BAYESIAN STATISTICAL MODELS

# Chapter 3 - Bayes, LM and GLM

Lecturer: Riccardo Corradin University of Milano-Bicocca

This chapter introduces how to specify and estimate linear model within a Bayesian paradigm, providing later extensions and related results.

## 1 Once again, the linear regression model

Linear models are one of the fundamental and most commonly used techniques in data analysis. Despite their simplicity, they are flexible models which can help a statistician in many situations. Let $Y_1, \dots, Y_n$, where the generic $y_i \in \mathbb{Y} \subseteq \mathbb{R}$, a set of response (dependent) variables. These variable are the target of our inference, something that we observe and we want to explain. We further denote by $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ a set of covariates (independent variables), with the generic $\boldsymbol{x}_i \in \mathbb{X} \subseteq \mathbb{R}^p$.

As usual in regression problems, we want to identify a function of the covariates and some parameter, say $\boldsymbol{\beta}$, $g(\boldsymbol{x}, \boldsymbol{\beta})$ which describes the response variable $Y$, ideally to either predict possible values of the response variable or to investigate how such a response depends on the covariate values. Here, at first we restrict our attention to the case where such a model is of the form

$$g(\boldsymbol{x}, \boldsymbol{\beta}) = \boldsymbol{x}^\intercal \boldsymbol{\beta} = x_1\beta_1 + x_2\beta_2 + \cdots + x_d\beta_d.$$

Even if the previous model appears to be simple, it can accommodate many behavior. In fact, linear models are a linear combination of covariates and parameters, but response variable and covariates can be possibly transformed non-linearly.

$$\beta_1 x_1 \sin(\beta_2 x_2) \qquad \text{non linearizable},$$
$$x_1^{\beta_1} \mathrm{e}^{\beta_2 x_2} \to \beta_1 \log(x_1) + \beta_2 x_2 \qquad \text{linearizable}.$$

In frequentist statistics, the goal is to find a model which is optimal with respect to some loss function. Hence, we usually assume that the response variables can be decomposed additively as

$$Y_i = \boldsymbol{x}_i^\intercal \boldsymbol{\beta} + \varepsilon_i,$$

where $\varepsilon_i$ is an error term for which (i) $\mathbb{E}[\varepsilon_i] = 0$, (ii) $\mathrm{var}(\varepsilon_i) = \sigma^2$, and (iii) $\mathrm{cov}(\varepsilon_i, \varepsilon_\ell) = 0$ for $i \neq \ell$. Note that (i) implies $\mathbb{E}[Y_i] = \boldsymbol{x}_i^\intercal \boldsymbol{\beta}$. Usually, the loss function is build to minimize some distance between the observed response variables $y_1, \dots, y_n$ and the model fitted values, of the kind

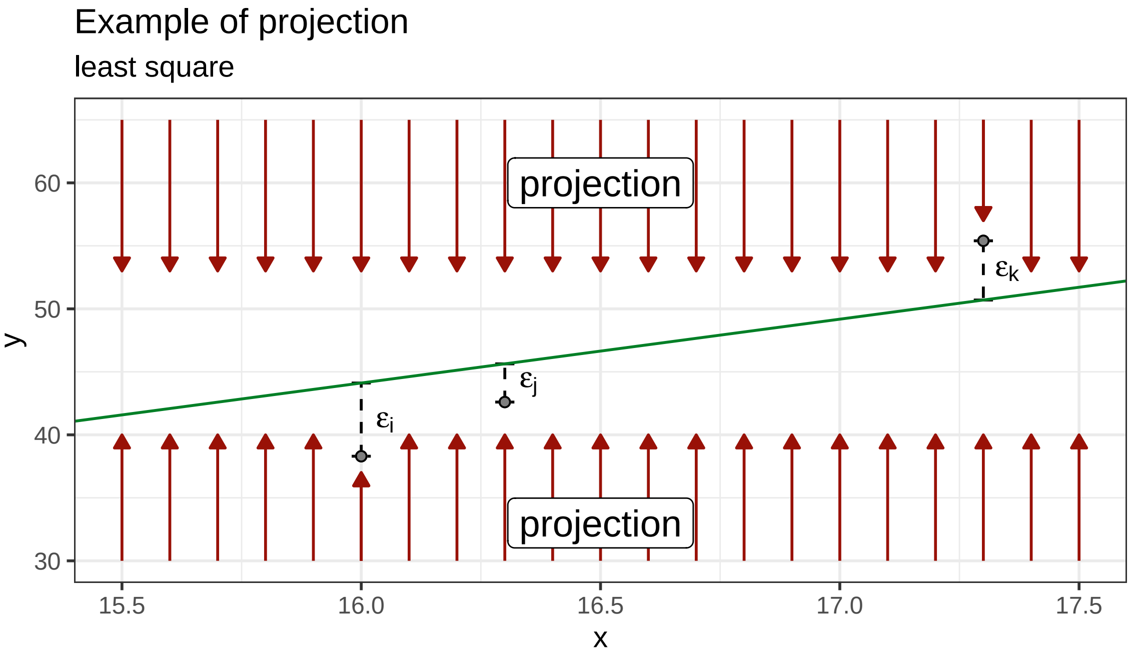
$$Q(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^\intercal \boldsymbol{\varepsilon} = \sum_{i=1}^n (y_i - \mathbb{E}[Y_i])^2 = \sum_{i=1}^n (y_i - \boldsymbol{x}_i^\intercal \boldsymbol{\beta})^2 = (\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta})^\intercal (\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}),$$

where $\boldsymbol{\varepsilon} = (\epsilon_1, \dots, \epsilon_n)^\intercal$, $\boldsymbol{y} = (y_1, \dots, y_n)^\intercal$ and $\mathrm{X}$ is the design matrix, whose $i$th row is $\boldsymbol{x}_i^\intercal$.

Assuming the design matrix $X$ being full rank $p$, it is easy to prove that the ordinary least square estimate corresponds to

$$\hat{\boldsymbol{\beta}}_{ML} = \min_{\boldsymbol{\beta} \in \mathbb{R}^d} Q(\boldsymbol{\beta}) = (\mathrm{X}^\intercal \mathrm{X})^{-1} \mathrm{X}^\intercal \boldsymbol{y}.$$

In practice we are minimizing the error committed by projecting the response variable on the regression hyperplane.



We can extend the specification of the model by setting a distributional assumption for the error term, specifically by setting

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathrm{I}_n).$$

Hence, the error are independent of each other and marginally $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. With the previous distributional assumption, maximum likelihood estimates equals OLS estimates. Once we assume a distribution, we are able to perform inference with our model.

## 1.1 Regularized linear regression models

One step further, a natural extension of the previous model includes a regularization term, that penalize the regression coefficient values. Ideally, we want to mitigate explosive behavior of the regression coefficients, especially for high dimensional data. Most commonly used, expressed with standardized response variable and marginally standardized covariates, are the following.

- Ridge regression, where the loss function is

  $$Q_R(\boldsymbol{\beta}) = (\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta})^\intercal(\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_2^2,$$

  with $||\cdot||_2$ denoting the Euclidean norm and $\lambda$ tuning parameter. The ridge regression coefficient estimate equals $\hat{\boldsymbol{\beta}}_R = (\mathrm{X}^\intercal\mathrm{X} + \lambda\mathrm{I}_d)^{-1}\mathrm{X}^\intercal\boldsymbol{y}$.

- Lasso, where the loss function is

  $$Q_L(\boldsymbol{\beta}) = (\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta})^\intercal(\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_1,$$

  with $||\cdot||_1$ denoting the absolute norm and $\lambda$ tuning parameter. There is no closed form for the lasso regression coefficient.

The previous strategies shrink their values toward the origin. This balances for unfriendly behaviour, overfitting, and allows for $p > n$ estimates.

# 2 Bayesian linear regression

Linear model can be specified also in a Bayesian perspective. Our model specification start from the probability model describing the conditional distribution of our data. Following the frequentist intuition, it is natural to assume the response variable

- symmetric around its expectation;
- with variance that does not depend on a specific covariate value;
- following a distribution that eventually leads to tractable inference.

The structural part of the model is then the usual linear model case

$$y = \boldsymbol{x}^\intercal \boldsymbol{\beta}.$$

Our first building block is the probabilistic model describing the data distribution conditioned on the parameters. It is natural to assume something symmetric around the structural part of the model, without any systematic asymmetry in the error we are committing. Here, we consider

$$Y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2 \sim N(\boldsymbol{x}^\intercal \boldsymbol{\beta}, \sigma^2), \qquad i = 1, \dots, n,$$

leading to a likelihood of the form

$$\mathrm{L}(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\intercal \boldsymbol{\beta})^2 \right\}.$$

The likelihood function drives empirical information in our posterior inference. Once again, measure how likely is the observed sample for a specific value of the parameter. The model specification is completed by selecting a suitable prior distribution on the parameter of interest, which in this case are the regression coefficients $\boldsymbol{\beta}$ and the dispersion term $\sigma^2$.

## 2.1 Objective prior specification

A first prior specification we consider here does not include any subjective information a priori. Hence, we consider a distribution which is uniform for $(\boldsymbol{\beta}, \log \sigma^2)$. We set

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

Note that, the previous prior corresponds to the Jeffreys noninformative prior for the linear regression model, assuming Gaussian likelihood. With many data points and few parameters it might be a good choice, as it gives nice results and it is easy to specify. On the counterpart, with few data points or many regression parameters, the likelihood is less peaked, and it is more important the prior specification.

