# Bayesian Computing

Casnici Davide

June 2023

## 1 Introduction

This report aims to elucidate the relationship between study hours and academic performance using a Bayesian approach. Relying on a small dataset, we model grades as a function of study hours. We explore the prior and posterior distributions, compute credible intervals, and provide predictions, allowing us to analyze the potential impact of study time on academic success.

## 2 Dataset Analysis

The dataset is composed by 25 datapoints in total, each one composed by the study hours of a student and the respective taken grade. The statistical details of the dataset are shown in Table 1.

| | **Hours** | **Scores** |
|---|---|---|
| **Count** | 25.000000 | 25.000000 |
| **Mean** | 5.012000 | 51.480000 |
| **Std** | 2.525094 | 25.286887 |
| **Min** | 1.100000 | 17.000000 |
| **25%** | 2.700000 | 30.000000 |
| **50%** | 4.800000 | 47.000000 |
| **75%** | 7.400000 | 75.000000 |
| **Max** | 9.200000 | 95.000000 |

Table 1: Dataset Descriptive Statistics

Upon examining the dataset, it becomes evident that the average study hours reported is relatively low, standing at a mere 5 hours. Additionally, the average grade achieved by the students is just over the midpoint, slightly exceeding 50 on a scale that extends up to a maximum of 100.

To gain a deeper understanding of the properties of our features, we examine Figure 1. This figure presents the histograms for both 'Hours' and 'Scores' variables, enabling a comprehensive analysis of their distributions.

An analysis of our dataset reveals a high linear correlation of 0.977 between the 'Hours' and 'Scores' variables, indicating a substantial linear relationship. This correlation is

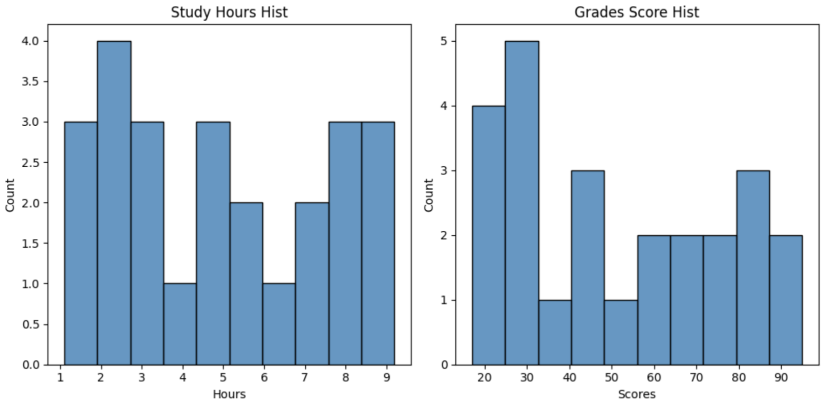

Figure 1: Features histogram

further accentuated in Figure 2. Such a strong linear association supports the selection of a linear model for further analysis, which will be elaborated upon in the subsequent section.

Looking closely at Figure 2, we can notice how I have divided the data points into three groups for ease of understanding. Even though a clear linear trend is visible across these groups, there's a lot of variability within each group. The way these points are spread within groups looks like a Gaussian, or ”bell curve”, pattern. This suggests that using a Bayesian linear regression model, where we assume the errors follow a normal distribution, could be a good fit for this data.

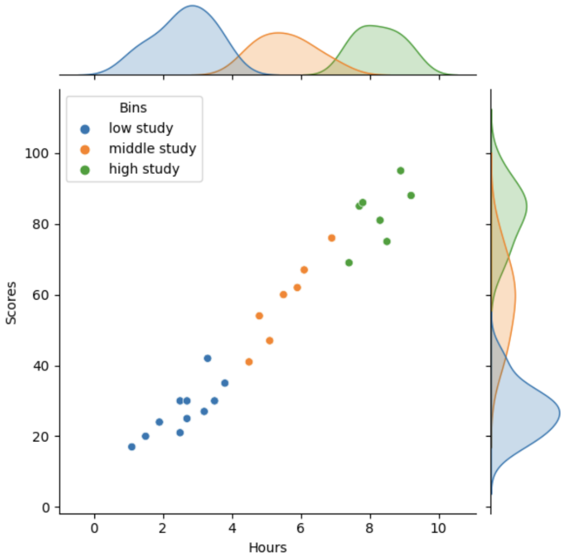

Figure 2: Features jointplot

# 3 Model and Prior Selection

## 3.1 Parameter Priors in the Context of Our Model

Following the discussion and motivations of the previous section, we therefor assume our model to be distributed as:

