Regression: The Classical and Bayesian Paradigm

Name- Srijita Dey

B.Sc. Department of Mathematics

Roll No. – 32

Supervisor's Name- Dr. Sucharita Roy

Registration No.- A01-2112-0501-22

Declaration

I hereby declare that the work presented in this dissertation titled “Regression: The Classical and Bayesian Paradigm”, is the result of my independent research conducted at the Department of Mathematics, St. Xavier’s College (Autonomous), Kolkata, in partial fulfillment of the requirements for the degree of Bachelor of Science (Honours) in Mathematics.

This dissertation was carried out under the guidance of Dr. Sucharita Roy, and to the best of knowledge, it has not been submitted elsewhere for any other degree or academic qualification.

I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.

Srijita Dey

Abstract

Regression analysis is a fundamental statistical tool which is used in predictive modelling. While regression analysis has distinct approaches each offer different advantages that depend on data characteristics and assumptions. This dissertation tries to draw a comparative analysis between Linear Regression (LR) and Bayesian Regression (BR) techniques to evaluate their performances based on different data scenarios. Linear regression aims at minimizing the sum of squared errors and assumes fixed parameters. Whereas Bayesian regression incorporates the concept of prior distributions to update beliefs about model parameters.

The study explores key theoretical foundations, implementation techniques of the different models and how they perform using different datasets. Key factors such as model interpretability, accuracy, confidence and credible intervals along are examined. Through extensive experimentation, results indicate that Linear regression is easy to perform but it lacks the concept of including prior knowledge and ability to interpret using past data trends where Bayesian regression comes into play. Bayesian regression is known for its ability to handle uncertainty and including the concept of prior knowledge.

The study helps us to develop a deeper understanding of regression methodologies, and the datasets help us to draw sufficient inference. This dissertation underscores the importance of Bayesian principles in enhancing traditional regression analysis and suggests for including probabilistic modelling in data-driven decision-making.

Acknowledgement

I take this opportunity to express my heartfelt gratitude to all those who have supported and guided me throughout the completion of this dissertation. First and foremost, I would like to extend my deepest gratitude to my professor and mentor, Dr. Sucharita Roy, for her constant encouragement, valuable feedback, constructive criticism, and unwavering support. Her guidance has been pivotal in shaping the direction and quality of this project work.

I am sincerely thankful to the college authorities for providing me with clear guidelines and a well-structured framework that greatly facilitated the smooth execution of this project. The discussions and continuous support from the institution have been valuable. I would also like to acknowledge the contribution of all those who, directly or indirectly, assisted me in completing this research.

Special thanks to the Mathematics faculty and curriculum, which provided the foundational knowledge and the necessary tools to conduct this study effectively. I am deeply grateful to my parents, classmates, and peers for their encouragement, valuable suggestions, and constructive feedback, which motivated me to improve the quality of my work consistently

Index

Introduction

Predictive modelling is a statistical technique used to forecast outcomes by analysing historical data and identifying patterns or past trends. Predictive modelling plays a significant role in various scientific and engineering disciplines, enabling data-driven decision-making and inference. Among the most widely used statistical methods for predictive analysis are Linear Regression (LR) and Bayesian methods, both of which offer distinct advantages and challenges in different application domains (Hastie, Tibshirani, & Friedman, 2009). While linear regression provides a straightforward parametric approach by establishing relationships between predictor variables and outcomes, Bayesian methods leverage probabilistic reasoning to incorporate prior knowledge and manage uncertainty (Gelman et al., 2013). The debate over which method is more effective remains central to statistical modelling, particularly in cases involving complex or noisy datasets (Murphy, 2012).

Linear regression is a classical technique based on the fundamental assumption that the relationship between independent and dependent variables can be represented by a linear function. For its simplicity and interpretability, it is a cornerstone of statistical analysis, particularly in fields where assumptions of normality and homoscedasticity hold (Seber & Lee, 2012). However, in real-world applications, data often exhibit non-linearity, multicollinearity and outliers, which can significantly be affected by the reliability of regression-based predictions (James et al., 2013).

Bayesian inference provides a probabilistic approach to modelling by updating of prior beliefs with new data using Bayes' theorem (Jaynes, 2003). Regression models produce point estimates whereas Bayesian models generate posterior probability distributions which gives room to more flexible and robust predictions in uncertain environments (Bernardo & Smith, 2009).

This dissertation conducts a comparative analysis of Bayesian methods and Linear Regression, focusing on their accuracy, robustness, and interpretability in predictive modelling. By evaluating both approaches using real-world datasets, the study aims to draw some inferences as to which method might be better.

I. Simple Linear Regression

Simple linear regression is a straightforward approach for predicting a quantitative response Y on the basis of a single predictor variable X. It assumes that there is approximately a linear relationship between X and Y. Mathematically, we can write this linear relationship as

$$Y=\alpha+\beta X \qquad (1.1)$$

We will sometimes describe the above equation by saying that we are regressing Y on X.

α and β are two unknown constants that represent the intercept and slope of the model. Together α and β are known as model coefficients or parameters. Once we have our training dataset, we produce estimates $\hat{\alpha}$ and $\hat{\beta}$ for the model.

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

In the above equation, $\hat{y}$ indicates a prediction of Y on the basis of X=x. Here we use the hat symbol,ˆ, to denote the estimated value for an unknown parameter or to denote the predicted value of the response.

1.1 Estimation of the Coefficients

In practice, α and β are unknown. So before we can use (1.1), to make predictions we must use the data to make estimate for the coefficients. Let (x_1, y_1), (x_2, y_2), …., (x_n, y_n) be the set of n sample observations, each of which consist of a measurement of X and a measurement of Y. Our goal is to obtain estimates of $\hat{\alpha}$ and $\hat{\beta}$ such that the linear model fits the available data well, i.e. $y_i=\hat{\alpha}+\hat{\beta}x_i$,i=1,2,…n. We want to find an intercept $\hat{\alpha}$ and a slope $\hat{\beta}$ such that the resulting line is as close as possible to the all the data points. The most common approach for measuring the closeness involves minimizing the least squares criterion.

Let $y_i=\hat{\alpha}+\hat{\beta}x_i$ be the prediction for Y based on the i^{th} value of X. Then $e_i=y_i-\hat{y}_i$ represents the ith residual which is the difference between the ith observed response value and the ith predicted response value. We define the residual sum of squares(RSS) as

RSS=$e_1^2+e_2^2+$ ……. $+e_n^2$

or equivalently as

RSS= $(y_1-\hat{\alpha}-\hat{\beta}x_1)^2+(y_2-\hat{\alpha}-\hat{\beta}x_2)^2+$ …. $+(y_n-\hat{\alpha}-\hat{\beta}x_n)^2$

The least squares approach chooses $\hat{\alpha}$ and $\hat{\beta}$ such that RSS is minimized. Using calculus, we can show that the minimizers are

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad (1.2)$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are the sample means. Therefore, the above $\hat{\alpha}$ and $\hat{\beta}$ defines the least squares coefficient estimates for simple linear regression.

1.2 Assessing the Accuracy of the Coefficient Estimates

The true relationship between X and Y takes the form Y=f(X)+ ϵ where ϵ is the random error term. If f is to be approximated by a linear function, then we write

$$Y=\alpha+\beta X+\epsilon \qquad (1.3)$$

The true relationship between X and Y is probably not linear, there may be other variables that cause variation in Y and there may be measurement error. We assume that the error term is

independent of X. The model given by the above equation defines the population regression line which is the best linear approximation for the true relationship between X and Y.

The difference between the population regression line and the least squares regression line may seem very subtle but it allows us to infer a lot about the relationship between the predictor and the response. The concept of these two lines is a natural extension of the standard statistical approach of using information from a sample to estimate characteristics of a large population. We explain this concept using an example. Suppose we are interested in estimating the population mean μ of some random variable Y. μ is unknown but we do know n sample observations of Y namely $y_1, y_2, \ldots, y_n$. The sample mean is $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\hat{\mu}$ $=\bar{y}$. In general sample mean and population mean are different but the sample mean provides a good estimate for the population mean. Similarly, $\hat{\alpha}$ and $\hat{\beta}$ in linear regression define the population regression line. We seek to estimate these unknown coefficients $\hat{\alpha}$ and $\hat{\beta}$ given in (1.2). These coefficient estimates define the least squares line.

The analogy between linear regression and estimation of mean of a random variable is an apt one based on the concept of bias. It might happen that on the basis of one particular set of observations $y_1, y_2, \ldots, y_n$, $\hat{\mu}$ overestimates μ and on the basis of another set of observations $\hat{\mu}$ underestimates μ. But if we take the average of a huge number of estimates of μ obtained from a huge number of sets of observations, then this average would be almost equal to μ. Hence an unbiased estimator does not necessarily underestimate or overestimate the true parameter.

This property of unbiasedness hold in case of least square estimates given in (1.2) as well. If we take the average of the estimates over a huge number of set of observations then the average of these estimates would be equal to the α and β.