**Proposition 2.1.** *Assume a priori $\pi(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$. Then a posteriori we have*

$$\pi(\boldsymbol{\beta}, \sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n}) = \pi(\boldsymbol{\beta} \mid \sigma^2, y_{1:n}, \boldsymbol{x}_{1:n}) \pi(\sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n}),$$

*where $\boldsymbol{\beta} \mid \sigma^2, y_{1:n}, \boldsymbol{x}_{1:n} \sim N(\hat{\boldsymbol{\beta}}_{ML}, (\mathrm{X}^\intercal \mathrm{X})^{-1}\sigma^2)$ is a multivariate Gaussian distribution and $\sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n} \sim IG(\frac{n-p}{2}, \frac{1}{2}(\boldsymbol{y} - \mathrm{X}\hat{\boldsymbol{\beta}}_{ML})^\intercal (\boldsymbol{y} - \mathrm{X}\hat{\boldsymbol{\beta}}_{ML}))$ is an inverse-gamma distribution.*

*Proof.* The posterior distribution is proportional to the prior distribution times the likelihood function. Hence, it suffices to identify whose distributional functions describes such a product. We first note that

$$\begin{aligned}
2(\mathrm{X}\hat{\boldsymbol{\beta}}_{ML} - \mathrm{X}\boldsymbol{\beta})^\intercal(\boldsymbol{y} - \mathrm{X}\hat{\boldsymbol{\beta}}_{ML}) &= 2\big[\mathrm{X}(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta})\big]^\intercal(\boldsymbol{y} - \mathrm{X}\hat{\boldsymbol{\beta}}_{ML}) \\
&= 2(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta})\mathrm{X}^\intercal\big[\boldsymbol{y} - \mathrm{X}(\mathrm{X}^\intercal\mathrm{X})^{-1}\mathrm{X}^\intercal\boldsymbol{y}\big] \\
&= 2(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta})\mathrm{X}^\intercal\big[I_n - \mathrm{X}(\mathrm{X}^\intercal\mathrm{X})^{-1}\mathrm{X}\big]\boldsymbol{y} \\
&= 2(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta})\big[\mathrm{X}^\intercal - \mathrm{X}^\intercal\mathrm{X}(\mathrm{X}^\intercal\mathrm{X})^{-1}\mathrm{X}\big]\boldsymbol{y} \\
&= 0.
\end{aligned}$$

We then have

$$\begin{aligned}
\pi(\boldsymbol{\beta}, \sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n}) &\propto \mathrm{L}(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta}, \sigma^2) \\
&= (2\pi\sigma^2)^{n/2} \exp\left\{\frac{1}{2\sigma^2}(\boldsymbol{y}_i - \mathrm{X}\boldsymbol{\beta})^\intercal(\boldsymbol{y}_i - \mathrm{X}\boldsymbol{\beta})\right\}\frac{1}{\sigma^2} \\
&= (2\pi)^{-n/2}(\sigma^2)^{-n/2-1} \exp\left\{\frac{1}{2\sigma^2}(\boldsymbol{y}_i \pm \mathrm{X}\hat{\boldsymbol{\beta}}_{ML} - \mathrm{X}\boldsymbol{\beta})^\intercal(\boldsymbol{y}_i \pm \mathrm{X}\hat{\boldsymbol{\beta}}_{ML} - \mathrm{X}\boldsymbol{\beta})\right\} \\
&\propto (\sigma^2)^{-\frac{n-p}{2}-1} \exp\left\{\frac{1}{2\sigma^2}(\boldsymbol{y}_i - \mathrm{X}\hat{\boldsymbol{\beta}}_{ML})^\intercal(\boldsymbol{y}_i - \mathrm{X}\hat{\boldsymbol{\beta}}_{ML})\right\} \\
&\qquad \times (\sigma^2)^{-\frac{p}{2}} \exp\left\{\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ML})^\intercal(\mathrm{X}^\intercal\mathrm{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ML})\right\}
\end{aligned}$$

□

With the previous prior specification we have the followings properties and summaries, which can be helpful while performing posterior inference with linear regression models. The expected value of $\sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n}$ is

$$\mathbb{E}[\sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n}] = \frac{1}{n-p-2}(\boldsymbol{y} - \mathrm{X}\hat{\boldsymbol{\beta}}_{ML})^\intercal(\boldsymbol{y} - \mathrm{X}\hat{\boldsymbol{\beta}}_{ML}),$$

which reminds the unbiased estimate for the error variance, but more conservative. The expected value of $\boldsymbol{\beta} \mid \sigma^2, y_{1:n}, \boldsymbol{x}_{1:n}$ equals

$$\mathbb{E}[\boldsymbol{\beta} \mid \sigma^2, y_{1:n}, \boldsymbol{x}_{1:n}] = \hat{\boldsymbol{\beta}}_{ML} = (\mathrm{X}^\intercal\mathrm{X})^{-1}\mathrm{X}^\intercal\boldsymbol{y}.$$

Hence, with an objective prior specification of the previous form, the posterior point estimate coincides with the frequentist maximum likelihood estimator. Our inference on the regression parameter values is driven by the empirical information of our data.

*Exercise* 2.2. We consider the following data

```
y <- c(-3.7, -5.0, 7.3, -4.3, -4.4, -7.3, 0.0, -5.7, 9.2, -4.0,
      -1.1, 1.5, -3.8, -7.4, 5.3, -8.8, -2.2, -5.4, 0.2, -2.6)
z <- c(1.1, 2.2, 3.6, 0.9, 1.9, 2.1, 2.7, 1.8, 4.0, 1.9,
      2.4, 3.0, 1.6, 1.0, 3.8, -0.3, 2.9, 2.0, 3.0, 2.4)
```

Consider a model of the form $y = \beta_1 + \beta_2 z + \beta_3 z^2$, so that the model is specified with a design matrix as

$$X = \begin{bmatrix} 1 & z_1 & z_1^2 \\ 1 & z_2 & z_2^2 \\ \vdots & \vdots & \vdots \end{bmatrix}.$$

- Write the function to sample from the posterior distribution of $(\boldsymbol{\beta}, \sigma^2)$ in R, using the explicit full conditionals of the previous vague prior, considering the first 3 observations.
- Repeat the previous points with the whole sample.

## 2.2 A more informative prior specification

Being Bayesian means we can resort to prior information, upon availability. Here we suppose a common scenario where we want to impose a stronger prior assumption, departing from the uniform case over $(\boldsymbol{\beta}, \log \sigma^2)$. Hence, we consider a prior specification of the form

$$\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\boldsymbol{\beta} \mid \sigma^2)\pi(\sigma^2).$$

In particular, we set

- $\boldsymbol{\beta} \mid \sigma^2 \sim N(\boldsymbol{\beta}_0, \sigma^2\Sigma_0)$, a multivariate Gaussian distribution, spreading over the support of $\boldsymbol{\beta}$ but having the variance dependent of the dispersion parameter $\sigma^2$;
- $\sigma^2 \sim IG(a_0, b_0)$, an inverse-gamma distribution, having support $\mathbb{R}_+$.

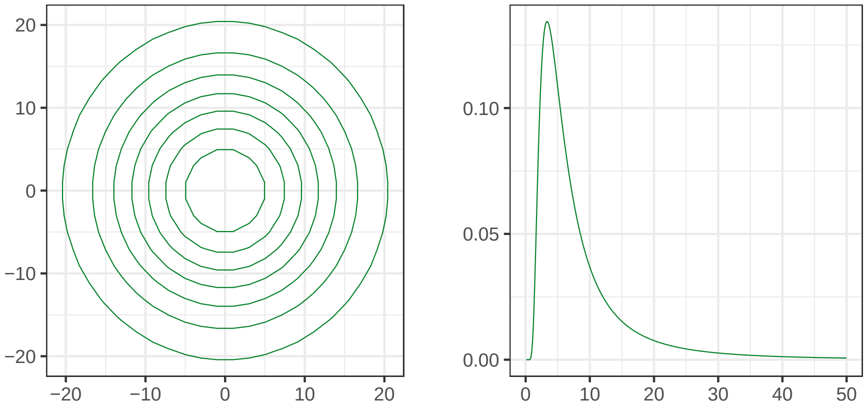


Figure 1: Graphical illustration of the previous prior specification. Left panel, prior distribution for the regression coefficient in a two-dimensional case. Right panel, prior distribution for the dispersion term.

With this specific prior setting, $\boldsymbol{\beta}_0$ describes where we center a priori our guess for the regression coefficients, usually set as $\boldsymbol{\beta}_0 = \mathbf{0}$ without any prior knowledge. However, if we have prior information on the regression coefficients, we can include that in the model specification. $\Sigma_0$, conditionally on $\sigma^2$, drives the prior dispersion of the regression coefficients distribution. $a_0$ and $b_0$ are shape and rate parameters, respectively. $a_0$ can be interpreted as the weight of our prior guess on $\sigma^2$.

- $\boldsymbol{\beta}_0$ is set usually equal to $\mathbf{0}$, except of when we have a strong prior opinion on the regression coefficient values.

- $\Sigma_0$ is commonly set as a diagonal matrix, with no correlation among different effects a priori. Any dependence among the regression coefficients is driven by the empirical information. In some scenarios it can be useful to set $\Sigma_0$ not diagonal, usually when we have strong prior opinion on the coefficients dependencies. Further, is usually set with large values on the diagonal, we do not want to be too concentrated a priori without any specific prior guess.

- To specify $a_0$ and $b_0$, we can look at the prior expectation, which is

  $$\mathbb{E}[\sigma^2] = \frac{b_0}{a_0 - 1}.$$

  For the shape parameter, $a_0 > 1$ and it weights the prior belief. Once we fix $a_0$, we can choose $b_0$ that guarantee the prior expectation we want for $\sigma^2$.

With the previous prior specification, $\boldsymbol{\beta} \mid \sigma^2 \sim N(\boldsymbol{\beta}_0, \sigma^2\Sigma_{\boldsymbol{\beta}})$ and $\sigma^2 \sim IG(a_0, b_0)$, we have the following posterior characterization.

**Proposition 2.3.** *Assume a priori $\boldsymbol{\beta} \mid \sigma^2 \sim N(\boldsymbol{\beta}_0, \sigma^2\Sigma_{\boldsymbol{\beta}})$ and $\sigma^2 \sim IG(a_0, b_0)$. Then a posteriori we have*

$$\pi(\boldsymbol{\beta}, \sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n}) = \pi(\boldsymbol{\beta} \mid \sigma^2, y_{1:n}, \boldsymbol{x}_{1:n})\pi(\sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n}).$$

*where $\boldsymbol{\beta} \mid \sigma^2, y_{1:n}, \boldsymbol{x}_{1:n} \sim N(\boldsymbol{\beta}_n, \sigma^2\Sigma_n)$ is a multivariate Gaussian distribution, $\sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n} \sim IG(a_n, b_n)$ is an inverse-gamma distribution, with*

$$\Sigma_n = \left[\Sigma_0^{-1} + (\mathrm{X}^\intercal\mathrm{X})\right]^{-1}, \qquad \boldsymbol{\beta}_n = \Sigma_n\left[\Sigma_0^{-1}\boldsymbol{\beta}_0 + (\mathrm{X}^\intercal\mathrm{X})\hat{\boldsymbol{\beta}}_{ML}\right]$$

$$a_n = a_0 + \frac{n}{2} \qquad b_n = b_0 + \frac{1}{2}\left(\boldsymbol{y}^\intercal\boldsymbol{y} - \boldsymbol{\beta}_n^\intercal\Sigma_n^{-1}\boldsymbol{\beta}_n + \boldsymbol{\beta}_0^\intercal\Sigma_0^{-1}\boldsymbol{\beta}_0\right)$$

*Remark* 2.4. The previous prior choice is conjugate to the Gaussian likelihood for linear regression, i.e., the posterior distribution is in the same family of the prior distribution.

