BSM2 - Linear models and Bayes

Lecturer: Riccardo Corradin

Introduction

Welcome back linear models!

Linear models are one of the fundamental and most commonly used techniques in data analysis. Despite their **simplicity**, they are a flexible model which can **help a statistician** in many situations.

- Let $Y_1, \ldots, Y_n$, 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 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ a set of covariates (idependent variables), with the generic $\boldsymbol{x}_i \in \mathbb{X} \subseteq \mathbb{R}^p$.

As usual in regression problems, we want to find a function of the covariates ans some parameter, say $\boldsymbol{\beta}$, $g(\boldsymbol{x}, \boldsymbol{\beta})$ which describes the response variable Y.

Here, at first, we restrict our attention to the case where such a model is of the form

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

Recall that: linear models are a linear combination of covariates and parameters, but response variable and covariates can be possibly transformed non-linearly.

$$\beta_1 x_1 \sin(\beta_2 x_2) \qquad \text{non linearizable,}$$

$$x_1^{\beta_1} \mathrm{e}^{\beta_2 x_2} \to \beta_1 \log(x_1) + \beta_2 x_2 \qquad \textit{linearizable.}$$

Introduction

In **frequentist** the goal is to find a model which is **optimal** with respect to some **loss function**.

Hence, we assume that the response variables can be decomposed additively as

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

where ε_i is an **error term** for which (i) $\mathrm{E}[\varepsilon_i] = 0$, (ii) $\mathrm{var}(\varepsilon_i) = \sigma^2$, and (iii) $\mathrm{cov}(\varepsilon_i, \varepsilon_\ell) = 0$ for $i \neq \ell$.

$\rightarrow$ (i) implies $\mathrm{E}[Y_i] = \boldsymbol{x}_i^\intercal \boldsymbol{\beta}$.

Usually, the loss function is build to **minimize some distance** between the **observed** response variables $y_1, \ldots, y_n$ and the model **fitted** values, of the kind

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

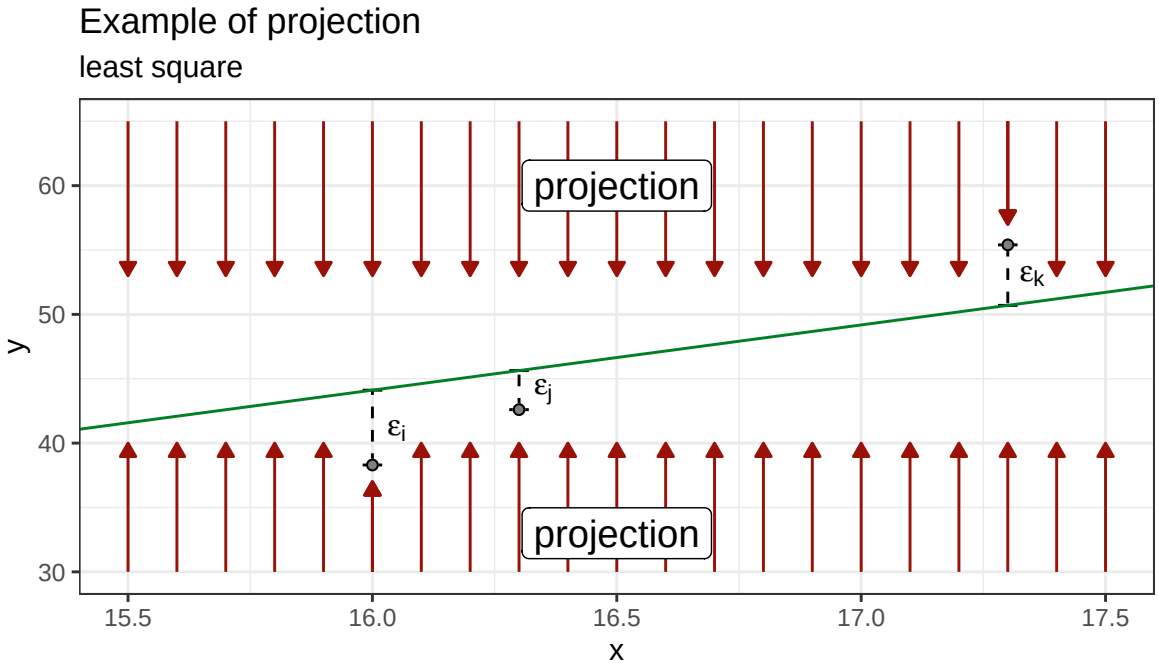
where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\intercal$, $\boldsymbol{y} = (y_1, \ldots, y_n)^\intercal$ and X is the design matrix, whose ith row is $\boldsymbol{x}_i^\intercal$.

Assuming the design matrix X **full rank** p, it is easy to prove that the OLS estimate corresponds to

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

Introduction

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 assuming a Gaussian distribution for the error terms, i.e.

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

- 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.

Introduction

A natural extension of the previous model include a **regularization** term, that penalize the regression coefficient values.

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 λ 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 λ tuning parameter. There is no closed form for the lasso regression coefficient.

The previous strategies allow for **regularization** of the regression coefficients, shrinking their values toward the origin.

$\rightarrow$ This balances for unfriendly behaviour, overfitting, and allows for $p > n$ estimates.

