University of Cape Town

Department of Statistical Sciences

Introduction to Bayesian Analysis

Assignment 1

Author:
Christopher Eason
Phillip Liu

Student Number:
ESNCHR001
LXXZHA003

April 14, 2025

Contents

Plagiarism Declaration:

1. I know that plagiarism is a serious form of academic misconduct.
2. I have read the document about avoiding plagiarism, am familiar with its contents and have avoided all forms of plagiarism mentioned there.
3. Where I have used the words of others, I have indicated this by the use of quotation marks.
4. I have referenced all quotations and properly acknowledged ideasborrowed from others.
5. I have not and shall not allow others to plagiarise my work.
6. I declare that this is my own work.
7. I am attaching the summary of the Turnitin match overview (when required to do so).

Signature:

ALL RELEVANT CODE IS INCLUDED IN SEPARATE QMD FILES

1 Question One: Linear Regression with Gibbs Sampling

Consider the linear regression model, $y_i = \boldsymbol{x_i}^T\boldsymbol{\beta} + e_i$ where e_i ∼N(0,σ^2) Assume that we have the following priors:

$$[\boldsymbol{\beta} \mid \sigma^2] \sim N_{k+1}(\tilde{\beta}, \sigma^2\mathbf{M})$$
$$[\sigma^2] \;\sim IG(a, b)$$

where $\tilde{\boldsymbol{\beta}} = \mathbf{0}$, $\mathbf{M} = \boldsymbol{I_{k+1}}$ and a = b = 1. Using the assumptions from eq. (A1), (clearly) show that the conditional posterior distributions are given as:

$$\left[\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{y}, \boldsymbol{X}\right] \sim \mathcal{N}_{k+1}\left(\boldsymbol{\mu_\beta}, \sigma^2\left(\mathbf{M} + \boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)$$
$$\left[\sigma^2 \mid \boldsymbol{y}, \boldsymbol{X}\right] \sim \mathcal{IG}\left(a + \tfrac{n}{2}, b + \tfrac{A_2}{2}\right)$$

where

$$\boldsymbol{\mu_\beta} = \left(\mathbf{M} + \boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\left(\boldsymbol{X}^T\boldsymbol{X}\right)\hat{\boldsymbol{\beta}} + \mathbf{M}\hat{\boldsymbol{\beta}}\right)$$
$$A_2 = \boldsymbol{y}^T\boldsymbol{y} + \hat{\boldsymbol{\beta}}^T\mathbf{M}\hat{\boldsymbol{\beta}} - \mu_{\boldsymbol{\beta}}^T\left(\mathbf{M} + \boldsymbol{X}^T\boldsymbol{X}\right)\mu_{\boldsymbol{\beta}}.$$

Note that $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$.
Hints for $\left[\sigma^2 \mid \boldsymbol{y}, \boldsymbol{X}\right]$

- $\left[\sigma^2 \mid \boldsymbol{y}, \boldsymbol{X}\right] = \int_{\boldsymbol{\beta}}\left[\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, \boldsymbol{X}\right] d\beta$
- $|a\boldsymbol{A}| = a^k|\boldsymbol{A}|$, where $\boldsymbol{A}$ is a k by k matrix.

Response Vector

$$\boldsymbol{y}_{n\times 1} = \begin{bmatrix} y_1 = \boldsymbol{x}_1^T\boldsymbol{\beta} + e_{11} \\ y_2 = \boldsymbol{x}_2^T\boldsymbol{\beta} + e_{12} \\ \vdots \\ y_{n_1} = \boldsymbol{x}_{n_1}^T\boldsymbol{\beta} + e_{1n_1} \end{bmatrix}_{n\times 1}$$

Predictor Variables

$$\boldsymbol{X}_{n\times(k+1)} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n_1 1} & x_{n_1 2} & \dots & x_{n_1 k} \end{bmatrix}_{n\times(k+1)}$$

Regression Coefficients

$$\boldsymbol{\beta}_{(k+1)\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}_{(k+1)\times 1}.$$

Mean for Beta Prior

$$\tilde{\boldsymbol{\beta}}_{(k+1)\times 1} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{(k+1)\times 1}.$$

Variance for Beta Prior

$$\boldsymbol{M}_{(k+1)\times(k+1)} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & & \dots & 1 \end{bmatrix}_{(k+1)\times(k+1)}$$

Error Vectors

$$\boldsymbol{e}_{n\times 1} = \begin{bmatrix} e_1 \sim N(0,\sigma^2) \\ e_2 \sim N(0,\sigma^2) \\ e_3 \sim N(0,\sigma^2) \\ \vdots \\ e_n \sim N(0,\sigma^2) \end{bmatrix}_{n\times 1}.$$

Complete Model

Given the error structure specified, the distribution of the response vector is:

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

Therefore,

$$f(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\sigma^2\mathbf{I}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\sigma^2\mathbf{I})^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right\}$$

Derivation of the Joint Likelihood

We start with a multivariate normal model for the data:

$$\boldsymbol{y} \sim \mathcal{N}_n\left(\boldsymbol{X}\boldsymbol{\beta},\, \boldsymbol{\Sigma}\right),$$

so that the joint likelihood is given by

$$\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\Sigma};\boldsymbol{y},\boldsymbol{X}) = \frac{1}{(2\pi)^{\frac{n}{2}}|(\sigma^2\mathbf{I})|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)^T(\sigma^2\mathbf{I})^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)\right\}$$

1.1 Deriving the Conditional Posterior Distribution for $\boldsymbol{\beta}$

Since

$$[\boldsymbol{\beta}|\sigma^2] \sim N_{k+1}(\tilde{\boldsymbol{\beta}}, \sigma^2\mathbf{M})$$

The prior distribution for $\boldsymbol{\beta}$ is given as:

$$\pi(\boldsymbol{\beta}|\sigma^2,\boldsymbol{y}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\sigma^2\mathbf{M}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^T(\sigma^2\mathbf{M})^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\right\}$$

Therefore:

$$\begin{aligned}
\pi(\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}) &\propto \mathcal{L}(\boldsymbol{\beta},\boldsymbol{\Sigma};\boldsymbol{y},\boldsymbol{X})\times\pi(\boldsymbol{\beta})\\
&\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X\beta}\right)^T\left(\sigma^2\boldsymbol{I}\right)^{-1}\left(\boldsymbol{y}-\boldsymbol{X\beta}\right)\right\}\times\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^T(\sigma^2\boldsymbol{M})^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\right\}\\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{y}-\boldsymbol{X\beta})^T(\boldsymbol{y}-\boldsymbol{X\beta})+(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^T\boldsymbol{M}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\right]\right\} \qquad (1)\\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{y}^T\boldsymbol{y}-2\boldsymbol{y}^T\boldsymbol{X\beta}+\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X\beta})+(\boldsymbol{\beta}^T\boldsymbol{M}^{-1}\boldsymbol{\beta}-2\tilde{\boldsymbol{\beta}}^T\boldsymbol{M}^{-1}\boldsymbol{\beta}+\tilde{\boldsymbol{\beta}}^T\boldsymbol{M}^{-1}\tilde{\boldsymbol{\beta}})\right]\right\}\\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left[\boldsymbol{\beta}^T(\boldsymbol{X}^T\boldsymbol{X}+\boldsymbol{M}^{-1})\boldsymbol{\beta}-2(\boldsymbol{y}^T\boldsymbol{X}+\tilde{\boldsymbol{\beta}}^T\boldsymbol{M}^{-1})\boldsymbol{\beta}+\boldsymbol{y}^T\boldsymbol{y}+\tilde{\boldsymbol{\beta}}^T\boldsymbol{M}^{-1}\tilde{\boldsymbol{\beta}}\right]\right\}
\end{aligned}$$

