Project Guidelines for SCI5020 – Principles of Statistical Inference – Part 2

Lecturers: Dr Monique Borg Inguanez, Dr Fiona Sammut and
Dr David Suda

Deadline for the Project (Part 2) is Strictly: noon, Friday 27th June, 2025.

The project for SCI5020 is made up of two parts. Each part accounts for 50% of the total mark. The details which follow describe the second part of the project.

You are required to work individually. Any questions related to this part of the project should be sent to Dr Borg Inguanez or Dr Suda by Monday 22nd June 2025.

Submission
A pdf document is to be uploaded in the assigned VLE area. Should you encounter issue when uploading then send by email to Dr Borg Inguanez and Dr Suda by the given deadline. Hard copies of the assignment will not be accepted.

Question 1 - [50 marks] (Dr Monique Borg Inguanez)

Find and download a dataset from the internet that includes **at least three quantitative variables** (more is better). Once you have selected a dataset, send the link via email to **Dr. Monique Borg Inguanez (monique.inguanez@um.edu.mt)** for approval, ensuring that each student works with a unique dataset.

Using the approved dataset, fit a **Bayesian multiple linear regression model** ($p < 2$) using **JAGS**.

Your submission should clearly include the following:

- A full specification of the Bayesian model:
  - The **likelihood** function
  - The **prior distributions** selected for the parameters
- A justification for the chosen priors, including any assumptions or reasoning behind them.
- The selected **burn-in period** for your MCMC chains, including a detailed explanation of how this was chosen.
- A presentation and interpretation of:
  - **Convergence diagnostics** (e.g., trace plots, Gelman-Rubin statistics)
  - **Accuracy diagnostics** (e.g., effective sample size, autocorrelation)
- A discussion of the **posterior distributions** obtained from the analysis based on the plots and summary statistics (which need to be presented in the text). Discuss also how the posterior distributions of the model parameters can be used to **predict new**

values of the response variable Y, given only values for the explanatory variables are available (prediction problem). Include both the interpretation of the parameter uncertainty and how it propagates into predictions for new observations.

- Also present the R script, including comments that explain what each section does in an appendix.

Question 2 - [50 marks] (Dr David Suda)

Let $\boldsymbol{x} = (x_1, \dots, x_n)'$ be a set of observations. Consider the following Bayesian model:

$$p(\mu, \sigma^2|\boldsymbol{x}) \propto \left[\prod_{i=1}^{n} N(x_i|\mu, \sigma^2)\right] N(\mu|m_0, v_0^2)\Gamma(\tau|a, b), \tau = \sigma^{-2}$$

We do not know explicitly the posterior here but we know through the theory of conjugate priors that:

$$p(\mu|\boldsymbol{x}, \sigma^2) \propto \left[\prod_{i=1}^{n} N(x_i|\mu, \sigma^2)\right] N(\mu|m_0, v_0^2) = N\left(\frac{n\tau\overline{\boldsymbol{x}} + v_0^{-2}m_0}{n\tau + v_0^{-2}}, (n\tau + v_0^{-2})^{-1}\right)$$

$$p(\tau|\boldsymbol{x}, \sigma^2) \propto \left[\prod_{i=1}^{n} N(x_i|\mu, \sigma^2)\right] \Gamma(\tau|a, b) = \Gamma\left(a + \frac{n}{2}, \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2} + b\right)$$

Do the following:

i. Decide on values for μ and σ^2 and simulate 100 values from a normal distribution with this mean and variance.

ii. Decide on hyperparameters m_0, v_0^2, a, b which yield a non-informative prior.

iii. Pick an initial value of τ. Use Gibbs' sampling to simulate 10000 values of μ and τ (excluding an appropriately chosen burn-in period).

iv. How can you use traceplot, the autocorrelation function and the effective sample size to determine whether the chain has converged, and whether the appropriate burn-in period has been eliminated? Use one or more of these techniques to determine whether your chain has converged and an adequate burn-in has been taken.

v. Make use of the *epdfPlot* command in the package *EnvStats* to determine the empirical pdf for μ, τ and σ^2 (of course, make sure that your chain has converged and do not include burn-in period).