$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2) \tag{1}$$

The likelihood function $L(\beta_0, \beta_1, \sigma^2|X, Y)$ for this model is then given by:

$$L(\beta_0, \beta_1, \sigma^2|X, Y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\beta_0 + \beta_1 x_i - y_i)^2}{2\sigma^2}\right) \tag{2}$$

where $n$ is the number of observations, $x_i$ and $y_i$ are the observed values of $X$ and $Y$ for the $i$-th observation, and $\beta_0$, $\beta_1$, and $\sigma^2$ are the parameters to be estimated.

As prior for my data I have decided to use the following distributions:

$$\pi(\beta_0) = \text{Gamma}(20, 1) = \frac{1^{20}}{\Gamma(20)} \beta_0^{20-1} e^{-1\cdot\beta_0} \propto \beta_0^{19} e^{-\beta_0} \tag{3}$$

To ensure that the intercept remains positive (given that grades range between 0 and 100), the intercept has been modeled using a Gamma distribution. The selected parameters for this distribution are alpha and beta, with respective values of 20 and 1.

This decision arises from personal experience, having observed instances where students achieved scores significantly higher than zero without dedicating study time to prepare for exams. This might be attributable to luck or the ability to assimilate information from lectures alone. To encapsulate this prior knowledge, the expected value of the prior has been set at 20 with a variance of 20 (controlled by the parameters). Although this lends the distribution certain rigidity, it still maintains enough flexibility to adapt to the observed data.

$$\pi(\beta_1) = \text{N}(5, 2.5^2) = \frac{1}{\sqrt{2\pi(2.5)^2}} e^{-\frac{(\beta_1-5)^2}{2\cdot(2.5)^2}} \propto e^{-\frac{(\beta_1-5)^2}{12.5}} \tag{4}$$

The Gaussian distribution, characterized by a mean of 5 and variance of $(2.5)^2$, was chosen as the prior for the slope parameter $\beta_1$, primarily due to two considerations. Firstly, the academic setting under observation typically shows that an increase in study hours leads to an improvement in grades. Setting the mean of the prior at 5 encapsulates this understanding that more study generally leads to better academic performance, while still allowing for the influence of other factors.

Secondly, employing a Gaussian distribution for the prior introduces a level of flexibility that can accommodate a range of scenarios. By opting for a relatively high variance $(2.5)^2$, the model caters to the variability in how study hours might influence different students' grades, acknowledging individual differences and other potential factors.

However, it's important to note that while this choice of prior introduces some level of prior knowledge into the model, the chosen distribution remains largely uninformative. This is evidenced by its Coefficient Of Variation (C.O.V.) of 1, indicating a significant degree of variability relative to the mean. In essence, this allows the observed data to primarily drive the results, while the prior serves to provide a sensible starting point based on our underlying assumptions.

$$\pi(\sigma) = \text{Gamma}(2, 5) = \frac{5^2}{\Gamma(2)}\sigma^{2-1}e^{-5\cdot\sigma} \propto \sigma^1 e^{-5\cdot\sigma} \tag{5}$$

The error term of our model, denoted as $\sigma$, is modeled with a Gamma distribution as the prior. Two key reasons underscore this choice.

The initial motivation for selecting a Gamma distribution as the prior for the error term $\sigma$ arises from a practical necessity. Within the PyMC computational framework that we are utilizing, the standard deviation parameter of a normal distribution, represented here by $\sigma$, is required to be positive. The Gamma distribution inherently fulfills this requirement, as it is defined exclusively over positive values. This prevents the need for imposing additional artificial constraints during model implementation.

Secondly, the specific parameterization of the Gamma distribution with shape=2 and scale=5 reflects our expectations about the data. The Gamma distribution with these parameters is left-skewed, indicating a belief that smaller values for the standard deviation are more likely. This fits with our intuitive understanding that most students' grades will be clustered around the mean. However, the long tail of the Gamma distribution also allows for the possibility of large deviations, accommodating potential outliers or unanticipated variability in the data.

