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.

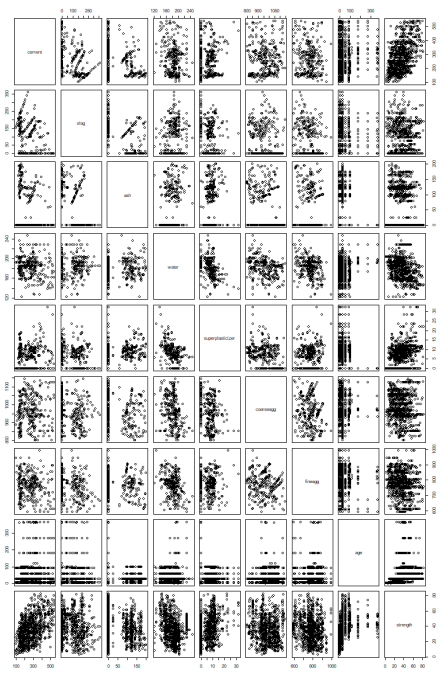

Figure 2: Pairwise scatterplot matrix of all variables in the Concrete Compressive Strength dataset.

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.

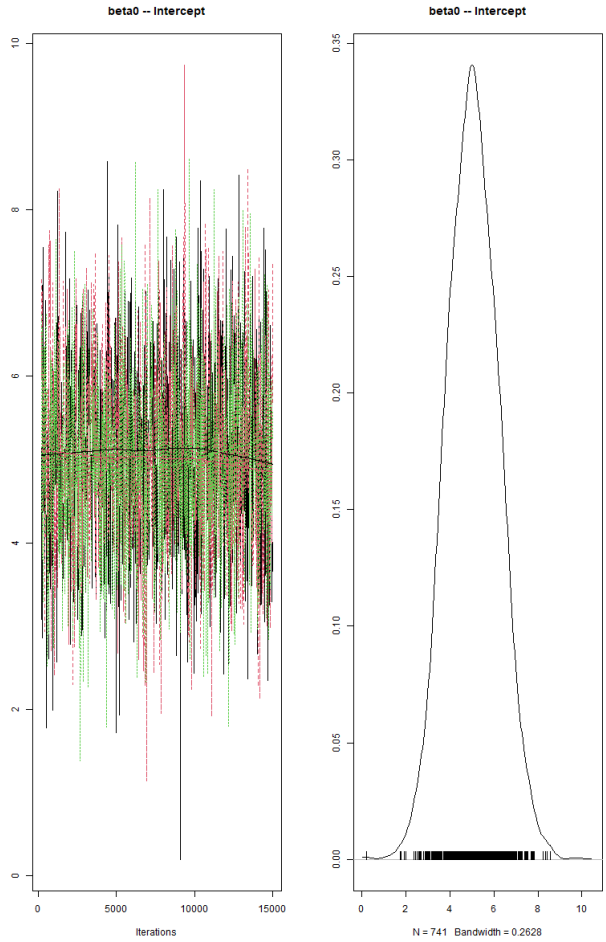

Figure 4: Trace and density plot for β_0 (Intercept) across 3 chains after iteration 200.

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.

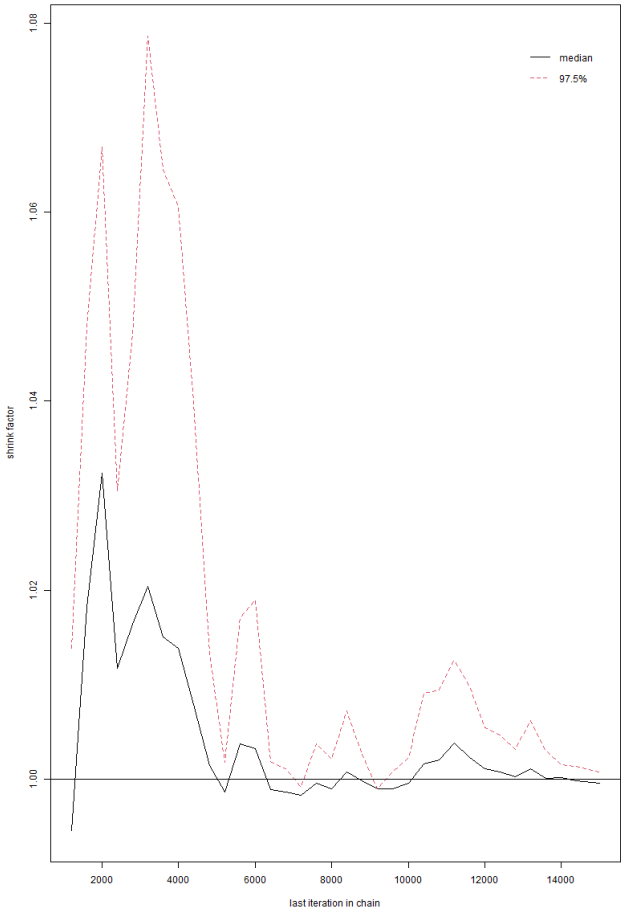

Figure 5: Gelman–Rubin diagnostic for β_0. Convergence is reached after approximately 6000 iterations

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.