We again bring into picture the analogy with the estimation of population mean μ of a random variable Y. We know that the average of $\hat{\mu}$'s over many datasets is close to the value of μ, but a single estimate of $\hat{\mu}$ many overestimate or underestimate μ. This brings us to the concept of standard error. The standard error of $\hat{\mu}$ tells us how far off is the value of $\hat{\mu}$ from the value of μ. Standard error of $\hat{\mu}$ is given by SE(μ^).

$$\text{Var}(\hat{\mu})=\text{SE}(\hat{\mu})=\frac{\sigma^2}{n}$$

σ is the standard deviation of each y_i from Y. This deviation shrinks with the value of n- the more observations we have, the smaller is the standard error of $\hat{\mu}$.

To compute the standard errors of $\hat{\alpha}$ and $\hat{\beta}$ we use the following formulae:

$$\text{SE}(\hat{\alpha})^2=\sigma\text{^}2\ \left(\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) \quad \text{and SE}\left(\hat{\beta}\right)^2=\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Where σ^2= Var (ϵ). For these formulas to be strictly valid, we assume that the errors ϵ_i for each observation are uncorrelated with common variance σ^2. SE($\hat{\beta}$) is smaller when the x_i are more spread out and we have more leverage in estimating the slope when such a situation occurs. SE($\hat{\alpha}$) will be the same as SE($\hat{\mu}$) if $\bar{x}$= 0 (in such a case $\hat{\beta}$ would be equal to $\bar{y}$).

In general, σ^2 is not known, it but can be estimated from the data. The estimate of σ is called residual standard error. It is given by the formula

$$\text{RSE} = \frac{RSS}{n-2}$$

Standard errors can be used to compute **confidence intervals**. A **95 % confidence interval** is defined as a range of values such that with 95 % probability, the range will contain the true unknown value of the parameter. The range is defined in terms of lower and upper limits computed from the sample of data. For linear regression, the 95 % confidence interval for β approximately takes the form

$$\hat{\beta} \pm 2 \cdot \text{SE}(\hat{\beta}).$$

That is, there is approximately a 95 % chance that the interval

$$[\hat{\beta} - 2 \cdot \text{SE}(\hat{\beta}) , \hat{\beta} + 2 \cdot \text{SE}(\hat{\beta})]$$

will contain the true value of β. Similarly, a confidence interval for α approximately takes the form

$$\hat{\alpha} \pm 2 \cdot \text{SE}(\hat{\alpha})$$

Standard errors can also be used to perform **hypothesis tests** on the coefficients. The most common hypothesis test involves testing the **null hypothesis** of

H_0 : There is no relationship between X and Y

versus the **alternative hypothesis**

H_a : There is some relationship between X and Y.

Mathematically, this corresponds to testing

$H_0 : \beta = 0$

versus

$H_a : \beta \neq 0$

since if $\beta = 0$ then the model reduces to $Y = \alpha + c$ and X is not associated with Y. To test the null hypothesis, we need to determine whether $\hat{\beta}$,our estimate for β, is sufficiently far from zero then we can be confident that β is non-zero. The term **'sufficiently far'** depends on the accuracy of $\hat{\beta}$—SE($\hat{\beta}$). If SE($\hat{\beta}$) is small, then even relatively small values of $\hat{\beta}$ may provide strong evidence that $\beta \neq 0$, and hence that there is a relationship between X and Y. In contrast, if SE($\hat{\beta}$) is large, then $\hat{\beta}$ must be large in absolute value in order for us to reject the null hypothesis.

Our test statistic T is given by

$$t = \frac{\hat{\beta} - 0}{SE(\hat{\beta})}$$

T measures the number of standard deviations such that $\hat{\beta}$ is away from 0. If there really is no relationship between X and Y, t has a t distribution with (n-2) degrees of freedom. The t-distribution has a bell-shaped graph and for values of n greater than approximately 30 it is quite similar to the normal distribution.

1.3 p-Value

The probability of observing any number equal $|t|$ or larger in absolute value, assuming $\beta_1 = 0$ is called *p-value*. We interpret the p-value as follows: a small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance,

in the absence of any real association between the predictor and the response. Hence, if we see a small p-value, then we can infer that there is an association between the predictor and the response. In such a case we *reject the null hypothesis*—that is, we declare a relationship to exist between *X* and *Y*—if the p-value is small enough. Typical p-value cutoffs for rejecting the null hypothesis are 5 or 1 %.

1.4 Examining the Accuracy of the Linear Model

When we have rejected the null hypothesis in favor of the alternative hypothesis, it is necessary to quantify the extent to which the model fits the data. The quality of a linear regression fit is typically assessed using two related quantities: the ***residual standard error* (RSE)** and the R^2 **statistic**.

Residual Standard Error (RSE):

From the model (1.3), we have seen that associated with each observation is an error term ϵ. Due to the presence of these error terms, even if we knew the true regression line (i.e. even if α and β were known), we would not be able to perfectly predict Y from X. The RSE is an estimate of the standard deviation of ϵ.In layman terms, it is the average amount that the response will deviate from the true regression line. It is computed using the formula

$$\text{RSE}=\sqrt{\frac{1}{n-2}RSS} =\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

$$\text{RSS}=\sum_{i=1}^{n}(y_i-\bar{y})^2$$

The RSE is considered a measure of the lack of fit of the model (1.3) to the data. If the predictions obtained using the model are close to the true outcome values—that is, if $y_i \approx \hat{y}_i$ for i = 1, 2,….., n—then RSE will be small, and we can conclude that the model fits the data very well. On the other hand, if $\hat{y}_i$ is very far from y_i for one or more observations, then the RSE may be quite large, which indicates that the model doesn't fit the data well.

R^2 statistic:

The RSE provides an absolute measure of lack of fit of the model (1.3) to the data. But as it is measured in the units of *Y*, it is not always clear what constitutes a good RSE. The R^2 statistic provides an alternative measure of fit. It takes the form of a proportion—the proportion of variance explained—and so it always takes on a value between 0 and 1, and is independent of the scale of *Y*.

$$R^2=1-\frac{RSS}{TSS}$$

where $\text{TSS}=\sum_{i=1}^{n}(y_i-\bar{y})^2$ is the total sum of squares.

TSS measures the total variance in the response Y and can be thought of as the amount of variability inherent in the response before the regression is performed. RSS measures the amount of variability that is left unexplained after performing the regression. Hence, TSS- RSS is the measure of the amount of variability in the response that is explained by performing the regression, and R^2 measures the proportion of variability in Y that can be explained

using X. An R^2 statistic that is close to 1 tells us that a large proportion of the variability in the response has been explained by the regression whereas a number near 0 indicates that the regression cannot explain much of the variability in the response. The R^2 statistic is a measure of the linear relationship between X and Y.

$$\text{Cor (X, Y)} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

gives a measure of the linear relationship between X and Y. This gives us an idea that we might be able to use r = Cor(X, Y) instead of R^2 in order to assess the fit of the linear model. It can be shown that in the simple linear regression setting, $R^2 = r^2$ which means the squared correlation and the R^2 statistic are identical.

1.5 Investigating the Relationship between Blood Sugar Levels and BMI using Linear Regression

Background

Diabetes is a chronic disease which is characterized by high blood sugar levels. One of the significant factors influencing blood sugar levels is BMI, which indicates obesity levels. This example aims to explore whether BMI can serve as a predictor for blood sugar levels.

Objective

Our main aim is to develop a predictive model between BMI and blood sugar levels using simple linear regression which in turn can be used to determine the statistical relationship between the two. We also assess the model performance using RSE and R^2 statistic.

Dataset

Given below is the dataset showing the BMI and their corresponding blood sugar levels corresponding to a randomly selected adult.

BMI	Blood Sugar Level
26.23988	129.2637227
38.91571	176.9695854
34.10387	155.6857173
31.17049	149.9438126
21.43241	120.1970954
21.43188	139.8274879
19.27784	118.8699727
37.05588	160.2947631
31.22453	159.009044
33.5776	149.2505216
18.45286	118.3389136
39.33802	159.4690859

36.31374	156.0529783
22.67146	130.0549791
22.00015	132.5081506
22.0349	128.0686641
24.69333	133.2161359
29.54464	145.3161666
27.50279	130.1796372
24.40704	127.5809614
31.46076	149.4050281
21.06886	132.4497996
24.42718	136.1450567
26.05996	123.8635683
28.03354	146.0865828
35.27387	160.686181
22.39282	122.284529
29.31316	151.9702517
31.03312	160.1407336
19.02191	125.7115864

Regression Model

The simple linear regression model is:

$Y=\alpha+\beta X+\epsilon$

where:

- Y denotes the blood sugar level
- X is BMI
- α is the intercept
- β is the slope (effect of BMI on blood sugar levels)
- ϵ is the error term

Data Analysis

Estimation of Coefficients

Using the least squares method, the estimated coefficients are

$\hat{\alpha}=65, \hat{\beta}=2.29$

Therefore, the regression equation is $\hat{Y}=77.55+2.29X$

Testing of Hypothesis

Null hypothesis is H_0: $\beta=0$ which implies there is no relationship between BMI and blood sugar level. The alternative hypothesis is H_a: $\beta \neq 0$ which implies there exists a relationship between BMI and blood sugar level.

t-statistic is calculated by t= $\frac{\hat{\beta}-0}{SE(\hat{\beta})}$

If the p-value is less than 0.05, we reject H_0 and conclude that BMI affects blood sugar levels in an average adult.

