Bayesian Linear Regression on Air Quality Data
STAT 844, Spring 2021

Aatrayee Bhattacharjee

{a56bhatt}@uwaterloo.ca
University of Waterloo
Waterloo, ON, Canada

ChinLin Chen

{c498chen}@uwaterloo.ca
University of Waterloo
Waterloo, ON, Canada

Abstract

In this study, we perform a literature survey and analysis on Bayesian linear regression. The main aim of this report is to understand the differences between the frequentist and bayesian approaches. In the literature survey, we compared different papers that focus on bayesian and frequentist methods. For the analysis part of the report, we used an air quality dataset to model ozone as a function of different explanatory variables. This dataset was fit with bayesian, multiple, LASSO and ridge regression models, and the MSE of each model was calculated. The multiple linear regression model performed the best for this dataset.

1 Contribution

The literature review portion of the project and the analysis portion of the project was performed by Aatrayee Bhattacharjee and ChinLin Chen.

2 Introduction

Air pollution is on the rise in today's world. The modernization and industrialization of our society has led to this. The number of factories keeps increasing, and they continue to emit harmful gases unfit for humans. Due to the increasing population, the number of vehicles and people relying on private cars keeps increasing. The more vehicles we use daily, the more fossil fuels and petroleum we burn, emitting dangerous air pollutants. Deforestation has led to the cutting down of trees, which release oxygen needed for human survival. After realizing the harmful effects of these activities, governments and environmental organizations worldwide have included practices to help reduce air pollution and improve air quality in towns and cities. However, the damage has already been done. The deadly effects of air pollution can be seen everywhere. Global warming is increasing at a tremendous rate, and it is leading to the melting of glaciers, increase of water at sea level and heatwaves in some parts of the world.

One of the direct effects of air pollution is acid rain, where harmful chemicals mix with the rain as they exist in excess. Acid rain is equally detrimental to humans, animals and plants.

This study will analyze different air pollutants in Seoul, South Korea and understand the relation between the other variables present in the dataset. The analysis will be performed using Bayesian Linear Regression, which will be compared to techniques such as Linear Regression, Ridge regression and LASSO regression.

The development of Statistical inference has given rise to two systems: the frequentist and Bayesian, thus giving rise to two different inferential statistical methods, the Null Hypothesis Test and the Bayesian statistical method. The Null Hypothesis Significance Test (NHT) is a falsification method that does not directly test the researcher's research hypothesis about experimental effects. Instead, the researcher needs to create a null hypothesis about the absence of an experimental effect, denoted by H_0, and use their research hypothesis as an alternative hypothesis. Then, the null hypothesis is rejected by applying statistical Bayesian inferential statistics to the sampled data in research. If the null hypothesis can be rejected, the alternative hypothesis cannot be rejected; if the null hypothesis cannot be rejected, only the alternative hypothesis can be rejected.

On the other hand, Bayesian statistical methods argue that in making statistical inference problems in addition to using the sample data obtained from sampling, it is also necessary to consider the sampling prior distributions that existed before. In other words, the researcher should first get a priori information about the subject of the study when making inference statistics. The researcher should first obtain a priori probabilities about the subject of the study and then correct the a priori distribution based on the experimental results of the sample. The posterior probabilities are obtained by correcting the prior probabilities for the experimental results obtained from the sample. In this way, the research hypothesis can be tested directly. The Bayesian statistical methods are based on Bayes' rule.

According to the "Bayes, Hume, Price, and Miracles", the Bayes' Law was first discovered by Thomas Bayes, and it was compiled and published after his death by his friend Richard Price [1]. It is generally accepted that the same conclusion as Bayes was independently reached and extended by Pierre-Simon Laplace twelve years after Bayes' death. Although Bayesian statistics had formally entered the statistical arena in the eighteenth century, it was not until the 1930s that Bayesian statistics was developed. The statistical community investigated the problem of inverse probability and prior distributions. By the 1980s, Bayesian statistics began to occupy an increasingly important place in statistics, with applications in agronomy, astronomy, biology, and other scientific fields.

This paper introduces the logic and basic computational process of Bayesian statistical methods and the literature review of Bayesian linear regression. It illustrates the implementation of Bayesian linear regression on an air quality dataset. Finally, the paper also describes the future trends of Bayesian linear regression and its far-reaching impact on the innovation of research methodology.

3 Literature Review: Bayesian Versus Frequentist

3.1 Frequentist Approach: Maximum likelihood estimation

In statistics, we can estimate unknown parameters in linear regression models using the Ordinary Least Squares (OLS) method. OLS selects the parameters of a linear function of a set of explanatory variables by means of the principle of least squares: minimizing the sum of squares of the residuals between the dependent variable (the value of the predicted variable) and the predictor variable observed in a given data set [YOI14].

Consider a dataset with x_n input vectors along with y_n targets. To fit a regression model, we would like to estimate the function f_n so that:

$$y_n = f(x_n; w) + \epsilon_n \tag{1}$$

where the ϵ is the additive noise in which each epsilons are independent and identically distributed and w is a vector of weights.

Frequentist (non-Bayesian) approach make some form of estimators for the labels y using linear regression model. To do so, we can define a Sum-of-squares error

$$E(w) = \frac{1}{2}\sum_{n=1}^{N} |y_n - \hat{y_n}|^2 \tag{2}$$

By minimizing this $E(w)$ function with respect to w can lead to a w^* estimate, which can be further used to perform predictions for the new incoming dataset. A smaller value for Sum-of-squares error resembles a good fit of the model to the data. However, this minimization of the loss function has a significant problem, which is complex models can often lead to over-fitting. Overfitting is simply and straightforwardly described as a model with very little training error and a large testing error. Over-fitting can produce overly optimistic model results: the "findings" that appears in the over-fitting model does not actually exist in the population, so it will not be replicated [Bab04]. However, the data set we work with are usually small and limited, and yet we would like to be able to use a flexible model that consists of many parameters. In the further sections, we will see how regularization can be applied to prevent over-fitting of the model.