Bayesian linear regression

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**. Here, we consider that

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

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**.
- The model specification is completed by selecting a suitable prior distribution.

Bayesian linear regression

A first prior specification can be to consider a **uniform prior** for $(\boldsymbol{\beta}, \log \sigma^2)$. Hence, we set

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

- 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.

The previous prior specification leads to a posterior distribution of the form

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

- $\boldsymbol{\beta} \mid \sigma^2, y_{1:n}, \boldsymbol{x}_{1:n}$ is a **multivariate Gaussian distribution**, $\sim N(\hat{\boldsymbol{\beta}}_{ML}, (\mathrm{X}^\intercal \mathrm{X})^{-1} \sigma^2)$,

$$\pi(\boldsymbol{\beta} \mid \sigma^2, y_{1:n}, \boldsymbol{x}_{1:n}) = \left(2\pi\sigma^2\right)^{-\frac{p}{2}} \det((\mathrm{X}^\intercal \mathrm{X})^{-1})^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ML})^\intercal \frac{(\mathrm{X}^\intercal \mathrm{X})^{-1}}{\sigma^2} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ML})}.$$

- $\sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n}$ is an **inverse-gamma**, $IG(\frac{n-p}{2}, \frac{1}{2}(\boldsymbol{y} - \mathrm{X}\hat{\boldsymbol{\beta}}_{ML})^\intercal (\boldsymbol{y} - \mathrm{X}\hat{\boldsymbol{\beta}}_{ML}))$,

$$\pi(\sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n}) = \frac{\left[\frac{1}{2}(\boldsymbol{y} - \mathrm{X}\hat{\boldsymbol{\beta}}_{ML})^\intercal (\boldsymbol{y} - \mathrm{X}\hat{\boldsymbol{\beta}}_{ML})\right]^{\frac{n-p}{2}}}{\Gamma\left(\frac{n-1}{2}\right)(\sigma^2)^{\frac{n-p}{2}+1}} e^{-\frac{(\boldsymbol{y} - \mathrm{X}\hat{\boldsymbol{\beta}}_{ML})^\intercal (\boldsymbol{y} - \mathrm{X}\hat{\boldsymbol{\beta}}_{ML})}{2\sigma^2}}.$$

Bayesian linear regression

Bayesian linear regression

Withe previous **prior specification** we have the followings.

- The expected value of $\sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n}$ is

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

  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

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

Suppose we are interested in **predictive inference** for a **future observation** with covariates $_{n+1}$.

- The expectation of $Y_{n+1} \mid \sigma^2, \boldsymbol{x}_{n+1}$ is given by

$$\mathrm{E}[Y_{n+1} \mid \sigma^2, \boldsymbol{x}_{n+1}] = \mathrm{E}[\mathrm{E}[Y_{n+1} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{x}_{n+1}] \mid \sigma^2, \boldsymbol{x}_{n+1}] = \boldsymbol{x}_{n+1}^{\mathsf{T}}\boldsymbol{\beta}.$$

- The variance of $Y_{n+1} \mid \sigma^2, \boldsymbol{x}_{n+1}$ is given by

$$\begin{aligned}\mathrm{var}[Y_{n+1} \mid \sigma^2, \boldsymbol{x}_{n+1}] &= \mathrm{E}[\mathrm{var}(Y_{n+1} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{x}_{n+1}) \mid \sigma^2, \boldsymbol{x}_{n+1}] \\ &\quad + \mathrm{var}(\mathrm{E}[Y_{n+1} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{x}_{n+1}] \mid \sigma^2, \boldsymbol{x}_{n+1}) \\ &= (1 + \boldsymbol{x}_{n+1}^{\mathsf{T}}(X^{\mathsf{T}}X)^{-1}\boldsymbol{x}_{n+1})\sigma^2.\end{aligned}$$

Bayesian linear regression

Example

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.

A more informative prior specification

A more informative prior specification

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

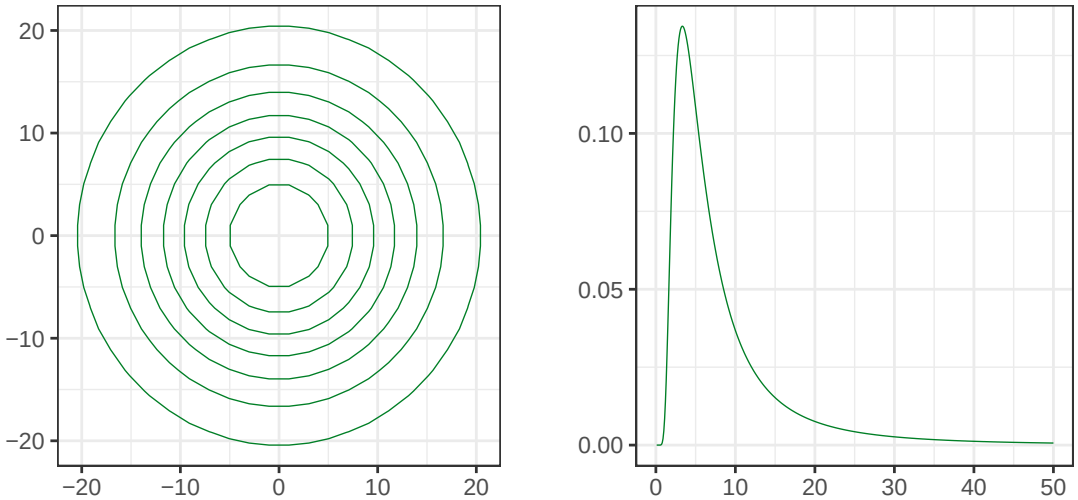
$$\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\boldsymbol{\beta} \mid \sigma^2)\pi(\sigma^2).$$