Model Performance:

To get an idea about the model performance we measure the RSE and R^2 statistic.

RSE= $\sqrt{\frac{1}{n-2}RSS}$ = $\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\bar{y})^2}$

R^2=1- $\frac{RSS}{TSS}$

Results:

In case of the above dataset the p-value is 0.003. Hence, we reject H_0 and confirm that there is a relationship between BMI and blood sugar levels.

R^2=0.82 which indicates there is a positive relationship BMI and blood sugar levels.
RSE=10.5

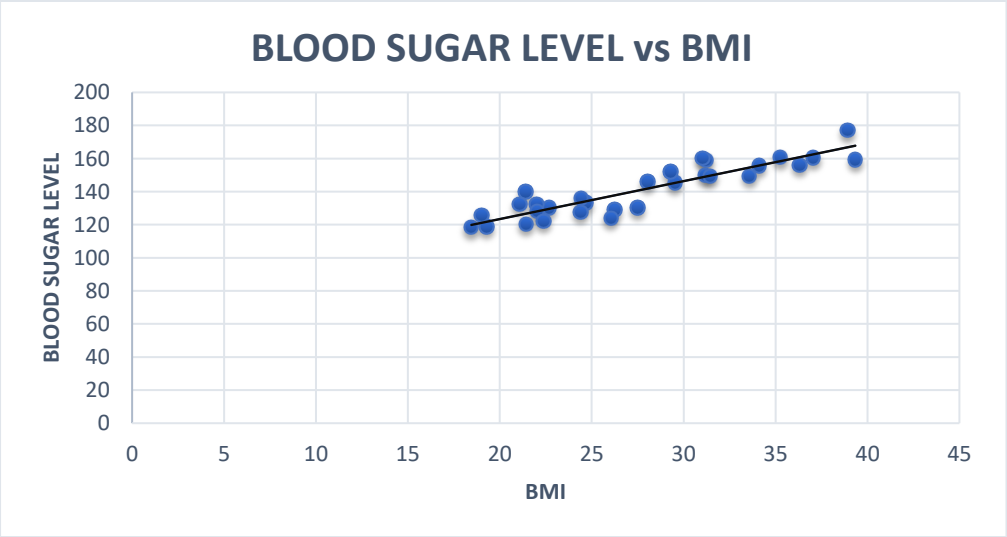

Figure 1: This illustration shows the Linear Regression graph between BMI and Blood Sugar Levels

The slope 2.29 indicates that for every one unit rise in BMI, the blood sugar level increases by 2.29 units on average.

II. Bayesian Regression

2.1 Posterior Distribution

Let $\theta = [\theta_1, \theta_2, \dots, \theta_n]^T$ denote all of the unknowns of the model, which we refer as parameters and $y = [y_1, y_2, \dots, y_n]^T$ the vector of observed data. Let *I* represent all relevant information that is currently available to the individual who is carrying out the analysis, in addition to y. Let us assume for the sake of simplicity that each element of θ is continuous.

Bayesian inference is based on the posterior probability distribution of θ after observing y. It is given by Bayes theorem:

$$p(\theta \mid y, I) = \frac{p(y \mid \theta, I)\pi(\theta \mid I)}{p(y \mid I)}$$

There are two key ingredients: **the likelihood function p(y | θ,I)** and **the prior distribution π(θ | I).**
The prior distribution represents the probability beliefs for θ held before observing the data y. Both likelihood function and prior distribution are dependent upon the current information *I*. Different individuals will have different information *I* based on the random sample, hence in general their prior distributions (and possibly their likelihood functions) might differ.

p(y | I) is called the normalizing constant which ensures that the right-hand side integrates to one over the parameter space. Though the dependence on *I* is of crucial importance, for notational convenience, from this point onwards we suppress this dependence and write

$$p(\theta \mid y) = \frac{p(y \mid \theta)\pi(\theta)}{p(y)}$$

where the normalizing constant is $p(y) = \int_\theta p(y \mid \theta)\pi(\theta)\, d\theta$. p(y) is also the marginal probability of the observed data given the model. Ignoring the normalizing constant gives

$$p(\theta \mid y) \propto p(y \mid \theta) \times \pi(\theta)$$

or, $$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

The use of the posterior distribution for inference is very appealing as it probabilistically combines the information on the parameters contained in the data and in the prior. The manner by which inference is updated from prior to posterior extends naturally to the sequential arrival of data.
Suppose that y_1 and y_2 represent the current totality of data. Then the posterior is

$$p(\theta \mid y_1, y_2) = \frac{p(y_1, y_2 \mid \theta)\pi(\theta)}{p(y_1, y_2)} \qquad (2.1)$$

Suppose in some previous occasion only y_1 was available to us. The posterior based on these data only is

$$p(\theta \mid y_1) = \frac{p(y_1 \mid \theta)\pi(\theta)}{p(y_1)}$$

After observing y_1 and before observing y_2, the "prior" for θ corresponds to the posterior $p(\theta \mid y_1)$, since this distribution takes into account the current beliefs concerning θ. We then

update using

$$p(\theta \mid y_1, y_2) = \frac{p(y_2 \mid y_1,\theta)\pi(\theta \mid y_1)}{p(y_2 \mid y_1)} \qquad (2.2)$$

Factorizing the right-hand side of (2.1) gives

$$p(\theta \mid y_1, y_2) = \frac{p(y_2 \mid y_1,\theta)}{p(y_2 \mid y_1)} \times \frac{p(y_1 \mid \theta)\pi(\theta)}{p(y_1)}$$

which equals the right-hand side of (2.2). Hence, we get a consistent inference based on y_1 and y_2 regardless of whether we produce the posterior in one or two stages.

In the case of conditionally independent observations,

$p(y_1, y_2 \mid \theta) = p(y_1 \mid \theta)p(y_2 \mid \theta)$ in (2.1)

and $p(y_2 \mid y_1, \theta) = p(y_2 \mid \theta)$ in (2.2)

Initially it appears that the Bayesian approach to inference is deceptively straightforward, but there are a number of important issues that must be taken into consideration when in practice. The first issue is prior specification. Second one is once the prior and likelihood ingredients have been decided, we have to summarize the multivariate posterior distribution. We will observe that this summarization requires integration over the parameter space, which might be of higher dimension.

2.2 Summarization of the Multivariable Posterior Distribution

The posterior distribution $p(\theta \mid y)$ is multivariate and marginal distributions for parameters of interest will be needed. The univariate marginal distribution for θ_i is

$$p(\theta_i \mid y) = \int_{\theta_{-i}} p(\theta \mid y)\ d\theta_{-i}$$

where θ_{-i} is the vector θ excluding θ_i, that is, $\theta_{-i} = [\theta_1, ..., \theta_{-i-1}, \theta_{-i+1}, ..., \theta_p]$. Reporting summaries of this distribution is very useful because moments and quantiles can be calculated easily using this. The posterior mean is given by

$$E[\theta_i \mid y] = \int_{\theta_i} \theta_i p(\theta_i \mid y)\ d\theta_i$$

The $100\times q\%$quantile, θ_i (q), with $0 < q < 1$, is obtained by solving

$$q = p[\theta_i \leq \theta_i(q)] = \int_{-\infty}^{\theta_i(q)} p(\theta_i \mid y)\ d\theta_i$$

2.3 Prior Choice

The specification of the prior distribution is a very crucial aspect of the Bayesian approach. An important first observation with respect to the choice of the prior is that for all θ for which $\pi(\theta)=0$,we have $p(\theta \mid y)=0$, regardless of any realization of the observed data, which clearly illustrates that great care should be taken in excluding parts of the parameter space a priori.

There are two types of prior specification- baseline priors and substantive priors. In baseline prior specification, we assume an analysis is required in which the prior distribution has "minimal impact" so that the information in the likelihood dominates the posterior. An alternative name for such an analysis is objective Bayes. Other labels have also been put forward for such prior specification which include reference, non-informative and non-subjective. There is a vast literature behind the construction of objective Bayesian procedures, with an aim often being to define procedures which have good frequentist properties.

An analysis with a baseline prior may be the only analysis performed or, alternatively, may provide an analysis with which other analyses in which substantive priors are specified can be taken into account for comparison. These substantive priors constitute the second type of specification in which the incorporation of contextual information is necessary. Once we have a candidate substantive prior, it is quite beneficial to simulate hypothetical data sets from the prior and examine these realizations to see if they conform to what is desirable. A popular label for analyses for which the priors are, at least in part, based on subject matter information is subjective Bayes.

2.4 Baseline Priors

At first look, it might seem that the specification of a baseline prior is straightforward since we can take

$$\pi(\theta) \propto 1$$

which will ensure that the posterior distribution is simply proportional to the likelihood $p(y \mid \theta)$. But there are two major problems when we use the above proportionality.