3.2 Regularization (non-Bayesian Technique)

To overcome the potential over-fitting problem we discussed above in the ordinary least square approach is to perform regression model with a penalty term. L1 regularization uses a penalty term to encourage the sum of absolute values of parameters to become smaller. L2 regularization, encourages the sum of squares of the parameters to become smaller [Ng04].

L1 regularization is also known as Lasso regression and it has the following form

$$minimize(SSE + \lambda \sum_{j=1}^{p} |\beta_j|) \tag{3}$$

By adding the L1 penalty function to the linear regression, the aim is to keep the model from having too many parameters, and the value of the penalty function will increase as the model has more parameters. Similar to the Ridge model, the Lasso model pushes variables that are correlated to each other and avoids the situation where one model parameter has a very large positive coefficient and another has a very large negative coefficient. The major difference with the Ridge model is that Lasso reduces the regression coefficient of non-influential variables to zero, which means that feature selection can be performed. However, when we remove variables, we also sacrifice the correctness of the model. So while Lasso can produce a clearer and cleaner model and improve the correctness of the model, it also reduce the generalisability of the model.

L2 regularization is also known as Ridge regression and it has the following form

$$minimize(SSE + \lambda \sum_{j=1}^{p} \beta_j^2) \tag{4}$$

The Ridge model retains all the variables and therefore cannot filter them as the Lasso model can, because Ridge can only regress the coefficients of non-influential variables to approximately zero (but not exactly equal to zero), so the final model may still have some unimportant parameters, which may affect the final correctness of the model. In the 2.5 section, we will discuss in detailed how regularization (L1 and L2) can be expressed as a Bayesian estimator with certain priors.

3.3 Basic Logic of Bayesian Inferential Statistics

The major difference between traditional hypothesis testing and Bayesian inferential statistics lies in the understanding of the concept of uncertainty. The former assumes that the data X obtained from an experiment are uncertain and random variables, while the parameter θ of interest to the researcher is a fixed value. The latter assumes that the data we observe experimentally are deterministic, while the parameter of interest, θ, is uncertain and random due to our lack of knowledge or understanding of a phenomenon.

In the framework of Bayesian statistics, researchers use prior distribution $p(\theta)$ to generalize knowledge already available to previous authors and to himself about a parameter of interest. Under the prior model combined with the data obtained from the experiment according to the likelihood principle (the likelihood function contains all the information in the sample) the researcher obtains the likelihood function $p(x|\theta)$. Taking the coin flipping problem as an example, if the researcher wants to investigate whether the coin is homogeneous (equal probability of heads and tails facing up), and in M coin toss experiments, N tosses facing up. The likelihood function $p(x|\theta)$ in this problem is the observed values x1, x2, ... xn obtained from the experiment are substituted into the probability density function of the Bernoulli distribution. In the end, we update the prior beliefs of the parameter with experimental data to obtain a posterior distribution of the parameter in question $p(\theta|x)$. The posterior distribution $p(\theta|x)$ is proportional to the product of the likelihood function $p(x|\theta)$ and the prior distribution $p(\theta)$ [CL08]. Bayes' theorem states that:

$$P(\theta|x) = \frac{P(x|\theta)P(x)}{\sum P(x|\theta)P(x)} \tag{5}$$

Eq. (1) is appropriate for the discrete set of hypotheses and discrete varying data. To handle continuous set of hypotheses and continuous varying data, we have to use integrals:

$$P(\theta|x) = \frac{P(x|\theta)P(x)}{\int d_\theta P(x|\theta)P(x)} \tag{6}$$

In order to use the Bayes' Theorem, we assumes that the data are conditionally independent, i.e.,

$$P(\theta_i|x) = P(\theta_i|x, \theta_j) \tag{7}$$

However, if the data are not conditionally independent, then we have to expand the combination rule. For two data, the formula for expanded version is:

$$P(\theta_i|X_1, X_2) \propto P(X_2|\theta_i, X_1)P(X_1|\theta_i)P(\theta_i) \tag{8}$$

and as the number of data increases, this will become more and more complicate and requires further expansion.

The principle of determining the prior distribution $p(\theta)$ is still highly controversial in the academic community. There are three general ways to obtain the prior distribution: namely,

an uninformative prior distribution represented by a uniform distribution; a prior distribution determined by the researcher combining his or her research experience and knowledge; and a likelihood function that is formally consistent with the posterior distribution for the sake of mathematical computational convenience of keeping the prior and posterior distributions consistent. Then the prior distribution is set as a conjugate distribution.

Also in the Bayesian statistical framework, the traditional method of hypothesis testing can actually be seen as a special form of Bayesian statistics in which the parameters of interest are taken to have different values before the experiment begins depending on whether they have equal possibilities, i.e. the prior distribution is an uninformative uniform distribution. If in experiment, the researcher uses the uninformative uniform distribution as the prior distribution, then Bayesian inferential statistics are functionally similar to traditional statistics.

3.4 Bayesian Regression

In the Bayesian view, we use probability distributions rather than point estimates for linear regression. Instead of estimating y as a single value, it is assumed to be drawn from a normal distribution. The Bayesian linear regression model is:

$$y \sim \mathcal{N}(\beta^T X,\, \sigma^2 I) \tag{9}$$

Specifically, we express the prior distribution $p(\theta)$ in the written form, where α is the hyper-parameter.

$$p(\theta|\alpha) \propto exp(-\alpha\Omega(\theta)) \tag{10}$$

Bishop demonstrated an example with a Gaussian distribution for $p(\theta|\alpha)$, and it can be written in form [BT03]:

$$p(\theta|\alpha) = (\frac{\alpha}{2\pi})^{M/2} exp(-\frac{\alpha}{2}||w||^2) \tag{11}$$

We can express the posterior distribution as the product of the likelihood function and the prior distribution using the Bayes' Theorem mentioned above.

$$p(\theta|t, \alpha, \sigma^2) \propto p(\theta|\alpha)L(\alpha) \tag{12}$$

$$where L(\alpha) = p(t|\alpha, \sigma^2)$$

We would like to find the point estimate of w, which maximizes the posterior distribution. We can obtain this by taking negative logarithm to the both side of the equation. By taking the log of both sides, we see that it equals to minimizing

$$\frac{1}{2\sigma^2}\sum_{n=1}^{N}|y(x_n; w) - t_n|^2 + \frac{\alpha}{2}\omega(w) \tag{13}$$

In the 2.5 section, we will in detailed how penalized regression techniques such as LASSO and Ridge can be expressed as a Bayesian estimator with certain priors.

