Principles of statistical inference project - Part 2

Carmel Gafaż

June 18, 2025

1 Question 1

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:

The selected dataset is the Concrete Compressive Strength dataset, which was downloaded from the University of California, Irvine. The response variable of the dataset is the Concrete compressive strength as a function of eight predictors, including ingredients and ageing time.

The following table outlines the variables of the dataset, which had no missing values:

Variable Name	**Role**	**Type**	**Units**
Cement	Feature	Continuous	kg/m^3
Blast Furnace Slag	Feature	Integer	kg/m^3
Fly `ash`	Feature	Continuous	kg/m^3
Water	Feature	Continuous	kg/m^3
Superplasticizer	Feature	Continuous	kg/m^3
Coarse Aggregate	Feature	Continuous	kg/m^3
Fine Aggregate	Feature	Continuous	kg/m^3
Age	Feature	Integer	day
Concrete Compressive Strength	Target	Continuous	MPa

Table 1: Variables in the Concrete Compressive Strength dataset

A first test on the dataset involved a pairwise and visual inspection of the relationships to determine the correlation of the predictors with the response and to start forming some ideas about multicollinearity. These results are shown below in Figure 1 and Figure 2.

	cement	slag	ash	water
cement	1.00000000	-0.27521591	-0.397467341	-0.08158675
slag	-0.27521591	1.00000000	-0.323579901	0.10725203
ash	-0.39746734	-0.32357990	1.000000000	-0.25698402
water	-0.08158675	0.10725203	-0.256984023	1.00000000
superplasticizer	0.09238617	0.04327042	0.377503146	-0.65753291
coarseagg	-0.10934899	-0.28399861	-0.009960828	-0.18229360
fineagg	-0.22271785	-0.28160267	0.079108491	-0.45066117
age	0.08194602	-0.04424602	-0.154370516	0.27761822
strength	0.49783192	0.13482926	-0.105754916	-0.28963338

	superplasticizer	coarseagg	fineagg	age
cement	0.09238617	-0.109348994	-0.22271785	0.08194602
slag	0.04327042	-0.283998612	-0.28160267	-0.04424602
ash	0.37750315	-0.009960828	0.07910849	-0.15437052
water	-0.65753291	-0.182293602	-0.45066117	0.27761822
superplasticizer	1.00000000	-0.265999148	0.22269123	-0.19270003
coarseagg	-0.26599915	1.000000000	-0.17848096	-0.00301588
fineagg	0.22269123	-0.178480957	1.00000000	-0.15609470
age	-0.19270003	-0.003015880	-0.15609470	1.00000000
strength	0.36607883	-0.164934614	-0.16724125	0.32887300

	strength
cement	0.4978319
slag	0.1348293
ash	-0.1057549
water	-0.2896334
superplasticizer	0.3660788
coarseagg	-0.1649346
fineagg	-0.1672412
age	0.3288730
strength	1.0000000

Figure 1: Correlation matrix of all variables in the Concrete Compressive Strength dataset.

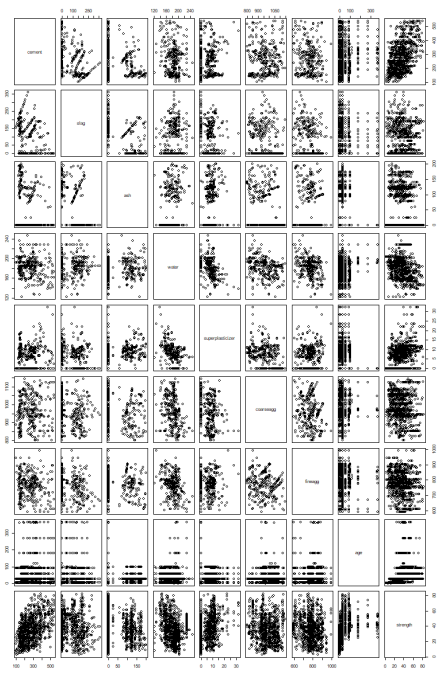

Figure 2: Pairwise scatterplot matrix of all variables in the Concrete Compressive Strength dataset.

From this result, the best correlation with strength is achieved with `cement`, followed by a moderate correlation with `superplasticiser` (+0.37) and `age` (+0.33). Water has a moderate negative correlation (-0.29). `Slag` and `ash`

```
Loading required package: carData
        cement           slag            ash          water
      7.488944       7.276963       6.170634       7.003957
superplasticizer     coarseagg       fineagg            age
      2.963776       5.074617       7.005081       1.118367
```

Figure 3: Variance Inflation Factor values for all predictors in the Concrete Compressive Strength dataset.

have a weak correlation. We can also see a strong negative correlation between `superplasticiser` and `water` (-0.66) and `water` and `fine aggregate` (-0.45).

Variance inflation factor analysis was then conducted to investigate multicollinearity, with the results of the analysis presented in Figure 3.