Thus, this choice of prior allows us to incorporate our expectations about the grade distribution while still being flexible enough to adapt to the actual data observed.

## 3.2 Full Posterior

$$\pi(\beta_0, \beta_1, \sigma \mid D) \propto L_D(\beta_0, \beta_1, \sigma)P(\beta_0)P(\beta_1)P(\sigma)$$

$$\pi(\beta_0, \beta_1, \sigma \mid D) \propto \sigma^{-n} \exp\left\{-\frac{1}{2}\sum\left(\frac{\beta_0 + \beta_1 x_i - y_i}{\sigma}\right)^2\right\} \beta_0^{19} \exp\{-\beta_0\} \exp\left\{-\frac{1}{2}\left(\frac{\beta_1 - 5}{2.5}\right)^2\right\} \sigma \exp\{-5\sigma\}$$

$$\pi(\beta_0, \beta_1, \sigma \mid D) \propto \beta_0^{19}\sigma^{1-n} \exp\left\{-\frac{1}{2}\sum\left(\frac{\beta_0 + \beta_1 x_i - y_i}{\sigma}\right)^2 - \beta_0 - \frac{1}{2}\left(\frac{\beta_1 - 5}{2.5}\right)^2 - 5\sigma\right\}$$

## 3.3 Full Conditionals

Considering $S_y$ and $S_x$ as the summation from $i$ to $n$ of $y_i$ and $x_i$ respectively, and the $SS_y$ and $SS_x$ as their squared sum, and lastly $\sum_{i=1}^{n} x_i y_i = CS_{xy}$, we have the following full conditionals.

Full conditional for $\beta_0$:

$$\begin{aligned}\pi(\beta_0 \mid -\beta_0) &\propto \beta_0^{19} \exp\{-\beta_0\} \exp\left\{-\frac{1}{2\sigma^2}\left(n\beta_0^2 + 2\beta_0\beta_1 S_x - 2\beta_0 S_y\right)\right\} \\ &\propto \beta_0^{19} \exp\left\{-\frac{1}{2\sigma^2}\left(n\beta_0^2 + 2\beta_0\beta_1 S_x - 2\beta_0(S_y - \sigma^2)\right)\right\} \\ &\propto \beta_0^{19} \exp\left\{-\frac{n}{2\sigma^2}\left(\beta_0^2 + 2\beta_0\left[\frac{\beta_1 S_x - S_y + \sigma^2}{n}\right] + ...\right)\right\}\end{aligned}$$

$$\propto \beta_0^{19} \exp\left\{-\frac{n}{2\sigma^2}\left(\beta_0^2 + \frac{\beta_1 S_x - S_y + \sigma^2}{n}\right)^2\right\}$$

Full conditional for $\beta_1$ (Let $q$ be 2(6.25)):

$$\pi(\beta_1 \mid -\beta_1) \propto \exp\left\{-\frac{1}{2}\left(\frac{\beta_1 - 5}{2.5}\right)^2\right\} \exp\left\{-\frac{1}{2\sigma^2}\left(2\beta_0\beta_1 n\bar{X} + n\beta_1^2 SS_x - 2\beta_1 CS_{xy}\right)\right\}$$

$$\pi(\beta_1 \mid -\beta_1) \propto \exp\left\{-\frac{1}{2(6.25)}\left(\beta_1^2 - 10\beta_1\right)\right\} \exp\left\{-\frac{1}{2\sigma^2}\left(nSS_x\beta_1^2 - 2\beta_1[CS_{xy} - n\bar{X}\beta_0]\right)\right\}$$

$$\pi(\beta_1 \mid -\beta_1) \propto \exp\left\{-\beta_1^2\left[\frac{nSS_x}{2\sigma^2} + \frac{1}{q}\right] + 2\beta_1\left[\frac{CS_{xy} - n\bar{X}\beta_0}{2\sigma^2} + \frac{10}{q}\right]\right\}$$

$$\pi(\beta_1 \mid -\beta_1) \propto \exp\left\{-\left[\frac{nSS_x}{2\sigma^2} + \frac{1}{q}\right]\left[\beta_1^2 - 2\beta_1\frac{\left[\frac{CS_{xy} - n\bar{X}\beta_0}{2\sigma^2} + \frac{10}{q}\right]}{\left[\frac{nSS_x}{2\sigma^2} + \frac{1}{q}\right]}\right]\right\}$$

$$\pi(\beta_1 \mid -\beta_1) \sim N\left(\frac{\left[\frac{CS_{xy} - n\bar{X}\beta_0}{2\sigma^2} + \frac{10}{q}\right]}{\left[\frac{nSS_x}{2\sigma^2} + \frac{1}{q}\right]}, \frac{1}{2\left[\frac{nSS_x}{2\sigma^2} + \frac{1}{q}\right]}\right)$$