In particular, we set

- $\boldsymbol{\beta} \mid \sigma^2 \sim N(\boldsymbol{b}_0, \sigma^2 \Sigma_0)$, a **multivariate Gaussian** distribution;
- $\sigma^2 \sim IG(a_0, b_0)$, an **inverse-gamma** distribution.

- $\boldsymbol{b}_0$ describes where we center a priori our guess for the regression coefficients, usually set as $\boldsymbol{b}_0 = \mathbf{0}$. However, if we have prior information on the regression coefficients, we can include that in the model specification.
- Σ_0, conditionally on σ^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 σ^2.

A more informative prior specification



About the prior specification

- $\boldsymbol{b}_0$ is set usually equal to $\mathbf{0}$.
- Σ_0 is commonly set as a diagonal matrix, with no correlation among different effects a priori.
  - $\rightarrow$ Any dependence among the regression coefficients is driven by the empirical information.
  - $\rightarrow$ In some scenarios it can be useful to set Σ_0 not diagonal.
- To specify a_0 and b_0 we can look at the prior expectation, which is

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

  - $\rightarrow$ $a_0 > 1$, and weight the prior belief.
  - $\rightarrow$ Once we fix a_0, we can choose b_0 that guarantee the prior expectation we want for σ^2.

A more informative prior specification

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

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

where

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

with

$$\begin{aligned}
\Sigma_n &= \left[\Sigma_0^{-1} + (\mathrm{X}^\intercal \mathrm{X}) \right]^{-1} \\
\boldsymbol{b}_n &= \Sigma_n \left[\Sigma_0^{-1} \boldsymbol{b}_0 + (\mathrm{X}^\intercal \mathrm{X}) \hat{\boldsymbol{\beta}}_{ML} \right] \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2} \left(\boldsymbol{y}^\intercal \boldsymbol{y} - \boldsymbol{b}_n^\intercal \Sigma_n^{-1} \boldsymbol{b}_n + \boldsymbol{b}_0^\intercal \Sigma_0^{-1} \boldsymbol{b}_0 \right)
\end{aligned}$$

Remark: the previous prior choice is conjugate to the Gaussian likelihood for linear regression.

A more informative prior specification

A more informative prior specification

A more informative prior specification

Some **remarks** on the posterior characterization.

- If $\boldsymbol{\beta}_{ML}$ is the **maximum likelihood estimator** of a linear model with Gaussian error $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathrm{I}_n)$, then

$$\mathrm{var}(\boldsymbol{\beta}_{ML}) = \sigma^2 (\mathrm{X}^\intercal \mathrm{X})^{-1}.$$

  Recall that the matrix term in the variance of $\boldsymbol{\beta} \mid y_{1:n}, \boldsymbol{x}_{1:n}, \sigma^2$ equals

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

  Such quantity is averaging the reciprocal of the **prior matrix term** Σ_0 and the **maximum likelihood** estimator variance **matrix term**, conditionally on σ^2.

- Similarly, we have that

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

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

A more informative prior specification

- A posteriori, the **shape parameter** of the inverse-gamma distribution becomes

$$a_n = a_0 + \frac{n}{2}.$$

We said that a_0 can be interpreted as sort of ***prior sample size***, i.e. how strongly we trust the prior guess on σ^2.

→ If we have a sample of size n and we want to weight the empirical measure on σ^2 a posteriori by $q \times 100\%$, we simply set

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

- For the **scale parameter**, a posteriori we recall that

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

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 $\frac{1}{2}$ balances the scale parameter posterior adjustment, i.e. $\frac{n}{2}$.

- Setting a prior on $(\boldsymbol{\beta}, \sigma^2)$ equals to express a prior guess on the **projection** of observations onto the **model space** and a prior on the **residual dispersion**.

A more informative prior specification

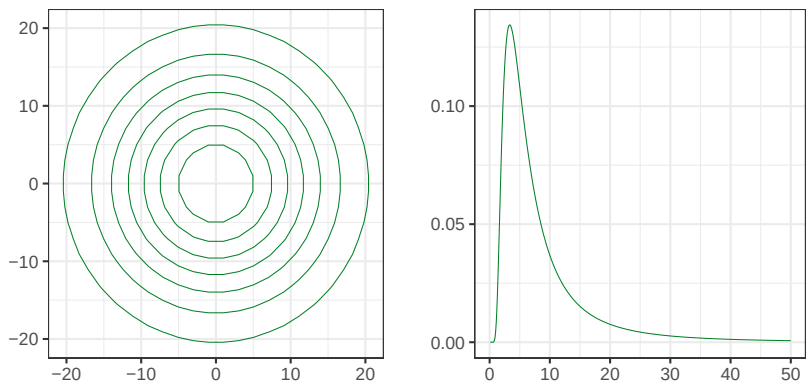
We consider a simple example to show the **effect of empirical information** on **posterior computation**.