The analysis shows that cement, `Slag`, `ash`, and `water` exhibit high collinearity (in the region of 7), while `superplasticiser` (2.96) and `age` (1.12) have moderate values.

Following this investigation, cement, `superplasticiser`, and `age` were selected as predictors.

The following code snippet illustrates the data preparation. Prior to this step, the datasets's column names were shortened so they could be manipulated easily.

```
script_dir <- getwd()
concrete_cleansed_path <- file.path(
script_dir, "project_2_code/concrete_cleansed.csv")

concrete_cleansed <- read.csv(concrete_cleansed_path)

jags_data <- list(
x_cement = concrete_cleansed$cement,
x_superplasticizer = concrete_cleansed$superplasticizer,
x_age = concrete_cleansed$age,
y_strength = concrete_cleansed$strength,
n = nrow(concrete_cleansed)
)
```

1.1 A full specification of the Bayesian model:

- **The likelihood function**
- **The prior distributions selected for the parameters**

The response variable in this excercise is `strength`, representing the concrete compressive strength in MPa. As we have seen in the previous section, the predictors included in the model are `cement`, `superplasticizer`, and `age`, all selected based on correlation and multicollinearity analysis.

The model assumes the following structure:

$$y_i \sim \mathcal{N}(\mu_i, \tau^{-1}) \quad \text{for } i = 1, \ldots, n$$
$$\mu_i = \beta_0 + \beta_1 \cdot x_{\text{cement},i} + \beta_2 \cdot x_{\text{superplasticizer},i} + \beta_3 \cdot x_{\text{age},i}$$

The prior distributions for the parameters are defined as follows:

$$\begin{aligned}
\beta_0 &\sim \mathcal{N}(0, 100) \\
\beta_1 &\sim \mathcal{N}(0, 100) \\
\beta_2 &\sim \mathcal{N}(0, 100) \\
\beta_3 &\sim \mathcal{N}(0, 100) \\
\tau &\sim \text{Gamma}(0.01, 0.01)
\end{aligned}$$

The variance of the likelihood is modeled via the precision parameter τ, where;

$$\tau = \frac{1}{\sigma^2}$$

The model was defined using the following statement:

```
model_description <- "
model {
        for (i in 1:n) {
                y_strength[i] ~ dnorm(mu[i], tau)
                mu[i] <- beta0 + (beta1 * x_cement[i]) + (beta2 *
                    x_superplasticizer[i]) + (beta3 * x_age[i])
        }
        beta0 ~ dnorm(0, 0.01)
        beta1 ~ dnorm(0, 0.01)
        beta2 ~ dnorm(0, 0.01)
        beta3 ~ dnorm(0, 0.01)
        tau ~ dgamma(0.01, 0.01)
} "
```

1.2 A justification for the chosen priors, including any assumptions or reasoning behind them.

The priors in this model are weakly informative; that is, they do not strongly influence the results, but they give only a little information about what values we expect for the parameters. The coefficients for β, for example, were given normal priors centred at 0 with a standard deviation of 10, which allows for a large range of possible values. These priors are wide enough to let the data have the biggest impact on the results, but they also help the model stay stable and avoid extreme or unrealistic values during sampling.

- **Regression Coefficients (β_0, β_1, β_2, β_3):** Each coefficient was assigned a normal prior $\mathcal{N}(0, 100)$, corresponding to a mean of 0 and a large variance of 100, reflecting a belief that the effect of each predictor is centred around

zero but allows for a wide range of plausible values. As discussed, these priors are considered weakly informative because they do not strongly constrain the parameter estimates but prevent extreme values that could destabilize the sampling process.

- **Precision Parameter (τ):** A Gamma$(0.01, 0.01)$ prior was used for the precision of the normal likelihood, which is a standard weak prior for precision in Bayesian models. It gives a very wide range of possible values for the standard deviation, which means we do not assume much about how spread out the data is, reflecting that we have little prior knowledge about the variation in the response.

The priors were selected to reflect minimal assumptions about the parameter values in the absence of strong domain knowledge, which is appropriate given the exploratory nature of this analysis.

1.3 The selected burn-in period for your MCMC chains, including a detailed explanation of how this was chosen.

In Bayesian analysis using Markov Chain Monte Carlo, the initial samples of each chain will, in most cases, not reflect the target posterior distribution, and the chain will settle into the posterior distribution after a number of samples. This early phase is known as the burn-in period. These initial samples are discarded before subsequent analysis, as they would bias the results.

Trace plots are the first tool that is used to determine a suitable burn-in period. In this analysis, we examine chain mixing, which the overlap of the chain samples can visually assess, and compactness within the chain samples. Additionally, chain variations should be consistent around a central value. Trace plots, however, are subjective in interpretation and can indicate early convergence. Another tool that is then used is the Gelman-Rubin convergence diagnostic, which utilises both between-chain and within-chain variance to assess convergence.