Let $\boldsymbol{A}=\boldsymbol{X}^T\boldsymbol{X}+\boldsymbol{M}^{-1},\ \mathbf{b}^T=-2(\boldsymbol{y}^T\boldsymbol{X}+\tilde{\boldsymbol{\beta}}^T+\boldsymbol{M}^{-1})$ and $\mathrm{c}=\boldsymbol{y}^T\boldsymbol{y}+\tilde{\boldsymbol{\beta}}^T\boldsymbol{M}^{-1}\tilde{\boldsymbol{\beta}}$ such that:

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left[\boldsymbol{\beta}^T(\boldsymbol{A})\boldsymbol{\beta}+\mathbf{b}^T\boldsymbol{\beta}+\mathrm{c}\right]\right\}$$

By completing the square we get

$$\begin{aligned}
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{\beta}-\boldsymbol{\mu_\beta})^T\boldsymbol{A}\ (\boldsymbol{\beta}-\boldsymbol{\mu_\beta})+\ \boldsymbol{y}^T\boldsymbol{y}+\tilde{\boldsymbol{\beta}}^T\boldsymbol{M}^{-1}\tilde{\boldsymbol{\beta}}-\boldsymbol{\mu_\beta}^T\ \boldsymbol{A}\ \boldsymbol{\mu_\beta}\right]\right\} \qquad (2)\\
&\propto \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}-\boldsymbol{\mu_\beta})^T\frac{\boldsymbol{A}}{\sigma^2}\ (\boldsymbol{\beta}-\boldsymbol{\mu_\beta})\right]\right\}
\end{aligned}$$

where $\boldsymbol{\mu_\beta}\ =-\frac{1}{2}\boldsymbol{A}^{-1}\mathbf{b}=(\boldsymbol{X}^T\boldsymbol{X}+\boldsymbol{M}^{-1})^{-1}(\boldsymbol{X}^T\boldsymbol{y}+\boldsymbol{M}^{-1}\tilde{\boldsymbol{\beta}})$

Substituting The Ordinary Least Squares Estimate For Beta

Recall the given:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}&=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\\
\boldsymbol{X}^T\boldsymbol{y}&=(\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{\beta}}
\end{aligned}$$

Substituting the OLS estimate for Beta and since $\boldsymbol{M}=\boldsymbol{I}_{k+1}\ \Rightarrow\ \boldsymbol{M}^{-1}=\boldsymbol{M}$ such that:

$$\begin{aligned}
\boldsymbol{A}&=\boldsymbol{X}^T\boldsymbol{X}+\boldsymbol{M}\\
\boldsymbol{\mu}_\beta&=(\boldsymbol{X}^T\boldsymbol{X}+\boldsymbol{M})^{-1}\left((\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{\beta}}+\boldsymbol{M}\tilde{\boldsymbol{\beta}}\right)
\end{aligned}$$

Therefore, the conditional posterior distribution for beta follows the same functional form as the normal distribution.

$$\boxed{\begin{aligned}
&\boldsymbol{\beta}\mid\sigma^2,\boldsymbol{y},\boldsymbol{X}\sim\mathcal{N}_{k+1}\left(\boldsymbol{\mu}_\beta,\left(\frac{\boldsymbol{A}}{\sigma^2}\right)^{-1}\right)\\
&\boldsymbol{\beta}\mid\sigma^2,\boldsymbol{y},\boldsymbol{X}\sim\mathcal{N}_{k+1}(\boldsymbol{\mu}_\beta,\sigma^2(\boldsymbol{X}^T\boldsymbol{X}+\boldsymbol{M})^{-1})
\end{aligned}}$$

1.2 Deriving the Conditional Posterior Distribution for σ^2

Using the hint provided we can define the conditional posterior probability for σ^2 as

$$\pi(\sigma^2 \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}) = \int_{\boldsymbol{\beta}} \pi(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}, \boldsymbol{X}) d\boldsymbol{\beta} = \int_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}, \boldsymbol{X}) \pi(\boldsymbol{\beta} \mid \sigma^2) \pi(\sigma^2) d\boldsymbol{\beta}$$

We assume a prior distribution for σ^2 using an inverse gamma distribution. Specifically,

$$\sigma^2 \sim \mathcal{IG}(a, b)$$

This implies the prior density function for σ^2 is given by:

$$\pi(\sigma^2) = \frac{b^a}{\Gamma(a)} (\sigma^2)^{-a-1} \exp\left(-\frac{b}{\sigma^2}\right)$$

Therefore,

$$\begin{aligned}
\pi(\sigma^2 \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}) &\propto \int_{\boldsymbol{\beta}} (\sigma^2)^{\frac{-n-k-1}{2} - a - 1} \exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^T \boldsymbol{M}^{-1} (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\right] - \frac{b}{\sigma^2}\right\} d\boldsymbol{\beta} \\
&\propto \int_{\boldsymbol{\beta}} (\sigma^2)^{\frac{-n-k-1}{2} - a - 1} \exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{\beta} - \boldsymbol{\mu_\beta})^T \boldsymbol{A} (\boldsymbol{\beta} - \boldsymbol{\mu_\beta}) + \boldsymbol{y}^T\boldsymbol{y} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{M}^{-1} \tilde{\boldsymbol{\beta}} - \boldsymbol{\mu_\beta}^T \boldsymbol{A} \boldsymbol{\mu_\beta}\right] - \frac{b}{\sigma^2}\right\} d\boldsymbol{\beta}
\end{aligned}$$

We recognise that part of the exponent has the same form as equation (1), therefore we use the same complete the square results in (2) where:

- $\boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{M}^{-1}$
- $\boldsymbol{\mu_\beta} = \left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{M}^{-1}\right)^{-1} (\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{M}^{-1}\tilde{\boldsymbol{\beta}})$

Now, observe that there are terms that do not depend on $\boldsymbol{\beta}$, so we can factor it out of the integral.

$$(\sigma^2)^{\frac{-n-k-1}{2} - a - 1} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y}^T\boldsymbol{y} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{M}^{-1} \tilde{\boldsymbol{\beta}} - \boldsymbol{\mu_\beta}^T \boldsymbol{A} \boldsymbol{\mu_\beta}\right) - \frac{b}{\sigma^2}\right\} \int_{\boldsymbol{\beta}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\mu_\beta})^T \boldsymbol{A} (\boldsymbol{\beta} - \boldsymbol{\mu_\beta})\right\} d\boldsymbol{\beta}$$

Let us focus on the integral first and evaluate it.

