 -0.5422
x5         -0.2069      0.1499 -0.3378
x6          0.2806      0.4760 -0.6857
x7         -1.0420      0.4178  0.5435      (**)
x8         -0.0221      0.1531 -0.7609

Posterior Mean of Sigma2: 0.6528
Posterior StError of Sigma2: 0.939
```

3.4.4 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 α, $\boldsymbol{\beta}$ and σ^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 $\sigma^2\mathbf{I}_m$. The *predictive distribution* on $\tilde{\mathbf{y}}$ is defined as the marginal in $\mathbf{y}$ of the joint posterior distribution on $(\tilde{\mathbf{y}}, \alpha, \boldsymbol{\beta}, \sigma^2)$.

Conditional on σ^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 α 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 σ^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

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

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

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

This decomposition of the variance makes perfect sense: 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 $\boldsymbol{\beta}$.

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

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

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

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

(following a straightforward but lengthy derivation that is very similar to the one we conducted at the end of Chap. 2, see (2.11)).

3.5 Markov Chain Monte Carlo Methods

Given the complexity of most models encountered in Bayesian modeling, standard simulation methods are not a sufficiently versatile solution. We now present the rudiments of a technique that emerged in the late 1980s as the core of Bayesian computing and that has since then revolutionized the field.

This technique is based on *Markov chains*, but we will not make many incursions into the theory of Markov chains (see Meyn and Tweedie, 1993), focusing rather on the practical implementation of these algorithms and trusting that the underlying theory is sound enough to validate them (Robert and Casella, 2004). At this point, it is sufficient to recall that a Markov chain $(\mathbf{x}_t)_{t\in\mathbb{N}}$ is a sequence of dependent random vectors whose dependence on the past values $\mathbf{x}_0, \ldots, \mathbf{x}_{t-1}$ stops at the value immediately before, $\mathbf{x}_{t-1}$, and that is entirely defined by its *kernel*—that is, the conditional distribution of $\mathbf{x}_t$ given $\mathbf{x}_{t-1}$.

The central idea behind these new methods, called *Markov chain Monte Carlo* (MCMC) algorithms, is that, to simulate from a distribution π (for instance, the posterior distribution), it is actually sufficient to produce a Markov chain $(\mathbf{x}_t)_{t\in\mathbb{N}}$ whose *stationary distribution* is π: when $\mathbf{x}_t$ is marginally distributed according to π, then $\mathbf{x}_{t+1}$ is also marginally distributed according to

π. If an algorithm that generates such a chain can be constructed, the ergodic theorem guarantees that, in almost all settings, the average

$$\frac{1}{T}\sum_{t=1}^{T} g(\mathbf{x}_t)$$

converges to $\mathbb{E}[g(\mathbf{x})]$, no matter what the starting value.[8]

More informally, this property means that, for large enough t, $\mathbf{x}_t$ is approximately distributed from π and can thus be used like the output from a more standard simulation algorithm (even though one must take care of the correlation between the $\mathbf{x}_t$'s created by the Markovian structure). For integral approximation purposes, the difference from regular Monte Carlo approximations is that the variance structure of the estimator is more complex because of the Markovian dependence. These methods being central to the cases studied from this stage onward, we hope that the reader will become sufficiently proficient with them by the end of the book! In this chapter, we detail a particular type of MCMC algorithm, the Gibbs sampler, that is currently sufficient for our needs. The next chapter will introduce a more universal type of algorithm.

3.5.1 Conditionals

A first remark that motivates the use of the Gibbs sampler[9] is that, within structures such as

$$\pi(x_1) = \int \pi_1(x_1|x_2)\tilde{\pi}(x_2)\,\mathrm{d}x_2\,, \tag{3.5}$$

to simulate from the joint distribution

$$\pi(x_1, x_1) = \pi_1(x_1|x_2)\tilde{\pi}(x_2) \tag{3.6}$$

automatically produces (marginal) simulation from $\pi(x_1)$. Therefore, in settings where (3.5) holds, it is not necessary to simulate from $\pi(x_1)$ when one can jointly simulate (x_1, x_2) from (3.6).

For example, consider $(x_1, x_2) \in \mathbb{N}\times[0,1]$ distributed from the joint density

$$\pi(x_1, x_2) \propto \binom{n}{x_1} x_2^{x_1+\alpha-1}(1-x_2)^{n-x_1+\beta-1}\,.$$

This is a joint distribution where

$$x_1|x_2 \sim \mathscr{B}(n, x_2) \quad \text{and} \quad x_2|\alpha,\beta \sim \mathscr{B}e(\alpha,\beta)\,.$$

[8] In probabilistic terms, if the Markov chains produced by these algorithms are *irreducible*, then these chains are both *positive recurrent* with stationary distribution π and *ergodic*, that is, asymptotically independent of the starting value $\mathbf{x}_0$.

[9] In the literature, both the denominations Gibbs *sampler* and Gibbs *sampling* can be found. In this book, we will use Gibbs sampling for the simulation technique and Gibbs sampler for the simulation algorithm.

Therefore, although

$$\pi(x_1) = \binom{n}{x_1} \frac{B(\alpha + x_1, \beta + n - x_1)}{B(\alpha, \beta)}$$

is available in closed form as the *beta-binomial distribution*, it is unnecessary to work with this marginal when one can simulate an iid sample $(x_1^{(i)}, x_2^{(i)})$ $(t = 1, \ldots, N)$ as

$$x_2^{(t)} \sim \mathscr{B}e(\alpha, \beta) \text{ and } x_1^{(t)} \sim \mathscr{B}(n, x_2^{(t)}) .$$

