Chapter 1

Bayesian Analysis of the Normal Regression Model

by Ioannis Ntzoufras,
Department of Statistics, Athens University of Economics and Business, Athens, Greece; E-mail: ntzoufras@aueb.gr.

Abstract

In this chapter, we implement Bayesian methods in regression models which are an essential tool in modern statistical science. They can be used for both interpretation of social or economic phenomena and prediction of future outcomes which is of major interest in risk analysis. Possible prior specifications are described in detail. Posterior inference is illustrated focusing on the conjugate case. Bayesian variable selection methods for the conjugate case are also illustrated while more advanced topics such as variable selection using MCMC and evaluation of the structural assumptions is briefly discussed accompanied with references for further reading. The chapter closes with a short discussion and conclusion.

1.1 Introduction

One of the most important elements of statistical inference is regression analysis inspired by the original work of Sir Francis Galton in the late years of the 19th century (Stanton, 2001). Regression models can be considered as the core of econometrics. They are frequently met in risk analysis either in their original form or using more realistic extensions which incorporate time dependence.

Regression models are based on the idea that a variable of major interest (*response*) exists for which we wish to find a way to explain or predict its behavior. To do so, we need to identify some external factors (*explanatory variables*) which help us to achieve this aim. Here we refer to *normal regression models* whose response (or equivalently the error term) is assumed to be normal.

In this chapter, we focus on different prior distributions and the corresponding posterior inference for normal regression models. Conjugate analysis is illustrated using a simple example. Details concerning model comparison and variable selection are also presented with emphasis given in the conjugate case. The chapter closes with a short discussion concerning certain extensions of the normal linear model.

1.2 The Normal Linear Model

The Usual Formulation

In normal regression models, the response variable Y is considered to be a continuous random variable defined in the whole set of real numbers. This response variable can be decomposed in two parts: the *systematic* and the *stochastic*. The first refers to the part of the response that can be accurately identified in a systematic way via explanatory variables (also called covariates or predictors) $X_1, X_2, \ldots, X_p$ which are assumed to be fixed within the model formulation. Although categorical covariates can be incorporated in the linear predictor using dummy variables, in this chapter we focus on numerical explanatory variables. The *stochastic* part refers to a random error which is assumed to follow a normal distribution with mean zero and variance σ^2 (also called *residual* variance). Although σ^2 is often neglected, it is of prominent importance for each regression model since it quantifies the uncertainty (and indirectly the precision) of our predictions. It actually refers to the variance of Y that cannot be explained or predicted by the systematic component of the model. Hence, assuming p covariates denoted by $X_1, \ldots, X_p$, the model can be written as

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma^2), \tag{1.1}$$

where $N(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2. Moreover, the β_j are referred to as the regression coefficients, specifically, β_0 is the intercept or the constant term. The systematic component of the above model formulation, given by $\eta = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$, is simply the linear combination of the covariates and is called *linear predictor*. An alternative formulation is the following

$$Y|X_1, \ldots, X_p \sim N(\mu, \sigma^2) \text{ with } \mu = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \, . \tag{1.2}$$

From this formulation, it is clear that the linear predictor is equal to the expected value of the response (under the assumed model).

Although (1.1) has a straightforward interpretation since the response variable is simply expressed as the sum of the linear combination of the covariates and an error term, expression (1.2) is more general and can be used to extend the model by simply changing the distribution of the response or the function connecting the linear predictor and the expected value of Y.

Note that, when a finite i.i.d. sample $(y_i, X_{i1}, \ldots, X_{ip})$ for $i = 1, \ldots, n$ of observations (or units or subjects) is given, a subscript i is added in (1.1) and (1.2) to denote that the corresponding expressions hold for every individual i of the sample. Hence, the model can be now written as

$$Y_i \;=\; \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i \text{ with } \varepsilon_i \sim N(0, \sigma^2)$$

or, equivalently,

$$\begin{aligned} Y_i \;&\sim\; N(\mu_i, \sigma^2) \text{ with } \mu_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} \\ &\quad \text{for } i = 1, \ldots, n\ . \end{aligned}$$

The induced model likelihood follows

$$f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{X}) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \Big)^2 \right),$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ is the vector of model coefficients, $\boldsymbol{y} = (y_1, \ldots, y_n)$ is the vector of the observed response data while the matrix $\boldsymbol{X}$ is the data matrix of dimension $n \times (p+1)$ with the ith row corresponding to the values of ith observation given by $(1, X_{i1}, X_{i2}, \ldots, X_{ip})$.

Multivariate Representation

The above model can be compactly rewritten using a multivariate normal distribution. This representation simplifies (especially in the conjugate case) the computations involved in the posterior distribution. We write

$$\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{X} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n) \tag{1.3}$$

where $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the d-dimensional normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix equal to $\boldsymbol{\Sigma}$. Note that the intercept β_0, corresponding to the column of 1's in $\boldsymbol{X}$, can be omitted if this is necessary without any implication to the multivariate representation. Excluding the intercept from the model assumes that the expected value of Y is equal to zero when all covariates are equal to zero. In the following we denote by P the number of columns of $\boldsymbol{X}$ to include both cases with or without the intercept corresponding to $\text{P} = p + 1$ and $\text{P} = p$, respectively. The likelihood can be now written as

$$f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{X}) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right). \tag{1.4}$$

Maximum Likelihood Estimation for Regression Models

Here we present the maximum likelihood estimates (MLE) which will be used for comparison reasons with corresponding Bayesian outcomes. A straightforward calculation shows that the global maximum is given by

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \;\text{ and }\; \widehat{\sigma}^2 = \frac{1}{n} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}). \tag{1.5}$$

Note that $\widehat{\boldsymbol{\beta}}$ is an unbiased estimate of $\boldsymbol{\beta}$ while $\widehat{\sigma}^2$ is only approximately unbiased estimate of σ^2. Alternatively, we may write $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$; where $\widehat{y}_i = \widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j X_{ij}$ is the *fitted* or *predicted value* for y_i, and $e_i := y_i - \widehat{y}_i$ is the *residual* value which can be considered as an observed value for the error term ε_i. The covariance matrix of $\widehat{\boldsymbol{\beta}}$ is given by $\text{Covar}(\widehat{\boldsymbol{\beta}}|\sigma^2, \boldsymbol{X}) = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \sigma^2$ and is estimated by

$$\widehat{\text{Covar}}(\widehat{\boldsymbol{\beta}}|\sigma^2, \boldsymbol{X}) = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \widehat{\sigma}^2\ . \tag{1.6}$$

In order to be able to obtain a unique MLE solution for a given set of data, we need to have $\det(\boldsymbol{X}^T \boldsymbol{X}) \neq 0$ which ensures the existence of the inverse matrix of $\boldsymbol{X}^T \boldsymbol{X}$ involved in $\widehat{\boldsymbol{\beta}}$. Therefore $\boldsymbol{X}$ must be of full rank (i.e. $\text{rank}(\boldsymbol{X}) = \text{P}$) and $\text{P} < n$. In the case of $n = \text{P}$ we may obtain estimates for $\boldsymbol{\beta}$ but not for σ^2 which is not identifiable. When a covariate can be expressed as the linear combination of the remaining covariates then no simple MLE solution can be obtained. Even in the case that one covariate is not an exact linear function of the remaining ones but can be efficiently described via a regression model with very low residual variance (i.e. $\det(\boldsymbol{X}^T \boldsymbol{X})$ is close to zero) we will still face problems regarding estimation of model parameters since the variance of the MLEs given in (1.6) will be inflated and thus $\widehat{\boldsymbol{\beta}}$ will be inaccurate and unstable. This situation is described in the literature as *multi-collinearity problem*.

Parameter Interpretation

One reason for the popularity of normal regression models is their straightforward interpretation of parameters. The effect of each covariate X_j is measured by β_j which accounts the expected increase of Y when X_j increases by one unit and the rest of covariates remain the same. The sign of β_j is also important since it defines whether the effect of X_j is positive or negative on Y. *Positive association* implies that changes of the explanatory variable X_j cause changes of the same direction for variable Y while *negative association* implies that changes of X_j cause changes of the opposite direction for Y.

The constant β_0 corresponds to the expected value of Y when all covariates are equal to zero. This often has no meaningful interpretation since the corresponding zero covariate values may be unrealistic in practice. Usually the constant term is included in the model unless practice supports setting it equal to zero or if it is simply not significant. A meaningful interpretation of β_0 can be obtained if we center each covariate X_j to its sample mean $\overline{X}_j$. In this case, the constant represents the expected value of Y for an individual with all covariates equal to the sample means i.e. it provides the expected response value for a typical individual of the sample.

Finally, it is of major interest to discriminate zero from non-zero coefficients which correspond to unimportant and important (respectively) determinants of Y. This is related to the variable selection problem which is a flourishing area of research, especially during the last decade, due to the large number

of available covariates. The size of available data has been tremendously increased over the last years due to the increasing storing and computational capabilities of personal computers and problems arising in sciences such as genetics. In Bayesian theory, posterior model probabilities, odds and Bayes factors are used to identify which effects should be removed or not from the model (for a concise review see Kass and Raftery, 1995) in contrast to the usual stepwise procedures using significance tests and p-values. Due to problems in the prior specification and in the computation of posterior model odds, alternative approaches have been also proposed in the literature such as the deviance information criterion (DIC), see Spiegelhalter et al. (2002), and the use of posterior p-values (see for example in Bayarri and Berger, 2000). More details concerning variable selection are provided in later in this chapter.

1.3 Prior Distributions for Normal Regression Models

1.3.1 The Conjugate Normal–Inverse Gamma Prior

The conjugate prior for $[\boldsymbol{\beta}, \sigma^2]$ in the normal regression model is the normal–inverse gamma distribution which is specified as

$$\boldsymbol{\beta}|\sigma^2, \boldsymbol{X} \sim N_{\mathrm{P}}(\boldsymbol{\mu}_\beta, \boldsymbol{V}\sigma^2) \;\text{ and }\; \sigma^2|\boldsymbol{X} \sim \mathrm{IG}(a, b), \quad a, b > 0; \tag{1.7}$$

where $\boldsymbol{V}$ is a $\mathrm{P} \times \mathrm{P}$ positive definite symmetric matrix controlling the prior variances and covariances of $\boldsymbol{\beta}$ given σ^2 and $IG(a, b)$ is the inverse gamma distribution with parameters a, b. Under this setup, the precision parameter $\tau = \sigma^{-2}$, which is frequently used in Bayesian inference, is assumed to a-priori follow the gamma distribution with mean a/b and variance a/b^2. Among the four hyper parameters, $a, b, \boldsymbol{\mu}_\beta$ and $\boldsymbol{V}$, the latter is the most difficult to elicit. This motivates Zellner's g-prior which is discussed next. However, there are methods to elicit $\boldsymbol{V}$ when prior information is available; see the Chapter of A. Daneshkhah in this book for further discussion.