$$\begin{aligned}
&\int_{\boldsymbol{\beta}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\mu_\beta})^T \boldsymbol{A} (\boldsymbol{\beta} - \boldsymbol{\mu_\beta})\right\} d\boldsymbol{\beta} \\
&= \int_{\boldsymbol{\beta}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu_\beta})^T \frac{\boldsymbol{A}}{\sigma^2} (\boldsymbol{\beta} - \boldsymbol{\mu_\beta})\right\} d\boldsymbol{\beta} \\
&= (2\pi)^{\frac{k+1}{2}} \left|\left(\frac{\boldsymbol{A}}{\sigma^2}\right)^{-1}\right|^{\frac{1}{2}} \int_{\boldsymbol{\beta}} \frac{1}{(2\pi)^{\frac{k+1}{2}} \left|\left(\frac{\boldsymbol{A}}{\sigma^2}\right)^{-1}\right|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu_\beta})^T \frac{\boldsymbol{A}}{\sigma^2} (\boldsymbol{\beta} - \boldsymbol{\mu_\beta})\right\} d\boldsymbol{\beta}
\end{aligned}$$

The following integral represents the density of a multivariate normal distribution integrated over all of $\boldsymbol{\beta}$:

$$\int_{\boldsymbol{\beta}} \frac{1}{(2\pi)^{\frac{k+1}{2}} \left|\left(\frac{\boldsymbol{A}}{\sigma^2}\right)^{-1}\right|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu_\beta})^T \frac{\boldsymbol{A}}{\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\mu_\beta})\right\} d\boldsymbol{\beta}$$

This is the integral of the probability density function of the multivariate normal distribution:

$$\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{y}, \boldsymbol{X} \sim \mathcal{N}_{k+1}(\boldsymbol{\mu_\beta}, \sigma^2 \boldsymbol{A}^{-1})$$

Since it integrates the full density, it evaluates to 1, ie:

$$\int_{\boldsymbol{\beta}} \frac{1}{(2\pi)^{\frac{k+1}{2}} \left|\left(\frac{\boldsymbol{A}}{\sigma^2}\right)^{-1}\right|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu_\beta})^T \frac{\boldsymbol{A}}{\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\mu_\beta})\right\} d\boldsymbol{\beta} = 1$$

Therefore, when substituiting back in the the full equation for the conditional posterior probability for σ^2 we have

$$\begin{aligned}
\pi(\sigma^2 \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}) &\propto (\sigma^2)^{\frac{-n-k-1}{2}-a-1} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y}^T\boldsymbol{y} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{M}^{-1}\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu_\beta}^T \boldsymbol{A}\boldsymbol{\mu_\beta}\right) - \frac{b}{\sigma^2}\right\} (2\pi)^{\frac{k+1}{2}} \left|\left(\frac{\boldsymbol{A}}{\sigma^2}\right)^{-1}\right|^{\frac{1}{2}} \\
&\propto (\sigma^2)^{\frac{-n-k-1}{2}-a-1} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y}^T\boldsymbol{y} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{M}^{-1}\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu_\beta}^T \boldsymbol{A}\boldsymbol{\mu_\beta}\right) - \frac{b}{\sigma^2}\right\} (\sigma^2)^{\frac{k+1}{2}} \left|\boldsymbol{A}^{-1}\right|^{\frac{1}{2}} \\
&\propto (\sigma^2)^{\frac{-n-k-1}{2}-a-1+\frac{k+1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y}^T\boldsymbol{y} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{M}^{-1}\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu_\beta}^T \boldsymbol{A}\boldsymbol{\mu_\beta}\right) - \frac{b}{\sigma^2}\right\}
\end{aligned}$$

Recall that we have

$$\boldsymbol{M} = \boldsymbol{I}_{k+1} \quad \Rightarrow \quad \boldsymbol{M}^{-1} = \boldsymbol{M}, \text{ and } \boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{M}^{-1} = \boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{M}$$

Thus, the expression becomes

$$\begin{aligned}
&\propto (\sigma^2)^{\frac{-n-k-1}{2}-a-1+\frac{k+1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y}^T\boldsymbol{y} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{M}\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu_\beta}^T (\boldsymbol{X}^T\boldsymbol{X}+\boldsymbol{M})\boldsymbol{\mu_\beta}\right) - \frac{b}{\sigma^2}\right\} \\
&\propto (\sigma^2)^{-\left(\frac{n}{2}+a\right)-1} \exp\left\{-\frac{1}{\sigma^2}\left[\frac{\left(\boldsymbol{y}^T\boldsymbol{y} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{M}\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu_\beta}^T (\boldsymbol{X}^T\boldsymbol{X}+\boldsymbol{M})\boldsymbol{\mu_\beta}\right)}{2} + b\right]\right\}
\end{aligned}$$

This has the functional form of a Inverse Gamma Distribution

Defining

$$A_2 = \boldsymbol{y}^T\boldsymbol{y} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{M}\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu}_\beta^T (\boldsymbol{X}^T\boldsymbol{X}+\boldsymbol{M})\boldsymbol{\mu}_\beta,$$

we can rewrite the conditional posterior distribution for σ^2 in the form:

$$\boxed{\sigma^2 \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta} \sim \mathcal{IG}\left(\tfrac{n}{2}+a, \tfrac{A_2}{2}+b\right)}$$

1.3 Trace Plots for Each Regression Coefficient

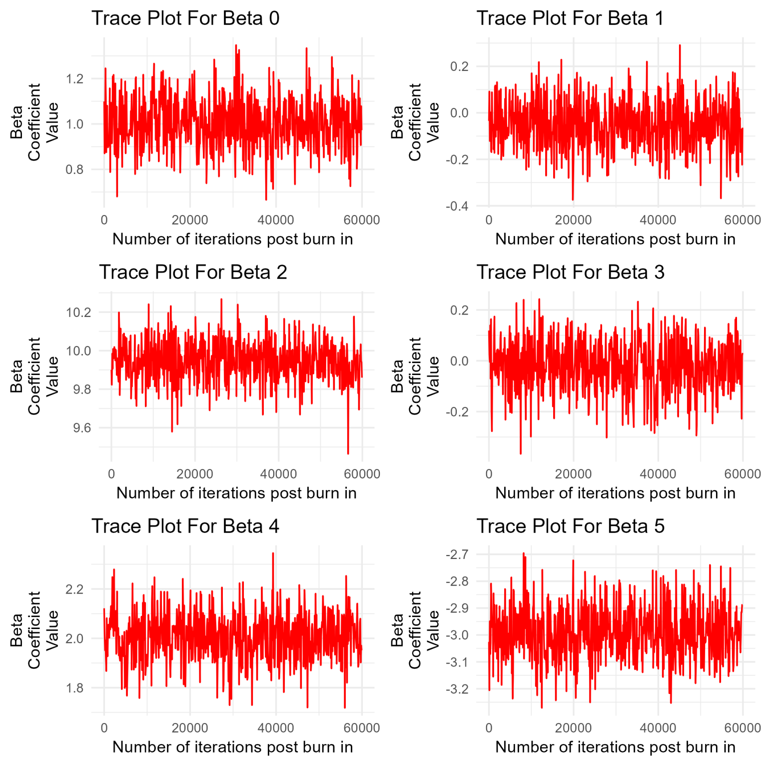

Figure 1: Trace Plot of the Beta Coefficients, 60000 posterior samples were used

Trace plots, such as those presented in Figure 1, are diagnostic tools used to evaluate the performance of Markov Chain Monte Carlo (MCMC) simulations by visualizing the sampled values of parameters across iterations. In this case, the trace plots were generated using Gibbs sampling from the marginal posterior distributions of a linear regression model's beta coefficients. To improve the quality of the samples, both burn-in and thinning were applied. Burn-in was used to discard initial iterations before convergence, and thinning to reduce autocorrelation by retaining every 2nd sample. These plots are crucial for assessing convergence—whether the chain has stabilized—and mixing, which reflects how well the sampler explores the parameter space.