Full conditional for $\sigma$:

$$Q(x, y, \beta_0, \beta_1) = n\beta_0 + 2\beta_0\beta_1 S_x + n\beta_1^2 SS_x - 2\beta_0 S_y - 2\beta_1 CS_{xy} + SS_y$$

$$\pi(\sigma \mid -\sigma) \propto \sigma \exp\{-5\sigma\}\sigma^{-N} \exp\left\{-\frac{1}{2\sigma^2}Q(x, y, \beta_0, \beta_1)\right\}$$

$$\pi(\sigma \mid -\sigma) \propto \sigma^{1-n} \exp\left\{-\frac{1}{2\sigma^2}\left(Q(x, y, \beta_0, \beta_1) + 10\sigma^3\right)\right\}$$

# 4 Convergence Analysis

The complexity of dealing directly with proportional probability distributions and the difficulty of sampling from these non-standard forms often calls for the utilization of advanced sampling techniques (indeed the ones seen during the course). In this study, Gibbs sampling could be used to approximate the parameters, followed by Markov Chain Monte Carlo (MCMC) sampling for drawing from the posterior distribution. However, due to time constraints and the involved nature of these methods rather than code this techniques completely from scratch, a more practical approach was adopted.

This work employed the PyMC framework, a Python library that abstracts away many of the low-level details of these sampling techniques. Instead of manually implementing the entire Gibbs sampling method, PyMC uses an advanced technique called Hamiltonian Monte Carlo sampling, which we have explored and discussed in a seminar during the course.

With PyMC, one only needs to specify the model's distributions and the priors on its parameters. The framework takes care of the rest, setting up and running the MCMC chains that will converge to the distributions of the parameters and the full posterior. In

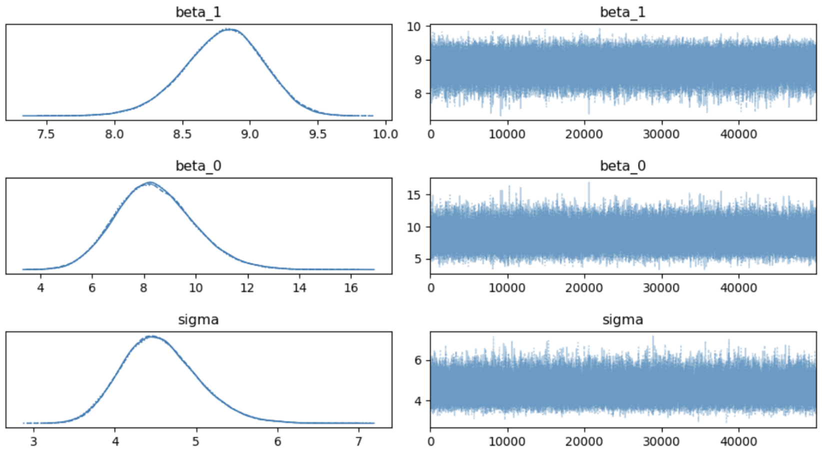

Figure 3: Chains traceplots

this case, three chains were constructed for each parameter, enhancing the robustness of the results, as displayed in Figure 3.

Without zooming a little it is difficult to clearly distinguish the distributions of the three chains, meaning that they converged to almost the same distribution. As we can notice from the traceplots, the chains performed well, with a good exploration of the parameters space. To create this chain, a tuning time (or more technically Burn-in) of 50 thousands steps have been used, then the markov chain continued for other 50 thousands steps. these values have been selected empirically doing some experiments.

The autocorrelation function of the three parameters – $\beta_1$, $\beta_0$, and $\sigma$ – shown in Figure 4 allows us to study the relationship between lagged values within the MCMC chains. By examining autocorrelation, we can understand how much the current sample in the chain depends on previous ones.