Zellner's g-prior

As already mentioned, the specification of $\boldsymbol{V}$ is not an easy task. Therefore, a special case of (1.7) is the popular Zellner's (1986) g-prior given by

$$\boldsymbol{\beta}|\sigma^2, \boldsymbol{X} \sim N_{\mathrm{P}}(\boldsymbol{\mu}_\beta, g(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2) \;\text{ and }\; f(\sigma^2|\boldsymbol{X}) \propto 1/\sigma^2\ . \tag{1.8}$$

This results from (1.7) if we set $\boldsymbol{V} = g(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $a \to 0, b \to 0$. Note that now $\boldsymbol{V}$ has the same structure as the variance-covariance matrix of the ML estimator for $\boldsymbol{\beta}$, see eq. (1.6). Although the limiting case of $a \to 0, b \to 0$ actually refers to the original setup of Zellner, reference to this prior is also given when the inverse gamma prior for σ^2 (with non-zero parameters) is used. In the case (1.8) the prior for σ^2 is improper and provides no information on the error variance; see also Jeffreys prior which follows.

The parameter g determines the amount of prior information relative to the empirical data. The information introduced by the prior can be measured by the ratio n/g and can be considered in terms of a effective sample size of the prior. Hence for $g = n$, the prior information will be equivalent to adding one observation in our analysis (i.e. $1/(n+1)$ of the posterior distribution will be due to the prior) while for $g = 1$, the prior information will be equivalent to adding n observations in our analysis (i.e. 50% of the posterior distribution will be due to the prior). The default choice of $g = n$ is usually adopted when no information is available since it has an interpretation of adding prior information equivalent to one data point (Kass and Wasserman, 1995; Fouskakis et al., 2009). The prior mean of $\boldsymbol{\beta}$ is usually set equal to zero in order to shrink values towards to zero especially the ones that are not important for the model.

This prior has been widely used in practice because it considerably simplifies posterior computations. Generally, it allows us for a sensible default prior choice reducing the number of unspecified prior covariance hyperparameters to one (i.e. only the specification of g); see Fernandez et al. (2000) for comparison between different values of g and Liang et al. (2008) for discussion and extensions concerning the g-priors. Finally, another reason for its popularity it its connection (for $g = n$) to the Bayesian Information Criterion (BIC) as is described later in this chapter.

Independent Coefficients

Another example for the case where no prior information is available, is by considering independent normal distributions, i.e. $\boldsymbol{V} = g\,\boldsymbol{I}_{\mathrm{P}}$ with g large in order to express prior ignorance (e.g., $g = 100$). Hence we can simply rewrite the prior as

$$\beta_j|\sigma^2, \boldsymbol{X} \;\sim\; N(\mu_{\beta_j}, g\,\sigma^2) \text{ for } j = 0, 1, \ldots, p, \tag{1.9}$$

where μ_{β_j} are the components of the prior mean vector $\boldsymbol{\mu}_\beta$.

1.3.2 The conditional conjugate Normal–Inverse Gamma Prior

Alternatively to the conjugate prior distributions described in the previous sections, we may consider a normal prior for $\boldsymbol{\beta}$ as in (1.7) but independent of σ^2 i.e. $\boldsymbol{\beta} \sim N_{\mathrm{P}}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$ and $\sigma^2 \sim \mathrm{IG}(a, b)$. This prior is not conjugate, and hence it is not analytically tractable. Nevertheless, this prior setup is conditionally conjugate since the posterior conditional distributions $f(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X})$ and $f(\sigma^2|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{X})$ have the same form as the prior distributions (i.e. multivariate normal and inverse gamma, respectively). Thus we can construct a simple but efficient Gibbs sampling algorithm in to obtain samples from the joint posterior distribution and accurately estimate it. The components of $\boldsymbol{\beta}$ can be obtained either in a single step by sampling $\boldsymbol{\beta}$ from a multivariate normal distribution or from sequential univariate steps by sampling each β_j from the corresponding conditional normal distributions.

An even simpler prior setup, which is well known for its connection to ridge regression, is defined by assuming a priori independence on all parameters, i.e.

$$\begin{aligned} f(\boldsymbol{\beta},\tau) &= \prod_{j=0}^{p} f(\beta_j) f(\tau), \\ \beta_j &\sim N(\mu_{\beta_j}, g_j) \text{ for } j = 0,\ldots,p \text{ and} \end{aligned} \tag{1.10}$$

and the usual inverse gamma prior for σ^2.

1.3.3 Non Conjugate Priors

Additional types of priors for $\boldsymbol{\beta}$ have been proposed in the related literature. For example, the Student t distribution or the Cauchy distribution for $\boldsymbol{\beta}$ can be used instead of the normal prior, but obvious differences in terms of posterior inference are seldom observed when no prior information is available.

Cauchy is introduced in hypothesis testing or model comparison literature by Zellner and Siow (1980). Jeffreys (1961) originally proposed Cauchy as a better alternative to normal priors since this prior satisfies certain consistency issues concerning model comparison and hypothesis testing. This prior setup has never become popular due to the analytical computation of Zellner's g-prior. Another interesting feature of the Cauchy distribution is that it can be written as a scale mixture of normal distributions with g following an inverse gamma distribution with parameters $1/2$ and $n/2$. This gave the motivation for Liang et al. (2008) to develop mixtures of g priors which are useful in variable selection problems.

Lately, the double exponential (DE) prior is discussed in the literature since it is directly connected with the LASSO method (Tibshirani, 1996). Hence we may use

$$\beta_j \sim DE\left(0, \frac{1}{\lambda}\right), \text{ for } j = 1,\ldots,p\,, \tag{1.11}$$

with $\lambda > 0$ and density $f(\beta_j|\lambda) = \frac{\lambda}{2}\exp(-\lambda|\beta_j|)$. The *shrinkage parameter* λ controls the prior variance which is equal to $2/\lambda^2$. The level of shrinkage towards zero of the posterior distribution of β_j is specified via λ since the prior distribution becomes more and more informative as λ increases.

Generally, the posterior distribution is not analytically available under this prior and MCMC must be used. Introducing the regression error variance σ^2 in the variance of the double exponential prior results in analytically tractable results which are discussed in Hans (2009). The use of different shrinkage parameters λ_j's and extending the hyrerchical structure of the model using hyperpriors of the shinkage parameters is currently under investigation by Lykou et al. (2010).

1.3.4 Jeffreys' Prior

Jeffreys' (1961) rule for obtaining flat non-informative prior distribution is well known in Bayesian inference, see Chapter 1 of Robert and Rousseau in this book. The following prior distributions are frequently used

$$f(\boldsymbol{\beta},\sigma^2) \propto 1/\sigma^2 \text{ or } f(\boldsymbol{\beta},\sigma) \propto 1/\sigma \text{ or } f(\boldsymbol{\beta},\tau) \propto 1/\tau$$

derived by independent a-priori treatment of the coefficients $\boldsymbol{\beta}$ (directly associated with the mean of the response data) and dispersion parameters σ^2 (or σ or τ) (see Bernardo and Smith, 1994, p.361). Jeffreys' rule is implemented here separately for $\boldsymbol{\beta}$ and σ^2 and the above prior is obtained by multiplying the two independent priors (see also in p. 328–330 of Bernardo and Smith, 1994).

1.4 General posterior inference for the normal linear model

Parameter Estimation

Inference for the model parameters $\boldsymbol{\beta}$ and σ^2 is based on the joint posterior distribution $f(\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\boldsymbol{X})$ or on the corresponding marginals $f(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X})$ and $f(\sigma^2|\boldsymbol{y},\boldsymbol{X})$. Point estimates can be obtained by measures of central tendency such as the posterior mean, median or mode. Although the posterior mean is frequently used as a point estimate, this might not be the optimal choice when the posterior distribution is skewed and in such cases, the posterior median is to be preferred. In normal regression models, the posterior distribution $f(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X})$ is usually symmetric (unless the prior is highly skewed and informative) due to the shape of the likelihood. On the other hand, due to difficulties in the analytical calculation of the posterior median, the posterior mean of σ^2 is often also reported as point estimate although the error variance σ^2 is skewed (for example the distribution is an inverse gamma in the conjugate case).

For conjugate prior (1.7), the posterior distribution is of a known form and the descriptive measures are analytically available. For the non-conjugate cases, Markov chain Monte Carlo (MCMC) methodologies are used to obtain a sample

$$\left(\boldsymbol{\beta}^{(1)},\sigma^{2(1)}\right), \left(\boldsymbol{\beta}^{(2)},\sigma^{2(2)}\right), \ \ldots\ , \left(\boldsymbol{\beta}^{(T)},\sigma^{2(T)}\right) \tag{1.12}$$

from the posterior distribution of model parameters $(\boldsymbol{\beta},\sigma^2)$. Therefore from an MCMC sample we can estimate the posterior measures of interest; for more details see in Ntzoufras (2009, Chap. 2).

Interval estimates are obtained by the so called credible intervals which can be considered as the Bayesian analogous to the confidence intervals. Credible intervals have a direct probability interpretation since the posterior probability that the parameter of interest lies within this interval is equal to $1-a$. When the posterior distribution is estimated using MCMC, then the appropriate credible intervals are estimated by the sample $a/2$ and $1-a/2$ quantiles calculated by the posterior sample (1.12). More information on credible intervals can be found in Chapter 1 of Robert and Rousseau in this volume.

Evaluation of Important Covariate Effects

As we have already mentioned in the previous section, tracing "good" covariates is an important issue especially when the number of potential covariates is large. In an initial rough analysis, we can examine the importance of each covariate by looking at the posterior distribution of β_j. Posterior distributions far away from the zero indicate an important contribution of X_j on the prediction of the response variable. Within this approach, we can calculate the following posterior tail-area probabilities:

$$\pi_j^0 = \min\Big\{P(\beta_j < 0|\boldsymbol{y}), P(\beta_j > 0|\boldsymbol{y})\Big\}\,. \tag{1.13}$$

When zero lies at the center of the posterior distribution, then π_j^0 will be close to $\frac{1}{2}$ indicating that there is no clear positive or negative effect of X_j on Y. When this probability is low (e.g., lower than 2.5%, 1%, or 0.5%), then we may conclude positive or negative association depending on the sign of the posterior location summaries. This procedure is vulnerable to problems caused by collinear variables causing inflation of the dispersion of the posterior distribution. Therefore important variables might not be traced in some cases. Such problems can be avoided if formal Bayesian variable selection methods are used instead. Nevertheless, such analysis can offer a first tool for tracing important variables. Detailed description of formal model comparison and evaluation based on posterior model odds and probabilities is given later in this Chapter.