The trace plots for the intercept term and the variables $\beta 1$ through $\beta 5$ display rapid, consistent fluctuations around stable mean values with no visible trends or drifts, indicating good convergence and mixing. This suggests that the Gibbs sampler has effectively explored the posterior distributions, and the resulting samples can be considered reliable for inference.

1.4 Density Plots for Each Regression Coefficient

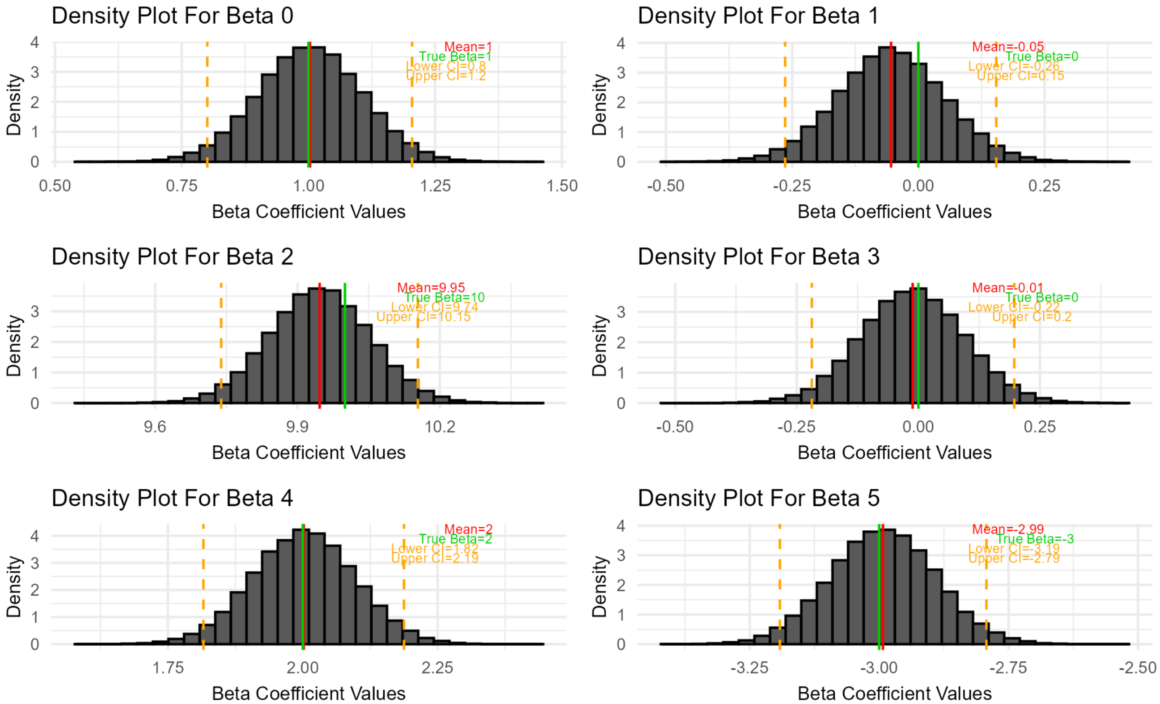

Figure 2: The Partial Dependency Plots for Beta Coefficients Where: Red Is The Average The Samples, Green Is The True Beta Value, Orange Is The Upper and Lower Credibility Intervals, 60000 posterior samples were used

Table 1: Gibbs Sampling Metric For Each Variable

Average_Samples	True_Beta	Credibility_Lower	Credibility_Upper	Range
1.003	1	0.800	1.204	0.404
−0.054	0	−0.264	0.154	0.418
9.947	10	9.739	10.154	0.415
−0.012	0	−0.219	0.197	0.416
2.002	2	1.816	2.188	0.372
−2.993	−3	−3.192	−2.793	0.399

The density plots presented in Figure 2 depict the conditional posterior distributions of the beta coefficients for the intercept and predictor variables $\beta 1$ through $\beta 5$ obtained through Gibbs sampling.

Interpretation of Sampling Average and True Coefficient

The density plot illustrate the uncertainty in the estimated coefficients, with the red line representing the sampled average (posterior mean) and the green line indicating the true beta coefficients.

Table 1 supports this visual interpretation:

- For $\boldsymbol{\beta_0}$, the sampled average is 1.003, while the true value is 1. The red and green lines align almost perfectly, suggesting that the estimate is highly accurate.
- For $\boldsymbol{\beta_1}$, the sample average is -0.054, compared to a true value of 0. Although slightly offset, the estimate is still reasonably close, indicating low bias.
- For $\boldsymbol{\beta_2}$, the sampled value is 9.947 and the true value is 10. The close alignment confirms that the model accurately recovers this large positive effect.
- $\boldsymbol{\beta_3}$ has a sampled average of -0.012, while the true coefficient is 0. This again shows a small deviation, suggesting the estimate is close to the true effect, which is negligible.
- For $\boldsymbol{\beta_4}$, the sampled value is 2.002 versus a true value of 2—nearly identical, indicating high accuracy in capturing a moderate positive effect.
- Lastly, for $\boldsymbol{\beta_5}$, the sample average is -2.993, compared with a true value of -3. The lines are closely aligned, indicating a strong and accurately recovered negative effect.

Overall, the red and green lines are relatively closely aligned across all coefficients, demonstrating that the Gibbs sampler has successfully estimated the true parameters of the model. Minor differences are within acceptable sampling variability and do not suggest any substantial bias or error in the estimation process.

Interpretation of 95% Credibility Interval

A **confidence interval** is constructed under the frequentist framework and is designed so that, in repeated sampling, a specified percentage (e.g. 95%) of such intervals will contain the true parameter. It does not imply that there is a 95% probability that the true parameter lies within a given interval.

In contrast, a **credibility interval** (or credible interval) comes from Bayesian inference. It represents the range within which the parameter lies with a given probability (e.g. 95%) given the observed data and the prior information. Thus, it directly expresses the degree of belief in the parameter's value being within that range.

The density plot also illustrates the 95% credibility intervals in orange. These intervals indicate the range within which the true coefficient likely falls, given the data and prior assumptions.

- $\boldsymbol{\beta_0}$ **:** The credibility interval for the intercept, $[0.800, 1.204]$, centres around 1.003. This indicates that, holding other variables constant, the baseline effect on the outcome is approximately 1. Such an intercept shifts the overall prediction upwards by roughly 1 unit.
- $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_3}$ **:** With average estimates close to zero (-0.054 for β_1 and -0.012 for β_3) and intervals $[-0.264, 0.154]$ and $[-0.219, 0.197]$ respectively, these coefficients have centres near zero. This suggests that these predictors likely have little to no practical impact on the response variable.
- $\boldsymbol{\beta_2}$ **:** The credibility interval $[9.739, 10.154]$ centres around 9.947, reflecting a strong positive effect. This substantial magnitude indicates that an increase in the predictor corresponding to β_2 leads to a significant increase in the response variable.
- $\boldsymbol{\beta_4}$ **:** The average value of 2.002, with the interval $[1.816, 2.188]$, signifies a moderate positive impact. The effect of the corresponding predictor is both meaningful and precise, as evidenced by the relatively narrow range of 0.372.
- $\boldsymbol{\beta_5}$ **:** Finally, β_5 has a credibility interval of $[-3.192, -2.793]$ centred around -2.993. This clearly shows a strong negative impact, where an increase in the predictor associated with β_5 results in a decrease in the response variable.