The following snippet shows the execution of the model and creation of the plot used in this analysis

```
n_chains <- 3
n_burnin <- 200
n_samples <- 15000
parameters_to_monitor <- c("beta0", "beta1", "beta2", "beta3",
    "tau")

initial_values <- list(
```

```
list(beta0 = -10, beta1 = 0.5, beta2 = 0.1, beta3 = 0.05, tau =
    0.5),
list(beta0 = 0,   beta1 = 1,   beta2 = 0.3, beta3 = 0.1,  tau = 1),
list(beta0 = 10,  beta1 = 2,   beta2 = 0.5, beta3 = 0.2,  tau = 2)
)

model <- jags.model(textConnection(model_description),
data = jags_data,
inits = initial_values,
n.chains = n_chains)

post <- coda.samples(model = model,
variable.names = parameters_to_monitor,
n.iter = n_samples,
thin = 20)

post_burned <- window(post, start = n_burnin)
print(summary(post_burned))
plot(post_burned[, "beta0"], main="beta0 -- Intercept")
```

1.4 A presentation and interpretation of:

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

The Figures below illustrate convergence diagnostics for all monitored parameters in the model. Trace plots are shown after discarding an initial burn-in of 200 samples, highlighting chain behaviour and visual mixing. Additionally, Gelman-Rubin diagnostic plots are presented both after a burn-in of 200 samples and after a more conservative burn-in of 6000 samples. These plots allow for a comparative evaluation of convergence, confirming that the longer burn-in more effectively removed transient behaviour and yielded PSRF values stabilising around 1.00 across all parameters.

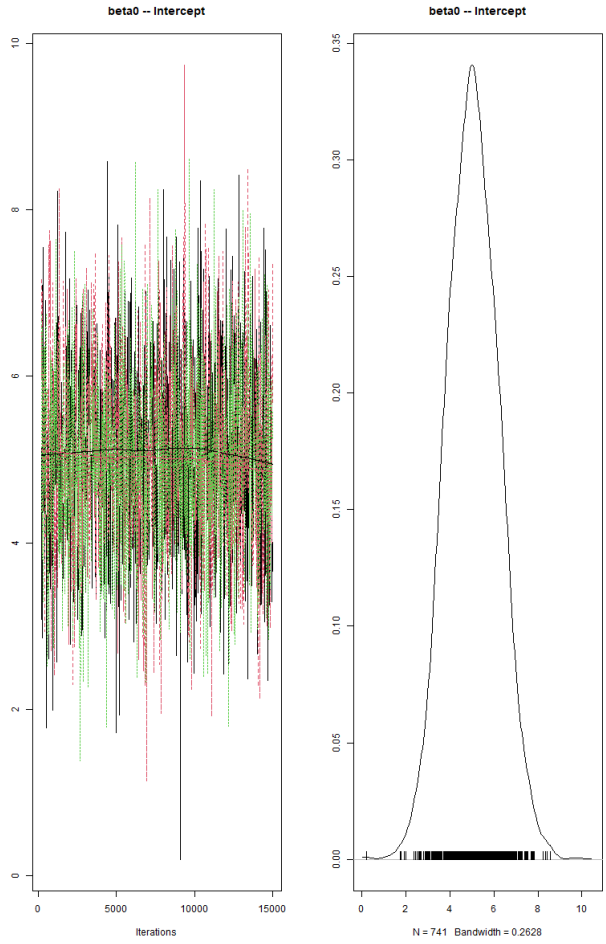

Figure 6: Trace and posterior density plots for the parameter β_0 (Intercept). The trace plot shows samples from three chains with good mixing and no signs of divergence or drift. All chains fluctuate around a common central value and remain tightly overlapped throughout the sampling period, a sign of stable convergence. The density plot shows a unimodal posterior distribution, peaking near 5.0 and spanning a broad range from approximately 2 to 8. The shape of the curve is slightly asymmetric near the peak, but remains smooth overall. The rug plot confirms dense and well-distributed sampling, supporting a reliable estimate of the intercept parameter. This indicates that the model has effectively identified a stable posterior distribution for β_0.