3.5 Connection between Bayesian and Regularization

In this section, we will be looking at the regularization in a Bayesian point of view. When the regression parameters have normal and independent Laplace (such as, double exponential) priors, the ridge and lasso estimates of the linear regression parameters can be interpreted as Bayesian posterior estimates [BL20].

In Bayesian approach, we would like to maximize the posterior distribution using the maximum a posteriori probability (MAP) estimate. By maximizing the posterior distribution, we have exactly answer the problem which is this setting of θ is the most likely ones given the data that we actually observed.

$$\begin{aligned}\hat{\theta}_{MAP} &= \underset{\theta}{argmax} P(\theta|y) \\ &= \underset{\theta}{argmax} \frac{P(y|\theta)P(\theta)}{P(y)} \\ &= \underset{\theta}{argmax} P(y|\theta)P(\theta) \\ &= \underset{\theta}{argmax} [\log P(y|\theta) + \log P(\theta)]\end{aligned} \tag{14}$$

With the assumption in Eq.9, we can compute the $P(y|\theta)$ as the multiplication of n probability density functions, each of which is a normal probability density function.

$$P(y|\theta) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(y_i - \theta^T x_i)^2}{2\sigma^2}} \tag{15}$$

To solve this, we can take the log of both side.

$$\log(P(y|\theta)) = \sum_{i=1}^{N} [\log \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2}(y_i - \theta^T x_i)^2] \tag{16}$$

Now we consider the prior ($P(\theta)$, we said that the prior distribution is our prior understanding about the world for the θs before even considering the data set. However, which prior should we pick? There have been many criticisms and nuances about the selection of the Bayesian priors.

First, let's pick the Gaussian prior as our prior and see where it leads to. We are going to assume that the prior distribution of the θ is normal with mean 0 and variance τ^2. When we discussed about the regularization (LASSO and Ridge) in 2.2, we are interested in minimizing the absolute values of θs. Similarly, with the Bayesian prior we pick here is trying to do the exact same thing. We are suggesting before even seeing the data that small absolute values of θ are more likely to appear. With the Gaussian prior, we can rewrite Eq.14 as:

$$\underset{\theta}{argmin} ||y - X\theta||_2^2 + \frac{\sigma^2}{\tau^2}||\theta||_2^2 \tag{17}$$

By replacing the $\frac{\sigma^2}{\tau^2}$ with λ, we obtain

$$\underset{\theta}{argmin} ||y - X\theta||_2^2 + \lambda||\theta||_2^2 \tag{18}$$

Notice that, this equation we obtain is exactly same as the equation (Eq.4) for L2 regularization (Ridge regression).

Similarly for L1 regularization (LASSO regression), we can obtain the same equation as Eq.3 by picking the Laplacian prior. In the LASSO regression, we had discussed above that it is capable of doing variable selection. Similarly, the Laplace distribution (double exponential distribution) has a similar property where at zero the distribution is non smooth. This enables the probability of $\theta = 0$ not equals to 0 [Tib96]. Indeed, with the prior [PC08]:

$$p(\theta) = \prod_{i=1}^{N} \frac{\lambda}{2} e^{-\lambda|\theta_i|} \tag{19}$$

With an independent prior $p(\sigma^2)$ on $\sigma^2 > 0$, we can express the posterior distribution that is conditioned on $\hat{y}$ as

$$p(\theta, \sigma^2|\tilde{y}) \propto p(\sigma^2)(\sigma^2)^{\frac{-(n-1)}{2}} exp(\frac{-1}{2\sigma^2}(\tilde{y} - X\theta)^T(\tilde{y} - X\theta) - \lambda \sum_{i=1}^{N} |\theta_i|) \tag{20}$$

Notice that, maximizing θ in this equation with any value with $\sigma^2 > 0$ will result in a LASSO estimate.

3.6 Bayesian vs Frequentist comparison

In this section, we will discuss how the following 4 studies compared the frequentist methods and Bayesian approaches. Researchers compared Bayesian and frequentist regularization methods with low information limits when variables' number is approximately equivalent to the number of observations on simulated and real data sets [CEAMR12]. By numerical results, they dominate the frequentist approach because they provide fewer prediction errors while selecting the most relevant variables in a parsimonious manner. The authors conclude that while there is no significant difference between the two methods from a forecasting perspective; from a variable selection perspective, Bayesian methods appear to be more parsimonious and relevant than the regularization methods.

Stegmueller performed a comparative research about how many countries requires multilevel modelling [Ste13]. The design of this research includes both linear and non-linear models as well as Bayesian framework consists of various levels of intraclass correlation. With the numerical results, the linear and non-linear models' estimates, and confidence intervals have high potential to be biased. With models that include cross-level interactions, there is more chance that the results are biased. In contrast, the confidence intervals and estimates generated by the Bayesian approach show a preferable and stronger properties.

In this study investigated by Yuan and Mackinnon, the researchers performed a mediation analysis study based on two different approaches: Bayesian and Frequentist [YM09]. The authors had concluded several advantages of Bayesian approach over conventional frequentist mediation analysis. First, the Bayesian approach allows the researchers to provide useful information of the world to the analysis before even seeing the data. This is a useful method to incorporate information that psychologists usually have before conducting experimental tests on the mediation process. The result of this study shows that it has increases the efficiency of the estimates. Second, the Bayesian approach can establish credible intervals for indirect effects. With this advantage, Bayesian approach is more appealing for research with rather small data sets.

Sarah had composed a study to investigate the impact of latent class separation on the Growth mixture modeling [Dep13]. The Maximum likelihood estimation(MLE) and Bayesian frameworks were compared in this study. Especially, the Bayesian framework were implemented in four different priors: accurate informative priors, weakly informative priors, partial

knowledge priors, and inaccurate priors. The result indicates that only the accurate informative priors and the partial knowledge priors showed optimal parameter recovery. The weakly informative priors, the inaccurate priors, and MLE performed poorly.