Overall, even though the ranges of the intervals are quite similar (approximately 0.37–0.42), the centres of the intervals vary. This variance in central values indicates substantial differences in the impact of the predictors: large positive effects for β_2, moderate positive for β_4, strong negative for β_5, and little to no effects for β_1 and β_3. Such insights are crucial for understanding the relative contribution of each predictor to the model.

Variable Selection using Credibility Intervals

In Bayesian analysis, the credibility interval of a parameter represents the range within which the true value of the parameter lies with a certain probability. In the context of variable selection, credibility intervals are used to assess the statistical significance of the predictors. For a given predictor's coefficient, if its credibility interval excludes zero, this indicates that the probability mass for the parameter's posterior distribution is concentrated away from zero. Consequently, there is strong evidence to suggest that the predictor has a non-negligible effect on the outcome.

For example, consider the coefficients β_2, β_4, and β_5 whose credibility intervals do not contain zero. This exclusion of zero implies that these predictors have a consistent influence on the dependent variable across the range of plausible parameter values derived from the data, thus making them statistically meaningful. In contrast, predictors such as β_1 and β_3 with credibility intervals that include zero indicate that there is a substantial probability that their true effects could be null. Such variables may not contribute significantly to the model's predictive power and hence could be considered for removal.

This approach to variable selection offers several advantages:

– **Balancing Complexity and Explanatory Power:** By focusing on predictors with high evidence of association (credibility intervals excluding zero), the model avoids overfitting through the inclusion of variables that do not contribute meaningfully to prediction.

– **Enhanced Interpretability:** A model with fewer predictors, each demonstrating a statistically significant impact, tends to be more interpretable, enabling clearer insights into the relationships within the data.

– **Robustness of Inference:** Credibility intervals account for uncertainty in parameter estimates by summarizing the entire posterior distribution. This leads to more robust decisions regarding which predictors are retained, as opposed to relying solely on point estimates or p-values from frequentist methods.

Overall, using credibility intervals for variable selection not only helps to identify the most influential predictors but also preserves a balance between model simplicity and explanatory power, ensuring that only variables with a demonstrable impact on the outcome are included in the final model.

2 Question Two: A Bayesian Search

2.1 Derivation of Equations 2 and 4

We are given that each cell i has a probability θ_i that the fisherman is there. This modelled by a Bernoulli distribution:

$$Y_i \sim \text{Bernoulli}(\theta_i)$$

which means:

- $P(Y_i = 1) = \theta_i$ (the fisherman is in cell i)
- $P(Y_i = 0) = 1 - \theta_i$

If the fisherman is in cell i ($Y_i = 1$) then the search outcome Z_i is also modelled as a Bernoulli random variable, with detection probability p_i :

$$Z_i | Y_i = 1 \sim \text{Bernoulli}(p_i)$$

which means:

- $P(Z_i = 1 | Y_i = 1) = p_i$ (the fisherman is detected in cell i)
- $P(Z_i = 0 | Y_i = 1) = 1 - p_i$ (the fisherman is missed in cell i)
- $P(Z_i = 0 | Y_i = 0) = 1$

Equation 2: The Posterior Probability for a Searched Cell

We want to find the updated probability that the fisherman is in cell i given that we searched that cell and did not detect him, that is we want:

$$\pi(Y_i = 1 | Z_i = 0)$$

By applying Bayes' Theorem:

$$\pi(Y_i = 1 | Z_i = 0) = \frac{P(Z_i = 0 | Y_i = 1) P(Y_i = 1)}{P(Z_i = 0)}$$

Likelihood:

$$P(Z_i = 0 | Y_i = 1) = 1 - p_i$$

Prior:

$$P(Y_i = 1) = \theta_i$$

Marginal Probability:

$$\begin{aligned} P(Z_i = 0) &= P(Z_i = 0 | Y_i = 1) P(Y_i = 1) + P(Z_i = 0 | Y_i = 0) P(Y_i = 0) \\ &= (1 - p_i)\theta_i + 1 - \theta_i \\ &= 1 - p_i \theta_i \end{aligned}$$

By substituting back in, we obtain the following equation:

$$\pi(Y_i = 1|Z_i = 0) = \frac{(1 - p_i)\theta_i}{1 - p_i\theta_i}$$

Equation 4: The Posterior Probability for Other Cells

We want to find the updated probability that the fisherman is in cell j given that we searched cell i and did not detect him, that is we want:

$$\pi(Y_j = 1|Z_i = 0)$$

By applying Bayes' Theorem:

$$\pi(Y_j = 1|Z_i = 0) = \frac{P(Z_i = 0|Y_j = 1)P(Y_j = 1)}{P(Z_i = 0)}$$

For any other cell j (where $j \neq i$) the search in cell i does not give direct information about cell j. Therefore the **Likelihood** for cell j remains:

$$P(Z_i = 0|Y_j = 1) = 1$$

and the **Prior** is:

$$P(Y_j = 1) = \theta_j$$

Marginal Probability:

$$\begin{aligned} P(Z_i = 0) &= P(Z_i = 0|Y_i = 1)P(Y_i = 1) + P(Z_i = 0|Y_i = 0)P(Y_i = 0) \\ &= (1 - p_i)\theta_i + 1 - \theta_i \\ &= 1 - p_i\theta_i \end{aligned}$$

By substituting back in we obtain that following:

$$\pi(Y_j = 1|Z_i = 0) = \frac{\theta_j}{1 - p_i\theta_i}$$

Using this posterior probability to update θ_j given that we did not detect the fisherman in cell i results in the following:

$$\theta_{j,new} = \frac{\theta_{j,old}}{1 - p_i\theta_{i,old}}$$

2.2 Occurrence Probability Updating Equation

The key idea is to view the update process in terms of the evolution of our beliefs over time:

- **Prior:** At a given time t, our belief about the fisherman being in cell i is represented by the prior $\theta_i^{(t)}$ (denoted as $\theta_{i,\text{old}}$).
- **Observation:** A search is conducted in cell i and no detection is made ($Z_i = 0$).

– **Bayesian Update:** Bayes' theorem is used to update the probability, resulting in the posterior probability

$$\pi(Y_i = 1 \mid Z_i = 0) = \frac{(1 - p_i)\theta_i^{(t)}}{1 - p_i\theta_i^{(t)}},$$

which corresponds to Equation (2).

– **New Prior:** In a sequential process, the posterior probability calculated after the search becomes the new prior for any subsequent searches. Consequently, we write

$$\theta_{i,\text{new}} = \theta_i^{(t+1)} = \frac{(1 - p_i)\theta_i^{(t)}}{1 - p_i\theta_i^{(t)}},$$

as shown in Equation (3).