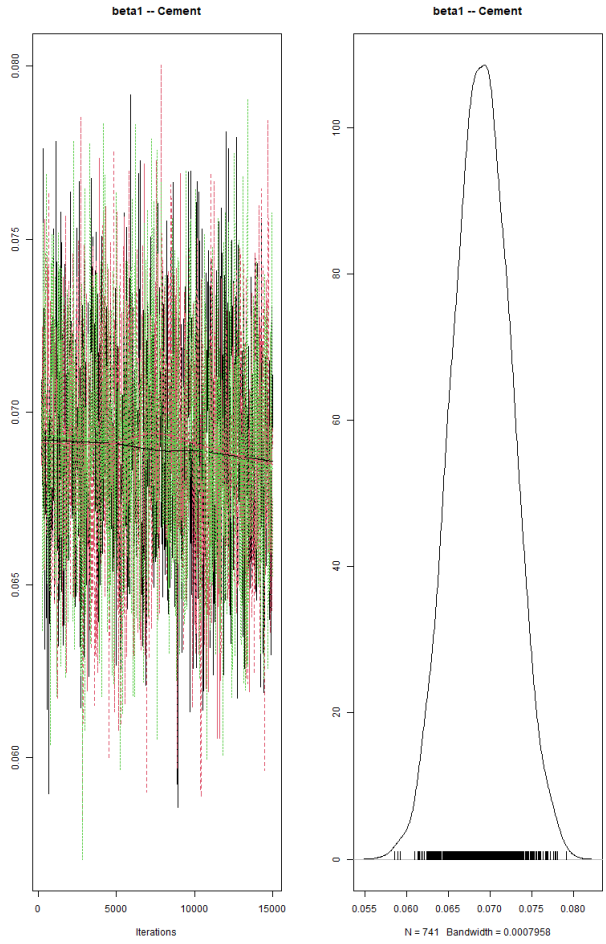

Figure 7: Trace and posterior density plots for the parameter β_1 (Cement). The trace plot on the left displays samples from three Markov Chain Monte Carlo (MCMC) chains. These chains show good mixing, with overlapping fluctuations around a stable mean after a burn-in period of 200 iterations. There is no visible drift or separation between the chains, and each chain explores the same region of the parameter space, indicating that convergence for β_1 has been achieved. The density plot on the right illustrates a smooth, unimodal posterior distribution, with the majority of the probability mass concentrated between approximately 0.060 and 0.078, peaking near 0.069. This suggests a strong positive effect of cement on concrete compressive strength.

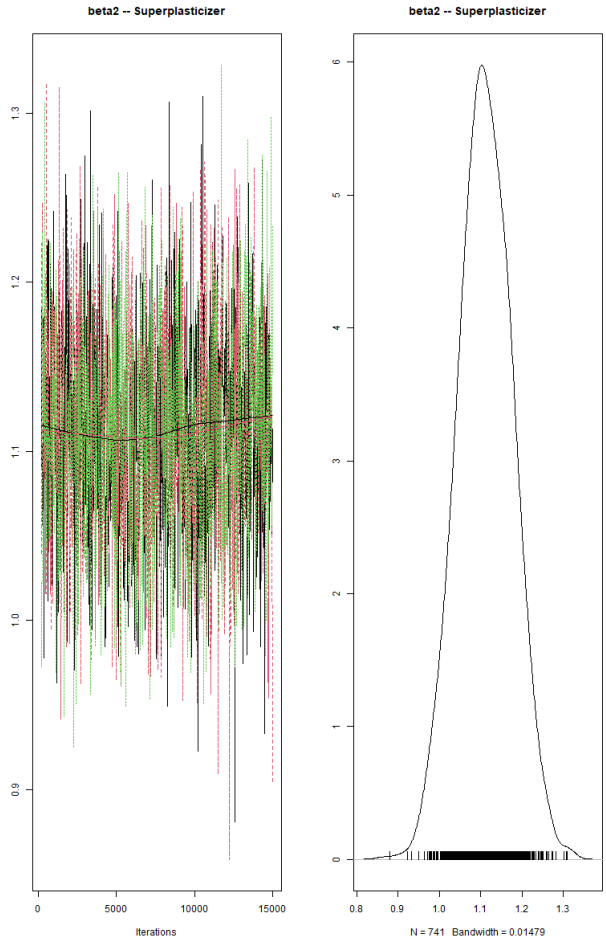

Figure 8: Trace and posterior density plots for the parameter β_2 (Superplasticizer). The trace plot shows three well-mixed chains with no indications of divergence, upward drift, or chain separation. The chains overlap consistently and fluctuate around a common mean, showing convergence and stationarity. The posterior density plot is unimodal, with the distribution between approximately 0.95 and 1.25, and a peak near 1.11. The curve is slightly asymmetric, with a marginally longer right tail, indicating slight skewness in the posterior. The rug plot confirms that sampling was dense and thorough across the high-probability region. These results indicate a stable and significant positive effect of superplasticizer on concrete compressive strength.