The first difficulty is that it provides an improper specification which means it does not integrate to a positive constant ($< \infty$) unless the range of each element of θ is finite. In some cases, this may not be a practical problem if the posterior corresponding to the prior is proper and does not exhibit any aberrant behaviour. A posterior arising from an improper prior can be justified using the limiting case of proper priors, though some statisticians are philosophically are not in support of this argument. Another justification for an improper prior is that such a choice may be thought of as approximating a prior that is "locally uniform" close to regions where the likelihood is non-negligible (so that the likelihood dominates) and decreasing to zero outside of this region. Great care is to be taken to ensure that the posterior corresponding to an improper prior choice is proper. For non-linear models, for example, improper priors should never be used. It is difficult to give general guidelines as to when a proper posterior will result from an improper prior.

2.5 Substantive Priors

The specification of substantive priors is very context specific, but we can give a number of general considerations. When specifying a substantive prior, it is necessary to have a clear knowledge of the meaning of the parameters of the model for which we are specifying priors and this can often be achieved by reparameterization.

To explain the above concept of reparameterization we take into account the follow example of linear regression.

Let $E[Y \mid x]=\gamma_0 + \gamma_1 x$.
It is easier to interpret if we reparameterize the above equation as

$$E[Y \mid z]=\beta_0 +\beta_1(z -\bar{z})$$

where $z = c \times x$ and c is chosen such that the units of z are convenient. Under this parameterization, β_0 is the expected response at $z = \bar{z}$. It is often easier to specify a prior for β_0 than for γ_0 which is average response at x =0, which may be meaningless at times. The slope parameter, β_1, is the change in expected response which corresponds to a c-unit increase in x (1-unit increase in z).

III. Bayesian Approach for Confidence Intervals

Bayesian inference is a statistical approach which aims at estimating a certain parameter (e.g., a mean or a proportion) from the population distribution, given the evidence provided by the observed (i.e., collected) data. Hence, the Bayesian approach for statistical inference is considered a more direct or natural approach to answer a research question, since it estimates the parameter of interest directly from the population distribution rather than estimating from the sampling distribution as the frequentist approach. In the Bayesian approach, parameters of interest are treated as random variables. Therefore, parameters can be described with probability distributions.

One of the main features of the Bayesian approach is the compromise of prior evidence with the observed data. Prior evidence and the observed data are represented with probability distributions that, in Bayesian terminology, are referred to as prior and likelihood distributions respectively. The prior distribution when combined with the likelihood distribution in order to update the previous knowledge, leads to the posterior distribution, which is formally represented as $p(\theta|y)$ where 'θ' represents the parameter of interest and 'y' represents the observed data. The outcome of a Bayesian analysis is the posterior distribution. The posterior distribution can be summarized by measures of central tendency (e.g., median, mean or mode) and measures of uncertainty (e.g., variance or standard deviation). One of the most used measures of uncertainty in Bayesian inference is the Bayesian credible interval (CrI), which is analogous to the confidence interval (CI) in the frequentist approach.

Once the posterior distribution that represents the updated knowledge about a parameter of interest is defined, obtaining the credible interval is straightforward. There are typically two types of Bayesian credible intervals: (1) equal tail interval; and (2) highest posterior density (HPD) interval. The following sections will focus on defining, explaining and interpreting such intervals.

3.1 Equal-tail Credible Interval:

The Bayesian equal-tail credible interval is a method which returns threshold values of the posterior distribution. This represents an interval with the probability of interest (e.g., 95%) of the distribution mass around the centre of the distribution (represented in figure given below).

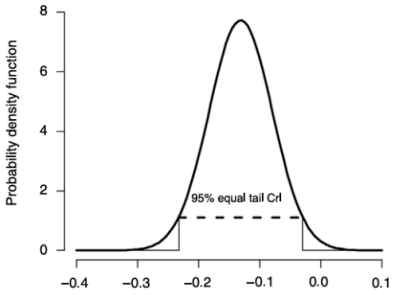

Figure 2:This illustration represents a 95% equal-tail CrI for a symmetric distribution.

In other words, the lower limit of the 95% equal tail CrI is the quantile that represents a probability of 0.025 (or the 2.5% per centile) of the posterior distribution, while the upper limit of the equal tail CrI is the quantile that represents a probability 0.975 (or the 97.5% percentile) of the posterior distribution. An advantage of estimating the equal tail Bayesian CrI is that this interval is easily calculated. However, a common concern related to the equal tail Bayesian CrI is that it might yield estimate values with lower probability inside the interval than those outside the interval when the posterior distribution is not symmetric (i.e., right or left skewed). When this happens, it will mean that some values would have a higher probability of representing the parameter when they lie outside the interval than compared to those values inside the interval. Graphically, this will yield a shift line connecting the lower and upper limits for this interval. As this situation is not desired, another method has been proposed so as to estimate Bayesian CrIs: the Highest posterior density (HPD) CrI.

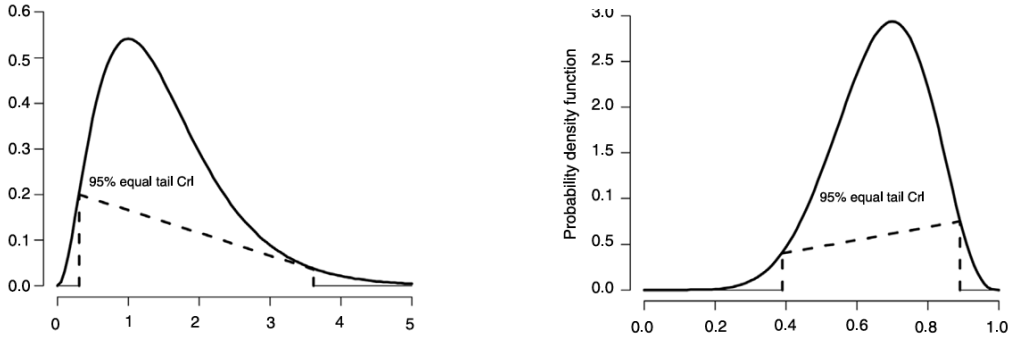

Figure 3: The above illlustrations represent a 95% equal-tails CrI for positive and negative skewed distributions.

3.2 Highest posterior density (HPD) Credible interval

The Bayesian HPD CrI method returns threshold values of the posterior distribution that

represent an interval with the probability of interest (e.g., 95%) of the distribution mass that is around the centre of the distribution. It is based on the assumption that all values inside the interval have higher probabilities of representing the parameter than all the values outside the interval. For example, for a 95% HPD CrI, the interval contains 95% of the mass of the posterior distribution around the centre of the distribution and all values inside the interval are more likely to represent the parameter than those values which lie outside the interval. Graphically, this is always represented by a straight line connecting the lower and upper limits for this interval.

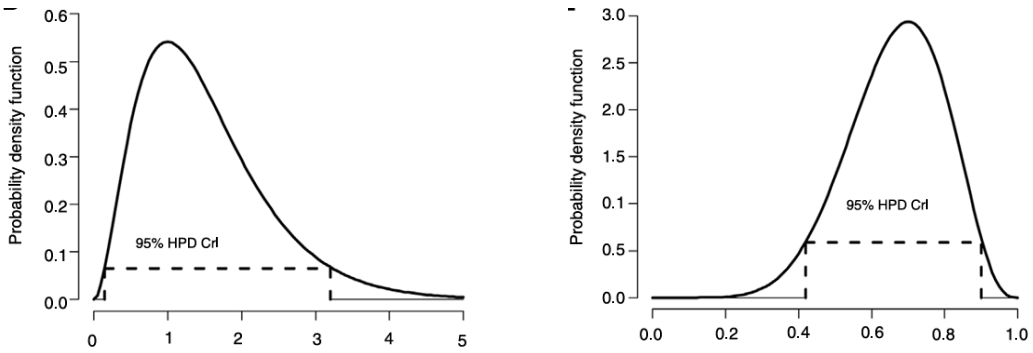

Figure 4: The above illustrations represent a 95% HPD CrI for positive and negative skewed distributions.

For symmetric posterior distributions, the HPD CrI is equivalent to the equal tail Bayesian CI. A disadvantage of the HPD CrI method is that the computation of the interval is more complex in comparison to the equal tail CrI method because the HPD CrI estimation requires numerical optimization.

3.3 Interpretation of Bayesian Credible Intervals (CrI)

Bayesian CrIs have a more natural interpretation in comparison to frequentist CIs. This happens mostly due to the fact that the Bayesian CrI estimates the most probable values of the parameter of interest directly from the computed posterior distribution which necessarily contains all knowledge and evidence about the population distribution at that moment. The interpretation of the Bayesian 95% CrI is as follows: there is a 95% probability that the true (unknown) effect estimate (represented by λ) would lie within the interval, given the evidence provided by the observed data. The way we judge if there is a statistical significance result when interpreting the Bayesian CrI is quite similar to the frequentist CI.