Figure 4 shows the trace plot for the parameter β_0, that is, the intercept after a burn-in of just 200 samples. The trance can visually satisfy the criteria for convergence.

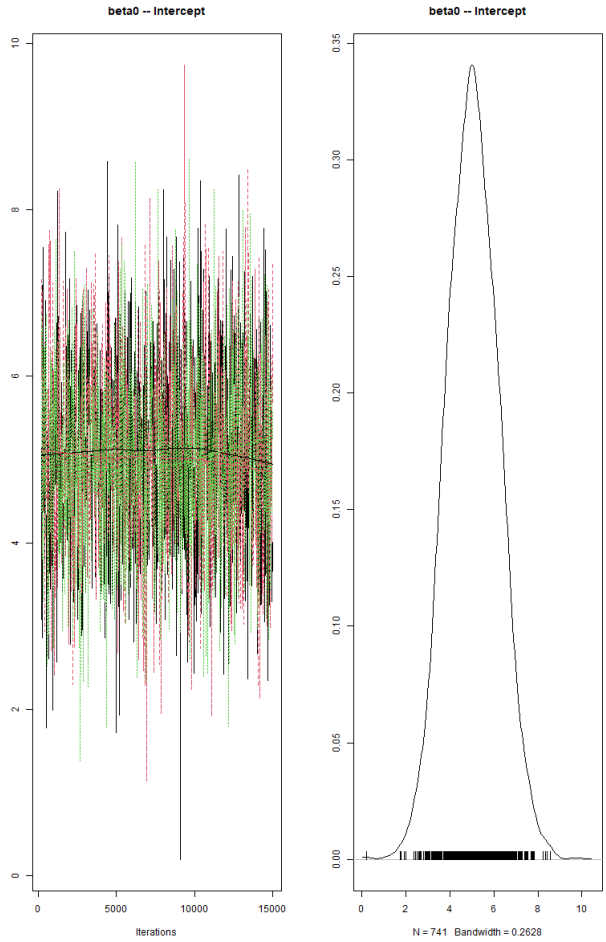

Figure 4: Trace and density plot for β_0 (Intercept) across 3 chains after iteration 200.

The Gelman-Rubin diagnostic was also examined. Figure 5 shows that the values for all parameters converge to 1.00 after iteration 6000, confirming chain convergence and the initial iterations were likely influenced by starting values.

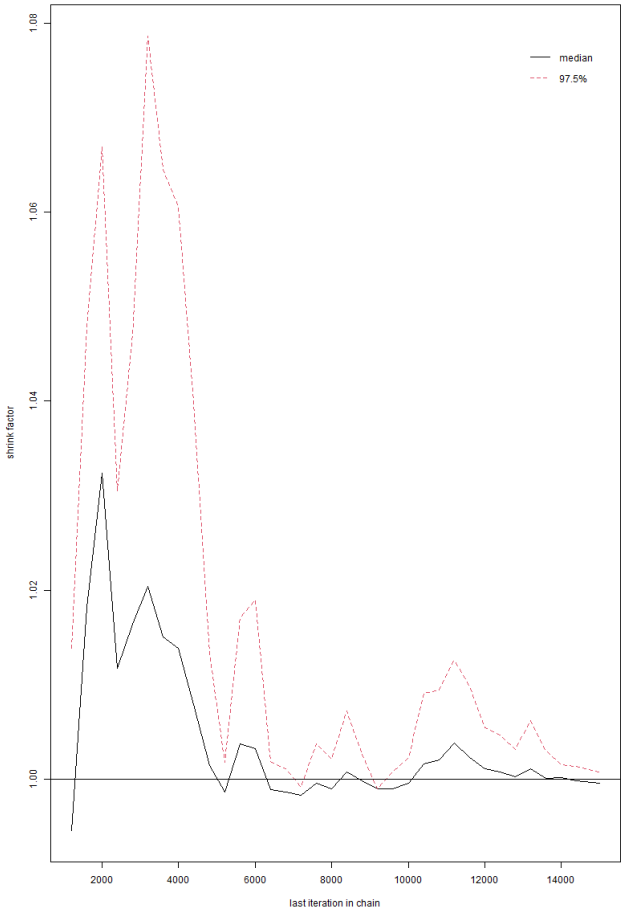

Figure 5: Gelman–Rubin diagnostic for β_0. Convergence is reached after approximately 6000 iterations

Based on these observations, a burn-in of 6000 iterations was selected to ensure that only samples drawn from the stationary portion of the chains were retained for posterior inference.

1.4 A presentation and interpretation of:

- **Convergence diagnostics (e.g., trace plots, Gelman-Rubin statistics)**
- **Accuracy diagnostics (e.g., effective sample size, autocorrelation)**

1.5 **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.**

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