BAREKENG: Journal of Mathematics and Its Applications
September 2026 Volume 20 Issue 3 Page 1985-1998
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss3pp1985-1998

MIXED-EFFECT MODELS WITH RESTRICTED MAXIMUM LIKELIHOOD (REML), BOOT-STRAPPED REML AND BAYESIAN INFERENCE IN APPLICATION OF GAPMINDER DATA

Asysta Amalia Pasaribu[1*], Kusman Sadik[2], Anang Kurnia[3]

[1]*Department of Statistics, School of Computer Science, Bina Nusantara University*
Jln. Raya Kb. Jeruk No. 27, Jakarta Barat, 11530, Indonesia

[1,2,3]*Study Program of Statistics and Data Science, School of Data Science, Mathematics, and Informatics, IPB University*
Jln. Raya Darmaga Kampus IPB, Kabupaten Bogor, Jawa Barat, 16680, Indonesia

*Corresponding author's e-mail: * **asysta.amalia@binus.ac.id***

Article Info

Article History:

Received: 14th July 2025
Revised: 14th December 2025
Accepted: 16th March 2026
Available online: 8th April 2026

Keywords:

Bayesian MCMC;
Bayesian theorem;
Boot-Strapped REML;
Gapminder data;
REML.

ABSTRACT

Mixed effects model combines fixed effects and random effects, allowing for the analysis of data with both fixed and random variations. This modeling approach is widely utilized across various fields. In R, the lme4 package is commonly employed to estimate mixed effects models using Restricted Maximum Likelihood (REML). There are several methods for estimating model parameters, including Bayesian inference, which has gained prominence with ongoing research advancements. Bayesian inference using Markov Chain Monte Carlo (MCMC) is among the most widely used Bayesian methods. Bayesian inference leverages probabilistic distributions to estimate parameters.to understand the general overview of life expectancy, serving as an indicator of survival time across different continents in the Gapminder dataset, it's essential to identify relevant variables after computing mixed effects predictions using Maximum Likelihood and REML estimation. This involves predicting life expectancy by integrating both random and fixed effects, determining relevant variables after estimating the Mixed Effects Model using REML Bootstrap estimation, and identifying influential variables after estimating the Mixed Effects Model using Bayesian MCMC inference. The methods employed include REML, Bootstrapped-REML, and Bayesian MCMC. The results indicate that all inference methods can be utilized to estimate parameters, with all predictor variables influencing life expectancy, except for the population variable. Further research is recommended to utilize data with more complex predictor variables.

How to cite this article:

A. A. Pasaribu, K. Sadik and A. Kurnia., "MIXED-EFFECT MODELS WITH RESTRICTED MAXIMUM LIKELIHOOD (REML), BOOT-STRAPPED REML AND BAYESIAN INFERENCE IN APPLICATION OF GAPMINDER DATA", *BAREKENG: J. Math. & App.*, vol. 20, no. 3, pp. 1985-1998, Sep, 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Dataset on the dependent variable is defined at multiple periods in time for each unit of analysis are referred to as longitudinal data. One kind of data set that can be utilized to develop mixed-effect models is longitudinal data. We acknowledge the presence of levels in the dataset using the mixed effect model. The notion of data levels is derived from concepts found in the literature on hierarchical linear modeling (HLM) [1] Each dataset appropriate in mixed-effect model contains at least of two data levels. Depending on the number of data levels, we classify the example data sets we study as either two-level or three-level data sets. Level 1, Level 2, and Level 3 are the designations for the maximum of three layers of data that we recognize. We present observations at the most in-depth level of the data in Level 1. The repeated measures conducted on the same unit of analysis are represented by Level 1 in the longitudinal data set. The hierarchy is represented by Level 2, which we express for the following level. The following level of the hierarchy is represented by Level 3. We estimate the fixed-effect parameters in the mixed-effect model. Restricted maximum likelihood (REML) is a method for estimating mixed-effect models. When there are random effects and their relative variance (variance components), REML estimation is used.