Let us consider two samples say $X_{11}, X_{12}, \ldots\ldots..X_{1m}$ and $X_{21}, X_{22}, \ldots\ldots X_{2n}$. Both are iid samples drawn from $N(\mu, \sigma^2)$ where both the parameters are known. We are to test that the group mean difference is zero, i.e.,$\overline{x_1} - \overline{x_2}$ =0. Let us suppose that a 95% CrI is composed of the following limits: −4.0 to −1.0 This means that there is a 95% probability that the population mean difference will lie between −4.0 and −1.0 given the chosen sample. This means that the interval

(-4.0,1.0) has a higher probability of representing the true unknown estimate which indicates that the mean of the intervention group will be lower compared to the comparison group, with at least a 95% probability since the interval contains all negative values. Let us assume another 95% CrI is composed of the following limits: 0.5 to 3.5. This CrI indicates that there

is a 95% probability that the population mean difference between the two groups will lie between 0.5 and 3.5, given the observed data. This means that the interval (0.5,3.5) has a higher probability of representing the true unknown estimate which indicates that the mean of the intervention group will be higher compared to the comparison group since it contains all positive values. Both scenarios would indicate a statistically significant result at a significance level of 5%, since both CrIs do not contain zero. However, in case of a 95% CrI composed of the following limits: −2.0 to 1.0, this would indicate that there is a 95% probability that the population mean difference would lie between −2.0 and 1.0, given the observed data. Since the most plausible values (i.e., −2.0 to 1.0) with higher probability of representing the true (unknown) estimate indicate that the mean of the intervention group could be either lower compared to the comparison group if the true unknown estimate lies between (-2.0,0.0) and higher compared to the intervention group if the true unknown estimate lies between (0.0,1.0). Hence this is a non-statistically significant result.

3.4 Predicting Systolic Blood Pressure(SBP) using Bayesian Regression

Background

Hypertension is a major risk factor for cardiovascular diseases in recent times. It is extremely crucial to understand the relationship between age and systolic blood pressure (SBP) in predicting and managing patient health. Ordinary Least Squares (OLS) provides point estimates and confidence interval but it cannot incorporate prior knowledge. Bayesian regression addresses these limitations by incorporating prior distributions and yielding posterior distributions that has the ability to quantify parameter uncertainty. We aim to represent a Bayesian regression model to predict SBP based on patient age and evaluates the resulting credible intervals.

Data Description

A synthetic dataset that represents 30 patients was used to model the relationship between age and SBP. The dataset includes patient id (to uniquely identify a patient), age of the patient and their SBP(mmHg).

Patient_ID	Age	SBP
1	25	115
2	30	118
3	35	122
4	40	125
5	45	130
6	50	135
7	55	140
8	60	145
9	28	116
10	33	120
11	38	124

12	43	128
13	48	132
14	53	138
15	58	142
16	63	147
17	27	114
18	32	119
19	37	123
20	42	127
21	47	131
22	52	137
23	57	141
24	62	146
25	26	113
26	31	117
27	36	121
28	41	126
29	46	129
30	51	134

Bayesian Regression Model

The relationship between SBP and Age is modelled given as follows:

$$Y = \alpha + \beta X + \epsilon$$

where Y is the SBP, X is the age, α is the intercept, β is the slope of the graph and ϵ is the error term. Assume $\epsilon \sim N(0,\sigma^2)$.

Prior distributions are:

$\alpha \sim N(91.33,(1.045)^2)$

$\beta \sim N(0.86,0.02^2)$

$\sigma^2 \sim$ Inverse Gamma (6.00,4.76)

After performing Bayesian regression, we get the following posterior estimates:

$\alpha \approx 91.33$

$\beta \approx 0.86$

$\sigma^2 \approx 0.95$

The credible interval for β (age effect on SBP) is:

$$P(0.82 \leq \beta \leq 0.9 | Data) = 0.95$$

Since the given CrI does not contain 0, hence we can conclude that SBP is significantly influenced by Age.

IV. Bayesian Regression versus Linear Regression

4.1 Dataset 1: Non-Adult Dental Age Assessment

Introduction:
Dental age estimation is a critical component of forensic science, orthodontics, and physical anthropology. Various methodologies have been developed to predict age from dental mineralization stages, among which linear regression using Correspondence Analysis(CAR) and Bayesian methods are commonly used. This chapter presents a comparative study between these two approaches, focusing on their accuracy, reliability, and practical implementation using Python.
This chapter focuses on the evaluation of factors influencing the quality (accuracy and reliability) of non-adult dental age assessment from radiographic stages of permanent teeth. For this purpose we have taken four distinct cross-sectional samples of healthy children from 3 known geographic region- France, Iran and Ivory Coast 3 and 1 additional sample of children whose grandparents originated from a different continent. As stated previously two different methods of calculation have been used to check if either one is more efficient than other or the method help us to draw some conclusions.

Data

AGE_CATEGORY	FRANCE		FRANCE		IRAN		IVORY COAST	
	F	M	F	M	F	M	F	M
<=48-54<	1	1	3	0	0	0	0	1
<=54-60<	1	0	1	3	0	0	0	0
<=60-66<	0	3	1	3	0	0	1	2
<=66-72<	2	1	4	1	1	0	4	4
<=72-78<	0	1	7	6	1	1	7	4
<=78-84<	1	3	5	7	9	4	9	11
<=84-90<	2	3	3	6	8	3	17	7
<=90-96<	10	9	11	5	13	12	13	10
<=96-102<	14	7	7	9	15	11	21	10
<=102-108<	13	12	16	10	16	9	20	4
<=108-114<	11	11	13	13	17	6	12	9
<=114-120<	14	18	24	8	20	10	5	8
<=120-126<	24	16	23	15	11	10	5	6
<=126-132<	21	15	20	8	10	11	6	10
<=132-138<	28	19	20	14	14	7	6	3
<=138-144<	26	13	18	16	12	9	9	3
<=144-150<	20	19	11	14	18	5	5	3

<=150-156<	15	17	11	6	16	10	2	2
<=156-162<	11	14	10	9	20	10	6	3
<=162-168<	13	7	11	5	9	6	2	4
<=168-174<	6	11	3	4	10	5	1	4
<=174-180<	2	4	6	2	10	3	2	0
<=180-186<	4	5	2	1	6	2	0	0
<=186-192<	3	3	4	5	13	2	0	0
<=192-198<	0	0	0	0	8	0	1	0
TOTAL	242	212	234	170	257	136	154	108

Cross sectional standardised orthopantomographs of the teeth of 262 children from The Ivory Coast (108 boys and 154 girls) aged 49 to 194 months, 393 children from Iran (136 boys and 257 girls) aged 69 to 197 months, and 454 children from France (212 boys and 257 girls) aged 45 to 192 months make up our three geographic population samples. All of the children had known chronological ages (measured in months; from the date of birth to the radiograph time) and, crucially, ethnicities (their grandparents came from the same nation and/or region). The three geographic populations consist of France, Iran and the Ivory Coast. An additional sampling was made for those children born in France but have one or more grandparents not originating from Europe.

Methodology:

Correspondence Analysis combined with Linear Regression

Correspondence Analysis (CA)

It is a statistical technique that is used to convert categorical data into numerical coordinates. Dental maturity stages are ordinal ranging from A to H as per Demirjian's method. Hence these stages need to be converted into numerical values before we apply regression on them. Under this transformation, teeth that are in similar stages are assigned similar values.

Inder traditional regression techniques, dental stages are treated as continuous variable but in reality, they are discrete. This method has the power to standardize the distance between different maturity stages which in turn ensures better numerical representation. CA can also identify patterns in categorical data by analyzing associations between different maturity stages

Linear Regression for Age Estimation

After applying CA each dental stage is represented by a numerical score which can summed up to get the GMI (Global Maturity Index).
The simple linear regression model is then applied:

$$\text{Age}=\beta_0+\beta_1\ (\text{GMI})+\epsilon$$

where Age is the estimated chronological age, GMI is the sum of dental maturity scores, β_0, β_1 are the regression coefficients and ϵ is the error term.

There are certain limitations of the CAR Method. This method is based on the assumption that age and dental maturity have linear relationship which might not be true. For advanced dental stages this method might overestimate the numerical coordinate during transformation.

Bayesian Predictions for Dental Age Estimation

Bayesian inference follows from **Bayes' Theorem** given by

$$P(A|B) = \frac{P(B|A)P(A)P(B)}{P(A|B)}$$

where P(A|B) is **Posterior probability** (Updated probability of age given observed data), P(B|A) is **Likelihood** (Probability of observing a specific dental maturity stage given a known age), P(A) is **Prior probability** (Initial assumption about the probability of different ages) and P(B) is **Marginal probability** (Normalization factor ensuring probabilities sum up to 1)

The prior probability P(A) is used to represent the initial likelihood of an individual belonging to a specific age category. We have assumed a uniform prior which means all age groups are assumed equally likely before considering the dental data. Then, we compute the Likelihood P(B|A) which represents the probability of observing a specific Dental Mineralization Sequence (DMS) for a given age. Each age group has a distinguishable pattern of tooth development which is used by Bayesian inference to estimate the likelihoods. We then apply Bayes theorem to get the Posterior probability P(A|B). The age category corresponding to the highest posterior probability is chosen as the estimated age.