*Proof.* We first recall that

$$\begin{aligned}(\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta})^\intercal(\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}) &= (\boldsymbol{y} \pm \mathrm{X}\hat{\boldsymbol{\beta}}_{ML} - \mathrm{X}\boldsymbol{\beta})^\intercal(\boldsymbol{y} - \pm\mathrm{X}\hat{\boldsymbol{\beta}}_{ML} - \mathrm{X}\boldsymbol{\beta})\\ &= (\boldsymbol{y} - \mathrm{X}\hat{\boldsymbol{\beta}}_{ML})^\intercal(\boldsymbol{y} - \mathrm{X}\hat{\boldsymbol{\beta}}_{ML}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ML})^\intercal(\mathrm{X}^\intercal\mathrm{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ML}).\end{aligned}$$

Then, we expand the exponent coming from the product of likelihood function and prior distributions,

$$\begin{aligned}&(\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta})^\intercal(\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\intercal\Sigma_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) =\\ &\qquad = \boldsymbol{\beta}^\intercal\left[\Sigma_0^{-1} + (\mathrm{X}^\intercal\mathrm{X})\right]\boldsymbol{\beta} - 2\boldsymbol{\beta}^\intercal\left[\Sigma_0\boldsymbol{\beta}_0 + (\mathrm{X}^\intercal\mathrm{X})\hat{\boldsymbol{\beta}}_{ML}\right] \pm \boldsymbol{\beta}_n^\intercal\Sigma_n^{-1}\boldsymbol{\beta}_n\\ &\qquad\quad + \boldsymbol{\beta}_0^\intercal\Sigma_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{y}^\intercal\boldsymbol{y}\cancel{-2\boldsymbol{y}^\intercal\mathrm{X}\boldsymbol{\beta}}\cancel{+2\boldsymbol{\beta}^\intercal\mathrm{X}^\intercal\mathrm{X}\boldsymbol{\beta}}\\ &\qquad = (\boldsymbol{\beta} - \boldsymbol{\beta}_n)^\intercal\Sigma_n^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_n) + \left[\boldsymbol{y}^\intercal\boldsymbol{y} - \boldsymbol{\beta}_n^\intercal\Sigma_n^{-1}\boldsymbol{\beta}_n + \boldsymbol{\beta}_0\Sigma_0^{-1}\boldsymbol{\beta}_0\right],\end{aligned}$$

Hence, the posterior distribution is identified by

$$\begin{aligned}
\pi(\boldsymbol{\beta}, \sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n}) &\propto \mathrm{L}(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta}, \sigma^2) \\
&= (2\pi\sigma^2)^{n/2} \exp\left\{ \frac{1}{2\sigma^2}(\boldsymbol{y}_i - \mathrm{X}\boldsymbol{\beta})^\intercal(\boldsymbol{y}_i - \mathrm{X}\boldsymbol{\beta}) \right\} \\
&\quad \times (\sigma^2)^{-a_0-1} \exp\left\{ \frac{b_0}{\sigma^2} \right\} (\sigma^2)^{-p/2} \exp\left\{ -\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\intercal \Sigma_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right\} \\
&= (\sigma^2)^{-(a_0 + \frac{n}{2})-1} \exp\left\{ \frac{1}{\sigma^2} \left[ b_0 + \frac{1}{2}\left( \boldsymbol{y}^\intercal \boldsymbol{y} - \boldsymbol{\beta}_n \Sigma_n^{-1} \boldsymbol{\beta}_n + \boldsymbol{\beta}_0 \Sigma_0^{-1} \boldsymbol{\beta}_0 \right) \right] \right\} \\
&\quad \times (\sigma^2)^{-p/2} \exp\left\{ -\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\beta}_n)^\intercal \Sigma_n^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_n) \right\},
\end{aligned}$$

with the parameter values given in the previous proposition.

□

The model has some connection with usual inferential procedures for linear regression. If $\hat{\boldsymbol{\beta}}_{ML}$ is the maximum likelihood estimator of a linear model with Gaussian error $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathrm{I}_n)$, then $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{ML}) = \sigma^2(\mathrm{X}^\intercal \mathrm{X})^{-1}$. Recall that the matrix term in the variance of $\boldsymbol{\beta} \mid y_{1:n}, \boldsymbol{x}_{1:n}, \sigma^2$ equals

$$\sigma^2 \Sigma_n = \sigma^2 \left[ \Sigma_0^{-1} + (\mathrm{X}^\intercal \mathrm{X}) \right]^{-1}.$$

The quantity in the right-hand term is averaging the reciprocal of the prior matrix term $\Sigma_0$ and the maximum likelihood estimator variance matrix term. Similarly, we have that

$$\boldsymbol{\beta}_n = \Sigma_n \left[ \Sigma_0^{-1} \boldsymbol{\beta}_0 + (\mathrm{X}^\intercal \mathrm{X}) \hat{\boldsymbol{\beta}}_{ML} \right],$$

is a weighted average the prior guess on the regression coefficients $\boldsymbol{\beta}_0$ and the maximum likelihood estimate $\hat{\boldsymbol{\beta}}_{ML}$, weighted by the matrix term in the prior variance of $\boldsymbol{\beta}$ and matrix term of the maximum likelihood estimator variance $(\mathrm{X}^\intercal \mathrm{X})$.

Regarding the dispersion term, a posteriori the shape parameter of the inverse-gamma distribution becomes $a_n = a_0 + n/2$. We said that $a_0$ can be interpreted as prior sample size, i.e. how strongly we trust the prior guess on $\sigma^2$. Hence, if we have a sample of size $n$ and we want to weight the empirical measure on $\sigma^2$ a posteriori by $q \times 100\%$, we can simply set

$$a_0 = \frac{n}{2} \frac{(1-q)}{q}.$$

Regarding the scale of the posterior distribution of the dispersion parameter, recall that its expression corresponds to

$$b_n = b_0 + \frac{1}{2}\left( \boldsymbol{y}^\intercal \boldsymbol{y} - \boldsymbol{\beta}_n^\intercal \Sigma_n^{-1} \boldsymbol{\beta}_n + \boldsymbol{\beta}_0 \Sigma_0 \boldsymbol{\beta}_0 \right).$$

We are adding to $b_0$ the response variable sum of squares, adjusted by the posterior shift of the regression coefficients, in quadratic form. The term $1/2$ balances the scale parameter posterior adjustment, i.e. $n/2$.

*Example* 2.5. We consider a simple example to show the effect of empirical information on posterior computation. We sample a set of covariates

$$Z_i \sim N(0,1), \qquad i = 1, \ldots, 50,$$

and

$$Y_i \sim N(25z_i + z_i^2, 4),\ i = 1, \ldots, 50.$$

We set $\boldsymbol{x}_i^\intercal = (1, z_i)$. We consider a model as

$$\begin{aligned} Y_i &\sim N(\boldsymbol{x}_i^\intercal \boldsymbol{\beta}, \sigma^2),\ i = 1, \ldots, 50, \\ \boldsymbol{\beta} \mid \sigma^2 &\sim N(\boldsymbol{\beta}_0, \Sigma_0), \\ \sigma^2 &\sim IG(a_0, b_0). \end{aligned}$$

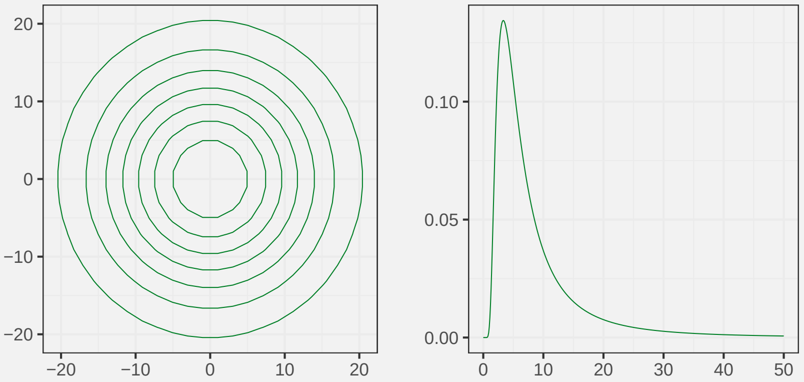
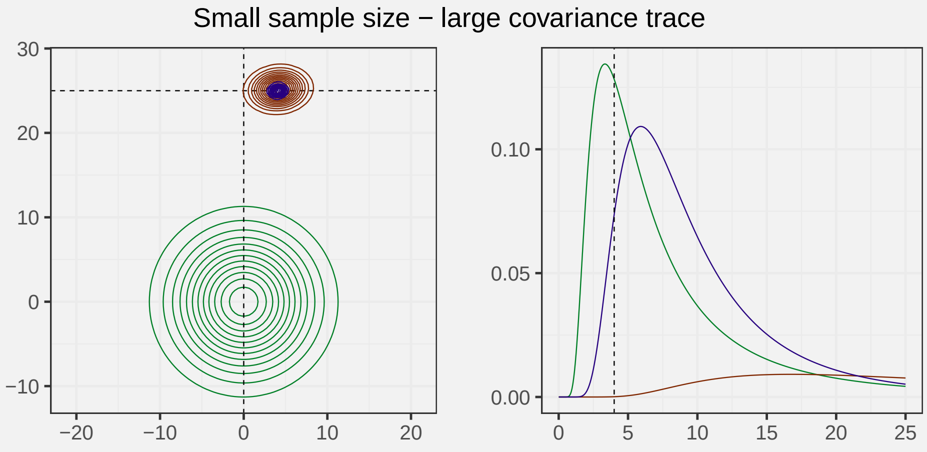
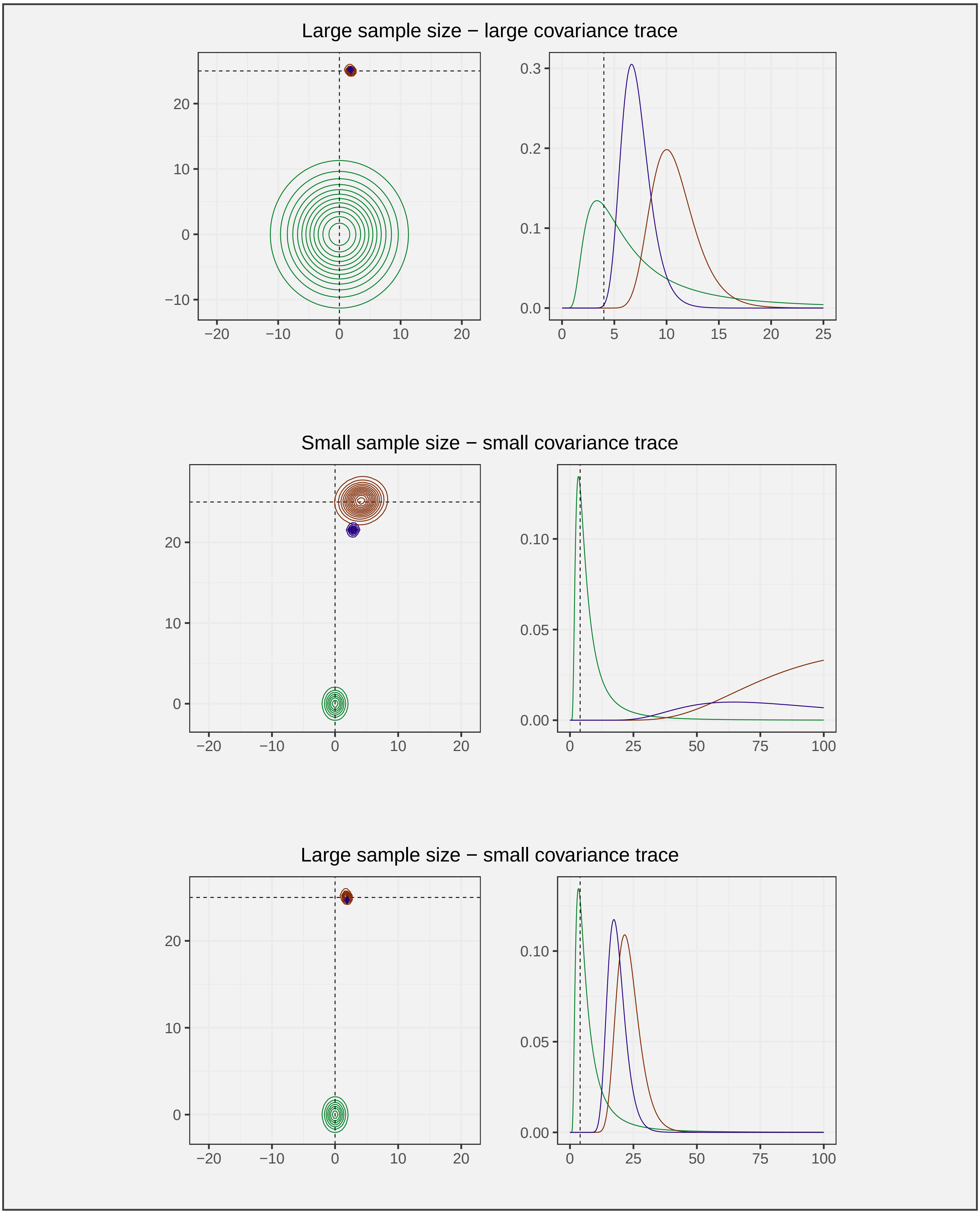