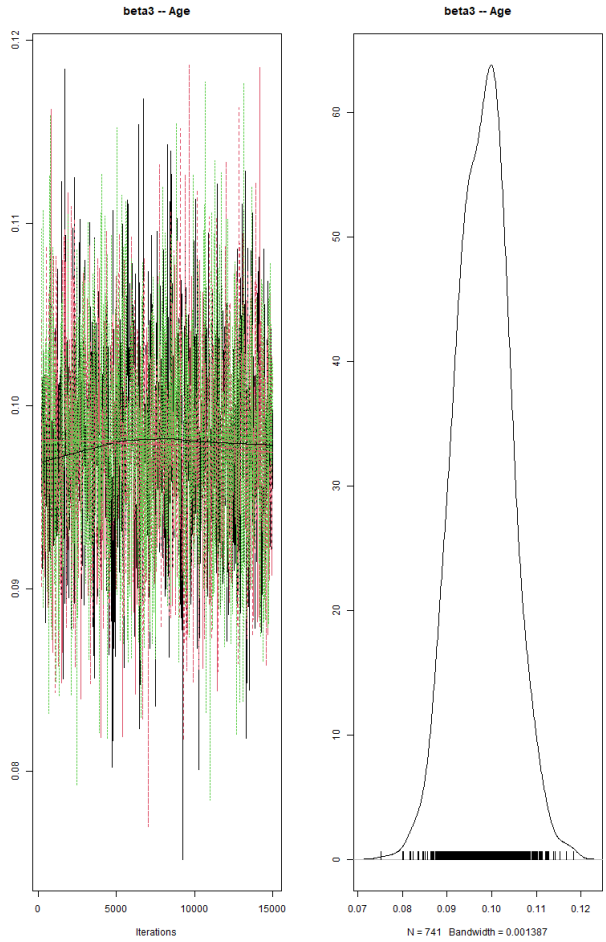

Figure 9: Trace and posterior density plots for the parameter β_3 (Age). The trace plot on the left displays three MCMC chains,exhibiting stable mixing with consistent fluctuations around a common central value, thus indicating convergence. The density plot on the right shows a unimodal posterior distribution concentrated between approximately 0.085 and 0.110. While the distribution has a clearly defined peak near 0.098, a slight asymmetry is visible near the mode, with a more gradual slope on the right side, indicating mild skewness. The rug plot confirms a dense and even sampling of values, supporting a reliable and stable estimate for the effect of age on compressive strength.

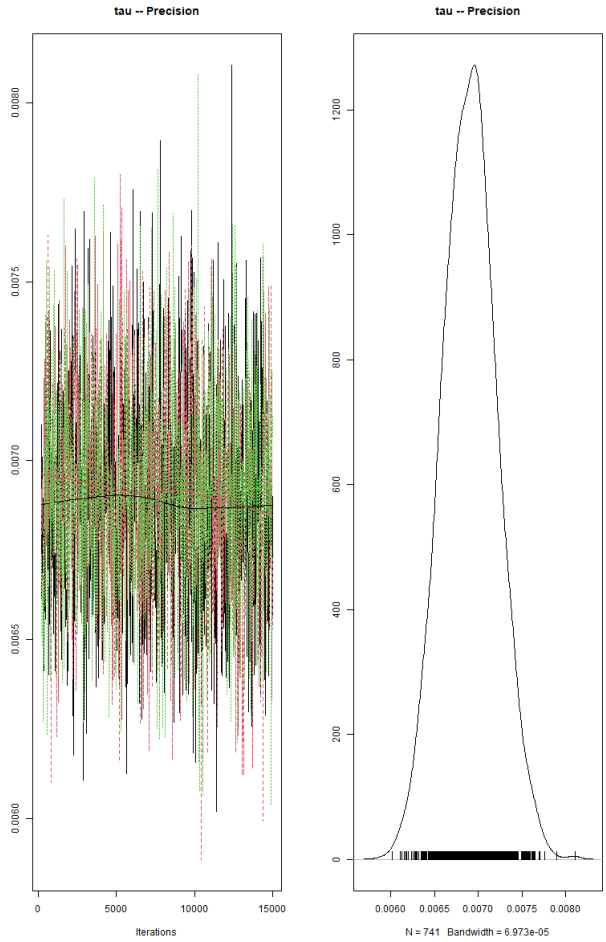

Figure 10: Trace and posterior density plots for the parameter τ (precision of the normal likelihood). The trace plot shows good chain mixing and stationarity across all three chains, with no divergence, drift, or separation after burn-in. Fluctuations are consistently centred around a stable mean, showing convergence. The posterior density plot reveals a narrow and sharply peaked unimodal distribution concentrated between approximately 0.0063 and 0.0078, peaking near 0.0069, showing a tightly constrained estimate of the model's residual precision. The rug plot below the density curve confirms dense and stable sampling from the posterior. This supports reliable inference for the precision parameter τ.