Interpretation

By "thinking in terms of priors and posteriors over time," we emphasise that:

1. The **Prior** reflects our initial belief about the location of the fisherman before a search is executed.
2. The **Posterior** is the updated belief after incorporating the result of the search (in this case, an unsuccessful detection).
3. Since the process is sequential, the **Posterior** from one search becomes the **Prior** for the next search. Thus, Equation (3) is not a separate result but directly follows from Equation (2), which represents the Bayesian updating rule applied each time a cell is searched and no detection is achieved.

This sequential updating allows the search strategy to continually refine the probability estimates for each cell based on the outcomes of previous searches.

2.3 Bayesian Search Process

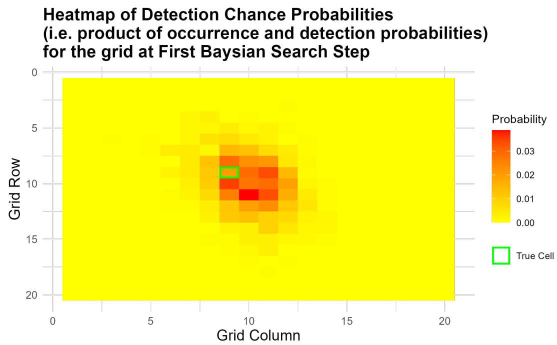

Figure 3: First Search Step

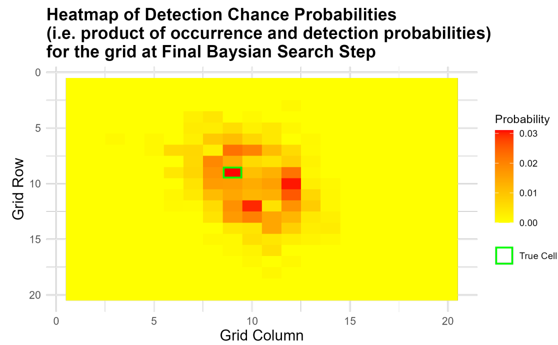

Figure 4: Final Search Step

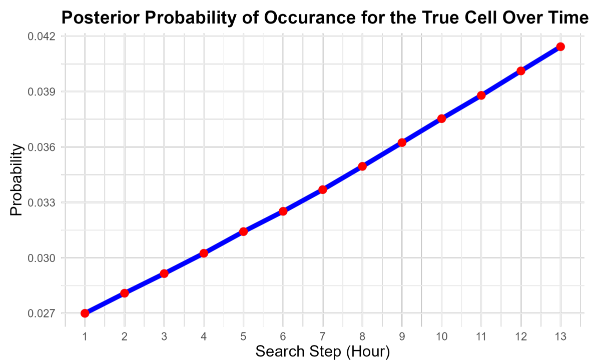

Figure 5: True Cell Posterior Probability Plot Over Time

Table 2: Search Pattern Analysis (True Cell: 169)

Hour	θ_i	p_i	Detection Chance	Searching Cell	True Cell Prob. Occ
1	0.0490	0.7906	0.0387	210	0.0270
2	0.0520	0.7006	0.0364	189	0.0281
3	0.0486	0.7471	0.0363	171	0.0292
4	0.0504	0.7409	0.0373	211	0.0302
5	0.0524	0.6439	0.0337	191	0.0314
6	0.0506	0.6889	0.0348	149	0.0325
7	0.0474	0.7595	0.0360	170	0.0337
8	0.0531	0.6692	0.0355	190	0.0350
9	0.0550	0.6277	0.0345	209	0.0362
10	0.0375	0.8632	0.0324	231	0.0375
11	0.0402	0.8211	0.0330	150	0.0388
12	0.0371	0.8522	0.0317	151	0.0401
13	0.0414	0.7499	0.0311	169	0.0414

In **Figure 3**, the heatmap shows the initial posterior probabilities for the 20×20 grid, reflecting expert prior beliefs about the fisherman's possible location. The redder cells represent higher probabilities of occurrence. Initially, the search algorithm focuses on the most promising area — a dense region of red and orange in the center.

Figure 4 shows the posterior distribution after several Bayesian updates, where search results (failures to detect) have refined the probabilities across the grid. Noticeably, the distribution becomes more dispersed, but there's a clearer, sharper peak near the true cell, as repeated searches eliminate less likely areas and concentrate belief around the most probable locations.

Figure 5 quantifies this progression by plotting the posterior probability of the true cell over time. It shows a consistent upward trend — starting at about 0.027 and rising steadily to over 0.042. This demonstrates how each unsuccessful search in other cells incrementally increases belief in the true location, reflecting Bayesian updating in action.

2.4 Implications of a Constant Detection Probability

If the detection probability p_i is constant across all cells (i.e. $p_i = p$), the search strategy simplifies significantly. Since p no longer varies spatially, the optimal search sequence depends only on the occurrence probabilities θ_i. The search crew should always target the cell with the highest current θ_i, as there is no additional benefit from prioritising cells with higher detectability.

However, by maintaining a constant detection probability p_i, the search strategy cannot adapt to areas where detection is inherently easier (e.g., calmer seas) or harder (e.g., regions with strong currents or poor weather). This may prolong the search in two ways:

– Failing to exploit high-probability cells with better visibility conditions (detection probability)

– Potentially wasting time repeatedly searching low-visibility areas where detection is consistently difficult

3 Question Three: A Twist on Linear Regression

Consider the linear regression model where the data is generated as:

$$Y_i = \begin{cases} \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i, & e_i \sim \mathcal{N}(0, \sigma_1^2), i \in \mathcal{I}_1 \\ \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i, & e_i \sim \mathcal{N}(0, \sigma_2^2), i \in \mathcal{I}_2 \end{cases}$$

where

- $\mathcal{I}$ is an index set
- $\sigma_1^2 < \sigma_2^2$

This implies that observations of the second index set have a higher variance than those of the first, potentially outliers in the data set.

As usual, let $\tau_i = \frac{1}{\sigma_i^2}$. If we assume that $\mathcal{I}_1$ is known, then without loss of generality we can let

$$\begin{aligned} \mathcal{I}_1 &= \{1, \ldots, n_1\}, \\ \mathcal{I}_2 &= \{n_1 + 1, \ldots, n\}. \end{aligned}$$

where the first n_1 is the number of 'standard' observations, and the rest of the observations are outliers.

3.1 Prior Distributions for Model Parameters

We begin by specifying the priors for the unknown parameters in our model:

- **Regression Coefficients:**

  $$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{T_0})$$

  This prior indicates that the p-dimensional coefficient vector $\boldsymbol{\beta}$ has a multivariate normal distribution with mean $\mathbf{0}$ and covariance-variance matrix $\boldsymbol{T_0}$.

- **Precision Parameters:** The noise or error variances in the model are governed by the precision (i.e., inverse variance) parameters. We impose the following priors:

  $$\tau_1 \sim \mathcal{G}(a, b)$$

  $$\tau_2 \mid \tau_1 \sim \mathcal{G}(a, b)\, I_{\{\tau_1 > \tau_2\}}$$

  Here, both τ_1 and τ_2 are assumed to follow Gamma distributions with shape a and rate b, but the second precision is additionally constrained to be smaller than τ_1.