Fitted values and predictions

The estimated expected values of Y are considered as the fitted values of the model under consideration. These expected values are given by the linear predictor $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ (see eq. 1.3) and can be described by the posterior distribution $f(\boldsymbol{\mu}|\boldsymbol{y}, \boldsymbol{X})$. The posterior means $E(\boldsymbol{\mu}|\boldsymbol{y}, \boldsymbol{X})$ and the corresponding credible intervals can be used to graphically compare them with the observed values $\boldsymbol{y}$ and by this way also assess the quality of the model fit.

In the Bayesian approach predictions are solely based on the (posterior) predictive distribution $f(\boldsymbol{y}_{\text{new}}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{X}_{\text{new}})$ in contrast to the fitted values $\widehat{y}_i$ used in the classical approach; where $\boldsymbol{y}_{\text{new}}$ are the new (future) response values (of size n_{new}) with explanatory variable values given by the (new) data matrix $\boldsymbol{X}_{\text{new}}$ (of dimension $n_{\text{new}} \times \text{P}$). This predictive distribution is defined as

$$f(\boldsymbol{y}_{\text{new}}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{X}_{\text{new}}) = \int f(\boldsymbol{y}_{\text{new}}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{X}_{\text{new}}) f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X}) d\boldsymbol{\beta} d\sigma^2, \tag{1.14}$$

where $f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X})$ is the joint posterior distribution of $\boldsymbol{\beta}$ and σ^2 while $f(\boldsymbol{y}_{\text{new}}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{X}_{\text{new}})$ is the model likelihood evaluated at the new response values $\boldsymbol{y}_{\text{new}}$ with design matrix $\boldsymbol{X}_{\text{new}}$ and, in regression models, is a multivariate normal density given by (1.3).

In the case that the predictive distribution is not analytically tractable, we can easily generate a random sample $\boldsymbol{y}_{\text{new}}^{(t)}$ from this distribution when a

posterior sample of the model parameters $(\boldsymbol{\beta}^{(t)}, \sigma^{2(t)}; t = 1, \ldots, T)$ is already available. We simply need to

- sample $\boldsymbol{y}_{\text{new}}^{(t)}$ from $N\big(\boldsymbol{X}_{\text{new}}\boldsymbol{\beta}^{(t)}, \boldsymbol{I}_{n_{\text{new}}}\sigma^{2(t)}\big)$

for $t = 1, 2, \ldots, T$.

Residual values

Residual values can be simply obtained by $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$ while standardized residuals by $\boldsymbol{e}^s = \boldsymbol{e}/\sigma$. Posterior samples of the residuals can be calculated using Monte Carlo or MCMC methods by setting $\boldsymbol{e}^{(t)} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}$ for all $t = 1, 2, \ldots, T$. Classical residual plots can be reproduced using posterior means of each residual. Other residual based statistics can be used to trace outliers or check model assumptions; see Spiegelhalter et al. (1996, pp. 40–47) and Ntzoufras (2009, Chapter 10) for more details.

1.5 Posterior Analysis Using Conjugate Priors

Here we adopt the conjugate multivariate normal–inverse gamma prior given in (1.7) and investigate the corresponding posterior distribution. We are mainly presenting results without providing computational details.

1.5.1 The General Normal–Inverse Gamma Prior Setup

The multivariate normal–inverse gamma prior distribution (1.7) is conjugate to the normal regression likelihood and will be denoted as

$$\boldsymbol{\beta}, \sigma^2 \sim \text{NIG}(\boldsymbol{\mu}_\beta, \boldsymbol{V}, a, b).$$

The resulting posterior is also a normal–inverse gamma,

$$\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X} \sim \text{NIG}(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{a}, \widetilde{b}),$$

or, more explicitly, $\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X} \sim N_{\text{P}}(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\Sigma}}\sigma^2)$ and $\sigma^2|\boldsymbol{y}, \boldsymbol{X} \sim \text{IG}(\widetilde{a}, \widetilde{b})$ with parameters

$$\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\Sigma}}\left(\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{V}^{-1}\boldsymbol{\mu}_\beta\right), \quad \widetilde{\boldsymbol{\Sigma}} = \left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{V}^{-1}\right)^{-1}, \tag{1.15}$$

$$\widetilde{a} = \frac{n}{2} + a \;\text{ and }\; \widetilde{b} = \frac{\text{SS}}{2} + b \tag{1.16}$$

$$\text{with } \text{SS} = \boldsymbol{y}^T\boldsymbol{y} - \widetilde{\boldsymbol{\beta}}^T\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{\beta}} + \boldsymbol{\mu}_\beta^T\boldsymbol{V}^{-1}\boldsymbol{\mu}_\beta.$$

The posterior mean $\widetilde{\boldsymbol{\beta}}$ can be expressed as the weighted average of the prior mean $\boldsymbol{\mu}_\beta$ and the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ given in (1.5)

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{W}\widehat{\boldsymbol{\beta}} + (\boldsymbol{I}_{\text{P}} - \boldsymbol{W})\,\boldsymbol{\mu}_\beta \;\text{ with }\; \boldsymbol{W} = \left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{V}^{-1}\right)^{-1}\boldsymbol{X}^T\boldsymbol{X}.$$

Using more algebra we can obtain the elegant expression of Atkinson (1978) for the posterior sum of squares given by

$$\text{SS} = \text{RSS} + \big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\mu}_\beta\big)^T\big[\,\big(\boldsymbol{X}^T\boldsymbol{X}\big)^{-1} + \boldsymbol{V}\big]^{-1}\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\mu}_\beta\big),$$

where

$$\text{RSS} = \big(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\big)^T\big(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\big) = \boldsymbol{y}^T\boldsymbol{y}-\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{y}^T\Big(\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\Big)\boldsymbol{y}$$

is the residual sum of squares in classical regression analysis. This expression is useful because it gives insight concerning the meaning of SS, which is equal to the traditional sum of squares and a measure of distance between the MLEs and the prior mean.

By integrating out σ^2, we obtain the marginal posterior distribution of $\boldsymbol{\beta}$ as a multivariate Student t distribution with parameters $\widetilde{\boldsymbol{\beta}}$, $\widetilde{\boldsymbol{\Sigma}}(\text{SS}+2b)/(n+2a)$ and $n+2a$ degrees of freedom. The density of a d–dimensional Student t distribution $\text{MSt}_d(\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)$ is in general defined for a variable y as

$$f_{St_d}(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\,\pi)^{d/2}}\det(\boldsymbol{\Sigma})^{-1/2}\left[1+\frac{1}{\nu}(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right]^{-(\nu+d)/2} \tag{1.17}$$

with mean $E(\boldsymbol{Y}) = \boldsymbol{\mu}$ (for $\nu > 1$) and variance $V(\boldsymbol{Y}) = \boldsymbol{\Sigma}\frac{\nu}{\nu-2}$ (for $\nu > 2$). The marginal posterior distribution of σ^2 is simply an inverse gamma distribution with the parameters $\widetilde{a}$ and $\widetilde{b}$ given by (1.16). Hence, summarizing, we have that

$$\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X} \sim \text{MSt}_{\text{P}}\Big(\widetilde{\boldsymbol{\beta}},\ \frac{\text{SS}+2b}{n+2a}\widetilde{\boldsymbol{\Sigma}},\ n+2a\Big) \ \text{ and } \ \sigma^2|\boldsymbol{y},\boldsymbol{X} \sim IG\big(\widetilde{a},\ \widetilde{b}\big)\ .$$

We often focus on the marginal posterior distributions $f(\beta_j|\boldsymbol{y},\boldsymbol{X})$, $j = 0,...,p$, which is univariate Student t distribution with

$$\beta_j|\boldsymbol{y},\boldsymbol{X} \sim \text{MSt}_1\Big(\widetilde{\beta}_j,\ \frac{\text{SS}+2b}{n+2a}\widetilde{\Sigma}_{jj},\ n+2a\Big)$$

following Nadarajah and Dey (2005, eq. 17, p.156) with β_j expressed as $\beta_j = \boldsymbol{\Delta}_j\boldsymbol{\beta}$ and $\boldsymbol{\Delta}_j$ is a indicator vector with j element equal to one and all other values equal to zero. Note that $Y \sim \text{MSt}_1(\mu,\sigma^2,\ \nu)$ is a non central scaled Student t distribution which can be written as a linear function, $Y = \mu + \sigma\mathsf{T}$, of a standard Student t random variable T with ν degrees of freedom.

From the above, we may report the posterior mean, standard deviation and 95% credible intervals (usually obtained by 2.5% and 97.5% posterior quantiles for simplicity and convenience) for each coefficient β_j, the error variance σ^2 and the corresponding standard deviation σ. A summary of these quantities is provided in Table 1.1. Note that the relation $\sigma = \sqrt{\sigma^2}$ can be used to calculate posterior quantiles of σ from the corresponding quantiles of σ^2 but it cannot be used to directly derive the posterior mean or variance of σ and hence the corresponding integrals must be calculated analytically.