We sample a set of covariates $Z_i \sim N(0,1)$, $i = 1, \ldots, 50$, and

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

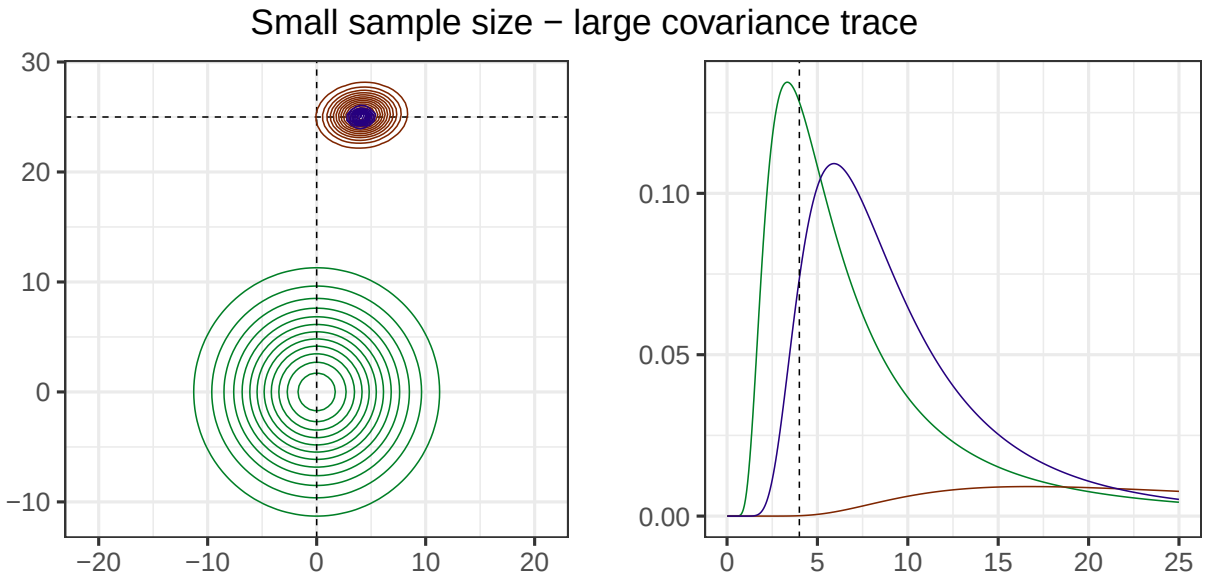
We set $\boldsymbol{x}_i^{\mathsf{T}} = (1, z_i)$. We consider a model as

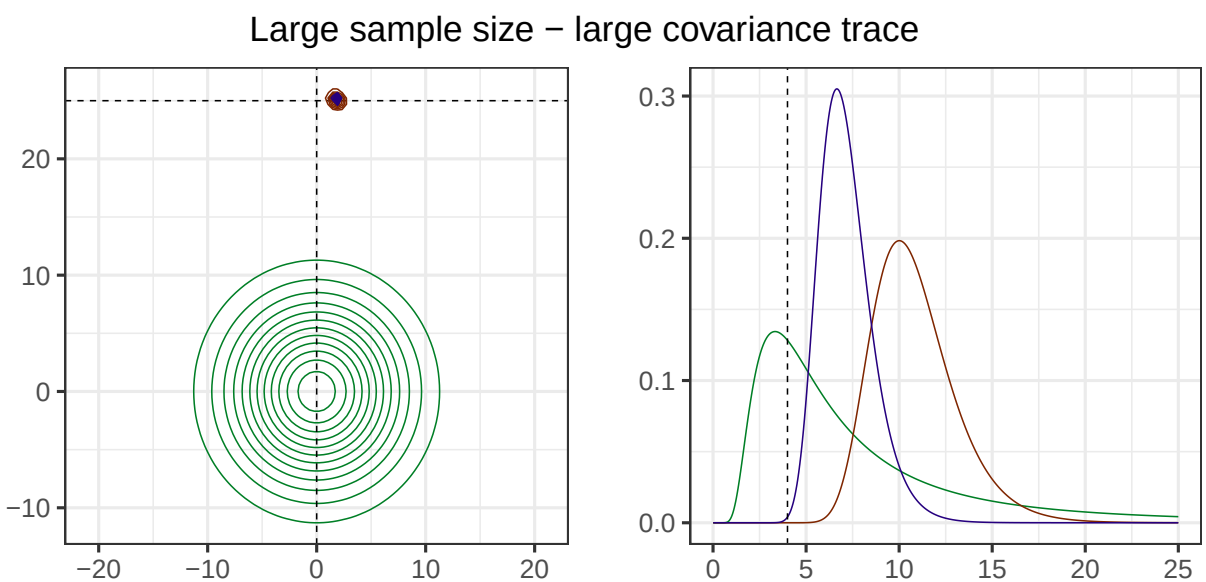
$$\begin{aligned} Y_i &\sim N(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}, \sigma^2), \ i = 1, \ldots, 50, \\ \boldsymbol{\beta} \mid \sigma^2 &\sim N(\boldsymbol{b}_0, \Sigma_0), \\ \sigma^2 &\sim IG(a_0, b_0). \end{aligned}$$