4 Dataset

The dataset used in this study is the Air Pollution in Seoul dataset from Kaggle [bap20], which contains 647,512 instances. The instances were collected hourly within three years, from 2017-2019, in 25 districts in Seoul, South Korea. The dataset is formed of data publicly made available by the Seoul Metropolitan Government. The dataset provides average values for the following pollutants: NO2, SO2, O3, CO, PM10, PM2.5. The dataset contains 11 features which are:

- **Measurement date**: Given in the format DD/MM/YYYY HH:MM. It provides information on the day and hour during which the measurement was taken.
- **Station code**: It represents the code of the area of the measurement location.
- **Address**: The exact address of the location of measurement of the air pollutants.
- **Latitude**: The latitude of the location where the measurement of the air pollutants took place.
- **Longitude**: The longitude of the location where the measurement of the air pollutants took place.
- **SO2**: The average amount of Sulphur Dioxide present in the air. It is emitted from processes like burning coal and combusting fossil fuels. It is measured in parts per million (ppm).
- **NO2**: The average amount of Nitrogen Dioxide present in the air. It is produced by natural resources like volcanic activities, bacterial reactions, and lightning during rains and thunderstorms and by some man-made activities like fossil fuel combustion. It is measured in parts per million (ppm).
- **O3**: The average amount of Ozone present in the air. It reaches the Earth's atmosphere due to ultraviolet radiations of oxygen and is usually formed due to a chemical reaction between sunlight and Nitrogen dioxide. It is measured in parts per million (ppm).
- **CO**: The average amount of Carbon Monoxide present in the air. It is mainly emitted by spark ignition combustion engine. It is measured in parts per million (ppm).
- **PM10**: Airborne particulate matter of a diameter of 10 microns or less. It is composed of dust from sources like wildfires and construction sites. It is measured in micrograms per cubic meter ($\mu g/m^3$). (https://ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health)
- **PM2.5**: Airborne particulate matter of a diameter of 2.5 microns or less. It is composed of emissions that result from combustions from fuels and oils. It is measured in micrograms per cubic meter ($\mu g/m^3$). (https://ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health)

Table 1 categorizes the values of air pollutant particles into Good, Normal, Bad and Very bad. Good indicates that the air pollutants are considerably low and the air quality is good. Normal means that the air pollutants are at an optimal level. Bad suggests that the air pollutants have a relatively high value and the air quality is unhealthy. Finally, very bad indicates that the air pollutants have an extremely high value and the air quality is very unhealthy.

Pollutant	Good	Normal	Bad	Very bad
SO2	0.02	0.05	0.15	1.0
NO2	0.03	0.06	0.20	2.0
CO	2.00	9.00	15.00	50.0
O3	0.03	0.09	0.15	0.5
PM10	30.00	80.00	150.00	600.0
PM2.5	15.00	35.00	75.00	500.0

Table 1: Air pollutants categories description

A map of the stations present in the dataset is shown in Figure 1, highlighted in blue. Some of the stations have a river nearby, so it will be interesting to find out if the presence of a waterbody influences the air quality of the stations surrounding it.