Several studies have used REML to estimate mixed-effect models in various fields such as Economy and Finance [2], [3], [4], Medicine and Biology [5], [6] Science Psychology [7], [8] Agriculture and Animal Science [9], [10], [11]. Further inference on mixed models can be used bootstrap in semiparametric likelihood. The bootstrap method presents a theory that applies to the construction of intervals and a corresponding testing procedure. Analysis of the method shows that the method is asymptotically consistent under general regularity conditions without any assumption of normal distribution on the stochastic components in the model. In addition to frequentist inference used in mixed-effect models, Bayesian inference can be used for estimation of mixed-effect models. We introduce the reader to further explore the literature on Bayesian methods using a complete probability model that accounts for not just our uncertainty in the value of an outcome variable is what Bayesian inference is all about y conditional on some unknown parameters θ, but also our a priori uncertainty about the parameters θ themselves. When it comes to regression models, besides the outcome variable y, we also have predictor variables, denoted x. The final is to update our beliefs about the parameters θ based on our model and data. Some studies that are interested in using mixed effect models with Bayesian inference are [12], [13].

In conducting parameter estimation with frequentist and Bayesian inference, there is a fundamental difference between the Bayesian framework and the frequentist framework, namely which quantity is assumed to be fixed. The frequentist paradigm is related to the probability of observed data with fixed parameters, meaning it does not have a probability distribution and frequentist inference is related to the order of the observed hypothesis data set (vector y). In contrast to Bayesian inference which is related to a certain N observation set and Bayesian inference is interested in the probability distribution of parameters. Several research have been widely using frequentist and Bayesian likes [14], [15].

Based on the above explanation, it is known that the objective of this study is to understand the general description of life expectancy that indicator of survival periods in several continents in Gapminder data, to create a model of life expectancy include random effects and fixed effects using Mixed-Effect Model, to estimate Mixed-Effect Model using maximum likelihood and Restricted Maximum Likelihood (REML) estimation and identify the influential variables, to estimate the Mixed-Effect Model using Bootstrapped REML estimation and to identify the influential variables and to estimate Mixed-Effect Model using Bayesian MCMC inference and then to identify the influential variables.

2. RESEARCH METHODS

2.1 Mixed-Effects Model

In the context of LMMs, it is crucial to distinguish between fixed and random components and the influence they have on a dependent variable. For these subjects, we define distinct subsections. In the context of a typical ANOVA or ANCOVA model, the idea of a fixed factor is most frequently employed. The term "fixed factor" refers to a category or classification variable for which the researcher has included all levels (or conditions) that are relevant to the research. A classification variable having levels that are essentially randomly selected from the population of levels under study is called a random factor. In order to examine variation in the dependent variable across levels of the random factors and to generalize the data analysis

results to a larger population of levels of the random factor, random factors are taken into consideration in an analysis. Mixed-effect models are statistical models for continuous outcome variables in which the residuals are normally distributed but may not be independent or have constant variance [16]. Studies with clustered data, like students in classrooms, or experimental designs with random blocks, like batches of raw materials for an industrial process, can produce data sets that can be suitably examined using this model. studies that use longitudinal or repeated measures, where participants are measured frequently throughout time or under various circumstances. Mixed-Effect models is defined as the following [17] [18].

$$\begin{aligned} \boldsymbol{Y_i} &= \boldsymbol{X_i\beta} + \boldsymbol{Z_i u_i} + \boldsymbol{\varepsilon_i}, \\ \boldsymbol{u_i} &\sim \boldsymbol{N(0, D)}, \\ \boldsymbol{\varepsilon_i} &\sim \boldsymbol{N(0, R_i)}, \end{aligned} \tag{1}$$

$$Y_i = \begin{pmatrix} Y_{1i} \\ Y_{2i} \\ \vdots \\ Y_{ni} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, u_i = \begin{pmatrix} u_{1i} \\ u_{2i} \\ \vdots \\ u_{qi} \end{pmatrix}, Z = \begin{pmatrix} Z_{1i}^{(1)} & Z_{1i}^{(2)} & \dots & Z_{1i}^{(q)} \\ Z_{2i}^{(1)} & Z_{2i}^{(2)} & \dots & Z_{2i}^{(q)} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n_i i}^{(1)} & Z_{n_i i}^{(2)} & \cdots & Z_{n_i i}^{(q)} \end{pmatrix}, \varepsilon_i = \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \vdots \\ \varepsilon_{qi} \end{pmatrix}.$$