Figure 2: Graphical illustration of the prior specification.

We consider two different sample sizes, $n = 3$ and $n = 100$, $\boldsymbol{\beta}_0^\intercal = (0,0)$, $a_0 = 2$, $b_0 = 10$, and two different prior specification varying $\Sigma_0 = \mathrm{diag}_2(25)$ with $\Sigma_0 = \mathrm{diag}_2(5)$.





Large sample size − large covariance trace

Small sample size − small covariance trace

Large sample size − small covariance trace

*Exercise* 2.6. We consider again the following data

```
y <- c(-3.7, -5.0, 7.3, -4.3, -4.4, -7.3, 0.0, -5.7, 9.2, -4.0,
       -1.1, 1.5, -3.8, -7.4, 5.3, -8.8, -2.2, -5.4, 0.2, -2.6)
z <- c(1.1, 2.2, 3.6, 0.9, 1.9, 2.1, 2.7, 1.8, 4.0, 1.9,
       2.4, 3.0, 1.6, 1.0, 3.8, -0.3, 2.9, 2.0, 3.0, 2.4)
```

and a model of the form $y = \beta_1 + \beta_2 z + \beta_3 z^2$, so that the model is specified with a design matrix as

$$X = \begin{bmatrix} 1 & z_1 & z_1^2 \\ 1 & z_2 & z_2^2 \\ \vdots & \vdots & \vdots \end{bmatrix}.$$

- Consider now the case with $\boldsymbol{\beta} \mid \sigma^2 \sim N(\boldsymbol{\beta}_0, \sigma^2\Sigma_0)$ and $\sigma^2 \sim IG(a_0, b_0)$. Write the model in STAN when $\boldsymbol{\beta}_0 = \mathbf{0}$, $\Sigma_0 = 10^2\mathrm{I}_p$, $a_0 = 3$ and $b_0 = 2$.

- Compare the inference you obtain with the non-informative prior case.

# 3 Posterior inference

Once we have specified a model, we are interested in the usual tasks, such as quantifying the posterior uncertainty of our estimates and performing hypothesis tests regarding model structural components.

## 3.1 Credible intervals

We are Bayesian, we want to provide also suitable uncertainty quantification summaries along with our point estimate. We can construct credible intervals for the parameters of interest. For the variance parameter, it is trivial, as the posterior distribution of $\sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n}$ is an inverse-gamma distribution with shape $a_n$ and rate $b_n$ paraemters.

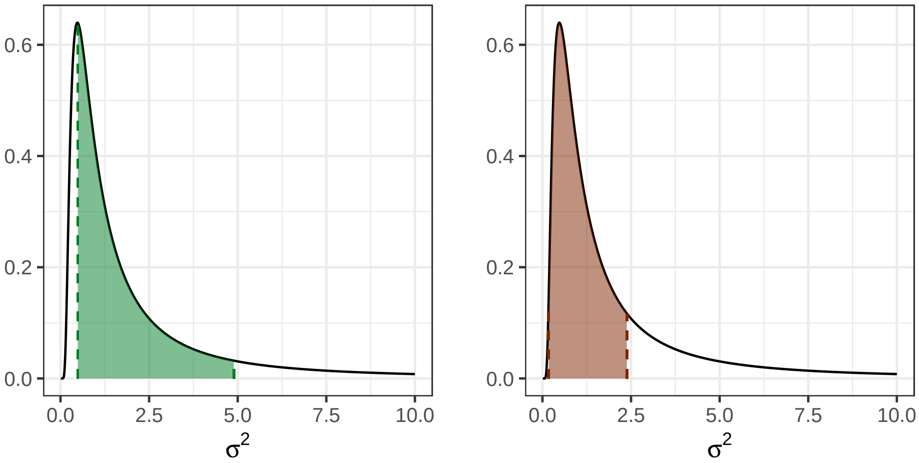


Figure 3: Left plot: equally tailed interval. Right plot: highest posterior density interval.

As emphasized in the previous figure, in extreme cases different approaches for constructing credible intervals lead to different subsets of the support. On the left panel, the probability mass kept on the tails is the same, however we are excluding values with high probability. On the right panel, the mass left on the right tail is way larger than the mass on the left tail, however we include all the values having high posterior probability.

For the regression coefficients, we have the hierarchical specification of the posterior distribution for which

$$\begin{aligned}\boldsymbol{\beta} \mid y_{1:n}, \boldsymbol{x}_{1:n}, \sigma^2 &\sim N(\boldsymbol{b}_n, \sigma^2\Sigma_n),\\ \sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n} &\sim IG(a_n, b_n).\end{aligned}$$

The distribution of $\boldsymbol{\beta}$ is conditioned on the specific value we are considering for $\sigma$. Hence, we cannot define immediately a credible region for $\boldsymbol{\beta}$ without considering specific values of $\sigma^2$. But we can marginalize out $\sigma^2$ from the joint distribution of $(\boldsymbol{\beta}, \sigma^2)$.

**Proposition 3.1.** *Once we marginalize out the dispersion parameter from $\pi(\boldsymbol{\beta}, \sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n})$, the resulting distribution is*

$$\boldsymbol{\beta} \mid y_{1:n}, \boldsymbol{x}_{1:n} \sim t_{2a_n}\left(\boldsymbol{\beta}_n, \tfrac{a_n}{b_n}\Sigma_n\right)$$

*Proof.* Here, we want to compute the following integral

$$\begin{aligned}\pi(\boldsymbol{\beta} \mid y_{1:n}, \boldsymbol{x}_{1:n}) &= \int_{\mathbb{R}_+} \pi(\boldsymbol{\beta}, \sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n}) \mathrm{d}\sigma^2 \\ &= \int_{\mathbb{R}_+} (2\pi)^{-p/2} |\sigma^2\Sigma_n|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)^\intercal(\sigma^2\Sigma_n)^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)\right\} \frac{b_n^{a_n}}{\Gamma(a_n)}(\sigma^2)^{-a_n-1}\mathrm{e}^{-b_n/\sigma^2}\mathrm{d}\sigma^2 \\ &\propto \int_{\mathbb{R}_+} (\sigma^2)^{-a_n-p/2-1} \exp\left\{-\frac{1}{\sigma^2}\left[b_n + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)^\intercal\Sigma_n^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)\right]\right\}\mathrm{d}\sigma^2.\end{aligned}$$

We can try to identify a known form from the previous integral. In particular, we have something that reminds us the density function of an inverse gamma distribution, that we are integrating, but we miss the normalization constant. We can multiply and divide for the term that we are missing, obtaining

$$\begin{aligned}&\int_{\mathbb{R}_+} \frac{\left[b_n + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)^\intercal\Sigma_n^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)\right]^{a_n+p/2}}{\Gamma(a_n+p/2)(\sigma^2)^{a_n+p/2+1}} \exp\left\{-\frac{1}{\sigma^2}\left[b_n + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)^\intercal\Sigma_n^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)\right]\right\}\mathrm{d}\sigma^2 \\ &\quad\times \left[b_n + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)^\intercal\Sigma_n^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)\right]^{-(a_n+p/2)} \Gamma(a_n+p/2) \\ &\quad\propto \left[b_n + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)^\intercal\Sigma_n^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)\right]^{-(a_n+p/2)} \\ &\quad= \left[1 + \frac{(\boldsymbol{\beta}-\boldsymbol{\beta}_n)^\intercal\left(\frac{b_n}{a_n}\Sigma_n\right)^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)}{2a_n}\right]^{-(2a_n+p)/2}\end{aligned}$$

The previous identify a multivariate t-Student distribution, up to a normalization constant which does not depend on $\boldsymbol{\beta}$.

□

As example, consider a set of synthetic data $Z_i \sim N(0,1)$, $Y_i \sim N(25z_i + z_i^2, 4)$, $i = 1, \dots, 50$. We consider a model as $Y_i \sim N(\boldsymbol{x}_i^\intercal \boldsymbol{\beta}, \sigma^2)$, with $\boldsymbol{x}_i^\intercal = (1, z_i)$, $i = 1, \dots, 50$, and priors $\boldsymbol{\beta} \mid \sigma^2 \sim N(\boldsymbol{\beta}_0, \Sigma_0)$, $\sigma^2 \sim IG(a_0, b_0)$. In the next figure we can appreciate the regression coefficients credible regions resulting from the previous marginalizations. Note that, given the symmetry of the marginal distribution, equally tailed interval and highest posterior density interval intervals coincide.