Integrals such as $\mathbb{E}[x_1/(x_1 + 1)]$ can then be approximated by

$$\frac{1}{N} \sum_{i=1}^{N} \frac{x_1^{(t)}}{x_1^{(t)} + 1} ,$$

using a regular Monte Carlo approach.

Unfortunately, even when one works with a representation such as (3.6) that is naturally associated with the original model, it is often the case that the mixing density $\tilde{\pi}(x_2)$ itself is neither available in closed form nor amenable to simulation. However, both *conditional posterior distributions*,

$$\pi_1(x_1|x_2) \quad \text{and} \quad \pi_2(x_2|x_1) ,$$

can often be simulated, and the following method takes full advantage of this feature.

3.5.2 Two-Stage Gibbs Sampler

The availability of both conditionals of (3.6) in terms of simulation can be exploited to build a transition kernel and a corresponding Markov chain, somewhat analogous to the derivation of the maximum of a multivariate function via an iterative device that successively maximizes the function in each of its arguments until a fixed point is reached.

The corresponding Markov kernel is built by simulating successively from each conditional distribution, with the conditioning variable being updated on the run. It is called the *two-stage Gibbs sampler* or sometimes the *data augmentation* algorithm, although both terms are rather misleading.[10]

[10] Gibbs sampling got its name from *Gibbs fields*, used in image analysis, when Geman and Geman (1984) proposed an early version of this algorithm, while data augmentation refers to Tanner's (1996) special use of this algorithm in missing-data settings, as seen in Chap. 6.

Algorithm 3.3 TWO-STAGE GIBBS SAMPLER

Initialization: Start with an arbitrary value $x_2^{(0)}$.
Iteration t: Given $x_2^{(t-1)}$, generate
1. $x_1^{(t)}$ according to $\pi_1(x_1|x_2^{(t-1)})$,
2. $x_2^{(t)}$ according to $\pi_2(x_2|x_1^{(t)})$.

Note that, in the second step of the algorithm, $x_2^{(t)}$ is generated conditional on $x_1 = x_1^{(t)}$, not $x_1^{(t-1)}$. The validation of this algorithm is that, for both generations, π is a stationary distribution. Therefore, the limiting distribution of the chain $(x_1^{(t)}, x_2^{(t)})_t$ is π if the chain is *irreducible*; that is, if it can reach any region in the support of π in a finite number of steps. (Note that there is a difference between the *stationary* distribution and the *limiting* distribution only in cases when the chain is not ergodic, as shown in Exercise 3.9.)

The practical implementation of Gibbs sampling involves solving two types of difficulties: the first type corresponds to deriving an efficient decomposition of the joint distribution in easily-simulated conditionals and the second one to deciding when to stop the algorithm. Evaluating the efficiency of the decomposition includes assessing the ease of simulating from both conditionals and the level of correlation between the $\mathbf{x}^{(t)}$'s, as well as the *mixing* behavior of the chain, that is, its ability to explore the support of π sufficiently fast. While deciding whether or not a given conditional can be simulated is easy enough, it is not always possible to find a manageable conditional, and more robust alternatives such as the *Metropolis–Hastings algorithm* will be described in the following chapters (see Sect. 4.2).

Choosing a stopping rule also relates to the mixing performances of the algorithm, as well as to its ability to approximate posterior expectations under π. Many indicators have been proposed in the literature (see Robert and Casella, 2004, Chap. 12) to signify convergence, or lack thereof, although none of these is foolproof. In the easiest cases, the lack of convergence is blatant and can be spotted on the raw plot of the sequence of the $\mathbf{x}^{(t)}$'s, while, in other cases, the Gibbs sampler explores very satisfactorily one mode of the posterior distribution but fails altogether to visit the *other* modes of the posterior: we will encounter such cases in Chap. 6 with mixtures of distributions. Throughout this chapter and the following ones, we give hints on how to implement these recommendations in practice.

Consider the posterior distribution derived in Exercise 2.11, for $n = 2$ observations,

$$\pi(\mu|\mathscr{D}_2) \propto \frac{e^{-\mu^2/20}}{\{1+(x_1-\mu)^2)(1+(x_2-\mu)^2\}} .$$

Even though this is a univariate distribution, it can still be processed by a Gibbs sampler through a data augmentation step, thus illustrating the idea behind (3.5). In fact, since $(j = 1, 2)$

$$\frac{1}{1+(x_j-\mu)^2} = \int_0^\infty e^{-\omega_j[1+(x_j-\mu)^2]}\,\mathrm{d}\omega_j\,,$$

we can define $\boldsymbol{\omega} = (\omega_1, \omega_2)$ and envision $\pi(\mu|\mathscr{D}_2)$ as the marginal distribution of

$$\pi(\mu, \boldsymbol{\omega}|\mathscr{D}_2) \propto e^{-\mu^2/20} \times \prod_{j=1}^{2} e^{-\omega_j[1+(x_j-\mu)^2]}\,.$$

For this multivariate distribution, a corresponding Gibbs sampler is associated with the following two steps:

1. Generate $\mu^{(t)} \sim \pi(\mu|\boldsymbol{\omega}^{(t-1)}, \mathscr{D}_2)$.
2. Generate $\boldsymbol{\omega}^{(t)} \sim \pi(\boldsymbol{\omega}|\mu^{(t)}, \mathscr{D}_2)$.

The second step is straightforward: the ω_i's are conditionally independent and distributed as $\mathscr{E}xp(1 + (x_i - \mu^{(t)})^2)$. The first step is also well-defined since $\pi(\mu|\boldsymbol{\omega}, \mathscr{D}_2)$ is a normal distribution with mean $\sum_i \omega_i x_i/(\sum_i \omega_i + 1/20)$ and variance $1/(2\sum_i \omega_i + 1/10)$. The corresponding R program then simplifies into two lines

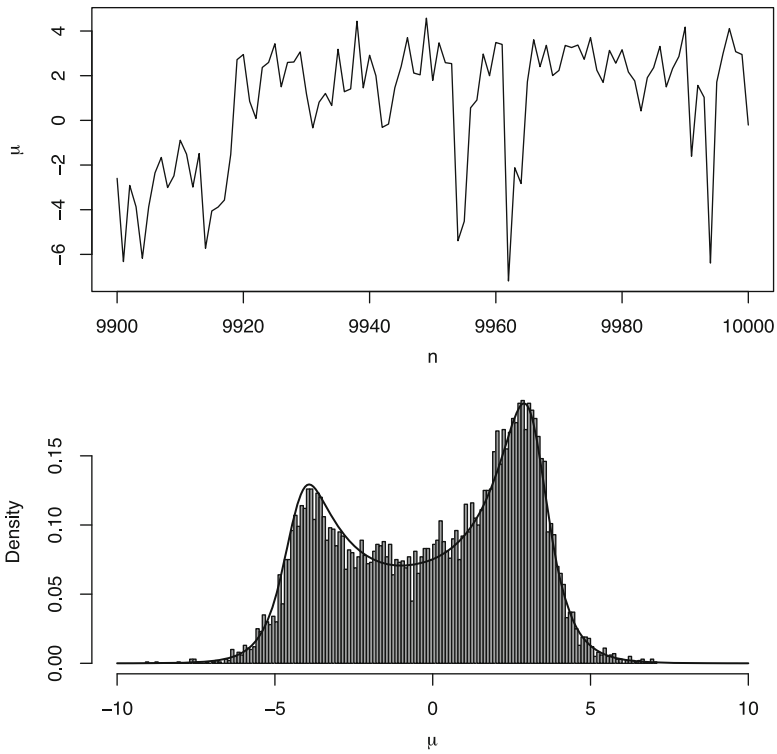

Fig. 3.4. (*Top*) Last 100 iterations of the chain $(\mu^{(t)})$; (*bottom*) histogram of the chain $(\mu^{(t)})$ and comparison with the target density for 10,000 iterations