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

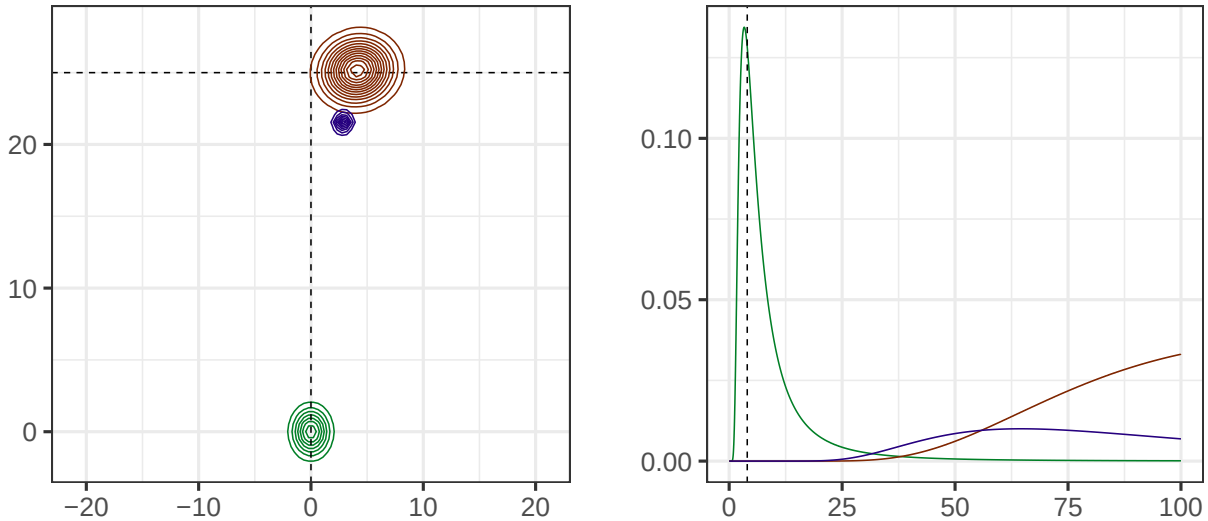
A more informative prior specification



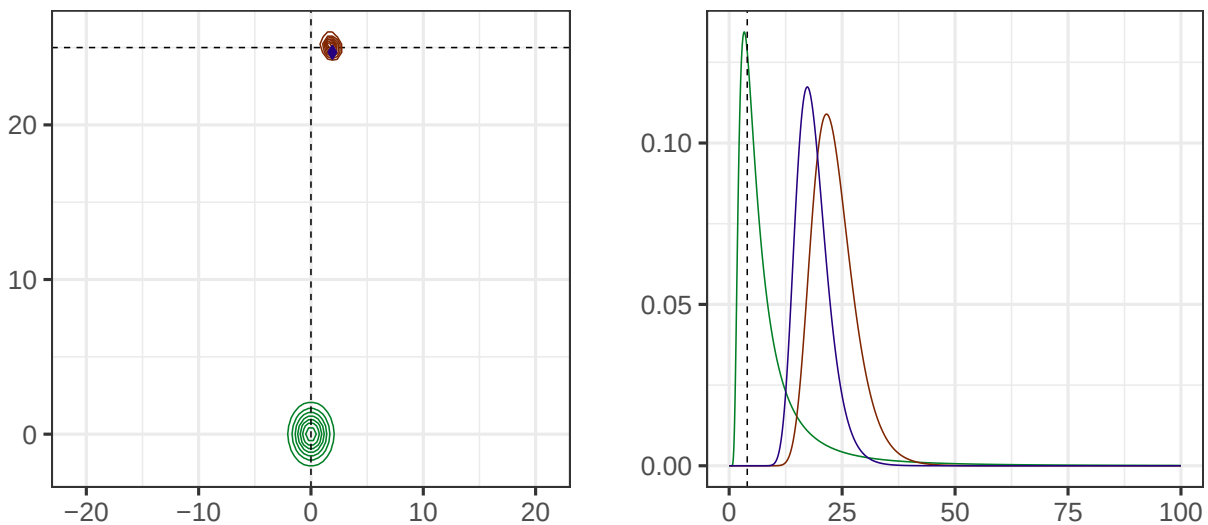


A more informative prior specification

Small sample size − small covariance trace



Large sample size − small covariance trace



A more informative prior specification

Example

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 $\boldsymbol{\beta} \mid \sigma^2 \sim N(\boldsymbol{b}_0, \sigma^2 \Sigma_0)$ and $\sigma^2 \sim IG(a_0, b_0)$. Write the model in STAN when $\boldsymbol{b}_0 = \mathbf{0}$, $\Sigma_0 = 10^2 \mathrm{I}_p$, $a_0 = 3$ and $b_0 = 2$.
- Compare the inference you obtain with the non-informative prior case.

Asymptotic behavior

Asymptotic behavior

Asymptotic analysis is a **tedious topic** in **Bayesian inference**.

→ We need to assume the existence of a true parameter generating our data.
→ Usually, is a frequentist analysis of posterior inference.

In general, under some **regularity conditions**, assuming a true parameter $\boldsymbol{\theta}_0 \in \Theta \subseteq \mathbb{R}^d$, the posterior distribution concentrates **nicely** in a neighborhood of such parameter.