In Eq. (1), $\boldsymbol{Y_i}$ represents the i-th subject's vector of continuous replies. We display the, $\boldsymbol{Y_i}$ vector's components as follows, using the notation for a single observation. The first column would simply be equal to 1 for every observation in a model with an intercept term. It should be noted that every element in a column of the $\boldsymbol{X_i}$ matrix will be identical to a time-invariant (or subject-specific) covariate. We assume that the $\boldsymbol{X_i}$ matrices are full rank for presentational purposes, meaning that no column (or row) is a linear combination of the others. The fixed effects contained in the vector may have aliasing (or parameter identifiability) issues as a result of $\boldsymbol{X_i}$ matrices generally not being full rank $\boldsymbol{\beta}$ [19], [20];

$$D = Var(u_i) = \begin{pmatrix} Var(u_{1i}) & cov(u_{1i}, u_{2i}) & \dots & cov(u_{1i}, u_{2i}) \\ cov(u_{1i}, u_{2i}) & Var(u_{2i}) & \dots & cov(u_{2i}, u_{qi}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(u_{1i}, u_{qi}) & cov(u_{2i}, u_{qi}) & \cdots & Var(u_{qi}) \end{pmatrix},$$

$$R_i = Var(\varepsilon_i) = \begin{pmatrix} Var(\varepsilon_{1i}) & cov(\varepsilon_{1i}, \varepsilon_{2i}) & \dots & cov(\varepsilon_{1i}, \varepsilon_{2i}) \\ cov(\varepsilon_{1i}, \varepsilon_{2i}) & Var(\varepsilon_{2i}) & \dots & cov(\varepsilon_{2i}, \varepsilon_{qi}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\varepsilon_{1i}, \varepsilon_{n_i i}) & cov(\varepsilon_{2i}, \varepsilon_{n_i i}) & \cdots & Var(\varepsilon_{n_i i}) \end{pmatrix}.$$

The $\boldsymbol{\beta}$ in Eq. (1) is a vector of p unknown fixed-effect parameters or regression coefficients connected to the p variables that were used to build the $\boldsymbol{X_i}$ matrix. The $n_i \times q$ matrix Z_i in Eq. (1) a design matrix that represents the known values of the q covariates, $Z^{(1)}, \dots, Z^{(q)}$, for the i-th subject. This matrix is very much like the $\boldsymbol{X_i}$ matrix in that it represents the observed values of covariates; however, it is usually has fewer columns that the $\boldsymbol{X_i}$ matrix. The $\boldsymbol{u_i}$ vector for the i-th subject in Eq. (1) represents a vector of q random effects associated with the q covariates in the Z_i matrix. [21]

2.2 Restricted Maximum Likelihood (REML) and Boot-Strapped REML

The basic concept of realizing that the variance estimator provided by Maximum Likelihood (ML) is biased leads to the creation of Restricted Maximum Likelihood (REML). If one wishes to summarize the statistical observations, two parameters must be estimated: μ (mean) and σ^2(variance), assuming that the observations follow a normal distribution. It turns out that the Maximum Likelihood (ML) model's variance estimator is biased; that is, the number it produces either overestimates or underestimates the true variance. In practice, we rarely consider the bias in the variance estimator when we use machine learning to solve linear regression models because we are typically more interested in the linear model's coefficients, such as the mean, and frequently aren't even aware that we are also estimating another fitting parameter, the variance, in parallel. Consider a straightforward one-dimensional scenario with a variable to show that ML does, in fact, provide a biased variance estimator $y = (y_1, y_2, \dots, y_N)$ following e.g. the Normal distribution.

$$L(\hat{y},\hat{\sigma}^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} e^{-\frac{(y_i-\hat{y})^2}{2\hat{\sigma}^2}}, \quad \hat{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \,; \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2. \tag{2}$$

In practice, we rarely consider the bias in the variance estimator when we use machine learning to solve linear regression models because we are typically more interested in the linear model's coefficients, such as the mean, and frequently aren't even aware that we are also estimating another fitting parameter, the variance, in parallel. Consider a straightforward one-dimensional scenario with a variable to show that ML does, in fact, provide a biased variance estimator.

$$\log L(\hat{y},\hat{\sigma}^2) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_1-\mu)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_2-\mu)^2}{2\sigma^2} - \cdots - \frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_n-\mu)^2}{2\sigma^2}. \tag{3}$$

Thus, the Eq. (1)5) can be written

$$\log L(\hat{y},\hat{\sigma}^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n}\frac{(y_i-\mu)^2}{2\sigma^2}. \tag{4}$$

Maximize the log-likelihood function of the first derivative of the function $\log L(\mu,\sigma^2)$ on μ stated as follows.

$$\frac{\partial}{\partial\mu}\left(-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n}\frac{(y_i-\mu)^2}{2\sigma^2}\right) = \frac{\partial}{\partial\mu}\left(-\frac{n}{2}\log(2\pi)\right) - \frac{\partial}{\partial\mu}\left(\frac{n}{2}\log(\sigma^2) - \frac{\partial}{\partial\mu}\left(\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{2\sigma^2}\right)\right). \tag{5}$$