```
> mu = rnorm(1,sum(x*omega)/sum(omega+.05),
+ sqrt(1/(.1+2*sum(omega))))
> omega = rexp(2,1+(x-mu)^2)
```

and the output of the simulation is represented in Fig. 3.4, with a very satisfying fit between the histogram of the simulated values and the target. A detailed zoom on the last 100 iterations shows how the chain $(\mu^{(t)})$ moves around, alternatively visiting each mode of the target.

When running a Gibbs sampler, the number of iterations should never be fixed in advance: it is usually impossible to predict the performance of a given sampler before producing a corresponding chain. Deciding on the length of an MCMC run is therefore a sequential process where output behaviors are examined after pilot runs and new simulations (or new samplers) are chosen on the basis of these pilot runs.

3.5.3 The General Gibbs Sampler

For a joint distribution $\pi(x_1, \ldots, x_p)$ with full conditionals $\pi_1, \ldots, \pi_p$ where π_j is the distribution of x_j conditional on $(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p)$, the Gibbs sampler simulates successively from all conditionals, modifying one component of $\mathbf{x}$ at a time. The corresponding algorithmic representation is given in Algorithm 3.4.

Algorithm 3.4 Gibbs Sampler

Initialization: Start with an arbitrary value $\mathbf{x}^{(0)} = (x_1^{(0)}, \ldots, x_p^{(0)})$.
Iteration t: Given $(x_1^{(t-1)}, \ldots, x_p^{(t-1)})$, generate

1. $x_1^{(t)}$ according to $\pi_1(x_1 | x_2^{(t-1)}, \ldots, x_p^{(t-1)})$,
2. $x_2^{(t)}$ according to $\pi_2(x_2 | x_1^{(t)}, x_3^{(t-1)}, \ldots, x_p^{(t-1)})$,

$\vdots$

p. $x_p^{(t)}$ according to $\pi_p(x_p | x_1^{(t)}, \ldots, x_{p-1}^{(t)})$.

Quite logically, the validation of this generalization of Algorithm 3.3 is identical: for each of the p steps of the t-th iteration, the joint distribution $\pi(\mathbf{x})$ is stationary. Under the same restriction on the irreducibility of the chain, it also converges to π for every possible starting value. Note that the order in which the components of $\mathbf{x}$ are simulated can be modified at each iteration, either deterministically or randomly, without putting the validity of the algorithm in jeopardy.

The two-stage Gibbs sampler naturally appears as a special case of Algorithm 3.4 for $p = 2$. It is, however, endowed with higher theoretical properties, as detailed in Robert and Casella (2004, Chap. 9) and Robert and Casella (2009, Chap. 7).

To conclude this section, let us stress that the impact of MCMC on Bayesian statistics has been considerable. Since the 1990s, which saw the emergence of MCMC methods in the statistical community, the occurrence of Bayesian methods in applied statistics has greatly increased, and the frontier between Bayesian and "classical" statistics is now so fuzzy that in some fields, it has completely disappeared. From a Bayesian point of view, the access to far more advanced computational means has induced a radical modification of the way people work with models and prior assumptions. In particular, it has opened the way to process much more complex structures, such as graphical models and latent variable models (see Chap. 6). It has also freed inference by opening for good the possibility of Bayesian model choice (see, e.g., Robert, 2007, Chap. 7). This expansion is much more visible among academics than among applied statisticians, though, given that the use of the MCMC technology requires some "hard" thinking to process every new problem. The availability of specific software such as BUGS has nonetheless given access to MCMC techniques to a wider community, starting with the medical field. New modules in R and other languages like Python are also helping to bridge the gap.

3.6 Variable Selection

3.6.1 Deciding on Explanatory Variables

In an ideal world, when building a regression model, we should include all relevant pieces of information, which in the regression context means including all predictor variables that might possibly help in explaining $\mathbf{y}$. However, there are obvious drawbacks to the advice of increasing the number of explanatory variables. For one thing, in noninformative settings, this eventually clashes with the constraint $p < n$. For another, using a huge number of explanatory variables leaves little information available to obtain precise estimators. In other words, when we increase the explanatory scope of the regression model, we do not necessarily increase its explanatory power because it gets harder and harder to estimate the coefficients.[11] It is thus important to be

[11] This phenomenon is related to the *principle of parsimony*, also called *Occam's razor*, which states that, among two models with similar explanatory powers, the simplest one should always be preferred. It is also connected with the *learning curve effect* found in information theory and neural networks, where the performance of a model increases on the learning dataset but decreases on a testing dataset as its complexity increases.

able to decide which variables—within a large pool of potential explanatory variables—should be kept in a model that balances good explanatory power with good estimation performance.

This is truly a *decision* problem in that all potential models have to be considered in parallel against a criterion that ranks them. This variable-selection problem can be formalized as follows. We consider a dependent random variable y and a set of p potential explanatory variables. At this stage, we assume that every subset of q explanatory variables could make a proper set of explanatory variables for the regression of y. The only restriction we impose is that the intercept (that is, the constant variable) is included in every model. There are thus 2^p models in competition and we are looking for a procedure that selects the "best" model, that is, the "most relevant" explanatory variables. Note that this variable-selection procedure can alternatively be seen as a two-stage estimation setting where we first estimate the indicator of the model (within the collection of models), which also amounts to estimating variable indicators, as detailed below, and we then estimate the parameters corresponding to this very model.