Theorem

Under certain regularity conditions, mainly on the likelihood smoothness and its behavior, let $\hat{\boldsymbol{\theta}}_{ML}$ be the maximum likelihood estimator of $\boldsymbol{\theta}$, and $\boldsymbol{\theta}_0$ the true value of the parameter. Then, for any prior $\pi(\boldsymbol{\theta})$ which is continuous and positive at $\boldsymbol{\theta}_0$,

$$\mathrm{P}\left(\lim_{n\to\infty} \int \left|\pi(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) - \frac{\sqrt{|nI(\boldsymbol{\theta}_0)|}}{(2\pi)^{d/2}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\intercal (nI(\boldsymbol{\theta}_0))(\boldsymbol{\theta}-\boldsymbol{\theta}_0)}\right| \mathrm{d}\boldsymbol{\theta}\right) = 1,$$

where $I(\boldsymbol{\theta}_0)$ denotes the Fisher information matrix of the generic $f(\boldsymbol{y}_i \mid \boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}_0$.

While the previous theorem is part of fundamental asymptotic results in Bayesian statistics, the purposes of this module are more related to modeling approaches.

If you are interested, further details on the assumptions and impact of the previous theorem are deferred to the end of this slides.

Asymptotic behavior

In out context, we have have $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ and

$$\begin{aligned}\ell(\boldsymbol{y}_{1:n} \mid \mathrm{X}, \boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta})^\mathsf{T} (\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}) \\ &= -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (\boldsymbol{y}^\mathsf{T}\boldsymbol{y} - 2\boldsymbol{\beta}^\mathsf{T}\mathrm{X}^\mathsf{T}\boldsymbol{y} + \boldsymbol{\beta}^\mathsf{T}\mathrm{X}^\mathsf{T}\mathrm{X}\boldsymbol{\beta})\end{aligned}$$

then we have

Asymptotic behavior

Posterior inference

Point estimates

Posterior point estimates can be obtained, as done in the first slide block, with a **decision theory approach** by assuming suitable **loss functions**.

- For the **regression coefficients**, the posterior distribution **conditionally on** σ^2 is a multivariate Gaussian distribution, i.e. symmetric.

  → All the estimates obtained with quadratic, linear and $0-1$ loss functions coincide with

  $$\hat{\boldsymbol{\beta}} = \mathrm{E}[\boldsymbol{\beta} \mid y_{1:n}, \boldsymbol{x}_{1:n}, \sigma^2] = \boldsymbol{b}_n = \Sigma_n \left[\Sigma_0^{-1}\boldsymbol{b}_0 + (\mathrm{X}^\intercal \mathrm{X})\hat{\boldsymbol{\beta}}_{ML}\right]$$

  with $\Sigma_n = [\Sigma_0^{-1} + (\mathrm{X}^\intercal \mathrm{X})]^{-1}$.

  Note that the previous **does not depend on** σ^2.

- For the **variance**, the posterior distribution is an inverse-gamma, not symmetric, with shape and scale parameters

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

  → With a quadratic loss function $\hat{\sigma^2} = \mathrm{E}[\sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n}] = \frac{b_n}{a_n - 1}$ if $a_n > 1$.

  → With a linear loss function, no closed form.

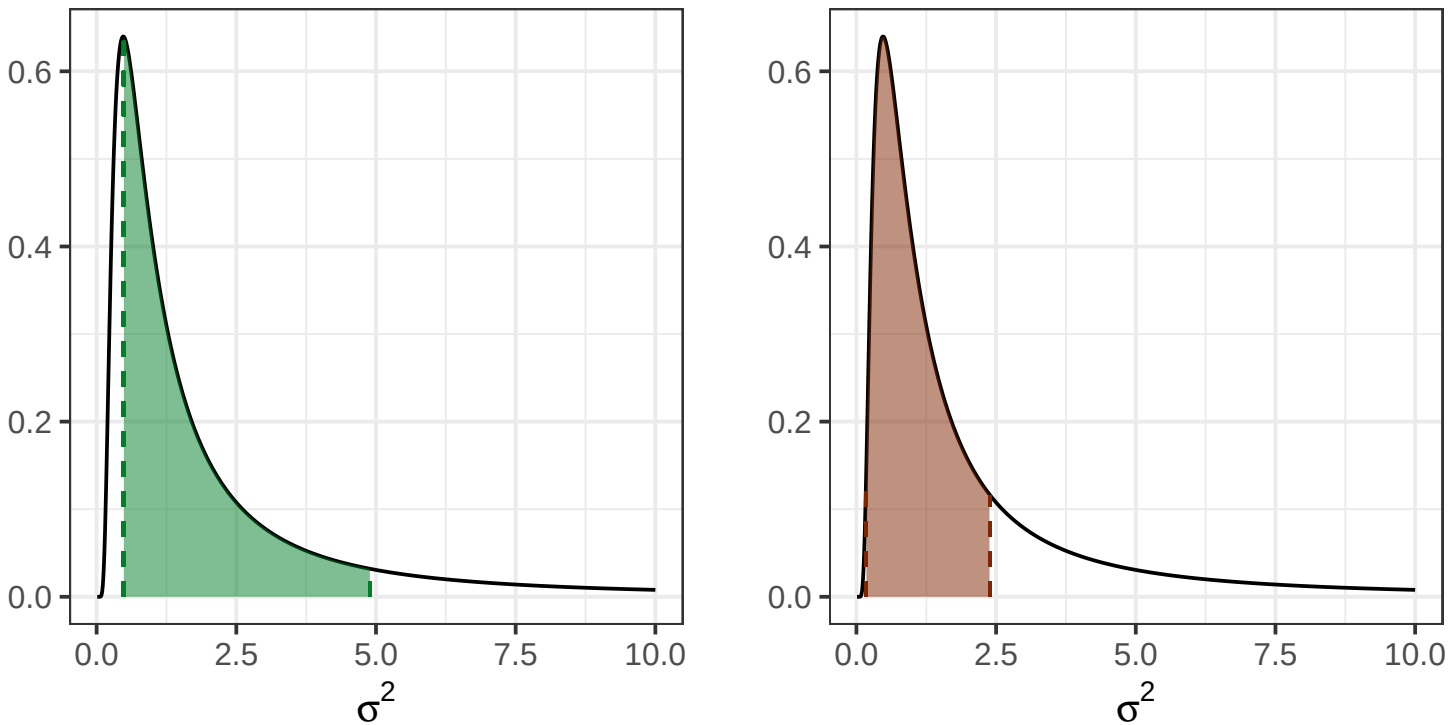
  → With a $0-1$ loss function $\hat{\sigma}^2 = \underset{\sigma^2 \in \mathbb{R}_+}{\arg\max}\left\{\pi(\sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n})\right\} = \frac{b_n}{a_n + 1}$.