Predictions are based on the predictive distribution $f(\boldsymbol{y}_{\text{new}}|\boldsymbol{y},\boldsymbol{X},\boldsymbol{X}_{\text{new}})$ as defined in (1.14); where $\boldsymbol{y}_{\text{new}}$ are the future (to be predicted) data under the

Table 1.1: Equations for typical posterior summaries in the conjugate case

Model Parameter	Posterior Summaries: Mean	Variance	q-th Quantile
β_j	$\widetilde{\beta}_j$	$\frac{\widetilde{b}}{\widetilde{a}-1}\widetilde{\Sigma}_{jj}$ *	$\widetilde{\beta}_j + t_{2\widetilde{a},q}\sqrt{\frac{\widetilde{b}}{\widetilde{a}}\widetilde{\Sigma}_{jj}}$ **
σ^2	$\frac{\widetilde{b}}{\widetilde{a}-1}$ †	$\Big(\frac{\widetilde{b}}{\widetilde{a}-1}\Big)^2\frac{1}{\widetilde{a}-2}$ ‡	$1/\Gamma_{\widetilde{a},\widetilde{b};1-q}$
σ	$\widetilde{b}^{1/2}\frac{\Gamma(\widetilde{a}-1/2)}{\Gamma(\widetilde{a})}$ ††	$\frac{\widetilde{b}}{\widetilde{a}-1} - \widetilde{b}\Big(\frac{\Gamma(\widetilde{a}-1/2)}{\Gamma(\widetilde{a})}\Big)^2$ †	$1/\sqrt{\Gamma_{\widetilde{a},\widetilde{b};1-q}}$

$\widetilde{\beta}$ and $\widetilde{\Sigma}$ are given by (1.15); $\widetilde{a}$ and $\widetilde{b}$ are given by (1.16)
* $\widetilde{\Sigma}_{jk}$ is the jth row and kth column element of matrix $\widetilde{\boldsymbol{\Sigma}}$ given by (1.15)
** $t_{\nu;q}$: q quantile of the Student t distribution with ν degrees of freedom
† for $\widetilde{a} > 1 \Leftrightarrow n > 2-2a$; ‡ for $\widetilde{a} > 2 \Leftrightarrow n > 4-2a$; †† for $\widetilde{a} > 1/2 \Leftrightarrow n > 1-2a$

design matrix $\boldsymbol{X}_{\text{new}}$. The resulting predictive distribution is a multivariate Student t distribution with parameters

$$\boldsymbol{Y}_{\text{new}}|\boldsymbol{y},\boldsymbol{X},\boldsymbol{X}_{\text{new}} \sim \text{MSt}_{n_{\text{new}}}\Big(\boldsymbol{X}_{\text{new}}\widetilde{\boldsymbol{\beta}},\ \frac{\text{SS}+2b}{n+2a}\big(\boldsymbol{I}_{n_{\text{new}}} + \boldsymbol{X}_{\text{new}}\widetilde{\boldsymbol{\Sigma}}\boldsymbol{X}_{\text{new}}^T\big),\ 2\widetilde{a}\Big)$$

where n_{new} is the size of the future data $\boldsymbol{y}_{\text{new}}$. For the case $a, b \to 0$, the posterior parameters of the inverse gamma now simplify to $\widetilde{a} = n/2$ and $\widetilde{b} = \text{SS}/2$.

1.5.2 Zellner's g-Prior Setup

For the Zellner's g-prior, we substitute $\boldsymbol{V} = g(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ resulting in

$$\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\boldsymbol{X} \sim \text{NIG}\Big(\,w\widehat{\boldsymbol{\beta}} + (1-w)\boldsymbol{\mu}_\beta,\ w(\boldsymbol{X}^T\boldsymbol{X})^{-1},\ \widetilde{a},\widetilde{b}\Big),$$

with weight $w = g/(g+1)$. The parameters $\widetilde{a}$ and $\widetilde{b}$ have similar definitions as in (1.16) but now the posterior sum of squares SS simplifies to

$$\text{SS} = \text{RSS} + \frac{1}{g+1}\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\mu}_\beta\big)^T\boldsymbol{X}^T\boldsymbol{X}\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\mu}_\beta\big).$$

If we additionally set the prior mean equal to zero, this further simplifies to

$$\text{SS} = \boldsymbol{y}^T\Big(\boldsymbol{I} - w\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\Big)\boldsymbol{y}. \tag{1.18}$$

The marginal posterior distributions are then given by

$$\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X} \sim \text{MSt}_{\text{P}}\Big(w\widehat{\boldsymbol{\beta}},\ w(\boldsymbol{X}^T\boldsymbol{X})^{-1}\widetilde{\sigma}^2,\ n\Big) \ \text{ and } \ \sigma^2|\boldsymbol{y},\boldsymbol{X} \sim IG\big(\,n/2,\ \text{SS}/2\,\big),$$

where $\widetilde{\sigma}^2 = \text{SS}/n$ tends to $\widehat{\sigma}^2 = RSS/n$ for large n or g. Finally, the predictive distribution can be expressed as

$$\boldsymbol{Y}_{\text{new}}|\boldsymbol{y},\boldsymbol{X},\boldsymbol{X}_{\text{new}} \sim \text{MSt}_{n_{\text{new}}}\Big(w\boldsymbol{X}_{\text{new}}\widehat{\boldsymbol{\beta}},\ \big(\boldsymbol{I}_{n_{\text{new}}} + w\boldsymbol{X}_{\text{new}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_{\text{new}}^T\big)\,\widetilde{\sigma}^2,\ n\Big)\ .$$

1.5.3 Jeffreys' Prior

The resulting posterior distribution under Jeffreys' prior is also a normal inverse gamma distribution, $\text{NIG}(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{a}, \widetilde{b})$, with parameters obtained by (1.15) and (1.16) if we set $\boldsymbol{V}^{-1} = 0$, $a = -\text{P}/2$ and $b = 0$. Using the above values, the posterior distribution is now given by

$$\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, \boldsymbol{X} \sim \text{NIG}\big(\widehat{\boldsymbol{\beta}}, (\boldsymbol{X}^T\boldsymbol{X})^{-1}, \tfrac{n-\text{P}}{2}, \tfrac{RSS}{2}\big)$$

while the marginals and the predictive distributions will be given by

$$\boldsymbol{\beta} | \boldsymbol{y}, \boldsymbol{X} \sim \text{MSt}_{\text{P}}\big(\widehat{\boldsymbol{\beta}},\ (\boldsymbol{X}^T\boldsymbol{X})^{-1}\widehat{\sigma}^2_U,\ n-\text{P}\big), \quad \sigma^2 | \boldsymbol{y}, \boldsymbol{X} \sim IG\big(\tfrac{n-\text{P}}{2}, \tfrac{RSS}{2}\big)\ .$$

and

$$\boldsymbol{Y}_{\text{new}} | \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{X}_{\text{new}} \sim \text{MSt}_{n_{\text{new}}}\Big(\boldsymbol{X}_{\text{new}}\widehat{\boldsymbol{\beta}},\ \Big(\boldsymbol{I}_{n_{\text{new}}} + \boldsymbol{X}_{\text{new}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T_{\text{new}}\Big)\widehat{\sigma}^2_U,\ n-\text{P}\Big)\ ,$$

where $\widehat{\sigma}^2_U = \frac{RSS}{n-\text{P}}$ is the classical unbiased estimator of σ^2.

1.5.4 An Illustrative Example from Operational Risk

The following data are taken from Mun (2007) [also see in Mun (2006)] and are monthly key risk indicators of a bank for a period of 50 months. The aim here is to assess how the monthly losses (expressed in million dollars) are affected by the following explanatory variables

- X_1: Cycle time and Timeliness of Transactions
- X_2: Transaction Volume
- X_3: Hiring and Training Costs
- X_4: Customer Satisfaction index
- X_5: IT Network Downtime

We start our analysis by presenting results for the linear model using the following three priors

1. Independent conjugate priors (1.9) with $\mu_{\beta_j} = 0$ for all $j = 0, 1, \ldots, p$, $g = n = 50$ and $a = b = 0.01$,
2. Zellner's g-prior with $\boldsymbol{\mu}_\beta = 0$, $g = n = 50$ and $a = b = 0.01$,
3. Jeffreys' prior.

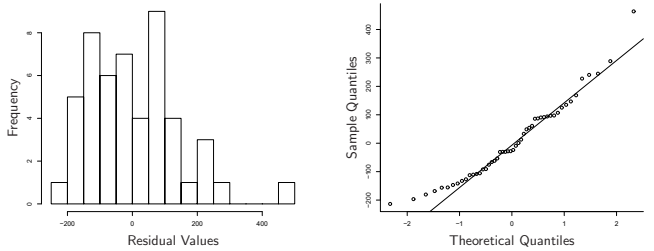
The resulting posterior summaries are presented in Table 1.2. Obviously, the posterior statistics for all three prior setups are similar indicating that they introduce only low prior information to our analysis.

The tail-area probabilities were calculated using equation (1.13). These probabilities are low for the variables "Volume" and "Costs" (X_2 and X_3) indicating nonzero effects. Although this is not a formal Bayesian evaluation

Table 1.2: Posterior summaries [mean ± standard deviation (tail-area probability of zero)] for the prior setups 1–3 using the original variables

	Prior Setup		
Parameter	Zellner's $g = 50$	Independent $g = 50$	Jeffreys'
β_0	56.82 ± 111.42 (0.30)	57.36 ± 105.82 (0.29)	57.96 ± 114.10 (0.30)
$\beta_1 \times 10^3$	-3.47 ± 3.61 (0.16)	-3.54 ± 3.44 (0.14)	-3.54 ± 3.69 (0.16)
$\beta_2 \times 10$	4.55 ± 2.60 (0.04)	4.65 ± 2.48 (0.03)	4.64 ± 2.66 (0.04) *
β_3	24.74 ± 14.46 (0.04)	25.22 ± 13.80 (0.03)	25.24 ± 14.81 (0.04) *
$\beta_4 \times 10^3$	-8.40 ± 104.02 (0.47)	-8.71 ± 99.28 (0.46)	-8.56 ± 106.52 (0.47)
β_5	16.23 ± 15.16 (0.13)	16.62 ± 14.42 (0.12)	16.56 ± 15.52 (0.13)
σ^2	22971.38 ± 4788.82	20528.11 ± 4279.48	23468.46 ± 5247.71
σ	150.78 ± 15.42	142.53 ± 14.58	152.29 ± 16.66

∗ Indicates coefficients with tail-area probability between 2.5% and 5%.



(a) Histogram of Residual Values (b) Quantile Plot of Residual Values

Figure 1.1: Plots for residual values evaluated at the posterior mean of $\boldsymbol{\beta}$ where it is evident that the assumption of normality is violated.

of the importance of each covariate, this is a first indication that these two covariates may be important determinants of "monthly impact losses".

A quick look at the residual values evaluated at the posterior mean (see Figure 1.1) provide indications that the assumption of normality may be violated. This is actually usual in financial data which are frequently highly skewed (as the data of this example). For this reason, we may consider the logarithmic transformations of the original data in order to correct for deviations from the regression assumptions such as normality and homoscedasticity of errors or linearity of effects. In the following we proceed presenting results on the log-transformed data which is a usual practice in financial and operational risk data. In this example, this transformation improved the corresponding picture of the residual values (it is not included here to save space).