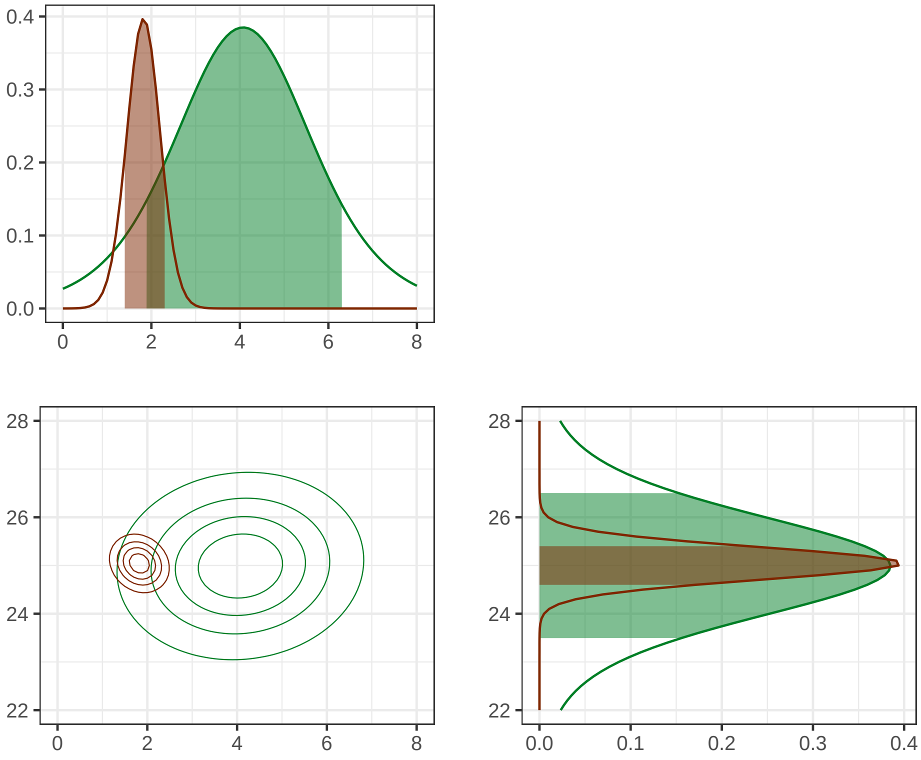


Figure 4: Posterior uncertainty quantification. Green lines/areas consider the first 3 observations, red ones with the whole sample.

## 3.2 Hypothesis testing

Once we have a posterior distribution, we can also evaluate if the empirical information supports some conjectures on the model structure. As usual, statistical tests for linear models are mainly used to answer two questions.

1. Is a single regression coefficient different from a specific value?
2. Is the whole model different from another model?

Within a Bayesian approach, we can also answer to this two inferential questions in simple ways, exploiting suitable functionals of the posterior.

1. For the first question, we want to test

$$H_0 : \beta_j = c \qquad vs \qquad H_1 : \beta_j \neq c,$$

   assuming both hypotheses having the same prior probability. Such assumption can be relaxed straightforwardly, by incorporating also the prior information. To test the previous hypothesis, we can resort to the Bayes factor. Under the previous setting, we have

$$\mathrm{BF}_{01} = \frac{m(y_{1:n} \mid \boldsymbol{x}_{1:n})\Big|_{\beta_j=c}}{m(y_{1:n} \mid \boldsymbol{x}_{1:n})}$$

where $m(y_{1:n} \mid \boldsymbol{x}_{1:n})$ denotes the marginal distribution of the data and

$$m(y_{1:n} \mid \boldsymbol{x}_{1:n})\Big|_{\beta_j=c}$$

the marginal distribution constraining the $j$th parameter to be equal to $c$. We can simplify a bit of things here. First, we need the marginal distribution, but luckily we can have it almost for free. In fact, recall that the joint distribution of response variable, regression coefficients and dispersion parameter is nothing but the product we used before in the posterior calculations, with

$$\begin{aligned}\mathcal{L}(y_{1:n}, &\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{x}_{1:n}) = (2\pi\sigma^2)^{-p/2}|\Sigma_0| \exp\left\{\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\beta}_n)^\intercal \Sigma_n^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_n)\right\} \\ &\times (\sigma^2)^{-a_n-1} \exp\left\{\frac{b_n}{\sigma^2}\right\} (2\pi)^{-n/2} \frac{b_0^{a_0}}{\Gamma(a_0)}.\end{aligned}$$

Hence, multiplying and dividing for the normalization constant missed to identify suitable density functions, we have

$$\begin{aligned}m(y_{1:n} \mid \boldsymbol{x}_{1:n}) &= \int_{\mathbb{R}_+} \int_{\mathbb{R}^p} \mathcal{L}(y_{1:n}, \boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{x}_{1:n}) \mathrm{d}\boldsymbol{\beta} \mathrm{d}\sigma^2 \\ &= \frac{|\Sigma_0|^{-1/2}}{|\Sigma_n|^{-1/2}} \frac{b_0^{a_0}}{b_n^{a_n}} \frac{\Gamma(a_n)}{\Gamma(a_0)} \frac{1}{(2\pi)^{n/2}},\end{aligned}$$

which corresponds to the marginal distribution of interest. Once we have such marginal, we can see that the constrained marginal

$$m(y_{1:n} \mid \boldsymbol{x}_{1:n})\Big|_{\beta_j=c}$$

equals to the same marginal distribution but considering as response variable a differentiated version of $y_i$, with

$$y_i^{(c)} = y_i - c x_{i,j}, \qquad i = 1, \dots, n,$$

with $\boldsymbol{x}_i^{(c)}$, $i = 1, \dots, n$, being the $i$th vector of covariates without the $j$th element. Hence, the $\mathrm{BF}_{01}$ equals

$$\mathrm{BF}_{01} = \frac{m\left(y_{1:n}^{(c)} \mid \boldsymbol{x}_{1:n}^{(c)}\right)}{m(y_{1:n} \mid \boldsymbol{x}_{1:n})} = \sqrt{\frac{|\Sigma_n^{(c)}|}{|\Sigma_n|}} \frac{b_n^{a_n}}{\left(b_n^{(c)}\right)^{a_n^{(c)}}} \frac{\Gamma\left(a_n^{(c)}\right)}{\Gamma(a_n)},$$

where$\boldsymbol{b}_n$, $\Sigma_n$, $a_n$, $b_n$ are the posterior distribution parameters of $\boldsymbol{\beta}, \sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n}$ and $\boldsymbol{b}_n^{(c)}$, $\Sigma_n^{(c)}$, $a_n^{(c)}$, $b_n^{(c)}$ are the restricted posterior distribution parameters of $\boldsymbol{\beta}, \sigma^2 \mid y_{1:n}^{(c)}, \boldsymbol{x}_{1:n}^{(c)}$.

2. For the second type of test, In general we can construct the Bayes factor to compare two different models, say $M_0$ and $M_1$. The Bayes factor is then defined as

$$\mathrm{BF}_{01} = \frac{\text{posterior odd}_{01}}{\text{prior odd}_{01}} = \frac{\frac{\Pr(M_0|y_{1:n},\boldsymbol{x}_{1:n})}{\Pr(M_1|y_{1:n},\boldsymbol{x}_{1:n})}}{\frac{\Pr(M_0)}{\Pr(M_1)}} = \frac{\Pr(M_0 \mid y_{1:n}, \boldsymbol{x}_{1:n})\Pr(M_1)}{\Pr(M_1 \mid y_{1:n}, \boldsymbol{x}_{1:n})\Pr(M_0)}.$$

With some algebraic manipulation, the previous can be rewritten as

$$\mathrm{BF}_{01} = \frac{m(y_{1:n} \mid M_0, \boldsymbol{x}_{1:n})}{m(y_{1:n} \mid M_1, \boldsymbol{x}_{1:n})},$$

where $m(y_{1:n} \mid M_0, \boldsymbol{x}_{1:n})$ and $m(y_{1:n} \mid M_1, \boldsymbol{x}_{1:n})$ denote the marginal distribution of the data under $M_0$ and $M_1$, respectively. Note that we can compare different models by simply looking at the marginal distribution of the data, which is measuring how likely are the data under such a model assumption.

*Exercise* 3.2. We consider again the following data

```
y <- c(-3.7, -5.0, 7.3, -4.3, -4.4, -7.3, 0.0, -5.7, 9.2, -4.0,
       -1.1, 1.5, -3.8, -7.4, 5.3, -8.8, -2.2, -5.4, 0.2, -2.6)
z <- c(1.1, 2.2, 3.6, 0.9, 1.9, 2.1, 2.7, 1.8, 4.0, 1.9,
       2.4, 3.0, 1.6, 1.0, 3.8, -0.3, 2.9, 2.0, 3.0, 2.4)
```

and a model of the form $y = \beta_1 + \beta_2 z + \beta_3 z^2$, so that the model is specified with a design matrix as

$$X = \begin{bmatrix} 1 & z_1 & z_1^2 \\ 1 & z_2 & z_2^2 \\ \vdots & \vdots & \vdots \end{bmatrix}.$$

A priori $(\boldsymbol{\beta}, \sigma^2) \sim NIG(\boldsymbol{\beta}_0, \Sigma_0, a_0, b_0)$ with $\boldsymbol{\beta}_0 = \mathbf{0}$, $\Sigma_0 = 10^2\mathrm{I}_p$, $a_0 = 3$ and $b_0 = 2$.

- Consider only the first three observation. Test if $\beta_2$ is different form $0$.
- Repeat the previous test with the whole sample.
- Test if the full model is different from the model with only the intercept term, considering only the first three observation.
- Repeat the previous test with the whole sample.

## 3.3 Predictive inference

We may be interested in predictive inference, say $n+1$, given what we observed and our updated belief, i.e. in the predictive distribution of a future observation integrating out the model parameters

$$\begin{aligned}&\mathcal{L}(y_{n+1} \mid \boldsymbol{x}_{n+1}, y_{1:n}, \boldsymbol{x}_{1:n}) \\ &\quad = \int_{\mathbb{R}_+}\int_{\mathbb{R}^p} f(y_{n+1} \mid \boldsymbol{x}_{n+1}, \boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta}, \sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n})\mathrm{d}\boldsymbol{\beta}\mathrm{d}\sigma^2.\end{aligned}$$

Such a distribution is available in a closed and simple form, as stated in the following proposition.

**Proposition 3.3.** *Consider a linear regression model with Gaussian likelihood and prior assumptions $\boldsymbol{\beta} \mid \sigma^2 \sim N(\boldsymbol{\beta}_0, \sigma^2\Sigma_{\boldsymbol{\beta}})$, $\sigma^2 \sim IG(a_0, b_0)$. Then,*

$$Y_{n+1} \mid \boldsymbol{x}_{n+1}, y_{1:n}, \boldsymbol{x}_{1:n} \sim t_{2a_n}\Big(\boldsymbol{x}_{n+1}^{\intercal}\boldsymbol{\beta}_n, \tfrac{b_n}{a_n}(1 + \boldsymbol{x}_{n+1}^{\intercal}\Sigma_n\boldsymbol{x}_{n+1}\Big).$$

*Proof.* We first compute the inner integral, with respect to $\boldsymbol{\beta}$. Note that

$$Y_{n+1} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{x}_{n+1} \sim N(\boldsymbol{x}_{n+1}^{\intercal}\boldsymbol{\beta}, \sigma^2), \qquad \boldsymbol{\beta} \mid \sigma^2, y_{1:n}, \boldsymbol{x}_{1:n} \sim N(\boldsymbol{\beta}_n, \sigma^2\Sigma_n).$$