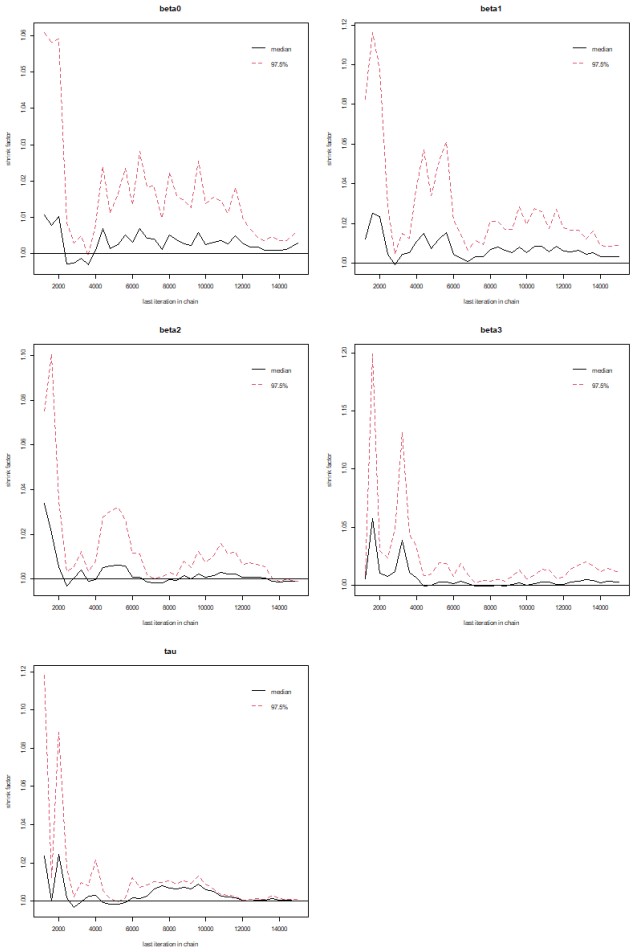

Figure 11: Gelman–Rubin diagnostic plots for all monitored parameters: intercept, cement, superplasticizer, age, and precision. The plots show the evolution of the shrink factor across increasing chain lengths. All parameters converge to Potential Scale Reduction Factor values close to 1.00 with their 97.5% confidence intervals stabilising below 1.05. Convergence for all parameters appears to occur around iteration 6000, after which the shrink factors remain flat and close to unity. This shows that the chains for all parameters have likely converged to the same posterior distribution and that no substantial between-chain variation remains.

To assess the reliability and efficiency of the Markov Chain Monte Carlo sampling process, we examined two key accuracy diagnostics: effective sample size and autocorrelation of the samples.

The effective sample sizes for all parameters were obtained using the sum-

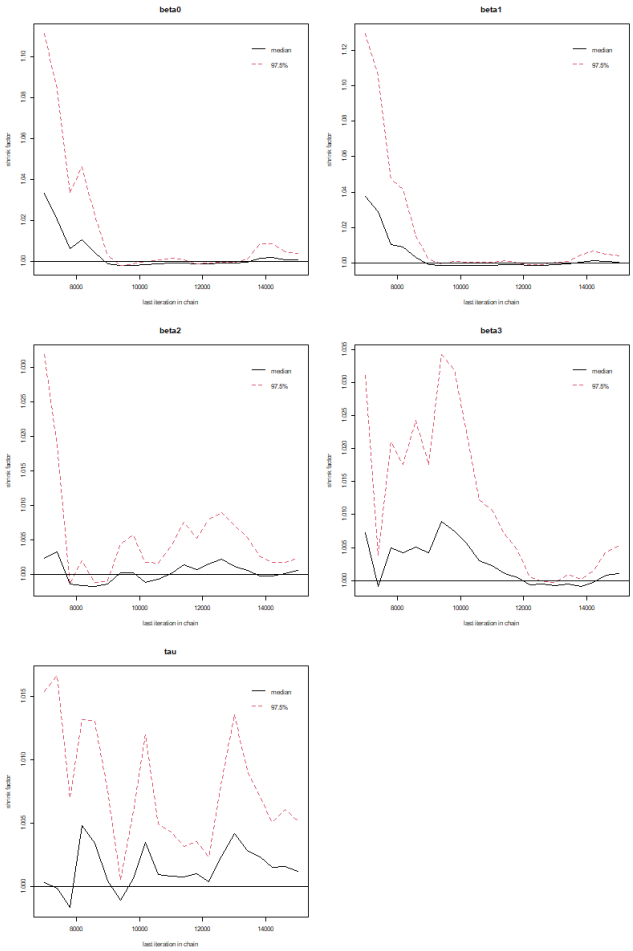

Figure 12: Gelman-Rubin diagnostic plots for all monitored parameters following the removal of the first 6000 iterations as burn-in. After burn-in, all parameters show Potential Scale Reduction Factor values stabilising near 1.00, with their 97.5% confidence intervals remaining below 1.05, indicating convergence. This confirms that discarding the initial 6000 iterations was effective in eliminating non-stationary behaviour, and that the retained samples are representative of the joint posterior distribution.

```
Iterations = 6000:15000
Thinning interval = 20
Number of chains = 3
Sample size per chain = 451

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean        SD  Naive SE Time-series SE
beta0 5.068960 1.1054650 3.005e-02      3.012e-02
beta1 0.068890 0.0035027 9.523e-05      9.781e-05
beta2 1.112023 0.0633109 1.721e-03      1.639e-03
beta3 0.097771 0.0060032 1.632e-04      1.572e-04
tau   0.006886 0.0003057 8.311e-06      8.262e-06

2. Quantiles for each variable:

          2.5%      25%      50%     75%    97.5%
beta0 2.995476 4.335459 5.045843 5.75967 7.325464
beta1 0.062154 0.066576 0.068897 0.07131 0.075520
beta2 0.984603 1.068858 1.109359 1.15179 1.243214
beta3 0.085775 0.093729 0.097781 0.10186 0.109553
tau   0.006284 0.006678 0.006876 0.00709 0.007479
```

H

Figure 13: Summary statistics for all model parameters after burn-in and thinning.

mary() function on the Markov Chain Monte Carlo output object.