Posterior summaries for the three prior setups (which are also considered for the original data) are presented in Table 1.3 while the 95% credible in-

tervals are depicted in Figure 1.2. A first comment is that now the posterior summaries of the three prior setup provide different estimates which is an indication that the choice of $g = 50$ for the Zellner's and the independence prior may be informative for the logarithmic data. More specifically, for Zellner's g-prior the posterior means of β are similar to those of Jeffreys' prior. In contrast, these priors yield different results for the standard deviations. Major differences are observed in the posterior means of the error variance which is four times the corresponding one obtained under the Jeffreys' prior. Moreover, if in Zellner's g-prior we consider the maximum likelihood estimate as the prior mean of the model coefficients, i.e. $\boldsymbol{\mu}_\beta = \widehat{\boldsymbol{\beta}}$, then the posterior results are the same as in Jeffreys' prior (these results are not presented in Table 1.3 to save some space). Although this approach is not fully Bayesian because we use data to specify a hyper parameter (here, the prior mean), it is frequently used in practice as a low information prior since it adds minimal information (equal to one datapoint in the case of $g = n$) to our posterior analysis. On the other hand, for the independent prior we observe differences in the posterior means of model coefficients $\boldsymbol{\beta}$ from the corresponding ones when Jeffreys' prior is used. For a comparison of the 95% credible intervals see Figure 1.2.

Table 1.3: Posterior summaries [mean ± standard deviation (tail-area probability of zero)] for the prior setups 1–3 using the log-transformed data.

	Prior Setup		
	Zellner's $g = 50$	Independent $g = 50$	Jeffreys'
β_0	5.565 ± 5.815 (0.16)	3.201 ± 2.200 (0.07)†	5.676 ± 3.087 (0.03)*
β_1	-0.442 ± 0.535 (0.20)	-0.241 ± 0.215 (0.12)	-0.450 ± 0.284 (0.05)*
β_2	0.543 ± 0.438 (0.10)	0.676 ± 0.198 (0.00)***	0.554 ± 0.232 (0.01)***
β_3	0.711 ± 0.511 (0.08)†	0.529 ± 0.208 (0.01)***	0.725 ± 0.271 (0.00)***
β_4	0.008 ± 0.133 (0.47)	-0.015 ± 0.065 (0.41)	0.008 ± 0.071 (0.45)
β_5	0.189 ± 0.533 (0.36)	0.261 ± 0.263 (0.15)	0.193 ± 0.283 (0.24)
σ^2	0.860 ± 0.179	0.215 ± 0.045	0.236 ± 0.053
σ	0.923 ± 0.094	0.461 ± 0.047	0.483 ± 0.053

Symbols indicate coefficients with tail-area probability between:
***less than 1%; *2.5% and 5%; †5% and 10%.

The fact that the Zellner's g-prior is informative for the choice of $g = n = 50$ is also depicted in Figure 1.3 where we observe that the credible intervals of $\boldsymbol{\beta}$ are stabilized only for $g \geq 5n$. The effect of the prior on the posterior distribution is even more evident for the error variance (and standard deviation) which are stabilized only for $g \geq 30n$.

Based on the posterior means of the Jeffreys' prior (which is considered here as the less informative prior) we may write our model using the following expression

$$\begin{aligned}\log(\text{Impact losses}) = 5.67 &- 0.45\log(\text{Cycle}) + 0.55\log(\text{Volume}) + 0.72\log(\text{Costs}) \\ &+ 0.08\log(\text{Customer Satisfaction}) + 0.193\log(\text{IT}) + \epsilon\end{aligned}$$

with $\epsilon \sim N(0, 0.483^2)$.

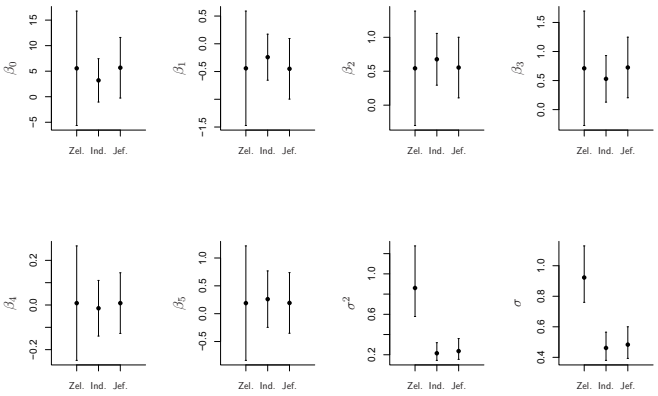


Figure 1.2: Posterior 95% credible intervals for model parameters under prior setups 1–3 using the logarithmic data.

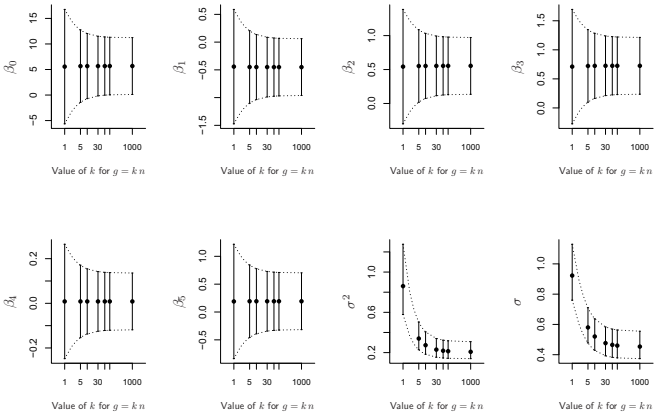


Figure 1.3: Posterior 95% Credible Intervals for Model Parameters for Zellner's g-prior with Different k with $g = k\,n$ for the Logarithmic Data.

The interpretation of the model coefficients here is more complicated since the effects are actually multiplicative. Detailed description of the interpretation of model coefficients is available in the electronic Appendix of this chapter available on the authors webpage.

Finally, from the tail-area probabilities (≤ 0.01) in Jeffreys' prior setup, we can see that the log-transformed variables "Transaction Volume" and "Costs" (X_2 and X_3) are possibly significant covariates. A similar conclusion can be drawn for the constant β_0 and the coefficient of the log-tranformed "Cycle Time" (X_1) with tail-area probabilities within $2.5\% - 5\%$. Finally, the posterior distributions of the coefficients for "Customer Satisfaction Index" and "IT Network Downtime" variables do not seem to differ from zero indicating that may be irelevant for the model. Note that the corresponding results for the Zellner's g-prior are different here since (as we have seen) this prior is informative and the posterior standard errors are much wider than in the case of Jeffreys' prior which is considered as noninformative. More formal approach of Bayesian variable selection and comparison is presented later in this Chapter.

1.6 Posterior Analysis Using MCMC

If a non-conjugate prior is selected, then the posterior distribution may not be tractable. In such cases, MCMC methods can be used to obtain a sample $(\boldsymbol{\beta}^{(t)}, \sigma^{2(t)})$ of the model parameters $(\boldsymbol{\beta}, \sigma^2)$ and estimate the posterior distribution using this sample. If the conditional conjugate prior is selected, then a Gibbs sampler can be used to obtain the posterior sample by generating values from the conditional posterior distributions which are also normal and inverse gamma for parameters $\boldsymbol{\beta}$ and σ^2 respectively. In all other cases, Metropolis-Hastings algorithm or Metropolis-within-Gibbs sampler can be used instead. The proposal distributions can be built from the posterior distributions obtained in the conjugate case. Extensive details of the use of MCMC methods can be found in the Chapter 1 by and Rousseau in this volume and in Gilks et al. (1996). Moreover, WinBUGS software (Spiegelhalter et al., 2003) can be used to generate samples (using MCMC) from the posterior distribution of normal (or even more complicated) models; for details and extensive WiNBUGS illustations of normal models, see in Ntzoufras (2009, Chap. 5 & 6).

1.7 Bayesian Variable Selection for the Normal Model

1.7.1 A Short Introduction to Bayesian Variable Selection

Posterior model probabilities and weights

In Bayesian paradigm, comparison and selection between a set of models $m \in \mathcal{M}$ is performed via the *posterior model probabilities* or *weights* $f(m|\boldsymbol{y})$ given by

$$f(m|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|m)f(m)}{\sum_{m' \in \mathcal{M}} f(\boldsymbol{y}|m')f(m')} \tag{1.19}$$

where $f(m)$ is the prior model probability (or weight), which has to be determined by the experimenter or risk manager, and $f(\boldsymbol{y}|m)$ is the *marginal likelihood* of model m given by

$$f(\boldsymbol{y}|m) = \int\int f(\boldsymbol{y}|\boldsymbol{\beta}_m, \sigma_m^2, m) f(\boldsymbol{\beta}_m, \sigma_m^2|m) d\boldsymbol{\beta}_m d\sigma^2 \tag{1.20}$$

in the normal linear regression model, i.e. it is the integral of the likelihood of model m over the prior distribution of this model. Hence, the marginal model likelihood does not depend on the model parameters, implying that parameter uncertainty for each model is accounted for. The subscript m in the model parameters denotes the corresponding model in which they are defined. To simplify things, we may assume the same error variance σ^2 (attached with the same prior) for all models under consideration.

Integral (1.20) is analytically tractable when the normal–inverse gamma prior (1.7) is adopted. For other priors, numerical or MCMC methods must be used instead.

Bayes Factor and Posterior model odds

Posterior model weights are sensitive to the number of models to be compared. For this reason, relative posterior model probabilities (called *posterior model odds*) are used for pairwise comparisons. If we wish to compare more than two models, then we may select a reference model $m_0 \in \mathcal{M}$ and calculate

$$PO_{k0} = \frac{f(m_k|\boldsymbol{y})}{f(m_0|\boldsymbol{y})} = \frac{f(\boldsymbol{y}|m_k)}{f(\boldsymbol{y}|m_0)} \times \frac{f(m_k)}{f(m_0)} = B_{k0} \times \frac{f(m_k)}{f(m_0)} \quad \text{for all} \quad m_k \in \mathcal{M}, \tag{1.21}$$

which is said to be the *posterior model odds* of model m_k versus model m_0. B_{k0} is the Bayes factor of model m_k versus model m_0 defined as the ratio of the "marginal" likelihoods $f(\boldsymbol{y}|m_k)$ and $f(\boldsymbol{y}|m_0)$. The Bayes factor can be expressed as the posterior odds divided by the prior odds of two compared models. Therefore, it quantifies the prior to posterior change of relative evidence for the two compared models. Note that the posterior probabilities can be directly obtained from posterior model odds using the expression

$$f(m_k|\boldsymbol{y}) = \frac{PO_{k0}}{\sum\limits_{m_\ell \in \mathcal{M}} PO_{\ell 0}} = \frac{1}{\sum\limits_{m_\ell \in \mathcal{M}} PO_{\ell k}} \, .$$

Bayes factors are of utmost importance within Bayesian model comparison and hypothesis testing and play a role equivalent to likelihood ratios in classical statistics. Moreover, if equal prior model probabilities are considered (which sometimes is the default choice when no information is available), posterior model odds become equal to Bayes factors. For this reason, frequently, Bayesian model comparison is solely based on Bayes factors.