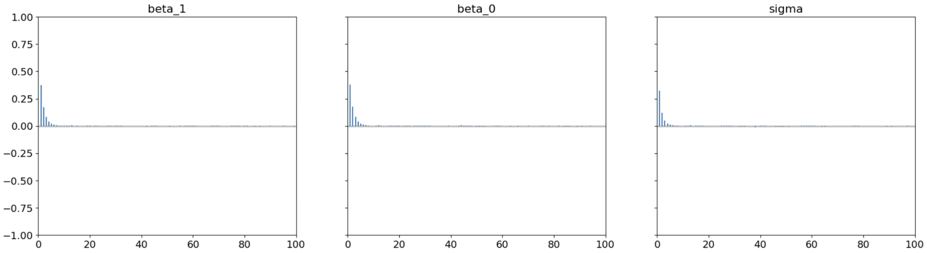

Figure 4: Parameters' chain Auto-Correlation Function

For each of our parameters, the autocorrelation function reveals a small amount of correlation in the first 5 to 6 lags, but this correlation decreases sharply and quickly approaches negligible levels. A correlation of less than 0.5 in the initial lags is acceptable and indicates a relatively efficient sampling process. If the autocorrelation were high, this would imply that our Markov chain is not exploring the parameter space efficiently, requiring more samples to achieve the same level of precision.

The observed autocorrelation patterns suggest that our MCMC sampler has done a good job of exploring the posterior distribution. The rapid decline in autocorrelation after the initial few lags indicates a good mixing of the chains. Therefore, the chains are likely to be representative of the underlying posterior distribution, but this is even more highlighted by the chains summary table reported below.

| | Mean | HDI_3% | HDI_97% | MCSE_Mean | ESS_Bulk | ESS_Tail | $\hat{R}$ |
|---|---|---|---|---|---|---|---|
| $\beta_1$ | 8.8 ± 0.29 | 8.246 | 9.339 | 0.001 ± 0.001 | 61644.106 | 63718.781 | 1.0 |
| $\beta_0$ | 8.4 ± 1.5 | 5.679 | 11.270 | 0.006 ± 0.004 | 61398.262 | 68257.383 | 1.0 |
| $\sigma$ | 4.5 ± 0.47 | 3.714 | 5.465 | 0.002 ± 0.001 | 72261.479 | 80354.488 | 1.0 |

Table 2: Posterior parameters estimation.

The results in Table 2 presents the MCMC process summary. The columns represent different aspects of the posterior distribution for each parameter ($\beta_1$, $\beta_0$, and $\sigma$) based on the samples generated by the PyMC framework.

The 'Mean' columns provide the average values for each parameter across all the posterior samples. We note that the slope ($\beta_1$) is estimated at around 8.8, while the intercept ($\beta_0$) has an average value of approximately 8.4. The standard deviation, $\sigma$, is more precisely 4.555, showing that there's a fair degree of uncertainty in the model's predictions.

The 'SD' column has been removed to let the table properly fit in the page, and the standard deviations of the samples have been putted right next to the mean values, as well as for the 'MCSE_SD' column.

The 'HDI_3%' and 'HDI_97%' columns provide the boundaries of the 94% Highest Density Interval (HDI). The HDI is the most credible range of values for the parameters. For $\beta_1$, 94% of the posterior samples fall within the interval 8.246 to 9.339. For $\beta_0$, this interval is between 5.679 and 11.270. And for $\sigma$, it is between 3.714 and 5.465. This suggests that, while our point estimates provide a single "best guess" for each parameter, there's a fair degree of uncertainty. The 'MCSE_Mean' is the Monte Carlo Standard Errors for the mean estimate. This provide san indication of the precision of the Monte Carlo estimates. The smaller this value, the better (for the relative standard deviation too).

The 'ESS_Bulk' and 'ESS_Tail' columns indicate the Effective Sample Size for the bulk of the distribution and the tails respectively. ESS is a measure of the number of "effective" samples - the higher this number, the better the estimate. It is also used as a diagnostic tool for the convergence of the chains. In our case, the ESS values are high, suggesting the chains have sampled well from the distributions.

Lastly, the $\hat{R}$ statistic, also known as the potential scale reduction factor, is an indicator of the convergence of the MCMC chains. A value of 1.0 indicates perfect convergence. Here, all parameters show $\hat{R}$ values of 1.0, which suggest good convergence and thus reliable results.