Uncertainty quantification

We are Bayesian, we want to provide also suitable **uncertainty quantification summaries** along with our point estimate.

$\rightarrow$ 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.



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

Uncertainty quantification

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

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

We can marginalize out σ^2 from the joint distribution of $(\boldsymbol{\beta}, \sigma^2)$.

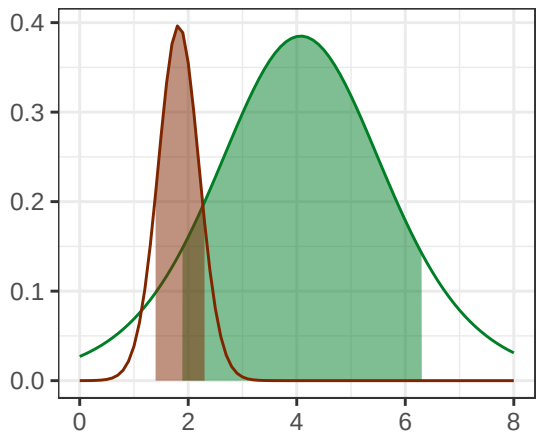
Uncertainty quantification

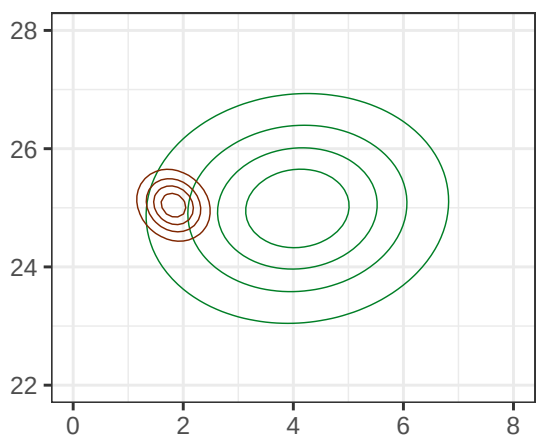
Uncertainty quantification

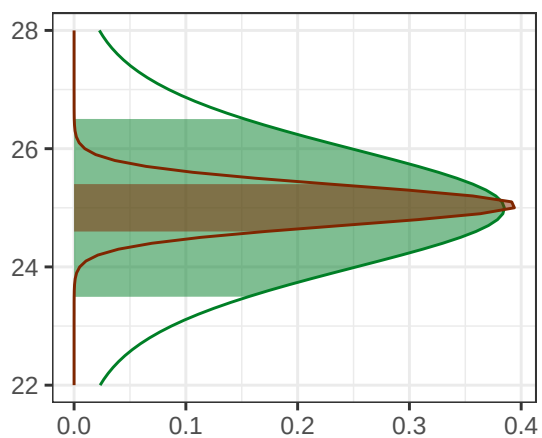
Synthetic data $Z_i \sim N(0,1)$, $Y_i \sim N(25z_i + z_i^2, 4)$, $i = 1, \ldots, 50$.

We consider a **model** as $Y_i \sim N(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}, \sigma^2)$, with $\boldsymbol{x}_i^\mathsf{T} = (1, z_i)$, $i = 1, \ldots, 50$, and **priors** $\boldsymbol{\beta} \mid \sigma^2 \sim N(\boldsymbol{b}_0, \Sigma_0)$, $\sigma^2 \sim IG(a_0, b_0)$.

Green lines/areas consider the first 3 observations, red ones with the whole sample.







Tests

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.

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.

We 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 jth parameter to be equal to c.

We can simplify a bit here ...

Tests

First, we need the marginal distribution, that we have almost for free.