Posterior model odds PO_{10} (and Bayes factors B_{10}) allow for a straightforward Bayesian model comparison and testing according to Jeffreys' interpretation. For $PO_{10} > 1$ (or $B_{10} > 1$) we have evidence in favor of model m_1 which

is can be denoted as: *negligible* ("not worth than a bare mention") if they lie within the interval $1-3$, *positive* if they lie within the interval $3-20$, *strong* if they take values between 20 and 150 and *very strong* for values greater than 150. For $PO_{10} < 1$ (or $B_{10} < 1$) we have evidence in favor of model m_0 using similar interpretation for $PO_{01} = 1/PO_{10}$ (or $B_{01} = 1/B_{10}$); for more details, see Kass and Raftery (1995).

Before closing this short subsection, we should mention that an important problem of Bayesian model comparison: marginal likelihoods and the resulting Bayes factors are sensitive on the dispersion parameters of the priors $f(\boldsymbol{\beta}_m|m)$. This problem is widely known as the *Lindley–Bartlett* or *Jeffreys paradox* (Lindley, 1957; Bartlett, 1957). Hence the specification of the prior parameters in variable selection problems becomes a very important issue which is partially confronted using Zellner's g-prior setup.

Posterior Probability of Variable Inclusion

In variable selection literature, the model indicator m is usually substituted by a vector of binary indicators $\boldsymbol{\gamma}$ of size P. Each γ_j corresponds to β_j with $\gamma_j = 1$ if X_j is included in the model (i.e. $\beta_j \neq 0$) and $\gamma_j = 0$ otherwise. We usually include the constant term in all models, hence $\gamma_0 = 1$ with prior probability equal to one.

As the size of the model space is given by $|\mathcal{M}| = 2^p$, even a moderate number of potential covariates results in a large number of models from which the best one has to be selected. For example for $p = 20$ covariates more than one million models have to be considered. In such cases, all posterior model weights will be low even if a small group of models is much better than the remaining ones. Alternatively, we may select a model based on the posterior inclusion probabilities

$$f(\gamma_j = 1|\boldsymbol{y}) = \sum_{\boldsymbol{\gamma}_{\setminus j} \in \{0,1\}^{p-1}} f(\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j}|\boldsymbol{y}). \tag{1.22}$$

In practice, this probability is the sum of posterior probabilities for all models which include covariate X_j in their linear predictor.

Selection of Models and Covariates

Selection of a single model. If we wish to select a single "best" model then we choose the one with the maximum posterior probability $f(m|\boldsymbol{y})$ or identically $f(\boldsymbol{\gamma}|\boldsymbol{y})$. This model is called the *maximum a posteriori (MAP) model*. Alternatively, we may use the posterior variable inclusion probabilities, to trace the *median probability (MP) model*, which is defined as the model with all covariates having $f(\gamma_j = 1|\boldsymbol{y}) > 0.5$. The MP model has better predictive performance than the MAP model under certain conditions; for details, see Barbieri and Berger (2004).

Reporting of a set of best models. An advantage of Bayesian model comparison is that we can evaluate posterior probabilities and hence also quantify

the uncertainty concerning the best fitted models. If we wish to report a group of best models, following the suggestions of Kass and Raftery (1995), we may report models m_k that are similar in terms of posterior evidence to the MAP model. Hence we may restrict attention and report models with posterior model odds $PO_{MAP,k} < 3$, i.e. models which have a posterior probability that is a least $1/3$ of the posterior probability of the MAP model.

Bayesian Model Averaging. In certain cases we do not wish to select a specific model but rather want to make inference or predictions by taking into account model uncertainty in our analysis. Hence we may obtain the model averaged posterior density of any quantity of interest ξ by considering the posterior distributions $f(\xi|m, \boldsymbol{y})$ weighted by the corresponding model weights $f(m|\boldsymbol{y})$. The set of models we include in the model averaging procedure may be remarkably reduced by considering only the ones with $PO_{MAP,k} < 3$ or by including models with covariates having posterior inclusion probabilities higher than 0.5.

1.7.2 Bayesian Variable Selection for the Conjugate Normal Model

The Marginal Likelihood in the Conjugate Case

In the case that we use the conjugate prior (1.7), the marginal likelihood (1.20) can be calculated analytically and is given by the density of a multivariate Student t distribution,

$$f(\boldsymbol{y}|m) = f_{St_n}\left(\boldsymbol{y}\,;\, \boldsymbol{X}_m\boldsymbol{\mu}_{\beta_m}, \frac{b_m}{a_m}\Big(\boldsymbol{I}_n + \boldsymbol{X}_m\boldsymbol{V}_m\boldsymbol{X}_m^T\Big),\, 2a_m\right), \tag{1.23}$$

where $f_{St_d}(\boldsymbol{y}\,;\, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ is the density function of a d-dimensional Student t distribution given by (1.17). Note that in the above equation we consider prior distribution (1.7) with a subscript m added to indicate each model. In practice we a-priori assume that $a_m = a$ and $b_m = b$ for all $m \in \mathcal{M}$.

Zellner's g-prior and the Bayes Information Criterion (BIC)

If we use the conjugate prior (1.7) with $a_m = b_m = a$ for all models, prior mean $\boldsymbol{\mu}_{\beta_m} = 0$ and $\boldsymbol{V}_m = g(\boldsymbol{X}_m^T\boldsymbol{X}_m)^{-1}$ for all m, then the marginal likelihood of model m is given by

$$f(\boldsymbol{y}|m) = f_{St_n}\left(\boldsymbol{y}\,;\, \boldsymbol{0},\, \boldsymbol{\Lambda}_m^{-1},\, 2a\right) \text{ with } \boldsymbol{\Lambda}_m = \left[\boldsymbol{I}_n - w\boldsymbol{X}_m\left(\boldsymbol{X}_m^T\boldsymbol{X}_m\right)^{-1}\boldsymbol{X}_m^T\right]. \tag{1.24}$$

Adopting uniform prior on the model space, i.e. $f(m) = 1/|\mathcal{M}|$ for all models (meaning that all models are a priory equally likely), then $f(m|\boldsymbol{y}) \propto f(\boldsymbol{y}|m)$. Considering the logarithm of the posterior model probabilities we obtain from (1.24)

$$\log f(m|\boldsymbol{y}) = C + \frac{1}{2}\log|\boldsymbol{\Lambda}_m| - \left(a + \frac{n}{2}\right)\log\left[2a + \boldsymbol{y}^T\boldsymbol{\Lambda}_m\boldsymbol{y}\right],$$

with the constant being equal to $C = \log\Gamma\big(a + n/2\big) - \log\Gamma(a) - \frac{n}{2}\log\pi + a\log(2a) - \log|\mathcal{M}|$.

The term $\boldsymbol{y}^T\boldsymbol{\Lambda}_m\boldsymbol{y}$ is equal to the posterior sum of squares SS_m of model m as given by (1.18). Moreover, following Sylvester's determinant theorem (Kollo and von Rosen, 2005, pp. 7–8, Prop. 1.1.1) we further obtain

$$-2\log f(m|\boldsymbol{y}) = \text{constant} + \big(2a+n\big)\log\big(2a + SS_m\big) + \mathrm{P}_m\log(g+1).$$

Further considering the usual improper prior for σ^2 assuming $a \to 0$, simplifies the above expression to

$$-2\log f(m|\boldsymbol{y}) = \text{constant} + n\log SS_m + \mathrm{P}_m\log(g+1).$$

Note that these posterior probabilities can be calculated without any problem since the constant terms (in which a is involved) cancel out in (1.19). Setting as usual, $g = n$, and considering that for a large sample size n $SS_m \approx RSS_m$ we may write

$$-2\log f(m|\boldsymbol{y}) = \text{constant} + n\log RSS_m + \mathrm{P}_m\log(n+1), \tag{1.25}$$

which is equivalent to the Bayes information criterion (BIC) defined as:

$$\mathrm{BIC}(m) = -2\log f(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}_m, \widehat{\sigma}^2, m) + \mathrm{P}_m\log n = K + n\log RSS_m + \mathrm{P}_m\log n, \tag{1.26}$$

where K is a constant term which is common for all models. BIC penalizes the twice maximized log-likelihood by a penalty term which is equal to $\log n$ multiplied by number of estimated parameters which stands for the model complexity, i.e. each estimated parameter adds to BIC a penalty term equal to $\log n$. In the expression of the log-posterior probability (1.25) the BIC penalty $\log n$ is naturally substituted by $\log(n+1)$ since information equivalent to one additional data point is inserted via the prior. Hence the Zellner's g-prior with $g = n$ is closely connected to BIC (with a slight modification on the penalty function). BIC is generally considered as a rough approximation of the log-marginal likelihood (and hence we can obtain a rough approximation of the log–Bayes factor) under a wide family of prior distributions; see Kass and Wasserman (1995) and Kass and Raftery (1995) for more details.

1.7.3 The Akaike Information Criterion (AIC) and Posterior Model Odds

The AIC statistic was introduced by Akaike (1973) as an approximation of the expected Kullback–Leibler distance between a true model and an estimated model. The AIC can be obtained from (1.26) by substituting $\log n$ with 2. The penalty induced by AIC is lower than the BIC penalty for a reasonable sample size ($n > 7$), and hence AIC supports less parsimonious models than BIC.

Model weights based on AIC can be obtained for large n by the expression

$$f(m|\boldsymbol{y}) = \frac{\exp\big[-\frac{1}{2}\mathrm{AIC}(m)\big]}{\sum_{m'\in\mathcal{M}}\exp\big[-\frac{1}{2}\mathrm{AIC}(m')\big]}, \tag{1.27}$$

where $AIC(m)$ is the AIC value for model m and is given by

$$\mathrm{AIC}(m) = -2\log f(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}_m, \widehat{\sigma}^2, m) + 2\mathrm{P}_m = K + n\log RSS_m + 2\mathrm{P}_m. \tag{1.28}$$

Model weights (1.27) correspond to posterior model probabilities obtained by the Zellner's g-prior with $g = n$ and prior model weights

$$f(m) \propto \exp\Big(\frac{\mathrm{P}_m}{2}\big\{\log(n+1) - 1\big\}\Big).$$

Similar arguments were used by (Burnham and Anderson, 2004) for the general approximate case.