Therefore, $Y_{n+1} = \boldsymbol{x}_{n+1}^{\intercal}\boldsymbol{\beta} + \varepsilon_{n+1}$, with $\varepsilon_{n+1} \sim N(0, \sigma^2)$, is a linear combination of Gaussian random variables. Further,

$$\mathbb{E}_{\boldsymbol{\beta}}[f(y_{n+1} \mid \boldsymbol{x}_{n+1}, \boldsymbol{\beta}, \sigma^2) \mid y_{1:n}, \boldsymbol{x}_{1:n}]$$

is the density function of a Gaussian distribution, with

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\beta}}[Y_{n+1} \mid -] &= \boldsymbol{x}_{n+1}^{\intercal}\boldsymbol{\beta}_n, \\ \mathrm{Var}_{\boldsymbol{\beta}}(Y_{n+1} \mid -) &= \sigma^2(1 + \boldsymbol{x}_{n+1}^{\intercal}\Sigma_n\boldsymbol{x}_{n+1}). \end{aligned}$$

Then, solving the outer integral, we have

$$\begin{aligned} &\int_{\mathbb{R}_+} \mathbb{E}_{\boldsymbol{\beta}}[f(y_{n+1} \mid \boldsymbol{x}_{n+1}, \boldsymbol{\beta}, \sigma^2) \mid y_{1:n}, \boldsymbol{x}_{1:n}]\pi(\sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n})\mathrm{d}\sigma^2 \\ &\quad\propto \int_{\mathbb{R}_+} (\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\frac{(y_{n+1} - \boldsymbol{x}_{n+1}^{\intercal}\boldsymbol{\beta}_n)^2}{1 + \boldsymbol{x}_{n+1}^{\intercal}\Sigma_n\boldsymbol{x}_{n+1}}\right\} (\sigma^2)^{-a_n-1} \exp\left\{-\frac{b_n}{\sigma^2}\right\} \mathrm{d}\sigma^2 \\ &\quad\propto \left[b_n + \frac{1}{2}\frac{(y_{n+1} - \boldsymbol{x}_{n+1}^{\intercal}\boldsymbol{\beta}_n)^2}{1 + \boldsymbol{x}_{n+1}^{\intercal}\Sigma_n\boldsymbol{x}_{n+1}}\right]^{-\frac{2a_n+1}{2}} \\ &\quad= \left[1 + \frac{1}{2a_n}\frac{(y_{n+1} - \boldsymbol{x}_{n+1}^{\intercal}\boldsymbol{\beta}_n)^2}{\frac{b_n}{a_n}(1 + \boldsymbol{x}_{n+1}^{\intercal}\Sigma_n\boldsymbol{x}_{n+1})}\right]^{-\frac{2a_n+1}{2}}, \end{aligned}$$

which identifies the distribution of interest.

□

*Exercise* 3.4. We consider again the following data

```
    y <- c(-3.7, -5.0, 7.3, -4.3, -4.4, -7.3, 0.0, -5.7, 9.2, -4.0,
           -1.1, 1.5, -3.8, -7.4, 5.3, -8.8, -2.2, -5.4, 0.2, -2.6)
    z <- c(1.1, 2.2, 3.6, 0.9, 1.9, 2.1, 2.7, 1.8, 4.0, 1.9,
           2.4, 3.0, 1.6, 1.0, 3.8, -0.3, 2.9, 2.0, 3.0, 2.4)
```

and a model of the form $y = \beta_1 + \beta_2 z + \beta_3 z^2$, so that the model is specified with a design matrix as

$$X = \begin{bmatrix} \boldsymbol{\beta}_0 1 & z_1 & z_1^2 \\ 1 & z_2 & z_2^2 \\ \vdots & \vdots & \vdots \end{bmatrix}.$$

- A priori: normal-inverse-gamma with $\boldsymbol{\beta}_0 = \mathbf{0}$, $\Sigma_0 = \mathrm{diag}_3(25)$, $a_0 = 2$, $b_0 = 5$.
  - Sample $1\,000$ realizations from the predictive distribution considering only the first three observations. Repeat the previous with the whole sample.

- Now, a normal-inverse-gamma with $\boldsymbol{\beta}_0 = \mathbf{0}$, $\Sigma_0 = \mathrm{diag}_3(5)$, $a_0 = 2$, $b_0 = 5$.
  - Sample $1\,000$ realizations from the predictive distribution considering only the first three observations. Repeat the previous with the whole sample.

Do you see any difference?

# 4 Shrinkage with Bayesian linear regression models

As mentioned before, regularization mitigates strange behavior of the regression coefficients, especially for high dimensional data regression problems. Let us fix the value of $\sigma^2$. Looking at the conjugate prior for $\boldsymbol{\beta}$ with $\sigma^2$ fixed, it is implicitly inducing a regularization on the regression coefficients. We set

$$Y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta} \sim N(\boldsymbol{x}_i^\intercal \boldsymbol{\beta}, \sigma^2), \quad i = 1, \dots, n,$$
$$\boldsymbol{\beta} \sim N\left(\boldsymbol{\beta}_0, \frac{\sigma^2}{\lambda} I_p\right).$$

Then, looking at the posterior distribution of $\boldsymbol{\beta}$, we have

$$\pi(\boldsymbol{\beta} \mid y_{1:n}, \boldsymbol{x}_{1:n}) \propto \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta})^\intercal(\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}) - \frac{\lambda}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\intercal(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right\}.$$

In particular, if we consider as point estimate the maximum of the posterior distribution, we identify the following quantity

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \mathrm{argmax}_{\boldsymbol{\beta} \in \mathbb{R}^p}\ \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta})^\intercal(\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}) - \frac{\lambda}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\intercal(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right\} \\ &= \mathrm{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p}\ (\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta})^\intercal(\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2^2.\end{aligned}$$

When $\boldsymbol{\beta}_0 = \mathbf{0}$, the previous expression is the usual ridge regression loss function. $\lambda$ is a penalty term which drives the regularization, i.e., how much we are shrinking the regression coefficients toward the origin. We can further relax the model specification by setting $\lambda \sim Gamma(\tau, \zeta)$.

Alternatively, we can consider another prior specification for $\boldsymbol{\beta}$ which is inducing a lasso regularization, by setting $\beta_j \sim Lap(\beta_{0,j}, \frac{\sigma^2}{\lambda})$, $j = 1, \dots, p$, i.e., independent components with Laplace distribution. We recall that the density function of a Laplace random variable is given by

$$\pi(\beta_j) = \frac{\lambda}{2\sigma^2} \exp\left\{-\frac{\lambda}{2\sigma^2}|\beta_j - \beta_{0,j}|\right\}.$$

Then, the joint posterior of $\boldsymbol{\beta}$ is proportional to the product of the likelihood term and the prior distribution

$$\pi(\boldsymbol{\beta} \mid y_{1:n}, \boldsymbol{x}_{1:n}) \propto \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta})^\intercal(\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}) - \frac{\lambda}{2\sigma^2}\sum_{j=1}^p |\beta_j - \beta_{0,j}|\right\}.$$

Looking at the maximum a posteriori estimate, we obtain

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \mathrm{argmax}_{\boldsymbol{\beta} \in \mathbb{R}^p}\ \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta})^\intercal(\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}) - \frac{\lambda}{2\sigma^2}\sum_{j=1}^p |\beta_j - \beta_{0,j}|\right\} \\ &= \mathrm{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p}\ (\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta})^\intercal(\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_1.\end{aligned}$$

When $\boldsymbol{\beta}_0 = \mathbf{0}$, the previous expression is the usual lasso regression loss function. $\lambda$ is a penalty term which drives the regularization, we can relax the model specification by setting $\lambda^2 \sim Gamma(\tau, \zeta)$. No closed form, we can use computational tools (e.g. STAN) to perform posterior inference.

*Exercise* 4.1. We consider again the following data

```
y <- c(-3.7, -5.0, 7.3, -4.3, -4.4, -7.3, 0.0, -5.7, 9.2, -4.0,
       -1.1, 1.5, -3.8, -7.4, 5.3, -8.8, -2.2, -5.4, 0.2, -2.6)
z <- c(1.1, 2.2, 3.6, 0.9, 1.9, 2.1, 2.7, 1.8, 4.0, 1.9,
       2.4, 3.0, 1.6, 1.0, 3.8, -0.3, 2.9, 2.0, 3.0, 2.4)
```

and a model of the form $y = \beta_1 + \beta_2 z + \beta_3 z^2$, so that the model is specified with a design matrix as

$$X = \begin{bmatrix} 1 & z_1 & z_1^2 \\ 1 & z_2 & z_2^2 \\ \vdots & \vdots & \vdots \end{bmatrix}.$$

- A priori: normal-inverse-gamma in the spirit of slide 38, with $\boldsymbol{\beta}_0 = \mathbf{0}$, $a_0 = 2$, $b_0 = 5$ and $\lambda \sim gamma(1, 1)$.
- Now, a Laplace-inverse-gamma in the spirit of slide 39, with $\boldsymbol{\beta}_0 = \mathbf{0}$, $a_0 = 2$, $b_0 = 5$ and $\lambda \sim gamma(1, 1)$.

Do you see any difference?

# 5 Beyond linear regression models

Generalized linear models (GLMs) are... a generalization of ordinary linear models. They extend the previous model strategies mainly in three directions: the response variable can be defined on a subspace of $\mathbb{R}$, e.g., $\{0, 1\}$ or $\mathbb{Z}_+$; the relation between the linear predictor and the response variable can be non-linear and the dispersion can be non-homogeneous when the covariates vary over their support.

Given a response variable $Y$ taking values in $\mathbb{Y} \subseteq \mathbb{R}$ and a set of covariates $\boldsymbol{x} \in \mathbb{X} \subseteq \mathbb{R}^p$, a GLM is composed by three main terms, which describes the response variable and its connection with the covariates.

a) a distributional assumption for the response variable $Y \sim f(y \mid -)$, which plays the role of likelihood term and depends on the data we are observing;

b) a linear predictor, which is defining a linear combination of the covariates with a set of parameters $\eta = \boldsymbol{x}^\intercal \boldsymbol{\beta}$, with $\boldsymbol{x} \in \mathbb{R}^p$;

c) a link function $g(\cdot)$, which is linking the linear predictor with the expectation of the response variable

$$\mathbb{E}[Y \mid \boldsymbol{x}, \boldsymbol{\beta}] = \mu = g^{-1}(\eta).$$

Regarding the distributional assumption, we consider distributions belonging to the exponential family. Specifically, we assume that the generic $Y \mid \boldsymbol{x} \sim EF(\theta, \psi)$ has density function of the form

$$f(y \mid \theta, \psi) = \exp\left\{\frac{y\theta - b(\theta)}{\psi} + c(y, \psi)\right\}, \qquad i = 1, \dots, n,$$

where $\theta$ is the natural parameter of the exponential family and $\psi$ is the scale parameter. The function $b(\cdot)$ and the parameter $\psi$ are common to all the observations. Further, all the functions $b(\cdot)$, $c(\cdot, \cdot)$, $g(\cdot)$ are assumed to be known. Mean and variance have a nice explicit form, with

$$\mathbb{E}[Y] = \frac{\mathrm{d}}{\mathrm{d}\theta} b(\theta) = \mu, \qquad \mathrm{Var}(Y) = \psi \times \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} b(\theta) = \psi V(\mu),$$

where $V(\mu)$ is called the variance function.