Overall, these parameter estimates and their corresponding diagnostics give us confidence in the robustness of the fitted Bayesian model and the validity of the posterior estimates.

# 5 Results

We now compare the Bayesian approach with the frequentist approach. In Figure 5, we display the fitted model obtained using the maximum likelihood estimation (MLE) approach, represented in red. The MLE approach solely relies on the observed data without incorporating any prior assumptions. It is evident that the MLE model fits the data slightly better, as it minimizes the least square error. However, it is crucial to note that the MLE approach may not be the most appropriate solution. As it relies solely on the observed data, it can lead to conclusions that may not be representative of the underlying reality due to the inherent randomness of the samples. On the other hand, the Bayesian approach incorporates our prior beliefs about the parameters governing the underlying distributions. This approach offers a more robust solution by accounting for both the observed data and our prior knowledge, allowing for a more comprehensive and reliable analysis. The Bayesian approach, incorporating both the data and prior beliefs, offers a robust framework that addresses the limitations of relying solely on observed data. This makes it advantageous, particularly in situations where incorporating prior knowledge is beneficial and when dealing with limited samples.

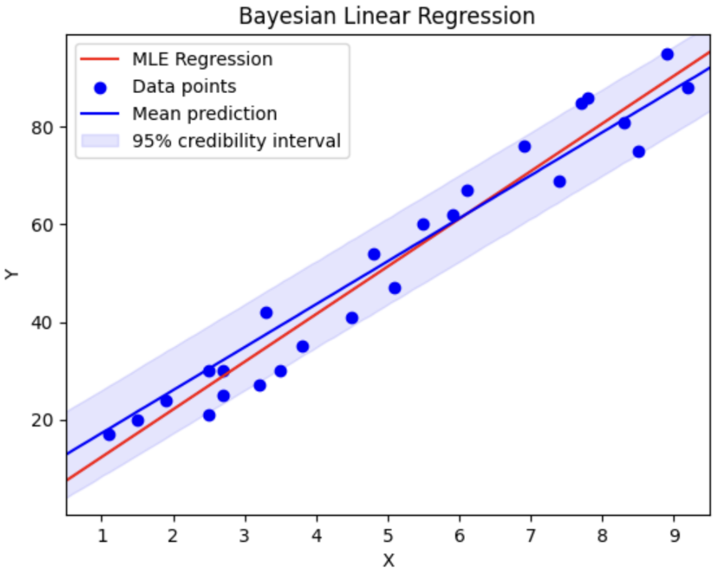

Figure 5: Fitted models

In the Bayesian setting, we have the advantage of obtaining a credibility interval, which represents the range of values that is likely to contain a sample from the model given a specific input. In our case, we have calculated a 95% credibility interval, which means that it has a 95% probability of containing the true sample from the model.

To calculate this interval, we sampled one hundred thousand values from the model for each input value. Since the distributions of the parameter samples are not perfectly Gaussian, we chose to use the median value rather than the mean value as a representative value for each parameter. This choice ensures that the sampled values are not overly influenced by extreme values in the distribution, providing a more robust estimation of the model's behavior (it is worth it to mention that i have tried both the approaches and the results were really similar).

To make a prediction for a new data point, I followed the same approach as before. I sampled one hundred thousand values from the model for a student who studied 6.5 hours. The resulting histogram of these samples is shown in Figure 6.

The mean value of the predicted scores for this student is 65.67, indicating the expected performance. Additionally, the 95% credibility interval, which represents the range of plausible values, is calculated as $[58.26, 78.08]$. This suggests that there is a high probability (95%) that the student's actual score falls within this range give the 6.5 hours of study.