Each of the 2^p models under comparison is in fact associated with a binary indicator vector $\boldsymbol{\gamma} \in \Gamma = \{0,1\}^p$, where $\gamma_j = 1$ means that the variable x_j is included in the model, denoted by $\mathfrak{M}_{\boldsymbol{\gamma}}$. This notation is quite handy since $\boldsymbol{\gamma}$=(1,0,1,0,0,...,1,0) clearly indicates which explanatory variables are in and which are not. We also use the notation

$$q_{\boldsymbol{\gamma}} = \mathbf{1}_p^{\mathsf{T}}\boldsymbol{\gamma}$$

for computing the number of variables included in the model $\mathfrak{M}_{\boldsymbol{\gamma}}$. We define $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$ as a sub-vector of $\boldsymbol{\beta}$ containing only the components such that x_j is included in the model $\mathfrak{M}_{\boldsymbol{\gamma}}$ and $\mathbf{X}^{\boldsymbol{\gamma}}$ as the sub-matrix of $\mathbf{X}$ where only the columns such that x_j is included in the model $\mathfrak{M}_{\boldsymbol{\gamma}}$ have been left. The model $\mathfrak{M}_{\boldsymbol{\gamma}}$ is thus defined as

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

Once again, and apparently in contradiction to our basic tenet that different models should enjoy completely different parameters, we are compelled to denote by σ^2 and α the variance and intercept terms common to *all models*, respectively. Although this is more of a mathematical trick than a true modeling reason, the prior independence of (α, σ^2) and $\boldsymbol{\gamma}$ allows for the simultaneous use of Bayes factors and an improper prior. Despite the possibly confusing notation, $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$ and $\boldsymbol{\beta}$ are completely unrelated in that they are parameters of different models.

3.6.2 *G*-Prior Distributions for Model Choice

Because so many models are in competition and thus considered in the global model all at once, we cannot expect a practitioner to specify one's own prior on every model $\mathfrak{M}_\gamma$ in a completely subjective and autonomous manner. We thus now proceed to derive *all* priors from a single global prior associated with the so-called *full model* that corresponds to $\boldsymbol{\gamma} = (1, \ldots, 1)$. The argument goes as follows:

(1) For the full model, we use Zellner's G-prior as defined in Sect. 3.4,

$$\boldsymbol{\beta}|\sigma^2 \sim \mathcal{N}_p(\tilde{\boldsymbol{\beta}}, g\sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}) \quad \text{and} \quad \pi(\alpha, \sigma^2) \propto \sigma^{-2}\,.$$

(2) For each (sub-)model $\mathfrak{M}_\gamma$, the prior distribution of $\boldsymbol{\beta}^\gamma$ conditional on σ^2 is fixed as

$$\boldsymbol{\beta}^\gamma|\sigma^2, \boldsymbol{\gamma} \sim \mathcal{N}_{q_\gamma}\left(\tilde{\boldsymbol{\beta}}^\gamma, g\sigma^2\left(\mathbf{X}^{\gamma\mathsf{T}}\mathbf{X}^\gamma\right)^{-1}\right),$$

where $\tilde{\boldsymbol{\beta}}^\gamma = \left(\mathbf{X}^{\gamma T}\mathbf{X}^\gamma\right)^{-1}\mathbf{X}^{\gamma T}\tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}$ and we use the same prior on (α, σ^2).

↯ This distribution is conditional on $\boldsymbol{\gamma}$; in particular, this implies that, while the variance notation σ^2 is common to all models, its distribution varies with $\boldsymbol{\gamma}$.

Although there are many possible ways of defining the prior on the model index[12] $\boldsymbol{\gamma}$, we opt for the uniform prior

$$\pi(\boldsymbol{\gamma}) = 2^{-p}\,.$$

The posterior distribution of $\boldsymbol{\gamma}$ (that is, the distribution of $\boldsymbol{\gamma}$ given $\mathbf{y}$) is central to the variable-selection methodology since it is proportional to the marginal density of $\mathbf{y}$ in $\mathfrak{M}_\gamma$. In addition, for prediction purposes, the prediction distribution can be obtained by averaging over all models, the weights being the model probabilities (this is called *model averaging*).

The posterior distribution of $\boldsymbol{\gamma}$ is

$$\begin{aligned}\pi(\boldsymbol{\gamma}|\mathbf{y}) &\propto f(\mathbf{y}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma}) \propto f(\mathbf{y}|\boldsymbol{\gamma})\\ &\propto (g+1)^{-(q_\gamma+1)/2}\left[\mathbf{y}^\mathsf{T}\mathbf{y} - \frac{g}{g+1}\mathbf{y}^\mathsf{T}\mathbf{X}^\gamma\left(\mathbf{X}^{\gamma\mathsf{T}}\mathbf{X}^\gamma\right)^{-1}\mathbf{X}^{\gamma\mathsf{T}}\mathbf{y}\right.\\ &\qquad\left. - \frac{1}{g+1}\tilde{\boldsymbol{\beta}}^{\gamma\mathsf{T}}\mathbf{X}^{\gamma\mathsf{T}}\mathbf{X}^\gamma\tilde{\boldsymbol{\beta}}^\gamma\right]^{-(n-1)/2}. \qquad (3.7)\end{aligned}$$

When the number of explanatory variables is less than 15, say, the exact derivation of the posterior probabilities for all submodels can be undertaken.

[12] For instance, one could instead use a uniform prior on the number q_γ of explanatory variables or a more parsimonious prior such as $\pi(\boldsymbol{\gamma}) = 1/q_\gamma$.

Indeed, $2^{15} = 32768$ means that the problem remains tractable. The following R code (part of the function `ModChoBayesReg`) is used to calculate those posterior probabilities and returns the top most probable models. The integrated likelihood for the null model is computed as `intlike0`.