To summarize the model specification, we have to specify the following three quantities

$$\underbrace{Y_i \mid \eta_i \sim EF(b(\theta_i), \psi)}_{\text{error structure}}, \qquad \underbrace{g(\mu_i) = \eta_i}_{\text{link function}}, \qquad \underbrace{\eta_i = \boldsymbol{x}_i^\intercal \boldsymbol{\beta}}_{\text{linear predictor}}, \qquad i = 1, \dots, n.$$

About the meaning of the previous expressions, the observations consist of independent random variables, where the generic $Y_i$ has distribution $EF(b(\theta_i), \psi)$, with

$$\mathbb{E}[Y_i] = \mu_i = \frac{\mathrm{d}}{\mathrm{d}\theta_i} b(\theta_i), \qquad i = 1, \dots, n.$$

The function $g(\mu_i) = \boldsymbol{x}^\intercal \boldsymbol{\beta}$, where $\boldsymbol{x}_i$ is a vector of constants and $\boldsymbol{\beta}$ a vector of parameters, relates the expectation of the response variable with the linear predictor. Some choices are better than others. If we set $g(\mu_i) = \theta_i$ so that $\eta_i = \theta_i$, we get the canonical link function. Many known distribution can be rewritten in this specific form.

– Bernoulli distribution (binary response);

– Poisson distribution (count response);

– Gamma distribution (positive real-valued response);

– ...

Further, for some distributional assumption is easy the explicit the quantities appearing in the previous exponential family representation.

| | $N(\mu, \sigma^2)$ | $Poi(\mu)$ | $Bin(m, \mu)/m$ | $Gamma(\alpha, \alpha/\mu)$ |
|---|---|---|---|---|
| SUPPORT | $(-\infty, +\infty)$ | $\{0, 1, 2, \dots\}$ | $\{0, \frac{1}{m}, \frac{2}{m}, \dots, 1\}$ | $(0, +\infty)$ |
| $\psi$ | $\sigma^2$ | 1 | 1 | $\alpha^{-1}$ |
| $\omega$ | 1 | 1 | m | 1 |
| $b(\theta)$ | $\theta^2/2$ | $\exp(\theta)$ | $\log(1 + \mathrm{e}^\theta)$ | $-\log(-\theta)$ |
| $c(y, \psi)$ | $-\frac{1}{2}\left(\frac{y^2}{\psi} + \log(2\pi\psi)\right)$ | $-\log y!$ | $\log\binom{m}{my}$ | $\alpha \log(\alpha y)$<br>$-\log(y)$<br>$-\log\Gamma(\alpha)$ |
| $\mu(\theta)$ | $\theta$ | $\exp(\theta)$ | $\mathrm{e}^\theta/(1 + \mathrm{e}^\theta)$ | $-1/\theta$ |
| CANONICAL LINK | IDENTITY | LOGARITHM | LOGIT | RECIPROCAL |
| $V(\mu)$ | 1 | $\mu$ | $\mu(1 - \mu)$ | $\mu^2$ |

Table 1: Some examples of distribution with related quantities.

*Example* 5.1. Consider the Poisson distribution. Then, the probability mass function in exponential form equals

$$\begin{aligned} f(y) &= \mathrm{e}^{-\mu}\frac{\mu}{y!} = \exp\left\{-y\log\mu - \mu - \log y!\right\} \\ &= \exp\left\{-y\theta - \mathrm{e}^{\theta} - \log y!\right\}, \end{aligned}$$

where is apparent that

$$\begin{gathered} b(\theta) = \mathrm{e}^{\theta}, \qquad \mu = \frac{\mathrm{d}}{\mathrm{d}\theta}b(\theta) = \mathrm{e}^{\theta}, \qquad V(\mu) = \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}b(\theta) = \mu = \mathrm{e}^{\theta} \\ \psi = 1, \qquad \omega = 1, \qquad c(\psi, y) = -\log y!. \end{gathered}$$

## 5.1 Classification models

Classification represents one of the fundamental approaches in statistical modeling. As illustration, suppose that we want to model an elector vote, with two possible choices - party A or B. This is a classical classification problem, whereas taking one of the possible outcomes as reference, e.g. A, we want to model the success probability of casting the vote for A. Usually, we have some covariates that we want to use for explaining the success probability.

Ideally, it is reasonable to assume the vote to be distributed such a random variable having support $\{0, 1\}$, as a Bernoulli distribution

$$Y_i \sim Be(\theta_i),$$

with $\theta_i$ being the success probability. The success probability is a function of a vector of covariates, multiplied by a suitable vector of parameters, of the form

$$\theta_i = g^{-1}(\boldsymbol{x}_i^{\intercal}\boldsymbol{\beta}).$$

Specifically, different functions $g$ lead to different classification models. In the following, we consider a specific model for binary classification. Furthermore, we remark that generalizations to more than two labels are straightforward.

Among the possible choices, the first GLM we study here from a Bayesian perspective is the probit regression model. Such a model is obtained considering either a binomial or Bernoulli distribution for the data combined with a specific link function. The link function is not the canonical one. Instead, we consider the so called probit function. The model specification we are considering is given by

$$\underbrace{Y_i \mid \theta_i \overset{ind}{\sim} Be(\theta_i)}_{\text{error structure}}, \qquad \underbrace{\theta_i = \Phi(\eta_i)}_{\text{inverse link function}}, \qquad \underbrace{\eta_i = \boldsymbol{x}_i^{\intercal}\boldsymbol{\beta}}_{\text{linear predictor}}, \qquad i = 1, \ldots, n,$$

where $\Phi(\cdot)$ denotes the cdf of a standard normal distribution. $Y_i \in \{0, 1\}$ binary observations. $\eta_i$ is the linear predictor, convoluting the covariates domain ($\mathbb{R}^p$) to a real space. $\Phi$ is mapping a real space into $(0, 1)$. The Bernoulli distribution takes as argument a $(0, 1)$ value, which is playing the role of success rate.

The choice of the link function is mainly dictated by aiming for posterior tractability. In fact, with the previous model specification, the likelihood becomes

$$\begin{aligned}\mathrm{L}(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\beta}) &= \prod_{i=1}^{n} \Phi(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^{y_i} \left[1 - \Phi(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\right]^{1-y_i} \\ &= \prod_{i=1}^{n} \left\{\mathbf{1}_{[y_i=1]} \Phi(\boldsymbol{x}_i^\intercal \boldsymbol{\beta}) + \mathbf{1}_{[y_i=0]} \left[1 - \Phi(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\right]\right\},\end{aligned}$$

with $\mathbf{1}_{[\cdot]}$ denoting the indicator function. Assuming, e.g., a multivariate Gaussian prior $\pi(\boldsymbol{\beta})$ for the regression coefficients, a posteriori we have

$$\pi(\boldsymbol{\beta} \mid y_{1:n}, \boldsymbol{x}_{1:n}) = \frac{\pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \Phi(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^{y_i} \left[1 - \Phi(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\right]^{1-y_i}}{\int_{\mathbb{R}^p} \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \Phi(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^{y_i} \left[1 - \Phi(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\right]^{1-y_i} \mathrm{d}\boldsymbol{\beta}}.$$

At first, we don't recognize a known posterior distribution (recently Durante (2019) identifies a unifies skew normal). The normalization constant is intractable, but maybe we can do a simple trick to overcome the problem. A common strategy to deal with intractable models is to introduce an ancillary quantity that augment the model space, and once we condition on such a quantity the model becomes tractable. Specifically, we introduce a set of suitable unobserved latent variables $\{v_1, \ldots, v_n\}$, where the generic $v_i \in \mathbb{R}$, $i = 1, \ldots, n$. With those latent variables, we can rewrite the likelihood and (hopefully) simplify the problem.

Consider the following generative model

$$v_i = \boldsymbol{x}_i^\intercal \boldsymbol{\beta} + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} N(0,1), \qquad i = 1, \ldots, n,$$

and then we apply a simple transformation of the unobserved $v_i$s, with

$$y_i = \mathbf{1}_{[v_i > 0]}.$$

The augmented likelihood function associated with the probit regression model can be then written as

$$\mathrm{L}(y_{1:n}, v_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \phi(v_i \mid \boldsymbol{x}_i^\intercal \boldsymbol{\beta}, 1) \Big[\mathbf{1}_{[v_i>0]} \mathbf{1}_{[y_i=1]} + \mathbf{1}_{[v_i \leq 0]} \mathbf{1}_{[y_i=0]}\Big], \tag{1}$$

where $\phi(\cdot \mid \mu, \sigma^2)$ denotes the density function of a Gaussian distribution with expectation $\mu$ and variance $\sigma^2$.

**Proposition 5.2.** *For the augmented likelihood of Equation (1), the following marginalization holds true*

$$\int_{\mathbb{R}^n} \mathrm{L}(y_{1:n}, v_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \Phi(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^{y_i} \left[1 - \Phi(\boldsymbol{x}_i^\intercal \boldsymbol{\beta}).\right]^{1-y_i}$$

*Hence, we recover the likelihood of the probit model.*

*Proof.* We have

$$
\begin{aligned}
\int_{\mathbb{R}^n} & \mathrm{L}(y_{1:n}, v_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\beta}) \mathrm{d}v_1 \ldots \mathrm{d}v_n \\
&= \prod_{i=1}^{n} \int_{\mathbb{R}} \phi(v_i \mid \boldsymbol{x}_i^\intercal \boldsymbol{\beta}, 1)[\mathbf{1}_{[v_i \geq 0]} \mathbf{1}_{y_1=1]} + \mathbf{1}_{[v_i<0]} \mathbf{1}_{[y_i=0]}] \mathrm{d}v_i \\
&= \prod_{i=1}^{n} \left[ \int_{0}^{\infty} \phi(v_i \mid \boldsymbol{x}_i^\intercal \boldsymbol{\beta}, 1) \mathbf{1}_{y_1=1]} \mathrm{d}v_i + \int_{-\infty}^{0} \phi(v_i \mid \boldsymbol{x}_i^\intercal \boldsymbol{\beta}, 1) \mathbf{1}_{y_1=0]} \mathrm{d}v_i \right] \\
&= \prod_{i=1}^{n} \left[ \Phi(\boldsymbol{x}_i^\intercal \boldsymbol{\beta}) \mathbf{1}_{[y_i=1]} + \left(1 - \Phi(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\right) \mathbf{1}_{[y_i=0]} \right],
\end{aligned}
$$

which corresponds to the quantity of interest. Note that the last equality holds since

$$
\int_{0}^{\infty} \phi(v_i \mid \boldsymbol{x}_i^\intercal \boldsymbol{\beta}, 1) \mathrm{d}v_i = \int_{-\boldsymbol{x}_i^\intercal \boldsymbol{\beta}}^{\infty} \phi(v_i \mid 0, 1) \mathrm{d}v_i = \Phi(\infty) - \Phi(-\boldsymbol{x}_i^\intercal \boldsymbol{\beta}) = \Phi(\boldsymbol{x}_i^\intercal \boldsymbol{\beta}).
$$