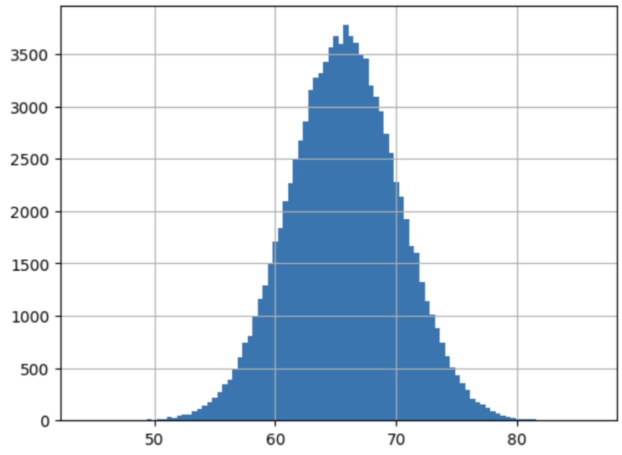

Figure 6: New prediction histogram

These results highlight the strength of Bayesian inference in not only providing point predictions but also quantifying uncertainty. This ability to quantify uncertainty is particularly valuable in practical applications where understanding and managing uncertainty is crucial, sometimes even more important than the precise prediction itself. Bayesian inference allows us to make more informed decisions by considering the full range of plausible outcomes and providing credible intervals that capture the uncertainty associated with our predictions.

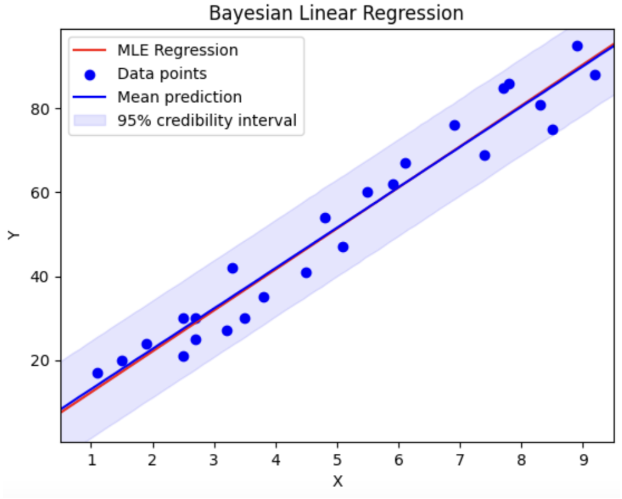

Figure 7: Non-informative-priors model

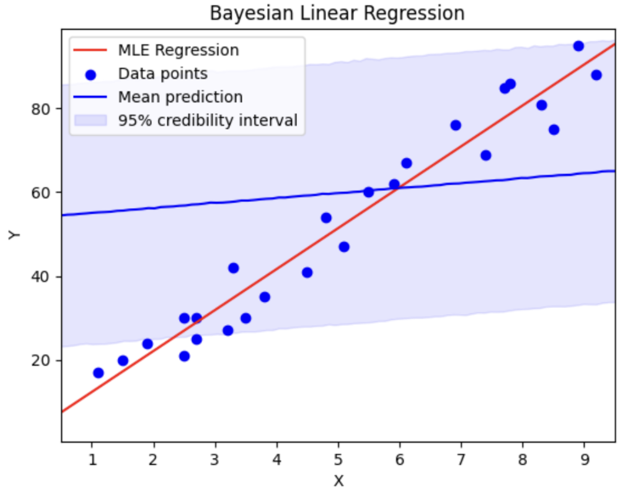

Figure 8: Badly-informative-priors model

Furthermore, we explored a Bayesian approach utilizing non-informative priors, specifically uniform priors for both $\beta_0$ and $\sigma$ spanning (0, 100), and a normal distribution for $\beta_1$ with mean 0 and standard deviation 33. As depicted in Figure 7, the results align closely with those obtained from the Maximum Likelihood Estimation (MLE) method, emphasizing the sensitivity of the Bayesian model to data when non-informative priors are used. As a result, the model's behavior predominantly reflects the data, mirroring the MLE solution.

Additionally, we experimented with poorly chosen informative priors (the same chosen in the first approach, but with less variance and less meaning-full values), as shown in

Figure 8. The misleading results (made extreme on purpose to highlight the issue) derived from these poorly selected priors underline the importance of careful prior selection in the Bayesian approach.

In sum, the Bayesian approach, while powerful and flexible, necessitates careful consideration, particularly in prior selection. While the integration of priors allows for the incorporation of prior knowledge, it introduces subjectivity, which can lead to misleading results if not judiciously managed.

Thus, it is of high importance to select appropriate and well-informed priors. A wisely chosen prior guides the model towards more plausible parameter values, improving the stability and interpretability of results. This balance between prior knowledge and observed data is a key strength of the Bayesian approach, albeit one that requires meticulous navigation.