In Eq. (5) is equal to zero with the following.

$$\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{2\sigma^2} = 0\,; \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\hat{\mu} = 0; \sum_{i=1}^{n} y_i - n\hat{\mu} = 0\,. \tag{6}$$

Thus, it can be defined as

$$\hat{\mu} = \bar{y},$$

The log-likelihood function's derivative to σ^2

$$\frac{\partial}{\partial\sigma^2}\left(-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n}\frac{(y_i-\mu)^2}{2\sigma^2}\right) \tag{7}$$

$$= \frac{\partial}{\partial\sigma^2}\left(-\frac{n}{2}\log(2\pi)\right) - \frac{\partial}{\partial\sigma^2}\left(\frac{n}{2}log(\sigma^2)\right) - \frac{\partial}{\partial\sigma^2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{2\sigma^2}. \tag{8}$$

The equation above is equated to zero, then it will $2\hat{\sigma}^2$, substitute μ with the estimator $\bar{y}$, so

$$-\frac{n}{2\hat{\sigma}^2} + \sum_{i=1}^{n}\frac{(y_i-\hat{\mu})^2}{2\hat{\sigma}^4} = 0, \tag{9}$$

Thus, we obtain

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\hat{\sigma}^2}. \tag{10}$$

Thus, Maximum likelihood estimators for the population mean μ unbiased and population variance are the maximum likelihood estimators for the normal distribution σ^2 biased, therefore the log likelihood function does not need to be broken down into two parts. Restricted maximum likelihood is one way to develop likelihood. Variance parameters can be estimated using residual likelihood σ^2. This solution is known as the REML estimator for σ^2. For the normal distribution to develop this idea depends on the fact that

$$\sum_{i=1}^{n}(y_i - \mu)^2 = \sum_{i=1}^{n}[(y_i - \bar{y}) + (\bar{y} - \mu)]^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \mu)^2. \tag{11}$$

Look again at the log likelihood function in the equation, then rewrite the log likelihood function for the normal distribution.

$$\log L(\hat{y}, \hat{\sigma}^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n}\frac{(y_i - \mu)^2}{2\sigma^2}. \tag{12}$$

In the equation above, the likelihood function can be written as

$$\log L(\hat{y}, \hat{\sigma}^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n}\frac{(y_i - \bar{y})^2}{2\sigma^2} - \frac{n(\bar{y} - \mu)^2}{2\sigma^2}. \tag{13}$$

In order to ensure robustness against model misspecification, including asymmetry and long tails in the distribution of errors and random effects, bootstrap-based inference for mixed effects is developed and simultaneous intervals are included. The following observations are used to generate bootstrap:

$$y^* = X\hat{\beta} + Zu^* + e^*, \tag{14}$$

Where e^* and u^* are bootstrapped replica of random components in the model. The generation of e^* and u^* relies on the bootstrap framework, which will be covered in more detail. Define $\delta^* = \hat{\delta}, V^* = \hat{V}$, $G^* = \hat{G}$, the following definitons are used so Bootstrapped-REML are defined as

$$\theta_j^* = k_j^T\hat{\beta} + l_j^T u_j^*, \qquad \widehat{\theta_j^*} = \theta_j(\hat{\delta}^*) = k_j^T\hat{\beta}^* + l_j^T\hat{u}_j^*. \tag{15}$$

2.3 Bayesian Markov Chain Monte Carlo Inference (MCMC)

When using a Bayesian method, a probability distribution $\pi(y|\theta)$,known as likelihood, is given for the data that has been observed. $y = (y_1, \ldots, y_2)$ given an unknown parameter vector θ. rior distribution $\pi(\theta|\eta)$ is assigned to θ where η denotes vector of hyperparameters. The prior distribution for θ denotes the information about θ previous in determine of the data y. If η is not known, a fully Bayesian approach would specify a hyperprior distribution for η is used as if η were known. Defined that η is known, inference concerning θ is based on the posterior distribution of θ which the Bayes' Theorem defines as.

$$\pi(\theta|y) = \frac{\pi(y,\theta)}{\pi(y)} = \frac{\pi(y|\theta)\pi(\theta)}{\int \pi(y|\theta)\,\pi(\theta)\,\mathrm{d}\theta}. \tag{16}$$