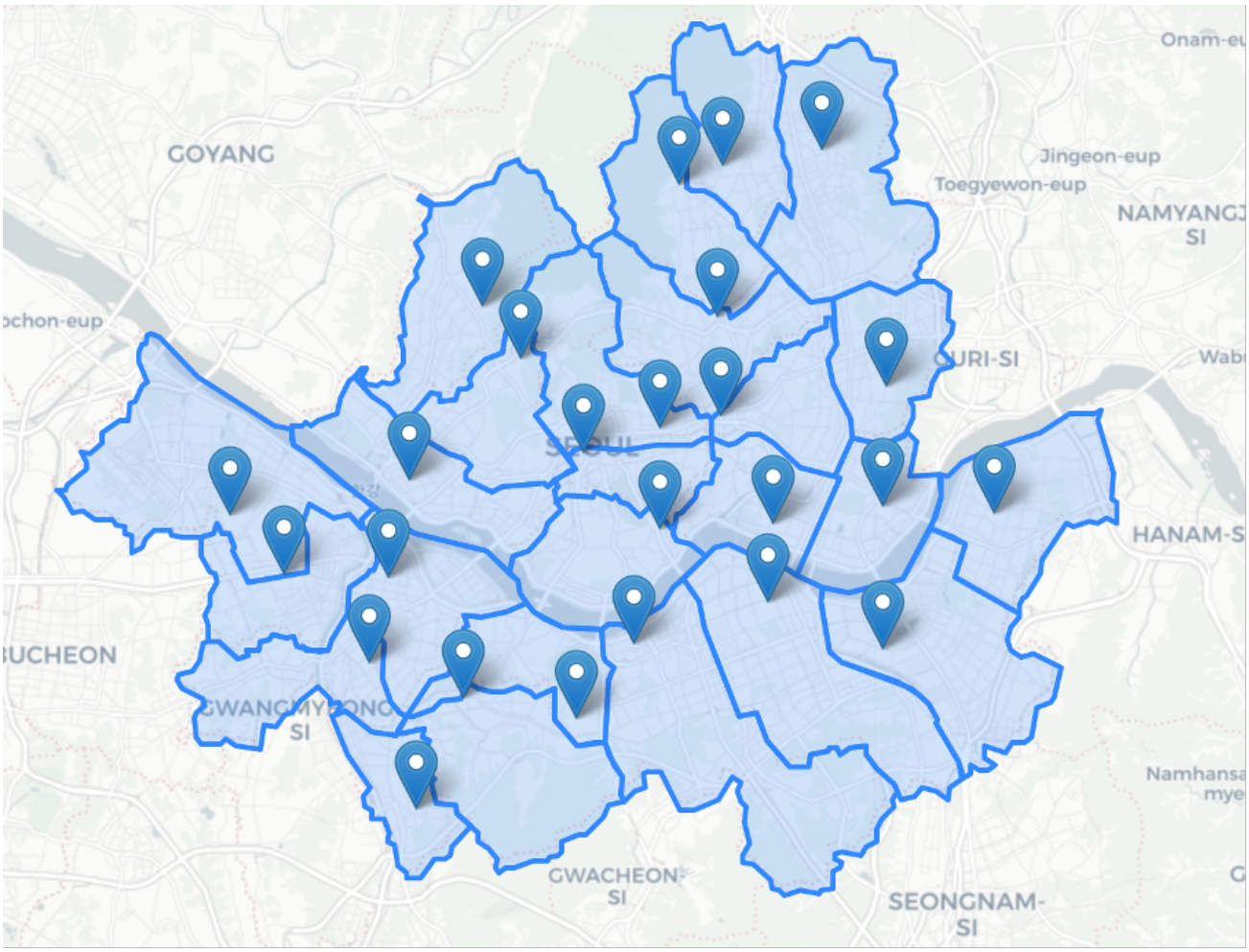

Figure 1: Map of stations from dataset

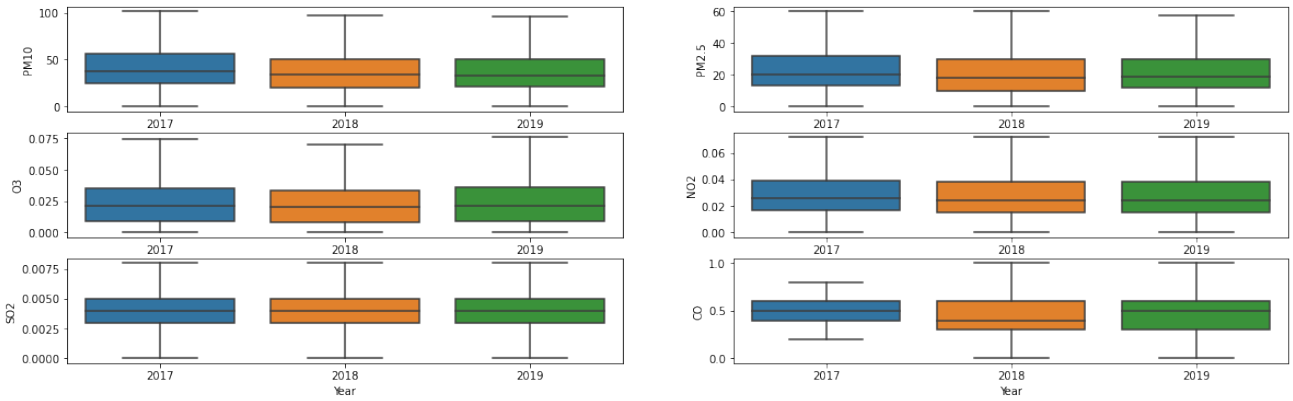

Figure 2: Boxplot of air pollutant values vs. Year

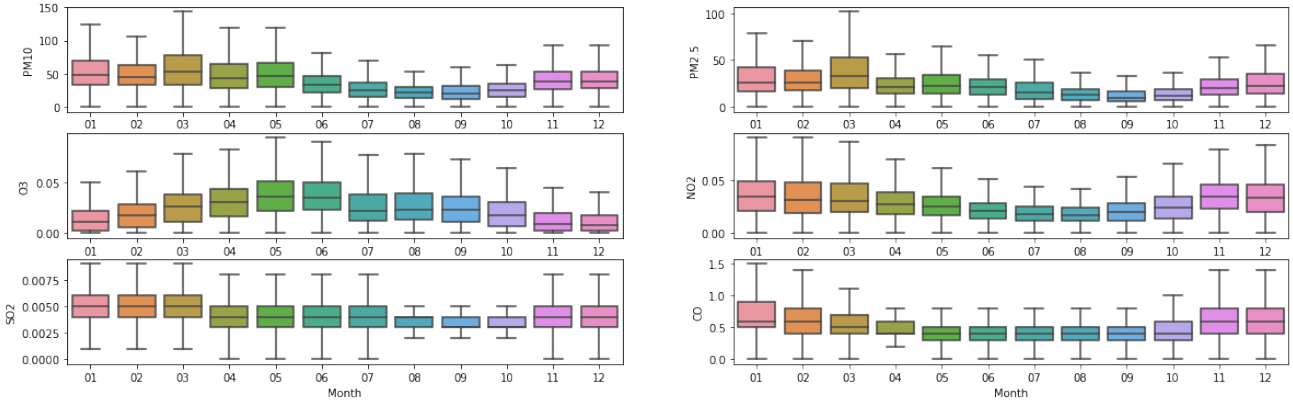

Figure 3: Boxplot of air pollutant values vs. Month

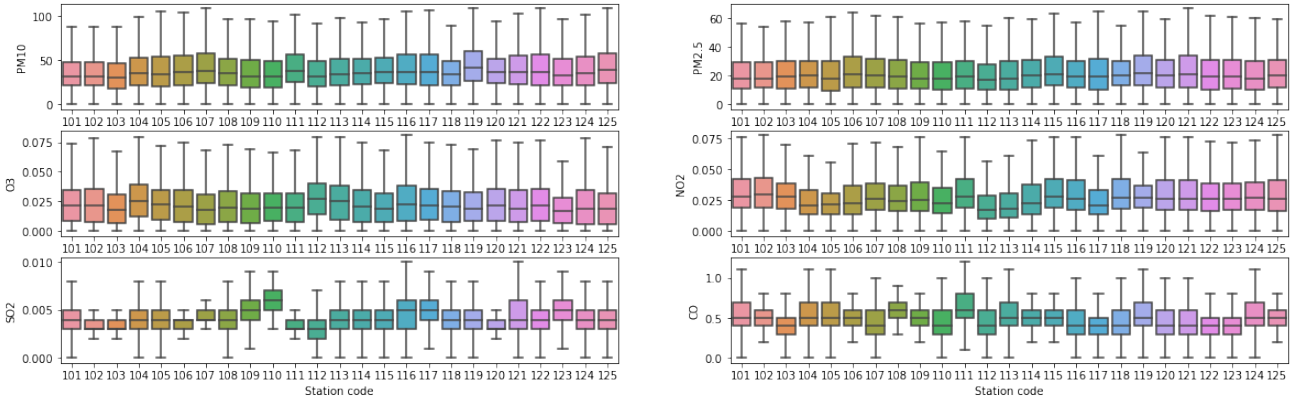

Figure 4: Boxplot of air pollutant values vs. Stations

Figure 2 is six box plots of the air pollutants versus the three years when the data was collected. No drastic change is seen over the three years for all the pollutants. This signifies that overall the average of the pollutants remained constant with minute changes throughout the period. If we want to learn more about the changes in the data pattern, we need to analyze the changes monthly instead. Figure 3 represents six box plots of air pollutants versus months. We can see that overall, the level of pollutants is lower from August-October.