Python Codes

Code for CAR Method

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.decomposition import TruncatedSVD  # Correspondence
Analysis
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score

# Step 1: Load Dataset
file_path =
"https://docs.google.com/spreadsheets/d/15wLuE25Na7pk6txG2bKGb-
AgDPFhYaqN7-0EcY6fGw4/export?format=csv"
df = pd.read_csv(file_path)

# Step 2: Data Cleaning
df.rename(columns={'Unnamed: 0': 'Age_Category'}, inplace=True)
df = df.loc[:, ~df.columns.str.contains('Unnamed')]

# Convert Age_Category from range labels (e.g., "<=48-54<") to
midpoints
df["Age_Category"] = df["Age_Category"].str.extract(r'(\d+)-
(\d+)').astype(float).mean(axis=1)
```

```
# Identify country columns (assumed as features for CA)
country_columns = ['FRANCE', 'FRANCE.1', 'IRAN', 'IVORY COAST']
df[country_columns] = df[country_columns].apply(pd.to_numeric,
errors="coerce")

# Drop NaN values
df = df.dropna()

# Step 3: Correspondence Analysis (CA) using SVD
n_components = 2  # Reduce to 2 dimensions
ca = TruncatedSVD(n_components=n_components)
ca_components = ca.fit_transform(df[country_columns])

# Step 4: Prepare Data for Linear Regression
X = ca_components  # Use CA-transformed features
y = df["Age_Category"]  # Target variable

# Split Data into Training and Testing Sets
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Step 5: Train Linear Regression Model
model = LinearRegression()
model.fit(X_train, y_train)

# Predict on Test Set
y_pred = model.predict(X_test)

# Step 6: Evaluate Model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"\n☑ Linear Regression Results After Correspondence Analysis:")
print(f"📉 Mean Squared Error (MSE): {mse:.2f}")
print(f"📈 R² Score: {r2:.2f}")

# Step 7: Visualizing Correspondence Analysis
plt.figure(figsize=(8, 6))
plt.scatter(X_test[:, 0], y_test, color="blue", label="Actual Ages",
alpha=0.6)
plt.scatter(X_test[:, 0], y_pred, color="red", label="Predicted Ages",
alpha=0.6)
plt.xlabel("First Correspondence Analysis Component")
plt.ylabel("Age (Months)")
plt.title("CA + Linear Regression: Predicted vs. Actual Age")
plt.legend()
plt.show()
```

Output

```
☑ Linear Regression Results After Correspondence Analysis:
📉 Mean Squared Error (MSE): 2075.83
📈 R² Score: 0.03
```

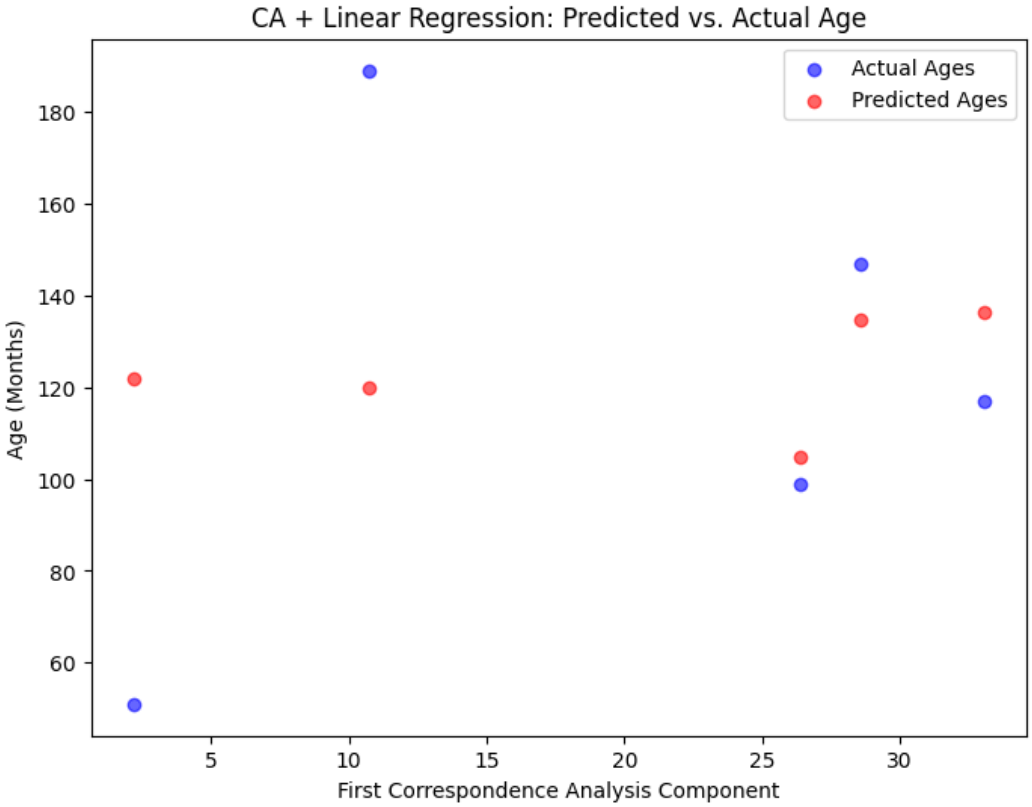

Figure 5: The above illustration represents the difference between predicted and actual ages using Correspondence Analysis along with Linear Regression.

Inference from CAR Graph

It is quite clear from the above graph that there is significant misalignment between the actual and predicted ages. We also observe there is a high MSE.

Bayesian Predictions

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from collections import defaultdict
from sklearn.model_selection import train_test_split, KFold
from sklearn.metrics import accuracy_score, confusion_matrix

# Step 1: Load Dataset
file_path =
"https://docs.google.com/spreadsheets/d/15wLuE25Na7pk6txG2bKGb-
AgDPFhYaqN7-0EcY6fGw4/export?format=csv"
df = pd.read_csv(file_path)

# Step 2: Data Cleaning
```

```
# Rename first column as Age Category
df.rename(columns={'Unnamed: 0': 'Age_Category'}, inplace=True)

# Drop unnecessary columns (Unnamed columns)
df = df.loc[:, ~df.columns.str.contains('Unnamed')]

# Convert Age_Category from range labels (e.g., "<=48-54<") to
midpoints
df["Age_Category"] = df["Age_Category"].str.extract(r'(\d+)-
(\d+)').astype(float).mean(axis=1)

# Convert country columns to numeric
country_columns = ['FRANCE', 'FRANCE.1', 'IRAN', 'IVORY COAST']
df[country_columns] = df[country_columns].apply(pd.to_numeric,
errors="coerce")

# Drop any remaining NaN values
df = df.dropna()

# Step 3: Compute Prior Probabilities (P(Age))
age_counts = df["Age_Category"].value_counts(normalize=True).to_dict()

# Step 4: Compute Likelihood P(DMS | Age) with Adaptive Smoothing
smoothing_factor = 1 / (len(df) + 1)  # Adjust smoothing based on
dataset size
dms_likelihood = defaultdict(lambda: defaultdict(lambda:
smoothing_factor))

for _, row in df.iterrows():
    age = row["Age_Category"]
    dms_sequence = tuple(row[country_columns])  # Convert country data
into a tuple
    dms_likelihood[age][dms_sequence] += 1  # Count occurrences

# Convert counts to probabilities
for age in dms_likelihood:
    total = sum(dms_likelihood[age].values())
    for dms in dms_likelihood[age]:
        dms_likelihood[age][dms] /= (total + smoothing_factor *
len(dms_likelihood[age]))  # Normalize

# Step 5: Bayesian Prediction Function
def bayesian_prediction(dms_sequence):
    posteriors = {}

    for age in age_counts:
        likelihood = dms_likelihood[age].get(dms_sequence,
smoothing_factor)
        prior = age_counts.get(age, smoothing_factor)
        posteriors[age] = likelihood * prior  # Compute posterior

    total_prob = sum(posteriors.values()) or 1  # Avoid division by
zero
    for age in posteriors:
        posteriors[age] /= total_prob  # Normalize

    return max(posteriors, key=posteriors.get)
```

```
# Step 6: K-Fold Cross-Validation for Model Evaluation
kf = KFold(n_splits=5, shuffle=True, random_state=42)
accuracies = []

for train_index, test_index in kf.split(df):
    train_data, test_data = df.iloc[train_index], df.iloc[test_index]

    # Apply Bayesian prediction on test data
    y_true = test_data["Age_Category"]
    y_pred = test_data[country_columns].apply(lambda row:
bayesian_prediction(tuple(row)), axis=1)

    # Evaluate Accuracy
    accuracy = accuracy_score(y_true, y_pred)
    accuracies.append(accuracy)

# Final Accuracy with Cross-Validation
print(f"\n☑ Bayesian Age Prediction Accuracy (Cross-Validation):
{np.mean(accuracies):.2f} ± {np.std(accuracies):.2f}")

# Step 7: Visualization (Confusion Matrix)
y_true = test_data["Age_Category"]
y_pred = test_data[country_columns].apply(lambda row:
bayesian_prediction(tuple(row)), axis=1)

conf_matrix = confusion_matrix(y_true, y_pred)
plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap="Blues",
xticklabels=np.sort(y_true.unique()),
yticklabels=np.sort(y_true.unique()))
plt.xlabel("Predicted Age")
plt.ylabel("Actual Age")
plt.title("Confusion Matrix of Bayesian Predictions")
plt.show()
```