The effective sample size tells us how many of the samples from the Markov Chain Monte Carlo simulation are as useful as completely independent values. This is important because, in Markov Chain Monte Carlo, the samples are usually correlated, meaning each one is slightly influenced by the one before it. So, even if we run many iterations, the actual amount of "independent information" is lower. In this analysis, each chain had about 451 effective samples for each parameter after discarding the burn-in and applying thinning (keeping every 20th sample). Since we ran three chains, that means we had over 1300 effective samples per parameter, meaning that the Markov Chain Monte Carlo sampling worked well and gave us enough reliable information to make accurate conclusions from the posterior distributions.

The time-series standard errors were small for all parameters (e.g., β_0: 0.030, β_1: 0.000098, β_2: 0.00164), confirming high sampling precision and low Monte Carlo error. These diagnostics collectively indicate that the model's parameter estimates are stable, reliable, and based on an efficient sampling process.

Autocorrelation plots are used to assess how much each sample in a Markov Chain Monte Carlo simulation depends on the samples that came before it. In a well-mixed Markov Chain Monte Carlo chain, we expect the correlation between successive samples to decline rapidly as the lag increases; that is, samples become effectively independent as we move further apart along the chain. In an autocorrelation plot, each vertical bar represents the correlation at a particular lag. A plot showing high autocorrelation across many lags indicates poor mixing and slow convergence, while a plot where autocorrelation drops quickly to near zero suggests good mixing and efficient sampling. Ideally, the autocorrelation should be close to zero after just a few lags, indicating that the chain is exploring

the posterior distribution effectively and that the resulting samples are reliable for inference.

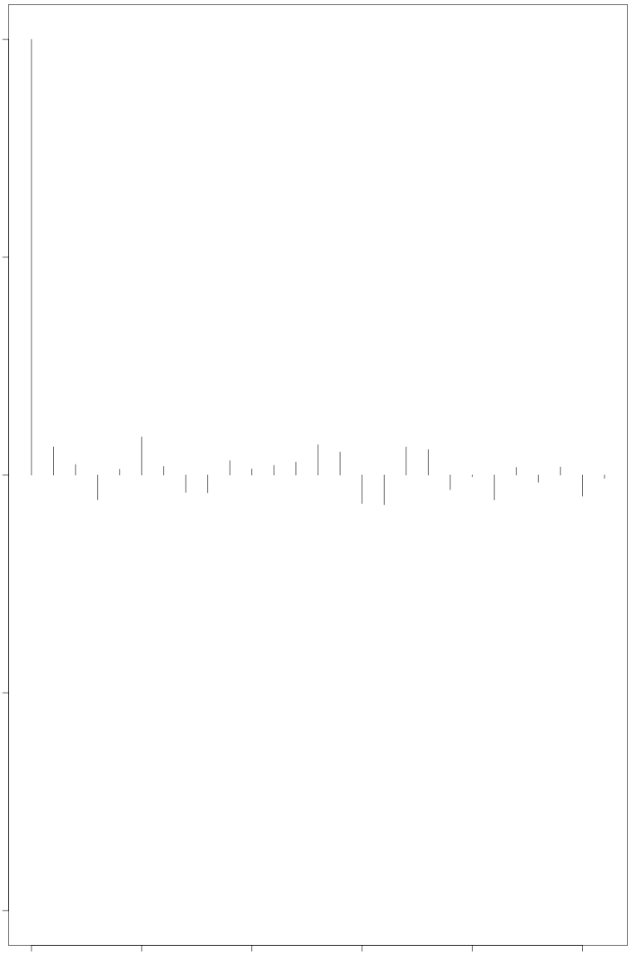

Figure 14: Autocorrelation plot for MCMC samples of β_0. The plot shows a rapid decline in autocorrelation after lag 1, with subsequent lags fluctuating closely around zero. This pattern indicates minimal serial correlation and suggests that the chain for β_0 mixes well, resulting in effectively independent samples suitable for posterior inference.