Most of the air pollutants except **O3** significantly rise from January to March and November to December. This is because the cold months of the year fall during that time, and people would most likely be using their automobiles and heaters to combat the cold. This might lead to the emission of more pollutants. **O3** is on the rise from May to June, and we can see that **O3** does not seem to follow the same monthly pattern as the other six pollutants. It will be interesting to understand the relation between **O3** and the other five pollutants in this study.

Since the dataset in our study also has information about the Stations and their location, we can learn more about which pollutants are prominent in a particular region. **PM10, PM2.5, O3** and **NO2** seem like they have similar levels in all the stations. **SO2** and **CO** have the most fluctuations across the stations. **SO2** is the highest in station 110 while **CO** is the highest in station 111. Our interactive map in Jupyter notebook finds that both Station 110 and 111 are located in crowded areas, away from the river. This answers our question of whether the presence of a river influences the pollutants. From the box plots, the level of air pollutants seems to be significantly higher in areas away from the river. Perhaps the existence of a river means people follow good practices to reduce harmful emissions causing air pollution.

5 Dataset pre-processing

Pre-processing is essential in this study as any discrepancies in the data that are not handled can lead to incorrect predictions. The steps taken to pre-process the data are as follows:

- The variable Measurement date is separated into Month, Year and Time to conveniently understand the trends of the air pollutants at a particular time and season.

- Upon checking for missing and unusual data in the dataset, it is found that though there are no missing values, all the air pollutants have a minimum value of -1. -1 is not a valid number to measure how average value of air pollutants. We replace these numbers with 0. We do not consider dropping the rows that contain -1 because dropping rows could lead to the loss of essential data. The maximum values are also exceptionally high, which indicates the presence of outliers. The outliers are detected using z-score, given by the formula:

$$z - score = \frac{x - \mu}{\rho} \tag{21}$$

  where x is the value of the data point, μ is the population mean, and ρ is the population standard deviation.

  If the z-score is greater than three, it signifies that the data point is a potential outlier.

  Here, we remove the rows containing outliers because only 0.7% of the data has outliers, and we will not be losing a lot of information due to this. We can also replace the outliers with the mean or median, but due to the nature of the dataset, replacing the outliers with the mean or median causes some discrepancies in the overall data summary, so we choose to drop the values containing outliers.

- Lastly, as we are performing linear regression techniques, removing any the collinearity is important to improve the accuracy of our prediction. We find problematic variables using Variance Inflation Factor (VIF). The VIF is for explanatory variable j calculated by:

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}, j = 1, ..., p \tag{22}$$

where $R^2_{X_j|X_{-j}}$ is the R^2 form of a regression of X_j onto every other predictor. When the VIF is 5 or greater, it shows that the predictor is problematic due to collinearity. Typically we prefer that there is no collinearity because having collinearity issues with our predictors means that the model's prediction is questionable. In the air quality dataset, we consider **O3** as the response variable and the other pollutants: **SO2, NO2, CO, PM10, PM2.5** are the explanatory variables. Table 3 shows the VIF for the different explanatory variables.

Explanatory variable	VIF
SO2	3.095747
NO2	2.448343
CO	1.937606
PM10	1.411393
PM2.5	1.411941

Table 2: VIF of explanatory variables

The VIF for the explanatory variables is between $1.41 - 3.09$ and does not exceed the value of 5. This shows that the exploratory variables do not have the issue of multicollinearity between them.

6 Results

6.1 Experimental Design

As mentioned in the previous section, the dataset needs to undergo pre-processing. This is implemented using Python built-in libraries, Numpy and Pandas. These libraries provide the necessary tools to convert data into the desired format. In addition to these libraries, we also use matplotlib and seaborn to plot the relationship between the variables and folium to produce an interactive map of the regions in the dataset to understand the data better.

For this analysis, we refer to [BL17], which introduces Bayesian linear regression and its use in clinical data. We have used the concepts from the report to implement it in the air quality dataset and have compared the Bayesian linear regression model to other regression techniques, namely:

- Multiple linear regression
- LASSO regression
- Ridge regression

We select these different regression models to understand how the dataset behaves with other models and also to understand and compare the differences between the Frequentist and Bayesian approaches. Bayesian linear regression falls under the Bayesian approach, while multiple linear regression, LASSO, and ridge regression fall under the frequentist approach category.

Here, we select Ozone (O3) as the response variable $\mathbf{y}$ and the remaining five variables, i.e., SO2, NO2, CO, PM10 and PM2.5, as the explanatory variables $\mathbf{X}$ to model O3 as a function of the explanatory variables. The dataset will be split into 70% training data and 30% testing data randomly. The dataset is split up randomly so that we can have an unbiased evaluation of the performance of the models.

We use R and some build-in libraries to help aid the model fitting and prediction process to build the models. Bayesian linear regression requires libraries like rstanarm, bayestestR and bayesplot to fit and plot the data relationship. In the case of multiple linear regression, the R stats package has built-in functions readily available. LASSO and ridge regression require glmnet to fit and predict data and learn more on the shrinkage estimators of the data.

Evaluating the performance of the models is one of the main aims of this study. We will use Mean Squared Error (MSE) to analyze the performance of the models. The formula of MSE is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{23}$$

where n is the number of data points, Y_i is the actual value of the response variable and $\hat{Y}_i$ is the predicted value of the response variable. MSE is a widely used technique to measure performance, which is why we use it in our study.