Response Vector

The vector of responses $\boldsymbol{y}$ is partitioned into two sub-vectors, corresponding to two different groups of observations:

$$\boldsymbol{y} = \left[\begin{array}{c} \boldsymbol{y}_1 \\ \hline \boldsymbol{y}_2 \end{array}\right]_{n \times 1},$$

with

$$\boldsymbol{y}_1 = \begin{bmatrix} y_1 = \boldsymbol{x}_1^T \boldsymbol{\beta} + e_{11} \\ y_2 = \boldsymbol{x}_2^T \boldsymbol{\beta} + e_{12} \\ \vdots \\ y_{n_1} = \boldsymbol{x}_{n_1}^T \boldsymbol{\beta} + e_{1n_1} \end{bmatrix}_{n_1 \times 1} \quad \text{and } \boldsymbol{y}_2 = \begin{bmatrix} y_{n_1+1} = \boldsymbol{x}_{n_1+1}^T \boldsymbol{\beta} + e_{2(n_1+1)} \\ y_{n_1+2} = \boldsymbol{x}_{n_1+2}^T \boldsymbol{\beta} + e_{2(n_1+2)} \\ \vdots \\ y_n = \boldsymbol{x}_n^T \boldsymbol{\beta} + e_{2n} \end{bmatrix}_{(n-n_1) \times 1}$$

In this setup, the first n_1 observations have errors $e_{1i} \sim \mathcal{N}(0, \sigma_1^2)$ and the remaining $n_2 = n - n_1$ observations have errors $e_{2i} \sim \mathcal{N}(0, \sigma_2^2)$.

Design Matrix

Similarly, the design matrix $\boldsymbol{X}$ is partitioned to reflect the grouping of the observations:

$$_n\boldsymbol{X}_{k+1} = \left[\begin{array}{c} \boldsymbol{X}_1 \\ \hline \boldsymbol{X}_2 \end{array}\right],$$

where

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n_1 1} & x_{n_1 2} & \dots & x_{n_1 k} \end{bmatrix}_{n_1 \times (k+1)} \quad \text{and } \boldsymbol{X}_2 = \begin{bmatrix} 1 & x_{(n_1+1)1} & x_{(n_1+1)2} & \dots & x_{(n_1+1)k} \\ 1 & x_{(n_1+2)1} & x_{(n_1+2)2} & \dots & x_{(n_1+2)k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}_{(n-n_1) \times (k+1)}.$$

Model Parameters and Error Structure

– **Regression Coefficients:** The p-dimensional vector of regression coefficients is given by

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}_{(k+1) \times 1}.$$

– **Error Vector:** The vector of errors is similarly partitioned:

$$\boldsymbol{e} = \left[\begin{array}{c} e_{11} \\ e_{12} \\ \vdots \\ e_{1n_1} \\ \hline e_{2(n_1+1)} \\ e_{2(n_1+2)} \\ \vdots \\ e_{2n} \end{array}\right]_{n \times 1},$$

with $e_{1i} \sim \mathcal{N}(0, \sigma_1^2)$ for $i = 1, \dots, n_1$ and $e_{2i} \sim \mathcal{N}(0, \sigma_2^2)$ for $i = n_1 + 1, \dots, n$.

The Complete Model

Combining the components above, the linear model is formulated as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

Given the error structure specified, the distribution of the response vector is:

$$\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X\beta}, \boldsymbol{\Sigma}),$$

where the covariance matrix $\boldsymbol{\Sigma}$ is block-diagonal to reflect the differing variances between groups:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 I_{n_1} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_2^2 I_{n_2} \end{bmatrix},$$

with I_{n_1} and I_{n_2} being the identity matrices of dimensions n_1 and $n_2 = n - n_1$, respectively.

Therefore,

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

3.2 Derivation of the Joint Likelihood

We start with a multivariate normal model for the data:

$$\boldsymbol{y} \sim \mathcal{N}_n\left(\boldsymbol{X\beta}, \boldsymbol{\Sigma}\right),$$

so that the joint likelihood is given by

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \boldsymbol{y}, \boldsymbol{X}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)\right\}$$

3.3 Deriving the Conditional Posterior Distribution for $\boldsymbol{\beta}$

Since

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{T_0})$$

the prior distribution for $\boldsymbol{\beta}$ is given as:

$$\pi(\boldsymbol{\beta}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{T_0}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{0})\boldsymbol{T}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{0})\right\}$$

Therefore:

$$\begin{aligned}
\pi(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Sigma}) &\propto \mathcal{L}(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Sigma}) \times \pi(\boldsymbol{\beta}) \\
&\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)\right\} \times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{0})\boldsymbol{T}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{0})\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right) + (\boldsymbol{\beta} - \boldsymbol{0})\boldsymbol{T}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{0})\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{y}^T\boldsymbol{\Sigma^{-1}y} - \boldsymbol{y}^T\boldsymbol{\Sigma^{-1}X\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{\Sigma^{-1}y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{\Sigma^{-1}\beta} + \boldsymbol{\beta}^T\boldsymbol{T}_0^{-1}\boldsymbol{\beta}\right]\right\}
\end{aligned}$$

*since ${}_1\boldsymbol{y}_n^T\boldsymbol{\Sigma}_n^{-1}\boldsymbol{X}_{k+1}\boldsymbol{\beta}_1$ is a scalar we can say that $\boldsymbol{y}^T\boldsymbol{\Sigma^{-1}X\beta} = (\boldsymbol{y}^T\boldsymbol{\Sigma^{-1}X\beta})^T = \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{\Sigma^{-1}y}$

$$\begin{aligned}
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{y}^T\boldsymbol{\Sigma^{-1}y} - 2\boldsymbol{y}^T\boldsymbol{\Sigma^{-1}X\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{\Sigma^{-1}\beta} + \boldsymbol{\beta}^T\boldsymbol{T}_0^{-1}\boldsymbol{\beta}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta^T(X^T\Sigma^{-1}X + T_0^{-1})\beta - 2y^T\Sigma^{-1}X\beta + y^T\Sigma^{-1}y}\right]\right\}
\end{aligned}$$

Let $\boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \boldsymbol{T}_0^{-1},\ \mathbf{b}^T = -2\boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$ and $\mathrm{c} = \boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$ such that:

$$\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}^T(\boldsymbol{A})\boldsymbol{\beta} + \mathbf{b}^T\boldsymbol{\beta} + \mathrm{c}\right]\right\}$$

By completing the square we get

$$\propto \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta} - \boldsymbol{\mu_\beta})^T\boldsymbol{A}(\boldsymbol{\beta} - \boldsymbol{\mu_\beta})\right]\right\}$$

where $\boldsymbol{\mu_\beta} = -\frac{1}{2}\boldsymbol{A}^{-1}\mathbf{b} = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \boldsymbol{T}_0^{-1})^{-1}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y})$

Covariance Matrix and Its Inverse

Recall that our covariance matrix has a block-diagonal structure:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 I_{n_1} & \mathbf{0} \\ \mathbf{0} & \sigma_2^2 I_{n_2} \end{bmatrix}_{n\times n}.$$

Thus, its inverse is

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} I_{n_1} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\sigma_2^2} I_{n_2} \end{bmatrix} = \tau_1 \begin{bmatrix} I_{n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \tau_2 \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{n_2} \end{bmatrix}.$$