Tests

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

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

and $\boldsymbol{x}_i^{(c)}$, $i = 1, \ldots, n$, being the ith **vector of covariates** without the jth element.

Hence, the 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

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

Tests

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

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

- 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.
  - $\rightarrow$ Which is measuring how likely are the data under such a model assumption.
  - $\rightarrow$ The construction works also with different model families.
  - $\rightarrow$ If the marginal is not available in a closed form (e.g. next slide block), we can use a numerical approximation of such quantity.

Example

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

- Consider only the first three observation. Test if β_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.

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

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

Predictive inference

Predictive inference

Example

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 with $\boldsymbol{b}_0 = \mathbf{0}$, $\Sigma_0 = \mathrm{diag}_3(25)$, $a_0 = 2$, $b_0 = 5$.
  - 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 $\boldsymbol{b}_0 = \mathbf{0}$, $\Sigma_0 = \mathrm{diag}_3(5)$, $a_0 = 2$, $b_0 = 5$.
  - 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?

Shrinkage with Bayesian regression models

Shrinkage with Bayesian regression models

Let us fix the value of σ^2. Looking at the conjugate prior for $\boldsymbol{\beta}$ with σ^2 fixed, it is implicitly inducing a **regularization** on the regression coefficients. We set

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

$$\boldsymbol{\beta} \sim N\left(\boldsymbol{b}_0, \frac{\sigma^2}{\lambda} I_p\right).$$

Then

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

Looking at the maximum a posteriori estimate, we have

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

→ When $\boldsymbol{b}_0 = \mathbf{0}$, the previous expression is the **usual ridge regression** loss function.

→ λ is a penalty term which drives the regularization, we can relax the model specification by setting $\lambda \sim \mathit{gamma}(\tau, \zeta)$.

Shrinkage with Bayesian regression models

We can set an **alternative prior specification** for $\boldsymbol{\beta}$ which is inducing a lasso regularization, by setting $\beta_j \sim Lap(b_{0,j}, \sigma^2/\lambda)$, with

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

Then, the joint posterior of $\boldsymbol{\beta}$ is proportional to

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

Looking at the maximum a posteriori estimate, we obtain

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

→ When $\boldsymbol{b}_0 = \mathbf{0}$, the previous expression is the **usual lasso regression** loss function.

→ λ 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.

Shrinkage with Bayesian regression models

Example

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

Do you see any difference?

Appendix

Assumptions for posterior asymptotic coverage

The theorem shown in the previous slides about posterior asymptotic works under the following assumptions.

(A1) The set $\{\boldsymbol{y} : f(\boldsymbol{y} \mid \boldsymbol{\theta}) > 0\}$ is the same for all $\boldsymbol{\theta} \in \Theta$.

(A2) The log-density $\log f(\boldsymbol{y} \mid \boldsymbol{\theta})$ is thrice differentiable w.r.t. $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$. Further, $\mathrm{E}\left[\mathrm{d}/\mathrm{d}\boldsymbol{\theta} \log f(\boldsymbol{y} \mid \boldsymbol{\theta})\right]$ and $\mathrm{E}\left[\mathrm{d}^2/\mathrm{d}\boldsymbol{\theta}^2 \log f(\boldsymbol{y} \mid \boldsymbol{\theta})\right]$ are both finite, and

$$\sup_{||\boldsymbol{\theta}-\boldsymbol{\theta}_0||<\delta} \left| \frac{\mathrm{d}^3}{\mathrm{d}\boldsymbol{\theta}^3} \log f(\boldsymbol{y} \mid \boldsymbol{\theta}) \right| \leq M(\boldsymbol{y}), \quad \text{and } \mathrm{E}[M(\boldsymbol{Y})] < \infty.$$

(A3) Interchange of integration and differentiation is justified at $\boldsymbol{\theta}_0$, so that

$$\mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}} \log f(\boldsymbol{y} \mid \boldsymbol{\theta})\right] = 0, \quad \mathrm{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}\boldsymbol{\theta}^2} \log f(\boldsymbol{y} \mid \boldsymbol{\theta})\right] = \mathrm{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}} \log f(\boldsymbol{y} \mid \boldsymbol{\theta})\right)^2\right].$$

The Fisher information matrix $I(\boldsymbol{\theta}) = \mathrm{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}} \log f(\boldsymbol{y} \mid \boldsymbol{\theta})\right)^2\right]$ is positive definite.

(A4) For any $\delta > 0$,

$$\mathrm{P}\left(\sup_{||\boldsymbol{\theta}-\boldsymbol{\theta}_0||<\delta} \frac{1}{n}\Big[\ell(\boldsymbol{y}_{1:n} \mid \boldsymbol{\theta}) - \ell(\boldsymbol{y}_{1:n} \mid \boldsymbol{\theta}_0)\Big] < -\epsilon\right) = 1$$

for some $\epsilon > 0$ and large n, where $\ell(\boldsymbol{y}_{1:n} \mid \boldsymbol{\theta})$ denotes the log-likelihood function of $\boldsymbol{y}_{1:n}$ at $\boldsymbol{\theta}$.

Further details and a sketch of proof can be found in "An Introduction to Bayesian Analysis: Theory and Methods" by Ghosh, Delampady, and Samanta.