6.2 Bayesian Linear Regression Model

The model used in this study is a multivariate regression where we define response variable O3 as a function of five explanatory variables. Therefore, the model can be represented by the following formula:

$$O3_i = \beta_0 + \beta_1 SO2_i + \beta_2 NO2_i + \beta_3 CO_i + \beta_4 PM10_i + \beta_4 PM2.5_i + \epsilon_i \tag{24}$$

where the residual ϵ_i is normally distributed with mean $\mu = 0$ and an unknown standard deviation σ

$$\epsilon_i \sim N(0, \sigma) \tag{25}$$

Bayesian linear regression models require likelihood, a probability distribution that defines the entire dataset and prior of all parameters in the dataset to be defined. Based on equation 24, we can say that our dataset is normally distributed. Therefore, the likelihood of our model can be defined as:

$$O3_i \sim N(\mu_i, \sigma) \tag{26}$$

where μ_i is the mean of the i^{th} value in the dataset, σ is the standard deviation of the dataset and is equal to the standard deviation in equation 25.

The mean in equation 25 can be rewritten as:

$$\mu_i = \beta_0 + \beta_1 SO2_i + \beta_2 NO2_i + \beta_3 CO_i + \beta_4 PM10_i + \beta_4 PM2.5_i \tag{27}$$

All the parameters require a prior distribution which is commonly chosen as a normal distribution. Multiple linear regression generally has a uniform prior, and all the values have the same probability.

6.3 Model Results

To perform Bayesian linear regression, the data was sampled into four chains, each having 2000 samples which is the default value for this model and are suitable with most datasets. When the Gelman-Rubin statistic $\hat{R} \approx 1$ means that the MCMC chain has converged. This convergence means that the samples were formed from the posterior. If there is no convergence, it signifies that the model has an error due to it being too complex or specifying the wrong type of prior, The median and median absolute deviation (MAD) from the MCMC simulation are mentioned in Table 3.

Explanatory variable	Median	MAD
SO2	5.945	0.018
NO2	-0.033	0.002
CO	-0.037	0.000
PM10	0.000	0.000
PM2.5	0.000	0.000

Table 3: Median and MAD of explanatory variables

We also check if the coefficients of the Bayesian regression model are significant. Table 4 contains the highest density interval (HDI), credible interval (CI) and equal-tailed interval (ETI). If these include have zero, the coefficient is considered non-significant.

LASSO Regression

Explanatory variable	95% HDI	95% CI	95% ETI
SO2	[5.91, 5.98]	[5.91, 5.98]	[5.91, 5.98]
NO2	[-0.04, -0.03]	[-0.04, -0.03]	[-0.04, -0.03]
CO	[-0.04, -0.04]	[-0.04, -0.04]	[-0.04, -0.04]
PM10	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]
PM2.5	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]

Table 4: HDI, CI and ETI of explanatory variables

Based on these interval values, exploratory variables PM10 and PM2.5 are not significant, so we will drop them from our final model. The modified Bayesian regression model with significant exploratory variables is:

$$O3_i = \beta_0 + \beta_1 SO2_i + \beta_2 NO2_i + \beta_3 CO_i + \epsilon_i \tag{28}$$

Multiple linear regression: When we check the significance of the multiple linear regression coefficients, all the five exploratory variables are significant, so we do not drop any variables from the multiple linear regression model. Therefore, the model can be written as:

$$O3_i = \beta_0 + \beta_1 SO2_i + \beta_2 NO2_i + \beta_3 CO_i + \beta_4 PM10_i + \beta_4 PM2.5_i + \epsilon_i \tag{29}$$

The model is similar to the bayesian linear regression model in Equation 28.

LASSO Regression: We perform 10-fold cross-validation on the training dataset to find the optimal lambda λ value that minimizes the MSE. The optimal λ value is found to be 0.03296221

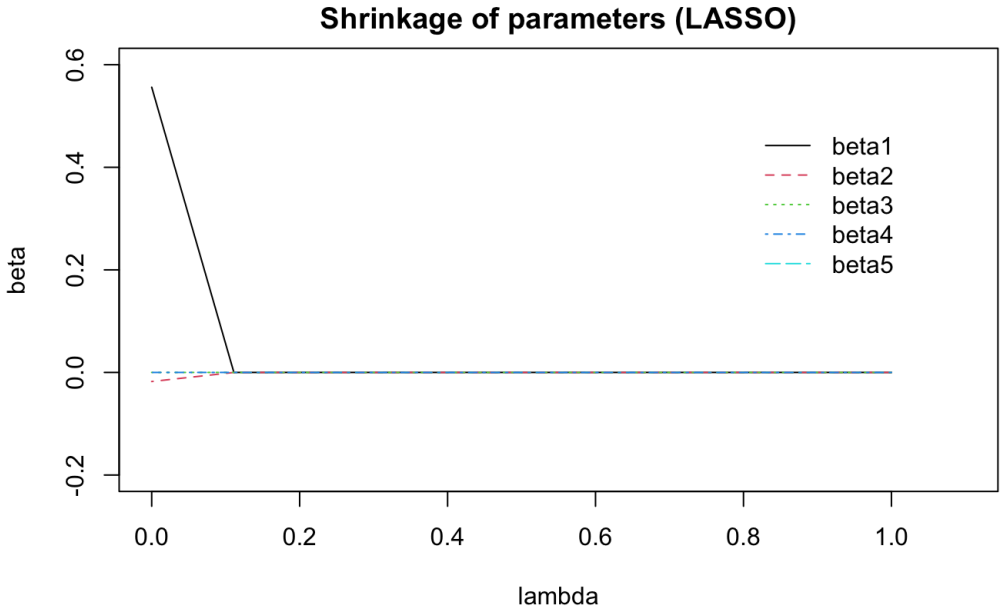

Figure 5: Shrinkage of parameters of LASSO model

Figure 5 shows the shrinkage of parameters of the LASSO regression model. We see that when the λ value is greater than 0.1, all the coefficients tend to 0. At $\lambda = 0.03296221$, none of the β coefficients are zero so we do not drop any parameters.