Output

```
Bayesian Age Prediction Accuracy (Cross-Validation): 1.00 ± 0.00
```

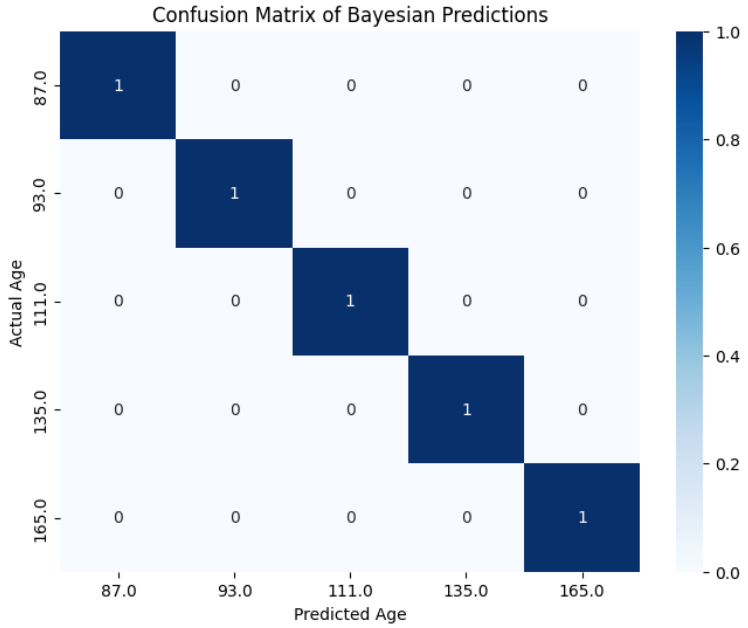

Figure 6: The above illustration represents the Confusion matrix between actual and predicted ages using Bayesian Predictions.

Inference from Confusion Matrix

The above confusion matrix of Bayesian predictions indicates a perfect classification of age predictions. Every actual age is correctly predicted with no misclassifications. Hence, we can conclude the Bayesian model is highly accurate for the given data.

Conclusion

In case of this dataset we see that the linear regression model fails to fit since there is misalignment between the actual and predicted ages as well as there is a high MSE. This is highly due to the fact that the relationship between dental maturity and age is not necessarily linear. Whereas, the confusion matrix in case of the Bayesian Predictions gives all the diagonal elements of the matrix which indicates 100% accuracy.

This happens because Bayesian models make use of the knowledge of prior probabilities and update them with observed data. It might happen that the relationship between dental maturity and age is not strictly linear. In such cases, Bayesian predictions make use of the available data and captures the patterns which can be used to predict the posterior distribution.

4.2 Dataset 2: NFL Weekly Playoff Probabilities (2002-2024)

Introduction

Our aim is to predict a team's playoff probability based solely on the week number using Linear Regression and Bayesian Ridge Regression. As the season progresses, playoff

probabilities should follow a predictable trend. The hypothesis is that as the season progresses, playoff probabilities should follow a predictable trend. We try to investigate whether week number alone is a sufficient predictor or if additional features are necessary for the payoff probability prediction.

Dataset

The raw data includes regular season NFL games from 2002 to 2024. We aim to estimate that given the probability a team qualifies for the playoff for a particular week what is the with probability that they will win the playoff. Several factors influence whether a team will qualify the playoff- week, opponent strength, etc. Here we mainly try to study the effect of the week on winning the probability. It would naturally occur to us that the later the team has their match higher is their chances of winning. To study this relationship, we mainly use two methods- Linear Regression and Bayesian Predictions. In the dataset the following have been specified

season: the year of the NFL season

week: The week of the season

team: The NFL team name

playoff probability: The probability of the team making into the playoff at that week

The following charts describe how playoff probability changes with respect to every year and a graphical representation of the playoff probabilities of every team over the years.

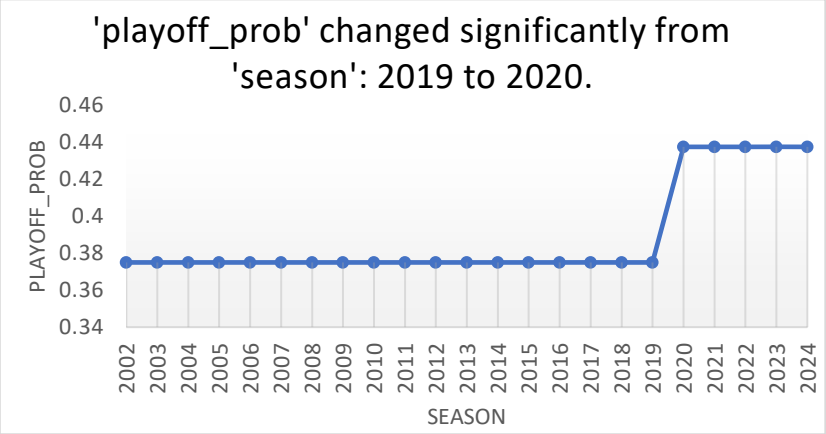

Figure 7: This graph depicts the playoff probability changes over the years and the significant rise from 2019 to 2020.

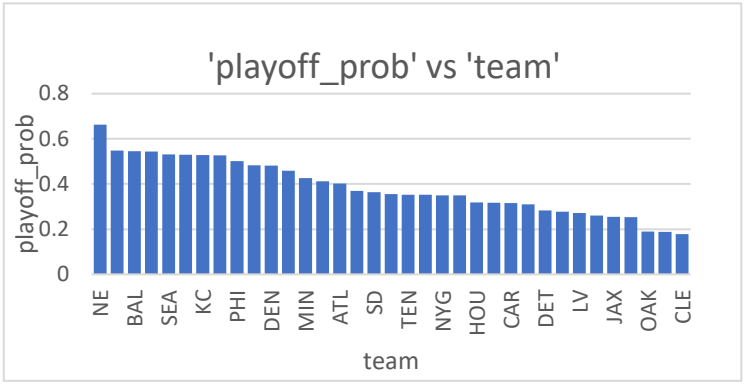

Figure 8:This illustration is a graph depicting the playoff probability of each team over the years.

Methodology

Linear Regression for Playoff Prediction

Linear Regression assumes a linear relationship between the independent variable (**Week**) and the dependent variable (**Playoff Probability)**

$y=mx+b+\epsilon$

where y is the predicted **playoff probability**, x is the **week number**, m (slope) represents the rate of change in probability per week, b(intercept) is the starting probability when x=0 and ϵ is the error term.

Bayesian Ridge Regression for Playoff Prediction

This method is a probabilistic approach to regression which is known for its ability to incorporate prior knowledge and uncertainty estimation. Although, it is an extension of linear regression, but this method helps in preventing overfitting by introducing regularization terms.

In the context of the NFL Weekly Playoff Probabilities dataset (2002-2024), Bayesian Ridge Regression is used to predict the playoff probability of a team based on a given week number.

In traditional Linear Regression, we solve for the coefficients w by minimizing the squared error:

$y=Xw+\epsilon$

where y is the playoff probability, x is the week number, w is the weights (coefficients to be learnt) and ϵ is the error term. We assume that the regression coefficients w come from a Gaussian distribution.

$$P(w)=N(0,\lambda^{-1}I)$$

This prior helps in regularizing the model by preventing overfitting of large values. Here λ is the precision parameter that controls the strength of regularization.

Likelihood Function

The observed playoff probabilities are assumed to follow Gaussian distribution given X.

$$P(y|X,w,\alpha)= N(Xw,\alpha^{-1}I)$$

Here α represents noise precision.

Posterior Distribution Estimation

We combine the prior distribution with the likelihood function using Bayes' theorem to obtain the posterior distribution of the regression coefficients.

$$P(w|X,\alpha,y)=\frac{P(y|X,w,\alpha)P(w|\lambda)}{P(y|X)}$$

This posterior distribution is also Gaussian distribution, and it gives a probabilistic estimate of w.