For clarity, we define:

$$\boldsymbol{C} = \begin{bmatrix} I_{n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}_{n\times n} \quad \text{and} \quad \boldsymbol{D} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{n_2} \end{bmatrix}_{n\times n}.$$

Then, we can write:

$$\boldsymbol{\Sigma}^{-1} = \tau_1\boldsymbol{C} + \tau_2\boldsymbol{D}.$$

By substituting this in

$$\boldsymbol{A} = \tau_1\boldsymbol{X}_1^T\boldsymbol{X}_1 + \tau_2\boldsymbol{X}_2^T\boldsymbol{X}_2 + \boldsymbol{T}_0^{-1}$$
$$\boldsymbol{\mu_\beta} = (\tau_1\boldsymbol{X}_1^T\boldsymbol{X}_1 + \tau_2\boldsymbol{X}_2^T\boldsymbol{X}_2 + \boldsymbol{T}_0^{-1})^{-1}(\tau_1\boldsymbol{X}_1^T\boldsymbol{y}_1 + \tau_2\boldsymbol{X}_2^T\boldsymbol{y}_2)$$

Therefore

$$\boxed{\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}, \tau_1, \tau_2 \sim \mathcal{N}_{k+1}(\boldsymbol{\mu_\beta}, \boldsymbol{A}^{-1})}$$

3.4 Deriving the Conditional Posterior Distribution for τ_1

We model the precision τ_1 (inverse variance) using a Gamma distribution:

$$\tau_1 \sim \mathcal{G}(a, b) I_{\tau_1 > \tau_2}$$

where a, b are scalars

This means the prior distribution for τ_1 is given by

$$\pi(\tau_1) = \left(\frac{b^a}{\Gamma(a)}\tau_1^{a-1}e^{-b\tau_1}\right) I_{\tau_1 > \tau_2}$$

Previously

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \boldsymbol{y}, \boldsymbol{X}) &= \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) \right\} \\
&= \frac{1}{(2\pi)^{\frac{n}{2}} (\sigma_1^2)^{\frac{n_1}{2}} (\sigma_2^2)^{\frac{n_2}{2}}} \exp\left\{ -\frac{1}{2} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) \right\} \\
&= \frac{\tau_1^{\frac{n_1}{2}} \tau_2^{\frac{n_2}{2}}}{(2\pi)^{\frac{n}{2}}} \exp\left\{ -\frac{1}{2} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) \right\},
\end{aligned}$$

where we have set $\tau_1 = \frac{1}{\sigma_1^2}$ and $\tau_2 = \frac{1}{\sigma_2^2}$.
By substituting $\boldsymbol{\Sigma}^{-1} = \tau_1 \boldsymbol{C} + \tau_2 \boldsymbol{D}$ into the likelihood, we obtain

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \boldsymbol{y}, \boldsymbol{X}) = \frac{\tau_1^{\frac{n_1}{2}} \tau_2^{\frac{n_2}{2}}}{(2\pi)^{\frac{n}{2}}} \exp\left\{ -\frac{1}{2} \left[\tau_1 (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{C} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \tau_2 (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{D} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right] \right\}$$

Since the matrices $\boldsymbol{C}$ and $\boldsymbol{D}$ effectively select the first n_1 and the remaining n_2 components respectively, we can write

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{C} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta})^T (\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}),$$

and

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{D} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta})^T (\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}).$$

Thus, the likelihood simplifies to

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \boldsymbol{y}, \boldsymbol{X}) &= \frac{\tau_1^{\frac{n_1}{2}} \tau_2^{\frac{n_2}{2}}}{(2\pi)^{\frac{n}{2}}} \exp\left\{ -\frac{1}{2} \left[\tau_1 (\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta})^T (\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}) + \tau_2 (\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta})^T (\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}) \right] \right\} \\
&= \mathcal{L}(\boldsymbol{\beta}, \tau_1, \tau_2; \boldsymbol{y}, \boldsymbol{X})
\end{aligned}$$

This final form clearly separates the contributions of the two groups of observations to the joint likelihood.

The posterior distribution for τ_1 combines the likelihood with the prior

$$\begin{aligned}
\pi(\tau_1 \mid \boldsymbol{\beta}, \tau_2, \boldsymbol{X}, \boldsymbol{y}) &\propto \mathcal{L}(\tau_1; \boldsymbol{\beta}, \tau_2, \boldsymbol{X}, \boldsymbol{y}) \times \pi(\tau_1) I_{\tau_1 > \tau_2} \\
&\propto \tau_1^{\frac{n_1}{2}} \exp\left\{ -\frac{1}{2} \tau_1 (\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta})^T (\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}) \right\} \times \left(\tau_1^{a-1} e^{-b\tau_1} \right) I_{\tau_1 > \tau_2} \\
&\propto \tau_1^{\frac{n_1}{2} + a - 1} \exp\left\{ -\left(\frac{(\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta})^T (\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta})}{2} + b \right) \tau_1 \right\} I_{\tau_1 > \tau_2}
\end{aligned}$$

This reveals the posterior as

$$\boxed{\tau_1 \mid \boldsymbol{\beta}, \tau_2, \boldsymbol{X}, \boldsymbol{y} \sim \mathcal{G}\left(\frac{n_1}{2} + a, \frac{(\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta})^T (\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta})}{2} + b \right) I_{\tau_1 > \tau_2}}$$

3.5 Deriving the Conditional Posterior Distribution for τ_2

We model the precision τ_2 (inverse variance) using a Gamma distribution:

$$\tau_2 \mid \tau_1 \sim \mathcal{G}(a, b) I_{\tau_1 > \tau_2}$$

where a, b are scalars

This means the prior distribution for τ_1 is given by

$$\pi(\tau_2 \mid \tau_1) = \left(\frac{b^a}{\Gamma(a)} \tau_2^{a-1} e^{-b\tau_2}\right) I_{\tau_1 > \tau_2}$$

The posterior distribution for τ_2 combines the likelihood with the prior

$$\begin{aligned}
\pi(\tau_2 \mid \boldsymbol{\beta}, \tau_1, \boldsymbol{X}, \boldsymbol{y}) &\propto \mathcal{L}(\tau_2; \boldsymbol{\beta}, \tau_1, \boldsymbol{X}, \boldsymbol{y}) \times \pi(\tau_2 \mid \tau_1) I_{\tau_1 > \tau_2} \\
&\propto \tau_2^{\frac{n_2}{2}} \exp\left\{-\frac{1}{2}\tau_2 (\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta})^T (\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta})\right\} \times \left(\tau_2^{a-1} e^{-b\tau_2}\right) I_{\tau_1 > \tau_2} \\
&\propto \tau_2^{\frac{n_2}{2}+a-1} \exp\left\{-\left(\frac{(\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta})^T (\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta})}{2} + b\right)\tau_2\right\} I_{\tau_1 > \tau_2}
\end{aligned}$$

This reveals the posterior as

$$\boxed{\tau_2 \mid \boldsymbol{\beta}, \tau_1, \boldsymbol{X}, \boldsymbol{y} \sim \mathcal{G}\left(\frac{n_2}{2} + a, \frac{(\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta})^T (\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta})}{2} + b\right) I_{\tau_1 > \tau_2}}$$