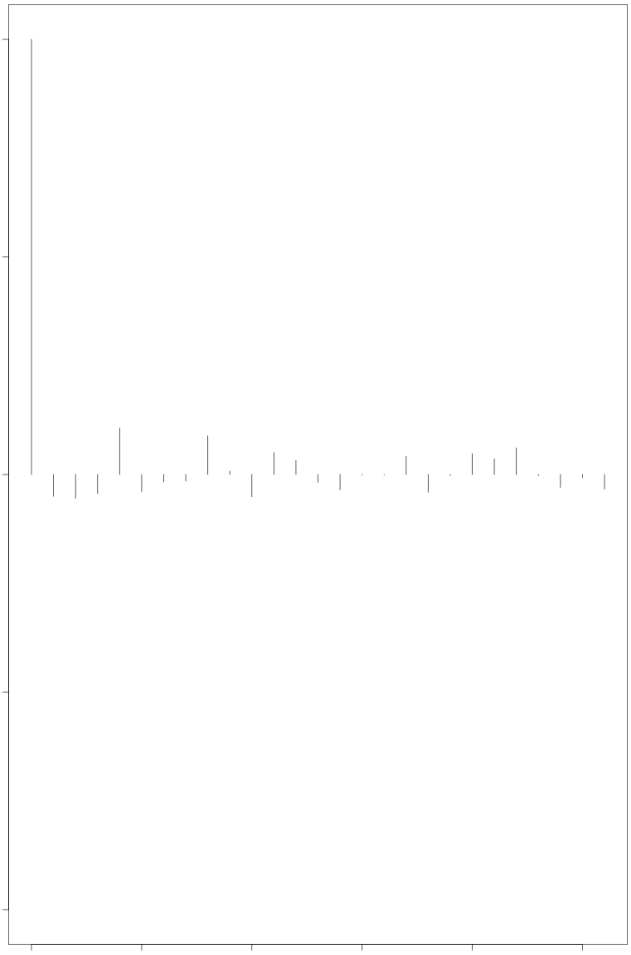

Figure 15: Autocorrelation plot for MCMC samples of β_1. As with β_0, the autocorrelation drops steeply after the first lag and remains near zero thereafter, confirming low intra-chain dependence and high sampling efficiency, implying that the posterior samples for β_1 are reliable and representative of the target distribution.

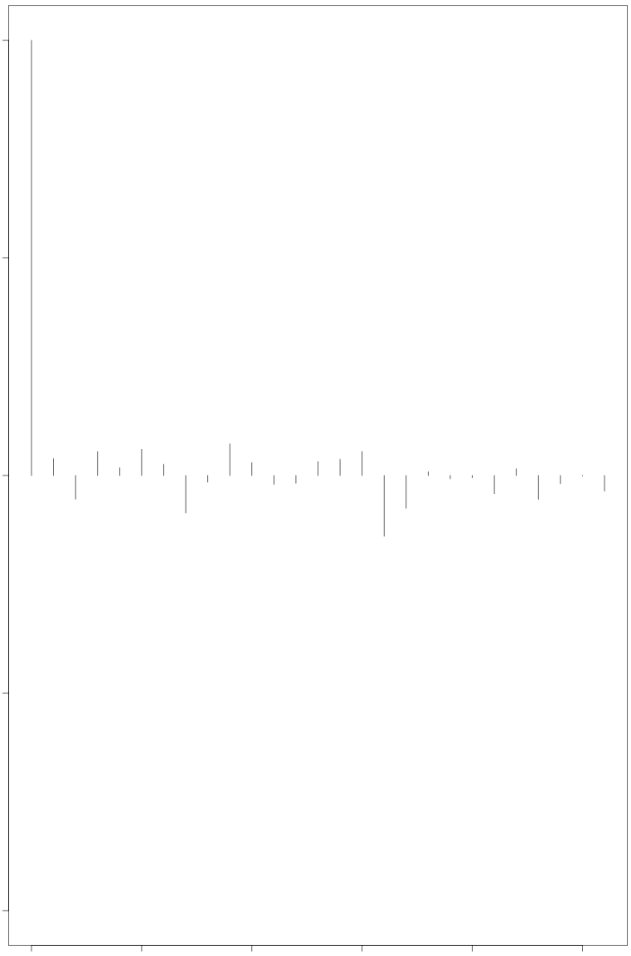

Figure 16: Autocorrelation plot for MCMC samples of β_2. The autocorrelation declines sharply after lag 1 and quickly levels off near zero for subsequent lags. This pattern indicates low dependence between successive samples and effective mixing of the chain, supporting the reliability of the posterior estimates for β_2.