```
intlike=rep(intlike0,2^p)
for (j in 2:2^p){
  gam=as.integer(intToBits(i-1)[1:p]==1)
  pgam=sum(gam)
  Xgam=X[,which(gam==1)]
  Ugam=solve(t(Xgam)%*%Xgam)%*%t(Xgam)
  betatildegam=b1=Ugam%*%X%*%betatilde
  betamlgam=b2=Ugam%*%y
  s2gam=t(y-alphaml-Xgam%*%b2)%*%(y-alphaml-Xgam%*%b2)
  kappagam=as.numeric(s2gam+t(b1-b2)%*%t(Xgam)%*%
  Xgam%*%(b1-b2)/(g+1))
  intlike[j]=(g+1)^(-pgam/2)*kappagam^(-(n-1)/2)
  }
intlike=intlike/sum(intlike)
modcho=order(intlike)[2^p:(2^p-9)]
probtop10=intlike[modcho]
```

The above R code uses the generic function `intToBits` to turn an integer `i` into the indicator vector `gam`. The remainder of the code is quite similar to the model choice code when computing the Bayes factors.

For the **caterpillar** data, we set $\tilde{\beta} = 0_8$ and $g = 1$. The models corresponding to the top 10 posterior probabilities are then given by

```
> ModChoBayesReg(y,X,g=1)

Number of variables less than 15
Model posterior probabilities are calculated exactly

   Top10Models  PostProb
1      1 2 3 7   0.0142
2    1 2 3 5 7   0.0138
3        1 2 7   0.0117
4    1 2 3 4 7   0.0112
5  1 2 3 4 5 7   0.0110
6      1 2 5 7   0.0108
7    1 2 3 7 8   0.0104
8    1 2 3 6 7   0.0102
9  1 2 3 5 6 7   0.0100
10 1 2 3 5 7 8   0.0098
```

In a basic $0-1$ decision setup, we would choose the model $\mathfrak{M}_\gamma$ with the highest posterior probability—that is, the model with explanatory variables x_1, x_2, x_3 and x_7—which corresponds to the variables

- altitude,
- slope,
- the number of pine trees in the area, and
- the number of vegetation strata.

The model selected by the procedure thus fails to correspond to the three variables identified in the R output at the end of Sect. 3.4. But interestingly, even under this strong shrinkage prior $g = 1$ (where the prior has the same weight as the data), all top ten models contain the explanatory variables x_1, x_2 and x_7, which have the most stars in this R analysis.

Now, the default or noninformative calibration of the G-prior corresponds to the choice $\tilde{\beta} = 0_p$ and $g = n$, which reduces the prior input to the equivalent of a *single* observation. Pushing g to a smaller value results in a paradoxical behaviour of the procedure which then usually picks the simpler model: this is another illustration of the *Jeffreys-Lindley paradox*, mentioned in Chap. 2.

For $\tilde{\boldsymbol{\beta}} = 0_p$ and $g = n$, the ten most likely models and their posterior probabilities are:

```
> ModChoBayesReg(y,X)

Number of variables less than 15
Models's posterior probabilities are calculated exactly

   Top10Models PostProb
1        1 2 7   0.0767
2          1 7   0.0689
3      1 2 3 7   0.0686
4        1 3 7   0.0376
5        1 2 6   0.0369
6    1 2 3 5 7   0.0326
7      1 2 5 7   0.0294
8          1 6   0.0205
9      1 2 4 7   0.0201
10           7   0.0198
```

For this different prior modelling, we chose the same model as the `lm` classical procedure, rather than when $g = 1$; however, the posterior probabilities of the most likely models are much lower for $g = 1$, which is logical given that the current prior is less informative. Therefore, the top model is not as strongly supported as in the informative case. Once again, we stress that the choice $g = 1$ is rather arbitrary and that it is used here merely for illustrative purposes. The default value we recommend is $g = n$.

3.6.3 A Stochastic Search for the Most Likely Model

When the number p of variables is large, it becomes impossible to compute the posterior probabilities for the whole series of 2^p models. We then need a tailored algorithm that samples from $\pi(\boldsymbol{\gamma}|\mathbf{y})$ and thus selects the most likely models, without computing first all the values of $\pi(\boldsymbol{\gamma}|\mathbf{y})$. This can be done rather naturally by Gibbs sampling, given the availability of the full conditional posterior probabilities of the γ_j's.

Indeed, if $\boldsymbol{\gamma}_{-j}$ $(1 \le j \le p)$ is the vector $(\gamma_1, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_p)$, the full conditional distribution $\pi(\gamma_j|\mathbf{y}, \boldsymbol{\gamma}_{-j})$ of γ_j is proportional to $\pi(\boldsymbol{\gamma}|\mathbf{y})$ and can be computed in both $\gamma_j = 0$ and $\gamma_j = 1$ at no cost (since these are the only possible values of γ_j).

Algorithm 3.5 Gibbs Sampler for Variable Selection

Initialization: Draw $\boldsymbol{\gamma}^0$ from the uniform distribution on Γ.
Iteration t: Given $(\gamma_1^{(t-1)}, \ldots, \gamma_p^{(t-1)})$, generate

1. $\gamma_1^{(t)}$ according to $\pi(\gamma_1|\mathbf{y}, \gamma_2^{(t-1)}, \ldots, \gamma_p^{(t-1)})$,
2. $\gamma_2^{(t)}$ according to $\pi(\gamma_2|\mathbf{y}, \gamma_1^{(t)}, \gamma_3^{(t-1)}, \ldots, \gamma_p^{(t-1)})$,

$\vdots$

p. $\gamma_p^{(t)}$ according to $\pi(\gamma_p|\mathbf{y}, \gamma_1^{(t)}, \ldots, \gamma_{p-1}^{(t)})$.

After a large number of iterations of this algorithm (that is, when the sampler is supposed to have converged or, more accurately, when the sampler has sufficiently explored the support of the target distribution), its output can be used to approximate the posterior probabilities $\pi(\boldsymbol{\gamma}|\mathbf{y}, X)$ by empirical averages based on the Gibbs output,

$$\hat{\mathbb{P}}^{\pi}(\boldsymbol{\gamma} = \boldsymbol{\gamma}^*|\mathbf{y}) = \left(\frac{1}{T - T_0 + 1}\right) \sum_{t=T_0}^{T} \mathbb{I}_{\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma}^*} ,$$

where the T_0 first values are eliminated as *burn-in*. (The number T_0 is therefore the number of iterations roughly needed to "reach" convergence.) The Gibbs output can also be used to approximate the inclusion of a given variable, $P^{\pi}(\gamma_j = 1|\mathbf{y}, X)$, as

$$\hat{\mathbb{P}}^{\pi}(\gamma_j = 1|\mathbf{y}) = \left(\frac{1}{T - T_0 + 1}\right) \sum_{t=T_0}^{T} \mathbb{I}_{\gamma_j^{(t)} = 1} ,$$

with the same asymptotic validation.

The following R code (again part of the function `ModChoBayesReg`) describes our implementation of the above variable-selection Gibbs sampler.

The code uses the null model with only the intercept α as a reference, based on the integrated likelihood `intlike0` as above. It then starts at random in the collection of models:

```
gamma=rep(0,niter)
mcur=sample(c(0,1),p,replace=TRUE)
gamma[1]=sum(2^(0:(p-1))*mcur)+1
pcur=sum(mcur)
```

and computes the corresponding integrated likelihood `intlikecur`

```
if (pcur==0) intlikecur=intlike0 else{ #integrated likelihood
 Xcur=X[,which(mcur==1)]
 Ucur=solve(t(Xcur)%*%Xcur)%*%t(Xcur)
 betatildecur=b1=Ucur%*%X%*%betatilde
 betamlcur=b2=Ucur%*%y
 s2cur=t(y-alphaml-Xcur%*%b2)%*%(y-alphaml-Xcur%*%b2)
 kappacur=as.numeric(s2cur+t(b1-b2)%*%t(Xcur)%*%
 Xcur%*%(b1-b2)/(g+1))
 intlikecur=(g+1)^(-pcur/2)*kappacur^(-(n-1)/2)
 }
```

It then proceeds according to Algorithm 3.5, proposing to change one variable indicator γ_j and accepting this move with a Metropolis–Hastings (defined and justified in Chap. 4) probability:

```
if (runif(1)<=(intlikeprop/intlikecur))
```

This modification is more efficient than directly simulating from the conditional as it avoids proposing the same value for γ_j twice.