Since AIC can be obtained as a posterior model probability, interpretation based on the arguments of Kass and Raftery (1995) can be used. In this lines, Burnham and Anderson (2004) suggest that all models with an AIC difference less than 2 to the best one should be reported as equally "good" and having substantial support (evidence) against the remaining ones.

1.7.4 Example (continued).

In this section we present results for Bayesian model selection according to the approaches presented for the operational risk example discussed previously in this Chapter. We will use the log-transformed data in combination with the conjugate prior setup. The five covariates ($p = 5$) result in $2^5 = 32$ candidate models. Hence using (1.23) we can evaluate analytically the marginal likelihoods for all models and the corresponding posterior model probabilities. Here we present the following prior setups

Prior 1: Zellner's g-prior (1.8) with $\boldsymbol{\mu}_\beta = \mathbf{0}$ and $g = n$,

Prior 2: Zellner's g-prior (1.8) with $\boldsymbol{\mu}_\beta = \widehat{\boldsymbol{\beta}}$ and $g = n$,

Prior 3: Independent normal prior setup (1.10) with prior means equal to the maximum likelihood estimates and variances equal to the corresponding standard errors multiplied by n (Empirical Independent Normal).

The first prior is not using any information from the response data $\boldsymbol{y}$ (i.e. the stochastic part of the model), and for this reason can be considered as a pure Bayesian approach. Although this choice is the most common one in Bayesian variable selection, as we have already seen earlier in this Chapter, for the data at hand, this prior is informative in terms of estimation of model parameters. For this reason, we also illustrate prior setups 2 and 3 which use information from the response data $\boldsymbol{y}$ through the MLE estimates and their standard errors. Although these choices can be thought as empirical Bayes approaches (since they use empirical information from data to specify the prior), this empirical information is reduced to a minimum because it only corresponds to one single data point. Finally, results based on BIC and AIC are additionally provided in the same Table. The corresponding posterior weights (1.27) are calculated for AIC while for BIC we use the same expression substituting AIC with BIC calculated by (1.26)

Table 1.4 provides the highest a-posteriori models for the prior setups described above. For the first prior setup, the four highest a-posteriori models are equivalent in terms of posterior evidence (i.e. with posterior odds, when compared to the MAP model, is lower than 3). The best model is the one including only the third variable (costs) while the second and the third models also include covariates X_2 or X_1 (cycle time and transaction volume) additionally to X_3. Finally, the fourth model includes only variable X_2. The support of four models with different covariates may imply that there is large uncertainty about which covariates must be finally added in the model and therefore Bayesian model averaging may be preferred. For prior setups 2 and 3, the picture is more clear since the model with all these three variables is the MAP with posterior weight 0.67 and 0.34 respectively. For the independent normal empirical Bayes setup, the model with covariates X_2 and X_3 is also very close to the MAP model. BIC provides similar results to prior setup 3 with the posterior weights of the two best models very close but with inverse sequence of support. Finally, AIC supports less parsimonious models than BIC as expected.

An even more clear picture is given in Table 1.5 with the posterior inclusion probabilities for each variable. We may summarize our finding by the following comments

- The first covariate (cycle time) is fully supported by the 2nd prior setup (posterior probability 0.988) but for the rest of the prior setups there is a lot of uncertainty since the corresponding posterior probabilities range from 0.32 to 0.72.
- The second covariate (transaction volume) has high posterior inclusion probability (> 0.88) for all priors used except the first one (posterior probability ~ 0.4).
- Covariate X_3 (costs) is supported in all prior setups (posterior probability > 0.98 for all setups except for the first one which was considerably lower and equal to 0.61).
- Covariates X_4 and X_5 (customer satisfaction index and IT network time) are not important determinants of "monthly impact losses" in all setups

Generally the first prior setup supports more parsimonious models than the rest of the setups while AIC supports less parsimonious models.

1.7.5 MCMC Based Variable Selection and Model Search Algorithms

When the number of covariates p is large, the size of the model space becomes enormous and therefore it is infeasible (or in some cases simply inefficient) to evaluate the marginal likelihoods for all 2^p models under consideration. In such cases, MCMC methods can be used to trace best models and variables without having to evaluate the whole model space. When the marginal likelihood is given in closed-form as in the case of the normal linear model, one can use

Table 1.4: Posterior model probabilities for equivalently best models (with posterior odds < 3 when compared to the MAP model) for prior setups described in the text using the log-transformed data.

Prior Setup	Model	Model Code (m)	Marginal log-likelihood	Posterior Weights	Posterior Odds (MAP vs. m)
Zellner's g-Prior[1]					
(Zero prior mean)	X_3	5	-77.35	0.225	1.00
	$X_2 + X_3$	7	-77.95	0.123	1.82
	$X_1 + X_3$	6	-78.34	0.083	2.69
	X_2	3	-78.39	0.079	2.83
Zellner's g-Prior[1]					
(MLE prior mean)	$X_1 + X_2 + X_3$	8	-44.44	0.668	1.00
Empirical					
Independent	$X_1 + X_2 + X_3$	8	-44.08	0.342	1.00
Normal[2]	$X_2 + X_3$	7	-44.28	0.280	1.22

Model Selection Criterion	Model	Model Code (m)	Approximate Marginal log-likelihood[3]	Posterior Weights[4]	Posterior Odds (MAP vs. m)
BIC	$X_2 + X_3$	7	-78.96	0.305	1.00
	$X_1 + X_2 + X_3$	8	-78.97	0.303	1.01
AIC	$X_1 + X_2 + X_3$	8	-75.15	0.308	1.00
	$X_1 + X_2 + X_3 + X_5$	24	-75.82	0.157	1.96
	$X_2 + X_3$	7	-76.10	0.119	2.58
	$X_1 + X_2 + X_3 + X_4$	16	-76.10	0.119	2.59

[1] $g = n$

[2] $\beta_j \sim N(\widehat{\beta}_j, n\widehat{\sigma}^2_{\widehat{\beta}_j})$; $\widehat{\beta}_j$ are MLE estimates from the full model and $\widehat{\sigma}_{\beta_j}$ is their standard error

[3] For BIC and AIC, the approximate log-marginal likelihoods are given by the AIC and BIC values multiplied by $-1/2$. BIC and AIC values are calculated from (1.26) and (1.28), respectively.

[4] Posterior weights for AIC are calculated using (1.27). For BIC we use the same expression substituting AIC with the corresponding BIC value.

the MCMC model composition, MC^3 (Madigan and York, 1995). Variants of MC^3 were used in normal linear models by Hoeting et al. (1996), Raftery et al. (1997) and Hoeting et al. (2002). MC^3 is a simple Metropolis algorithm which facilitates us to explore the large model spaces. Let us denote by $j(m, m')$ for all $m, m' \in \mathcal{M}$ the probability of proposing model m' given the current model m. Then, the algorithm can be summarized by:

1. For any current model m, we propose model m' with probability $j(m, m')$.
2. Calculate and store the marginal likelihood $f(m'|\boldsymbol{y})$ of the proposed model m'.

Table 1.5: Posterior variable inclusion probabilities for the prior setups described in the text using the log-transformed data.[1]

	Prior Setup			Model Selection Criteria	
	Zellner's g-Prior[2]		Empirical		
	Prior Mean		Indep.		
Variable (in log scale)	Zero	MLE	Normal[3]	BIC	AIC
X_1: Cycle time	0.317	0.988	0.579	0.539	0.718
X_2: Transaction Volume	0.398	0.988	0.885	0.884	0.937
X_3: Costs	0.608	0.996	0.980	0.970	0.981
X_4: Customer Satisfaction Index	0.145	0.133	0.137	0.141	0.290
X_5: IT Network Downtime	0.150	0.215	0.170	0.175	0.348

[1] Posterior inclusion weights are calculated using (1.22) and corresponding posterior weights (the highest posterior weights are presented in Table 1.4).
[2] $g = n$
[3] $\beta_j \sim N(\widehat{\beta}_j, n\widehat{\sigma}^2_{\beta_j})$; $\widehat{\beta}_j$ are MLE estimates from the full model and $\widehat{\sigma}_{\beta_j}$ is their standard error.

3. Accept the proposed model with probability

$$\alpha = \min\left(1, \frac{f(m'|\boldsymbol{y})}{f(m|\boldsymbol{y})} \times \frac{j(m', m)}{j(m, m')}\right).$$

4. Store the current value of m as the currently visited model.

5. Repeat steps 1–4 until a sufficient number of models is visited.

Posterior model weights can be estimated by considering the marginal likelihoods of the visited and proposed models stored in step 2. Alternatively, the relative frequencies of visited models obtained by the MCMC output provide accurate estimates of the posterior model weights. Gibbs versions of this algorithm can be also adopted using $\boldsymbol{\gamma}$ instead of m and updating each γ_j iteratively.

In the non-conjugate cases, MC^3 can be still used by substituting the marginal likelihoods with their Laplace or BIC based approximations. Further methods have been developed over the last 20 years (for the normal model originally but they are also used for more complicated models) including the stochastic search variable selection of George and McCulloch (1993), the Kuo and Mallick (1998) sampler, and the Gibbs variable selection of Dellaportas et al. (2002). Here we should also add the reversible jump MCMC (Green, 1995), which is a general model comparison algorithm, but has been widely used for variable selection.

MC^3 can be directly implemented in the statistical computing software `R` using the `BMA` package of Raftery et al. (2009). A more recent package which incorporates latest advances in the Bayesian variable selection research is the `BAS` package developed by Clyde et al. (2009), which implements Bayesian model averaging for linear models using stochastic or deterministic sampling without replacement from the posterior distributions. The user can choose among the simple Zellner's g-prior and mixtures of g-priors including the Cauchy prior of

Zellner and Siow (1980), and the prior of Liang et al. (2008). It further allows to select between the usual uniform and the beta-binomial prior for models under consideration.

Variable selection can be also implemented in WinBUGS. Details can be found in Dellaportas et al. (2000), Ntzoufras (2002) and Ntzoufras (2009, Chap. 11). The WinBUGS `jump` interface, recently developed by Dave Lunn, can be also used for implementing variable selection for normal and generalized linear models; for more details, see Lunn et al. (2005, 2006) and the manual for this interface.