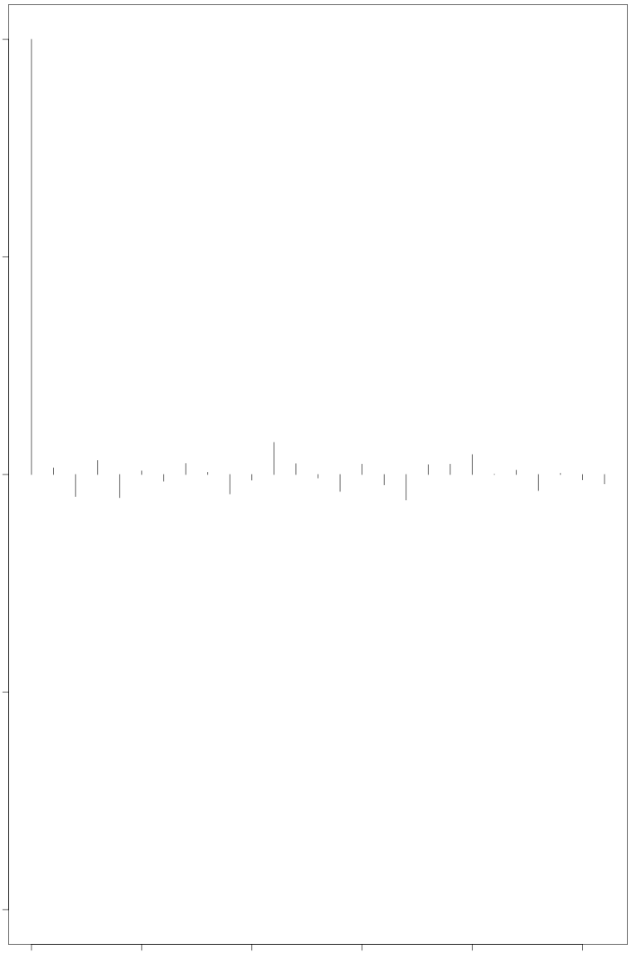

Figure 17: Autocorrelation plot for MCMC samples of β_3. The autocorrelation rapidly decays after lag 1 and remains close to zero for all subsequent lags, indicating minimal correlation between samples. This suggests that the MCMC chains for β_3 are mixing efficiently and producing effectively independent samples.

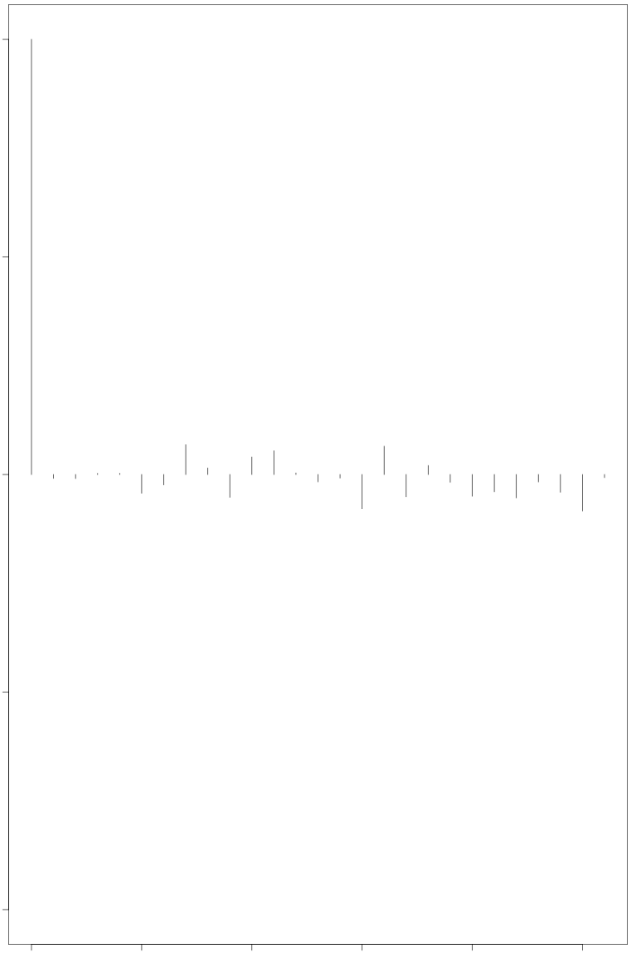

Figure 18: Autocorrelation plot for MCMC samples of τ (precision parameter). The autocorrelation drops steeply after lag 1 and fluctuates around zero for all subsequent lags, suggesting low serial correlation. This indicates that the MCMC sampler produced nearly independent draws, supporting efficient posterior estimation for the precision.

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.

The Bayesian linear regression model produced posterior distributions for all parameters; the intercept β_0, and the coefficients for cement (β_1), superplasticizer (β_2), and age (β_3), along with the residual precision parameter τ. The posterior summaries indicate how each predictor is associated with the concrete compressive strength after accounting for uncertainty and prior beliefs.

Table 2 presents the posterior mean, standard deviation, and the 95% credible interval for each parameter. None of the 95% credible intervals include 0, which suggests that all predictors have a statistically meaningful association with the response variable. The posterior distributions reveal meaningful associations between the predictors and the response variable. The cement coefficient (β_1) has a positive mean effect with a narrow credible interval, indicating a consistent and stable contribution of cement to compressive strength. The effect of superplasticizer (β_2) remains the most prominent, with its posterior centred substantially away from zero, suggesting a strong and reliably positive influence. The age coefficient (β_3) is modest but tightly concentrated, implying that even small increases in curing time yield measurable improvements in strength. The intercept term (β_0) exhibits greater variability, reflecting the model's baseline uncertainty when predictors are near zero. Finally, the residual precision parameter (τ) results in a relatively narrow distribution, indicating a well-estimated error structure and providing confidence in the model's predictive consistency. Collectively, these posterior characteristics reinforce the significance of each predictor while quantifying their uncertainty, supporting robust and interpretable inference.

Parameter	**Posterior Mean**	**SD**	**95% Credible Interval**
β_0 (Intercept)	4.672	11.23	[2.714, 7.294]
β_1 (Cement)	0.070	0.035	[0.062, 0.076]
β_2 (Superplasticizer)	1.116	0.101	[0.988, 1.243]
β_3 (Age)	0.098	0.007	[0.087, 0.110]
τ (Precision)	0.00689	0.00037	[0.00632, 0.00750]

Table 2: Posterior means, standard deviations, and 95% credible intervals.

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