```
for (t in 1:(niter-1)){ #iteration index
 mprop=mcur
 j=sample(1:p,1)
 mprop[j]=abs(mcur[j]-1)
 pprop=sum(mprop)
 if (pprop==0) intlikeprop=intlike0 else{ #integrated
   likelihood Xprop=X[,which(mprop==1)]
   Uprop=solve(t(Xprop)%*%Xprop)%*%t(Xprop)
   betatildeprop=b1=Uprop%*%X%*%betatilde
   betamlprop=b2=Uprop%*%y
   s2prop=t(y-alphaml-Xprop%*%betamlprop)%*
    %(y-alphaml-Xprop%*%betamlprop)
   kappaprop=as.numeric(s2prop+t(betatildeprop-betamlprop)%*
    %t(Xprop)%*%Xprop%*%
   (betatildeprop-betamlprop)/(g+1))
   intlikeprop=(g+1)^(-pprop/2)*kappaprop^(-(n-1)/2)
   }
 if (runif(1)<=(intlikeprop/intlikecur)){
```

```
    mcur=mprop
    intlikecur=intlikeprop
    }
  gamma[t+1]=sum(2^(0:(p-1))*mcur)+1
  }
gamma=gamma[20001:niter] #20,000 burnin steps
res=as.data.frame(table(as.factor(gamma)))
odo=order(res$Freq)[length(res$Freq):(length(res$Freq)-9)]
modcho=res$Var1[odo]
probtop10=res$Freq[odo]/(niter-20000)
```

In this setting of **caterpillar**, handling only eight (potential) explanatory variables means that it is possible to compute all of the 2^8 probabilities $\pi(\boldsymbol{\gamma}|\mathbf{y})$ and to thus deduce the normalizing constant in (3.7). We can therefore compare these exact values with the approximations produced by the Gibbs sampler. Using $T_0 = 20{,}000$ and $T_0 = 80{,}000$, i.e. a total of 10^5 simulations, we obtain the following results for the top five models:

```
          Models PostProb  Gibbs estimates
                           of the PostProb