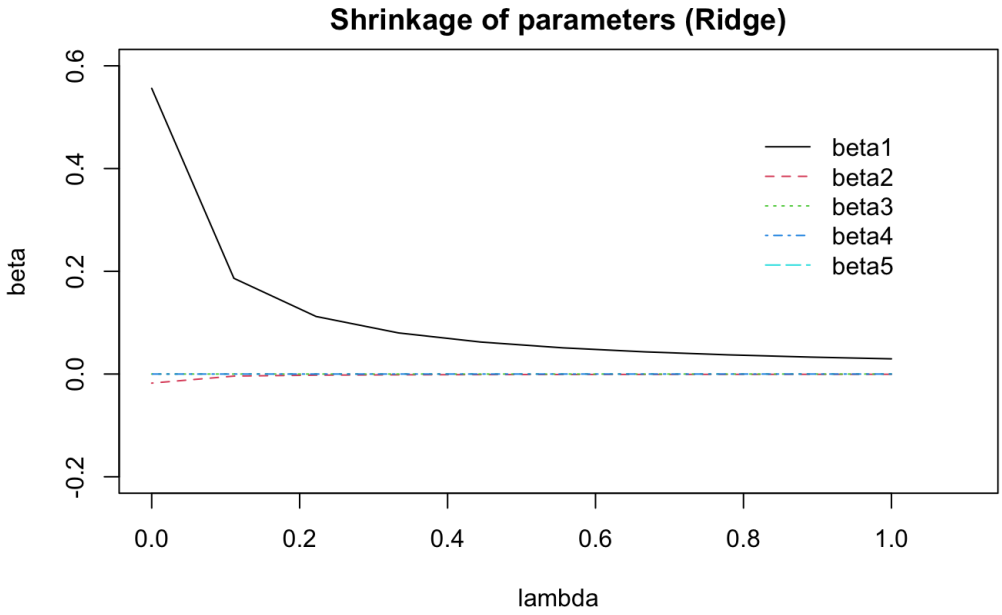

Figure 6: Shrinkage of parameters of Ridge model

Ridge Regression: We perform 10-fold cross-validation on the training dataset to find the optimal lambda λ value that minimizes the MSE. The optimal λ value is found to be 0.1362002

Figure 6 shows the shrinkage of parameters of the ridge regression model. As expected

from a ridge regression model, the β coefficients never tend to zero.

Model performance: We use the predict function in R on the test dataset to calculate the MSE for each model, the results of which are shown in Table 5.

Model	MSE
Bayesian Linear Regression	0.003575877
Multiple Linear Regression	0.0007371189
LASSO Regression	0.0007565461
Ridge Regression	0.0007565461

Table 5: Mean Squared Error of the different models

From Table 5 we see that the Multiple Linear Regression performs best with our dataset.

7 Conclusion

This study explores Bayesian linear regression techniques and compares its results with other frequentist models like multiple linear regression, LASSO regression, and Ridge Regression. Although the Bayesian linear regression has the worst MSE among the other models, the MSE is still not that low, which means that the Bayesian linear regression does perform well and will probably perform better on datasets that are more fit for Bayesian linear regression techniques. Furthermore, we learn the different vital components of a bayesian linear regression, which coefficients are considered significant and what likelihood distribution is considered in our dataset.

We also perform a literature survey on Bayesian regression and find that all of the four studies concluded that the Bayesian approach performed better in general than the frequentist approach. It is not surprising as, in the Bayesian approach, we can provide thoughtful priors. And the benefit of having useful priors enables the posterior to provide a more precise and less fluctuating result than frequentist approaches. However, when providing an uninformative prior or inaccurate prior, the Bayesian framework will perform poorly.

8 Future Work

For future work, we would like to implement the following:

- Since the dataset did not perform the best with the Bayesian linear regression model. We would like to explore more datasets that will better fit the Bayesian linear regression and provide a better prediction.
- Since R hosts many different libraries of the Bayesian linear regression model. We would like to explore the features and working of the libraries and compare the functioning of all libraries to gain a deeper understanding of which library would work best for a particular function
- We only compared regression models that have multiple variables. In the future, we would like to explore bi-variate regression models to understand if the Bayesian linear regression model performs better for bi-variate models compared to multi-variate regression models.

References

[Bab04] Michael A Babyak. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic medicine*, 66(3):411–421, 2004.

[bap20] bappe. Air pollution in seoul, 2020. [Online; accessed August 8, 2021].

[BL17] Scott A Baldwin and Michael J Larson. An introduction to using bayesian linear regression with clinical data. *Behaviour research and therapy*, 98:58–75, 2017.

[BL20] Adel Bedoui and Nicole A Lazar. Bayesian empirical likelihood for ridge and lasso regressions. *Computational Statistics & Data Analysis*, 145:106917, 2020.

[BT03] Christopher M Bishop and Michael E Tipping. Bayesian regression and classification. *Nato Science Series sub Series III Computer And Systems Sciences*, 190:267–288, 2003.

[CEAMR12] Gilles Celeux, Mohammed El Anbari, Jean-Michel Marin, and Christian P Robert. Regularization in regression: comparing bayesian and frequentist methods in a poorly informative situation. *Bayesian Analysis*, 7(2):477–502, 2012.

[CL08] Bradley P Carlin and Thomas A Louis. *Bayesian methods for data analysis*. CRC Press, 2008.

[Dep13] Sarah Depaoli. Mixture class recovery in gmm under varying degrees of class separation: frequentist versus bayesian estimation. *Psychological methods*, 18(2):186, 2013.

[Ng04] Andrew Y Ng. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78, 2004.

[PC08] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[Ste13] Daniel Stegmueller. How many countries for multilevel modeling? a comparison of frequentist and bayesian approaches. *American Journal of Political Science*, 57(3):748–761, 2013.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[YM09] Ying Yuan and David P MacKinnon. Bayesian mediation analysis. *Psychological methods*, 14(4):301, 2009.

[YOI14] WB Yahya, OR Olaniran, and SO Ige. On bayesian conjugate normal linear regression and ordinary least square regression methods: A monte carlo study. *Ilorin Journal of Science*, 1(1):216–227, 2014.