1.8 Discussion: Extending the Normal Regression Model

The normal linear model can easily accommodate categorical covariates in its setup by suitably defining the data matrix $\boldsymbol{X}$ using dummy variables. Generally model parameters, especially when mixed types of covariates and interactions are used, must be carefully interpreted since underlying relationships may change for different levels (or combinations) of the categorical covariates. For a detailed description of the topic, see Chapter 6 of Ntzoufras (2009).

Econometric and financial data are highly skewed resulting in violation of the "normality" of errors assumption which is essential for the normal regression model. Violations of this assumption can be handled by (a) transforming the response variable or (b) changing the assumed distribution of the response variable or (c) changing the error distribution. The first approach is the simplest one and was also illustrated in the operational risk example presented in this Chapter. Usually transforming the response variable resolves many of the violations of the normal model. Although the normal linear model holds for the transformed data, neither the assumed distribution for the original response is normal nor the association between the covariates and the response is linear. Comparison between models with transformed responses is not straightforward and the interpretation usually becomes cumbersome. For example, when the logarithmic transformation is used, the distribution of the response is now log-normal while the association between the expected values of the response and the covariates (and the error term) is not linear but multiplicative. In the second approach, the distribution of the response is changed which will also lead to a modified error distribution. This action does not ensures the linear relationship between the response variable and the error terms. Finally, changing the distribution of the error term does not affects the linearity and may lead to more robust models if distributions with fatter tails (such as the Student's t distribution) or asymmetric distributions are adopted.

Another obvious extension of the normal model is achieved via the generalized linear models (GLM) setup introduced by McCullagh and Nelder (1989). Interpretation of model parameters is similar but it is not solely based on expectations but in suitable transformations of the location parameter of each distribution used to describe the response variable. In the Bayesian analysis of GLMs, we directly work using MCMC methods since posterior distribu-

tions are rarely conjugate. Further details concerning the Bayesian analysis of GLMs can found in Dey et al. (2000) and Ntzoufras (2009, Chapters 7–8).

Finally, hierarchical and random effect models can be also considered as natural extensions of the normal linear model. In these models random parameters are added in the linear predictor to incorporate additional structural properties in the model. This structure may be used to change the response distribution or add dependence between different values of the response variables (as for example when repeated measures for the same individuals are collected). An excellent description of Bayesian hierarchical models can be found in Gelman and Hill (2006) while a short introduction with simple examples is available in Chapter 9 of Ntzoufras (2009).

1.9 Conclusion

To sum up, the Bayesian paradigm is a strong tool for statistical inference. It can be efficiently used for any type of statistical model. Here we have illustrated its use to the normal linear regression model. We have presented a variety of possible prior setups but we focus on the posterior analysis of the conjugate case. The variable selection problem, which is an important problem in modeling, is also introduced and described in detail. An operational risk example was used to illustrate the presented theory. A brief discussion about the non-conjugate case was also added. The chapter closes with a short discussion concerning certain extensions of the normal linear model.

Bibliography

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle", in B. Petrov and F. Csaki, eds., *Proceedings of 2nd International Symposium on Information Theory*, Academiai Kiado, Budapest, pp. 267–281.

Atkinson, A. (1978), "Posterior probabilities for choosing a regression model", *Biometrika* **65**, 39–48.

Barbieri, M. and Berger, J. (2004), "Optimal predictive model selection", *Annals of Statistics* **32**, 870–897.

Bartlett, M. (1957), "Comment on D.V. Lindley's statistical paradox", *Biometrika* **44**, 533–534.

Bayarri, M. and Berger, J. (2000), "P-values for composite null models (with discussion)", *Journal of the American Statistical Association* **95**, 1127–1142.

Bernardo, J. and Smith, A. (1994), *Bayesian Theory*, Wiley, Chichester, UK.

Burnham, K. and Anderson, D. (2004), "Multimodel inference: Understanding AIC and BIC in model selection", *Sociological Methods Research* **33**, 261–304.

Clyde, M., Littman, M. and Ghosh, J. (2009), 'Package `BAS` version 0.45'. available at `http://www.stat.duke.edu/~clyde/BAS/`.

Dellaportas, P., Forster, J. and Ntzoufras, I. (2000), "Bayesian variable selection using the Gibbs sampler", in D. Dey, S. Ghosh, and B. Mallick, eds., *Generalized Linear Models: A Bayesian Perspective*, Marcel Dekker, New York, pp. 271–286.

Dellaportas, P., Forster, J. and Ntzoufras, I. (2002), "On Bayesian model and variable selection using MCMC", *Statistics and Computing* **12**, 27–36.

Dey, D., Ghosh, S. and Mallick, B. (2000), *Generalized Linear Models: A Bayesian Perspective*, Marcel Dekker, New York.

Fernandez, C., Ley, E. and Steel, M. (2000), "Benchmark priors for Bayesian model averaging", *Journal of Econometrics* **100**, 381–427.

Fouskakis, D., Ntzoufras, I. and Draper, D. (2009), "Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care", *Annals of Applied Statistics* **3**, 663–690.

Gelman, A. and Hill, J. (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, New York.

George, E. and McCulloch, R. (1993), "Variable selection via Gibbs sampling", *Journal of the American Statistical Association* **88**, 881–889.

Gilks, W., Richardson, S. and Spiegelhalter, D. (1996), *Markov Chain Monte Carlo in Practice*, Interdisciplinary Statistics, Chapman & Hall, Suffolk, UK.

Green, P. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination", *Biometrika* **82**, 711–732.

Hans, C. (2009), 'Bayesian Lasso regression', *Biometrika* **96**, 835–845.

Hoeting, J., Madigan, D. and Raftery, A. (1996), 'A method for simultaneous variable selection and outlier identification in linear regression', *Computational Statistics and Data Analysis* **22**, 251–270.

Hoeting, J., Raftery, A. and Madigan, D. (2002), 'A method for simultaneous variable and transformation selection in linear regression', *Journal of Computational and Graphical Statistics* **11**, 485–507.

Jeffreys, H. (1961), *Theory of Probability*, 3rd. ed., Oxford University Press.

Kass, R. and Raftery, A. (1995), "Bayes factors", *Journal of the American Statistical Association* **90**, 773–795.

Kass, R. and Wasserman, L. (1995), "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion", *Journal of the American Statistical Association* **90**, 928–934.

Kollo, T. and von Rosen, D. (2005), *Advanced Multivariate Statistics with Matrices*, Mathematics and Its Applications, Springer, New York, USA.

Kuo, L. and Mallick, B. (1998), "Variable selection for regression models", *Sankhyā B* **60**, 65–81.

Liang, F., Paulo, R., Molina, G., Clyde, M. and Berger, J. (2008), 'Mixtures of g priors for Bayesian variable selection', *Journal of the American Statistical Association* **103**, 410–423.

Lindley, D. (1957), "A statistical paradox", *Biometrika* **44**, 187–192.

Lunn, D. J., Best, N. and Whittaker, J. (2005), *Generic Reversible Jump MCMC Using Graphical Models*, Technical Report, No EPH-2005-01, Department of Epidemiology and Public Health, Imperial College, London, UK, available at `https://www1.imperial.ac.uk/resources/8b3cf549-039e-4f96-8bec-cab969a0695ceph-2005-01.pdf`.

Lunn, D. J., Whittaker, J. C. and Best, N. (2006), "A Bayesian toolkit for genetic association studies", *Genetic Epidemiology* **30**, 231–247.

Lykou, A., Ntzoufras, I. and Whitaker, J. (2010), 'Bayesian variable selection using LASSO and related methods', *Postdoc Research Project (financed by Athens University of Economics and Business)* **(in progress)**.

Madigan, D. and York, J. (1995), 'Bayesian graphical models for discrete data', *International Statistical Review* **63**, 215–232.

McCullagh, P. and Nelder, J. (1989), *Generalized Linear Models*, Monographs on Statistics and Applied Probability, Vol. 37, 2nd ed., Chapman & Hall, Cambridge, UK.

Mun, J. (2006), *Modeling Risk: Applying Monte Carlo Simulation, Real Options Analysis, Forecasting, and Optimization*, 2nd ed., Wiley.

Mun, J. (2007), *Advanced Forecasting Techniques and Models : Regression & Econometrics*, Short Example Series Using Risk Calculator, Real Options Valuation INC, Dublin, California, USA; available at `http://www.realoptionsvaluation.com/download.html`.

Nadarajah, S. and Dey, D. K. (2005), 'Multitude of multivariate t-distributions', *Statistics* **39**, 149–181.

Ntzoufras, I. (2002), "Gibbs variable selection using BUGS", *Journal of Statistical Software* **7**, 1–19.

Ntzoufras, I. (2009), *Bayesian Modeling Using WinBUGS*, Wiley Series in Computational Statistics, Hoboken, NJ.

Raftery, A., Hoeting, J., Volinsky, C., Painter, I. and Yeung, K. (2009), 'Package `BMA` version 3.12'. available at `http://www2.research.att.com/~volinsky/bma.html`.

Raftery, A., Madigan, D. and Hoeting, J. (1997), "Bayesian model averaging for linear regression models", *Journal of the American Statistical Association* **92**, 179–191.

Spiegelhalter, D., Best, N., Carlin, B. and van der Linde, A. (2002), "Bayesian measures of model complexity and fit (with discussion)", *Journal of the Royal Statistical Society B* **64**, 583–639.

Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W. (1996), *BUGS 0.5: Bayesian Inference Using Gibbs Sampling Manual*, MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK.

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2003), *WinBUGS User Manual,* Version 1.4, MRC Biostatistics Unit, Institute of Public Health and Department of Epidemiology and Public Health, Imperial College School of Medicine, UK, available at `http://www.mrc-bsu.cam.ac.uk/bugs`.

Stanton, J. (2001), "Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors", *Journal of Statistics Education* **9**(3).

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society B* **58**, 267–288.

Zellner, A. (1986), "On assessing prior distributions and Bayesian regression analysis using g-prior distributions", in P. Goel and A. Zellner, eds., *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, North-Holland, Amsterdam, pp. 233–243.

Zellner, A. and Siow, A. (1980), Posterior odds ratios for selected regression hypothesis (with discussion), in J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds., *Bayesian Statistics*, Vol. 1, Oxford University Press, pp. 585–606 & 618–647 (discussion).