1        1 2 7   0.0767  0.0740
2          1 7   0.0689  0.0675
3      1 2 3 7   0.0686  0.0668
4        1 3 7   0.0376  0.0376
5        1 2 6   0.0369  0.0370
```

The comparison is quite comforting for the Gibbs sampler as the differences are truly minor! Rather naturally, as the number of variables grows, the number of simulations needed to provide a good approximation grows as well. Once more, we recommend running the code several times (with different random sequences) to ensure the stability of the approximation.

3.7 Exercises

3.1 Show that the matrix $\mathbf{Z}$ is of full rank if and only if the matrix $\mathbf{Z}^{\mathsf{T}}\mathbf{Z}$ is invertible (where $\mathbf{Z}^{\mathsf{T}}$ denotes the transpose of the matrix $\mathbf{Z}$, which can be produced in R using the `t(Z)` command). Apply to $\mathbf{Z} = [\mathbf{1}_n \quad \mathbf{X}]$ and deduce that this cannot happen when $p + 1 > n$.

3.2 Show that solving the minimization program

$$\min_{\beta} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

requires solving the system of equations $(\mathbf{X}^\mathsf{T}\mathbf{X})\beta = \mathbf{X}^\mathsf{T}\mathbf{y}$. Check that this can be done via the R command `solve(t(X)%*%(X),t(X)%*%y)`.

3.3 Show that the variance of the maximum likelihood estimator of β in the regression model is given by $\mathbb{V}(\hat{\boldsymbol{\beta}}|\sigma^2) = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$.

3.4 For the model

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

a conjugate prior distribution is as follows: the conditional distribution of $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\beta}|\sigma^2 \sim \mathcal{N}_p(\tilde{\boldsymbol{\beta}}, \sigma^2\mathbf{M}^{-1})\,,$$

where $\mathbf{M}$ is a (p, p) positive definite symmetric matrix, and the marginal prior on σ^2 is an inverse Gamma distribution

$$\sigma^2 \sim \mathcal{IG}(a, b), \qquad a, b > 0\,.$$

Taking advantage of the matrix identities

$$\begin{aligned}\left(\mathbf{M} + \mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1} &= \mathbf{M}^{-1} - \mathbf{M}^{-1}\left(\mathbf{M}^{-1} + (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)^{-1}\mathbf{M}^{-1} \\ &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} - (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\left(\mathbf{M}^{-1} + (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)^{-1}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\end{aligned}$$

and

$$\begin{aligned}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{M} + \mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{M} &= \left(\mathbf{M}^{-1}(\mathbf{M} + \mathbf{X}^\mathsf{T}\mathbf{X})(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)^{-1} \\ &= \left(\mathbf{M}^{-1} + (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)^{-1}\,,\end{aligned}$$

establish that

$$\boldsymbol{\beta}|\mathbf{y}, \sigma^2 \sim \mathcal{N}_p\left((\mathbf{M} + \mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\{(\mathbf{X}^\mathsf{T}\mathbf{X})\hat{\boldsymbol{\beta}} + \mathbf{M}\tilde{\boldsymbol{\beta}}\}, \sigma^2(\mathbf{M} + \mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right) \tag{3.8}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$ and

$$\sigma^2|\mathbf{y} \sim \mathcal{IG}\left(\frac{n}{2} + a, b + \frac{s^2}{2} + \frac{(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^\mathsf{T}\left(\mathbf{M}^{-1} + (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)^{-1}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})}{2}\right) \tag{3.9}$$

where $s^2 = (\mathbf{y} - \hat{\boldsymbol{\beta}}\mathbf{X})^\mathsf{T}(\mathbf{y} - \hat{\boldsymbol{\beta}}\mathbf{X})$ are the correct posterior distributions. Give a $(1 - \alpha)$ HPD region on $\boldsymbol{\beta}$.

3.5 The regression model of Exercise 3.4 can also be used in a predictive sense: for a given $(m, p + 1)$ explanatory matrix $\tilde{\mathbf{X}}$, *i.e.*, when predicting m unobserved variates $\tilde{y}_i$, the corresponding outcome $\tilde{\mathbf{y}}$ can be inferred through the *predictive distribution* $\pi(\tilde{\mathbf{y}}|\sigma^2, \mathbf{y})$. Show that $\pi(\tilde{\mathbf{y}}|\sigma^2, \mathbf{y})$ is a Gaussian density with mean

$$\mathbb{E}^\pi[\tilde{\mathbf{y}}|\sigma^2, \mathbf{y}] = \tilde{\mathbf{X}}(\mathbf{M} + \mathbf{X}^\mathsf{T}\mathbf{X})^{-1}(\mathbf{X}^\mathsf{T}\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{M}\tilde{\boldsymbol{\beta}})$$

and covariance matrix

$$\mathbb{V}^\pi(\tilde{\mathbf{y}}|\sigma^2,\mathbf{y}) = \sigma^2(\mathbf{I}_m + \tilde{\mathbf{X}}(\mathbf{M}+\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\tilde{\mathbf{X}}^\mathsf{T})\,.$$

Deduce that

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

3.6 Show that the marginal distribution of $\mathbf{y}$ associated with (3.8) and (3.9) is given by

$$\mathbf{y} \sim \mathscr{T}_n\left(2a, \mathbf{X}\tilde{\boldsymbol{\beta}}, \frac{b}{a}(\mathbf{I}_n+\mathbf{X}\mathbf{M}^{-1}\mathbf{X}^\mathsf{T})\right)\,.$$

3.7 Show that the matrix $(\mathbf{I}_n + g\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T})$ has 1 and $g+1$ as only eigenvalues. (*Hint:* Show that the eigenvectors associated with $g+1$ are of the form $\mathbf{X}\boldsymbol{\beta}$ and that the eigenvectors associated with 1 are those orthogonal to $\mathbf{X}$.) Deduce that the determinant of the matrix $(\mathbf{I}_n + g\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T})$ is indeed $(g+1)^{p+1}$.

3.8 Under the Jeffreys prior, give the predictive distribution of $\tilde{\mathbf{y}}$, m dimensional vector corresponding to the (m,p) matrix of explanatory variables $\tilde{\mathbf{X}}$.

3.9 If (x_1,x_2) is distributed from the uniform distribution on

$$\left\{(x_1,x_2);\,(x_1-1)^2+(x_2-1)^2\le 1\right\}\cup\left\{(x_1,x_2);\,(x_1+1)^2+(x_2+1)^2\le 1\right\}\,,$$

show that the Gibbs sampler does not produce an irreducible chain. For this distribution, find an alternative Gibbs sampler that works. (*Hint:* Consider a rotation of the coordinate axes.)

3.10 If a joint density $g(y_1,y_2)$ corresponds to the conditional distributions $g_1(y_1|y_2)$ and $g_2(y_2|y_1)$, show that it is given by

$$g(y_1,y_2) = \frac{g_2(y_2|y_1)}{\int g_2(v|y_1)/g_1(y_1|v)\,\mathrm{d}v}.$$

3.11 Considering the model

$$\eta|\theta \sim \mathcal{B}\mathrm{in}(n,\theta)\,, \quad \theta \sim \mathcal{B}e(a,b),$$

derive the joint distribution of (η,θ) and the corresponding full conditional distributions. Implement a Gibbs sampler associated with those full conditionals and compare the outcome of the Gibbs sampler on θ with the true marginal distribution of θ.

3.12 Take the posterior distribution on (θ, σ^2) associated with the joint model

$$x_i|\theta, \sigma^2 \sim \mathcal{N}(\theta, \sigma^2), \quad i = 1, \ldots, n,$$
$$\theta \sim \mathcal{N}(\theta_0, \tau^2), \quad \sigma^2 \sim I\mathcal{G}(a, b).$$

Show that the full conditional distributions are given by

$$\theta|\mathbf{x}, \sigma^2 \sim \mathcal{N}\left(\frac{\sigma^2}{\sigma^2 + n\tau^2}\,\theta_0 + \frac{n\tau^2}{\sigma^2 + n\tau^2}\,\bar{\mathbf{x}},\ \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right)$$

and

$$\sigma^2|\mathbf{x}, \theta \sim I\mathcal{G}\left(\frac{n}{2} + a, \frac{1}{2}\sum_i (x_i - \theta)^2 + b\right),$$

where $\bar{x}$ is the empirical average of the observations. Implement the Gibbs sampler associated with these conditionals.