Python Code for performing Linear Regression and Bayesian Ridge Regression

```
# Install necessary libraries
!pip install kagglehub[pandas-datasets]

# Import libraries
import kagglehub
from kagglehub import KaggleDatasetAdapter
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression, BayesianRidge
from sklearn.metrics import mean_squared_error, r2_score

# Load the dataset from Kaggle
file_path = "nfl_weekly_playoff_probabilities_2002_2024.csv"
df = kagglehub.load_dataset(
    KaggleDatasetAdapter.PANDAS,
    "justinstockssmith/nfl-weekly-playoff-probabilities-2002-2024",
    file_path
)

# Print column names to verify correct labels
print(df.columns)

# Data preprocessing
# Adjust column names if they are different
df = df.dropna(subset=['week', 'playoff_prob'])  # Ensure no missing
values
X = df[['week']]  # Independent variable (Week of the season)
y = df['playoff_prob']  # Dependent variable (Playoff probability)

# Split data into training and testing sets (80% train, 20% test)
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Initialize and train the Linear Regression model
linear_model = LinearRegression()
linear_model.fit(X_train, y_train)

# Make predictions on the test set
y_pred_linear = linear_model.predict(X_test)
```

```
# Evaluate Linear Regression
mse_linear = mean_squared_error(y_test, y_pred_linear)
r2_linear = r2_score(y_test, y_pred_linear)
print(f"Linear Regression - MSE: {mse_linear:.4f}, R2:
{r2_linear:.4f}")

# Initialize and train the Bayesian Regression model
bayesian_model = BayesianRidge()
bayesian_model.fit(X_train, y_train)

# Make predictions on the test set
y_pred_bayes = bayesian_model.predict(X_test)

# Evaluate Bayesian Regression
mse_bayes = mean_squared_error(y_test, y_pred_bayes)
r2_bayes = r2_score(y_test, y_pred_bayes)
print(f"Bayesian Regression - MSE: {mse_bayes:.4f}, R2:
{r2_bayes:.4f}")

# Plot Linear Regression line
plt.figure(figsize=(10, 5))
plt.scatter(X_test, y_test, color="blue", label="Actual Playoff
Probabilities", alpha=0.5)
plt.plot(X_test, y_pred_linear, color="red", linewidth=2, label="Linear
Regression Line")
plt.xlabel("Week of Season")
plt.ylabel("Playoff Probability")
plt.title("Linear Regression - NFL Playoff Probability Prediction")
plt.legend()
plt.show()

# Compare actual vs predicted playoff probabilities for both models
plt.figure(figsize=(10, 5))
plt.scatter(y_test, y_pred_linear, alpha=0.5, label="Linear Regression
Predictions", color="blue")
plt.scatter(y_test, y_pred_bayes, alpha=0.5, label="Bayesian Regression
Predictions", color="red")
plt.plot([0, 1], [0, 1], "--", color="black")  # Perfect prediction
line
plt.xlabel("Actual Playoff Probability")
plt.ylabel("Predicted Playoff Probability")
plt.title("Comparison of Linear & Bayesian Regression Predictions")
plt.legend()
plt.show()
```

Output

```
Linear Regression - MSE: 0.1241, R2: -0.0005
Bayesian Regression - MSE: 0.1241, R2: -0.0005
```

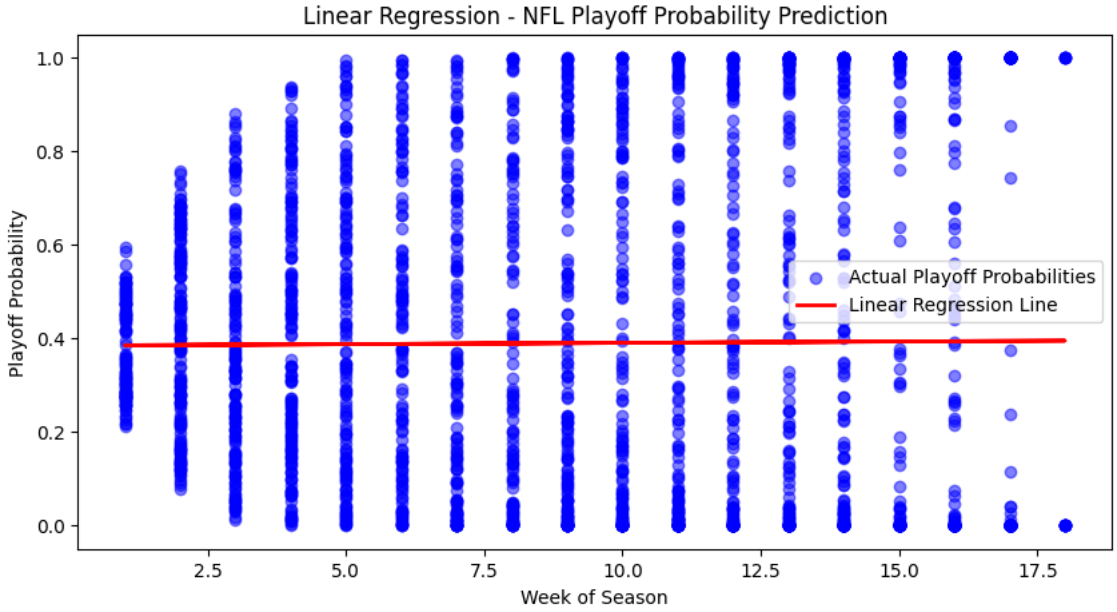

Figure 9: This picture illustrates the playoff probability prediction using Linear Regression versus the actual one.

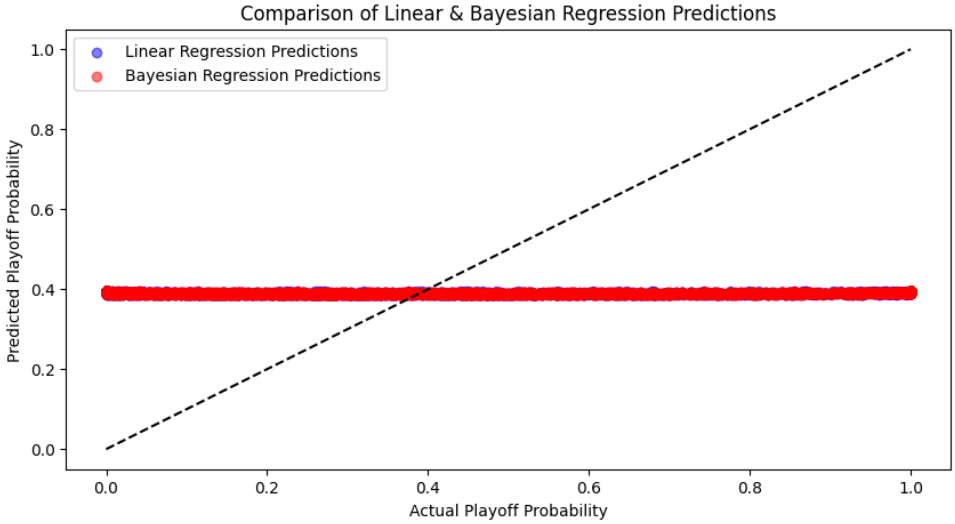

Figure 10: This illustration depicts how Bayesian and Linear regression give the same predictions which is different from the actual one depicted by the line in black.

Inference

R^2 = -0.005 in case of both the models. Hence, we can say that neither of the models could capture meaningful trends of the dataset.

Possible reasons behind the failure of both the models:

- Weak Predictor Variable- NFL playoff probabilities depend on several factors not just week. Playoff probabilities are influenced by team strategy changes, strength of the opponent and the team's performance as well.
- Absence of a Linear Relationship- Linear Regression assumes a linear relationship between week and playoff probability. But in real life scenarios, playoff probabilities might change non-linearly as well. It might happen that early in the season the probabilities fluctuate and stabilize on a later stage. In such cases a non-linear model might work better.
- No impact of Bayesian Ridge Regression- In this case there is only one independent variable week which is also a week predictor. Hence this method fails to capture the trends of the previous year's available. This method is mostly applied to avoid overfitting but here there is only feature so there is no risk of overfitting. Consequently, Bayesian Ridge Regression performs exactly like linear regression.
- Lack of additional information- NFL playoff probabilities are team-specific, but the model does not include team-level data. The model should display margin of victory/defeat of a team, head-to-head matches and their divisional standings and most importantly a teams win-loss record.

Conclusion

This study conducted a comparative analysis between Bayesian methods and Linear regression for predictive modelling. In case of Dental Age Estimation, we see that Bayesian predictions outperforms Linear regression. This is because the former due to its probabilistic framework can incorporate prior knowledge of the data and it can also analyse based on the past trend. Hence the predictions made based on Bayesian statistics are more accurate. Linear regression in general assumes a linear relationship between the dependent and independent variable which might not always be the case. If sufficient data is not available, then it might happen that both the models fail to deliver correctly as we have seen with the NFL Weekly Playoff Probability (2002-2024).

Overall, we can conclude that Bayesian methods are more accurate when we deal with probabilistic reasoning and uncertainty quantification. Linear regression can be used when it is already predefined, or we can strongly infer based on the given data that there is a linear relationship between the dependent and independent variable.

References

1. Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., ... & Gabry, J. (2020). *Bayesian and frequentist regression methods*. Springer.
2. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer.
3. Hespanhol, L., Vallio, C. S., Costa, L. M., & Saragiotto, B. T. (2019). Understanding and interpreting confidence and credible intervals around effect estimates. *Brazilian Journal of Physical Therapy, 23*(4), 290–301. https://doi.org/10.1016/j.bjpt.2018.12.006
4. J. Braga ,Y. Heuze , O. Chabadel ,N. K. Sonan ,A. Gueramy (2003). Non-adult dental age assessment: correspondence analysis and linear regression versus Bayesian predictions.