□

Hence, in force of the augmentation, we can resort to what we studied about linear regression to perform inference with a probit model. We set a priori $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \Sigma_0)$. The posterior distribution is still a multivariate Gaussian distribution, $\boldsymbol{\beta} \mid v_{1:n}, \boldsymbol{x}_{1:n} \sim N(\boldsymbol{b}_n, \Sigma_n)$, with

$$
\Sigma_n = (\Sigma_0^{-1} + \mathrm{X}^\intercal \mathrm{X})^{-1}, \qquad \boldsymbol{b}_n = \Sigma_n (\Sigma_0^{-1} \boldsymbol{\beta}_0 + \mathrm{X}^\intercal \boldsymbol{v}).
$$

where $\mathrm{X}$ denotes the design matrix of the model. The model is now tractable, and we saw in the previous result how to produce posterior inference in this scenario. However, the posterior distribution is conditioned on covariates and augmented variables, and we do not observe the latter. We notice that the distribution of $v_i \mid y_i, \boldsymbol{x}_i, \boldsymbol{\beta}$ is a truncated Gaussian distribution, with

$$
f(v_i \mid y_i, \boldsymbol{x}_i, \boldsymbol{\beta}) \propto \begin{cases} \phi(v_i \mid \boldsymbol{x}_i^\intercal \boldsymbol{\beta}, 1) \mathbf{1}_{[v_i > 0]}, & \text{if } y_i = 1, \\ \phi(v_i \mid \boldsymbol{x}_i^\intercal \boldsymbol{\beta}, 1) \mathbf{1}_{[v_i \leq 0]}, & \text{if } y_i = 0. \end{cases}
$$

We can perform posterior inference implementing a Gibbs sampler which sequentially updates the augmented variables $v_i$s and the regression coefficients $\boldsymbol{\beta}$.

Commonly, we are interested into performing predictive inference. Assuming we observe the covariates for a future $n+1$, $\boldsymbol{x}_{n+1}$, but not the response variable, are interested into describing the behavior of $Y_{n+1}$. In the previous chapters we saw cases where the predictive distribution of $Y_{n+1}$ was available in closed form. Here, we can use the sampled values from the posterior distribution. Note that

$$
\begin{aligned}
f(y_{n+1} \mid \boldsymbol{x}_{n+1}, \boldsymbol{x}_{1:n}, y_{1:n}) &= \int_{\mathbb{R}^p} f(y_{n+1} \mid \boldsymbol{x}_{n+1}, \boldsymbol{\beta}) \pi(\boldsymbol{\beta} \mid y_{1:n}, \boldsymbol{x}_{1:n}) \mathrm{d}\boldsymbol{\beta} \\
&\approx \frac{1}{R} \sum_{r=1}^{R} f(y_{n+1} \mid \boldsymbol{x}_{n+1}, \boldsymbol{\beta}_r), \qquad \boldsymbol{\beta}_r \sim \pi(\boldsymbol{\beta}_r \mid y_{1:n}, \boldsymbol{x}_{1:n}).
\end{aligned}
$$

We can obtain a predictive sample, e.g., by sampling from each kernel function, $Y_{n+1} \mid \boldsymbol{\beta}_r \sim f(y_{n+1} \mid \boldsymbol{x}_{n+1}, \boldsymbol{\beta}_r)$. We can use the previous sample to perform predictive inference, such as point estimates and predictive intervals.

*Exercise* 5.3. Let us consider the following response variable and covariates

```
set.seed(123); betatrue <- c(-1, -2, 2)
z <- round(rnorm(100, 0, 1), digits = 1)
X <- cbind(rep(1, 100), z, z * z)
tempprobs <- pnorm(X \%*\% betatrue)
y <- sapply(tempprobs[,1], function(x) rbinom(1,1,x))
```

Consider a GLM with Bernoulli distribution for the response variable and logit link function, where the linear predictor is given by

$$\eta_i = \beta_1 + \beta_2 z_i + \beta_3 z_i^2.$$

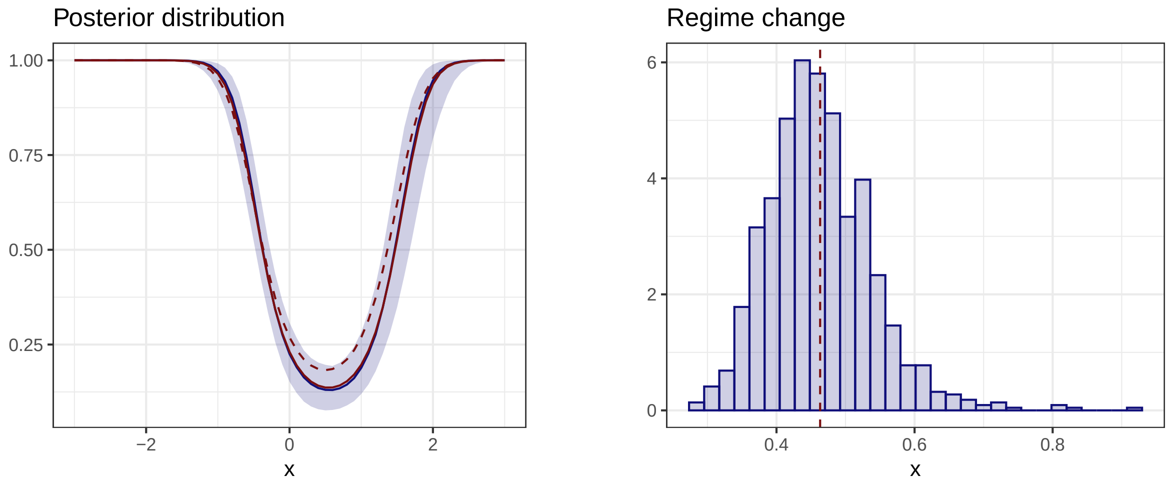
- Write down the Gibbs sampler to perform posterior inference with the Bayesian probit model.
- Produce a sample from the posterior distribution of size $1\,000$, after discarding $200$ observation as burn-in phase.
- Plot the marginal posterior distributions of the regression coefficients.
- Test if $\beta_2$ is significantly greater than $0$.
- Perform predictive inference for the $n+1$ observation, with $z_{n+1} = 2$.

Once we have a sample from the posterior distribution, we can consider functionals of the regression parameters, and perform inference on them. In the previous example, we can consider two functionals answering the following questions.

- How the model behaves over the covariate support? we can plot the posterior distribution over the model space (left plot) with its uncertainty quantification.
- Given the quadratic form, where is the model changing regime? We can plot the change point (right plot), with its uncertainty quantification, and perform tests on it. Note that

$$\frac{\theta}{1-\theta} = \mathrm{e}^{\beta_1 + \beta_2 x + \beta_3 x^2},$$

so that the change happens at $x_0 = -\frac{\beta_2}{2\beta_3}$.



## 5.2 COUNT RESPONSES AND POISSON REGRESSION MODEL

We now consider a GLM for count data, when data have no clear upper bounds. In this situation, a realistic assumption is a Poisson distribution to describe the data behavior.

*Example* 5.4. Let us consider a fabric production, where we are producing linen sheets of specific requested lengths. We are interested into modelling the expected number of defects $Y_i$ as function of the produced length $z_i$, for the generic $i$th produced piece.

A suitable model assumption is to consider some function of the produced length of the form

$$\mathbb{E}[Y_i \mid \boldsymbol{x}_i, ..] = \lambda_i = \alpha_1 x_i^{\alpha_2} = \mathrm{e}^{\log \alpha_1 + \alpha_2 \log z_i},$$

where $\alpha_1$ is a scalar term multiplying the length, i.e., the baseline expected number of defects when the length is equal to $1$, $\alpha_2$ is a stress parameter, as far the production is getting longer we can expect an increased number of defects, and the dispersion of $y_i$ increases as far $z_i$ is increasing. Ideally, this is an example of Poisson regression model with canonical link function.

Being more formal, we have as usual a sequence of observations and covariates, $\{y_i, \boldsymbol{x}_i\}$, for $i = 1, \ldots, n$. The data are now count data, assumed to be described by a Poisson distribution. The link function is the canonical one, which in this case corresponds to the logarithm. The model specification we are considering is the following

$$\underbrace{Y_i \mid \lambda_i \overset{ind}{\sim} Poi(\lambda_i)}_{\text{error structure}}, \qquad \underbrace{\lambda_i = \exp(\eta_i)}_{\text{inverse link function}}, \qquad \underbrace{\eta_i = \boldsymbol{x}_i^\intercal \boldsymbol{\beta}}_{\text{linear predictor}}, \qquad i = 1, \ldots, n.$$

$Y_i \in \mathbb{N}$ non-negative discrete observations. $\eta_i$ is the linear predictor, convoluting the covariates domain ($\mathbb{R}^p$) to a real space. The exponential function is mapping a real space into $\mathbb{R}_+$. The Poisson distribution takes as argument a $\mathbb{R}_+$ value, which is playing the role of expected count value.

We remark that the expectation of $Y_i$ can be written as

$$\mathbb{E}[Y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}] = \mathrm{e}^{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}} = \prod_{j=1}^{p} \mathrm{e}^{x_{i,j}\beta_j},$$

hence, the generic $\beta_j$ coefficient has an exponential-multiplicative effect on the expected count, as far $x_{i,j}$ increases by a unit value. Under the previous model assumption, the likelihood function becomes

$$\mathrm{L}(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\mathrm{e}^{-\lambda_i} \lambda_i^{y_i}}{y_i!} = \prod_{i=1}^{n} \frac{\mathrm{e}^{-\mathrm{e}^{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}}} \left(\mathrm{e}^{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}}\right)^{y_i}}{y_i!}.$$

Given a vector of observed covariates, each term contributing in the likelihood function is the pmf of a Poisson distribution with expectation and variance equal to $\mathrm{e}^{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}}$. As far as the magnitude of the product between covariates and regression coefficients increases, we expect higher counts and dispersion. We do not recognize any suitable augmentation for the previous likelihood. Inference can be done resorting to computational approaches, such as implementing the model in STAN and sampling with an Hamiltonian Monte Carlo. The previous model can be extended in many directions, e.g., zero-inflated model and over-dispersed model (mixture of Poisson).

*Exercise* 5.5. Let us consider the following response variable and covariates

```
set.seed(123); betatrue <- c(-1, -2, 2)
z1 <- round(rnorm(100, 0, 1), digits = 1)
z2 <- round(rnorm(100, 0, 1), digits = 1)
X <- cbind(rep(1, 100), z1, z2)
tempprobs <- exp(X \%*\% betatrue)
y <- sapply(tempprobs[,1], function(x) rpois(1,x))
```

Consider a GLM with Poisson distribution for the response variable and log link function, where the linear predictor is given by

$$\eta_i = \beta_1 + \beta_2 z_{i,1} + \beta_3 z_{i,2}.$$

- Write down the STAN model to perform posterior inference.
- Produce a sample from the posterior distribution of size $1\,000$, after discarding $1\,000$ observation as burn-in phase.
- Plot the marginal posterior distributions of the regression coefficients.