Function of r $\pi(y) = \int \pi(y,\theta)\,\pi(\theta)\,\mathrm{d}\theta$ indicates the data's marginal likelihood y. This is devoid of θ and may be set to a scaling constant which does not impact the shape of the posterior distribution [22], [23] [24]. Consequently, the posterior distribution is frequently written as

$$\pi(\theta|y)\ \alpha\ \pi(y|\theta)\pi(\theta). \tag{17}$$

Bayesian methods [25], [26] allow for the incorporation of preexisting ideas into the model and offer a means of formalizing the process of updating the previous information by learning from the data. Bayesian approaches, as opposed to frequentist ones, offer believable intervals on parameters and probability values on hypotheses that make sense. Furthermore, complex models that are challenging to fit using traditional techniques, such repeated measures, missing data, and multivariate data, may be handled by Bayesian methods. One major challenge when using Bayesian techniques to calculate the posterior $\pi(\theta|y)$ the process

typically entails high-dimensional integration, which is typically unmanageable in closed form. As a result, the posterior distribution might not have closed-form expressions even while the likelihood and the previous distribution do. Markov Chain Monte Carlo (MCMC) generates a sample of values. $\{\theta^{(g)}, g = 1, \dots, G\}$ from a convergent Markov chain summaries of the $\theta^{(g)}$ values may be used to summarize the posterior distribution of the parameters of interest. For example, we might use the sample mean to estimate the posterior mean.

$$E(\widehat{\theta}_I|y) = \frac{1}{G}\sum_{i=1}^{G}\theta_i^{(g)}, \tag{18}$$

and the sample variance to estimate the variance.

$$Var(\widehat{\theta}_I|y) = \frac{1}{G-1}\sum_{i=1}^{G}\left(\theta_i^{(g)} - E(\widehat{\theta}_I|y)\right)^2. \tag{19}$$

to determine whether the sample chains have reached the stationary distribution, or the posterior distribution, MCMC methods necessitate the use of diagnostics. Examining the trace plot, which is a plot of the parameter value at each iteration versus the number of iterations, is a simple method to determine whether the chain has converged. It shows how well the chain is mixing or moving about the parameter space.

3. RESULTS AND DISCUSSION

3.1 Gapminder Data

This study uses secondary data, namely Gapminder raw data. This data can be obtained through the link https://www.gapminder.org/data/. The Swedish nonprofit Gapminder is independent and unaffiliated with any political, religious, or commercial groups. Gapminder dispels harmful myths and advances an understandable, fact-based worldview. The unit of analysis of this study is life expectancy in the population in each continent. The response variable used in this study is life expectancy at birth, in years. While the independent variables used in this study are countries consisting of 142, continents consisting of 5, year, population, and GDP per capita (US$, inflation-adjusted). The study's independent and dependent variables are displayed in Table 1.

Table 1. Description of Variables

Variable	Description	Type	Effects
Y	life expectancy at birth, in years	Numeric	Fixed Effects
X_1	Country	Factor	Fixed Effects
X_2	Continent	Factor	Fixed Effects
X_3	Year	Integer	Random Effects
X_4	Population	Integer	Fixed Effects
X_5	GDP per capita (US$, inflation-adjusted)	Numeric	Fixed Effects

***Data source**: https://www.gapminder.org/data/*

3.2 Analysis Method

Method analysis used descriptive analysis and inference. The inference method used is restricted maximum likelihood (REML), Furthermore, there is a development of the REML method with a machine learning approach, namely Boot-Strapping REML. This study not only uses a frequentist approach but this study also uses a Bayesian approach. The Bayesian inference used is Bayesian MCMC. The software used in this study is RStudio. The package used for the REML estimation method is lme4 with the lmer function with its default configuration. The *bootMer* function for parametric bootstrapping with 2000 simulations and the lme4 package are utilized for the Boot-Strapping REML estimate approach. To use four CPUs to simulate spherical random effects, set the u option to TRUE. The package for Bayesian MCMC inference is STAN v2.32.2 with the rstanarm function.

The steps taken in this research are as follows.

1. Collect data from the Gapminder website to obtain life expectancy data in various countries and continents.
2. Determine the response variables and predictor variables using the data in Table 1. The predictor variables used in this study consist of fixed and random effects variables presented in Table 1.
3. The model used in this study is the Mixed-Effect Model with frequentist and Bayesian estimation.
4. Conduct descriptive analysis of Gapminder data with Histogram and Boxplot.
5. Conducting life expectancy data modeling with a Mixed-Effect Model to determine the predictor variables that have a significant effect on life expectancy. The significance level used in this study is 5%. The steps in building the model as follows

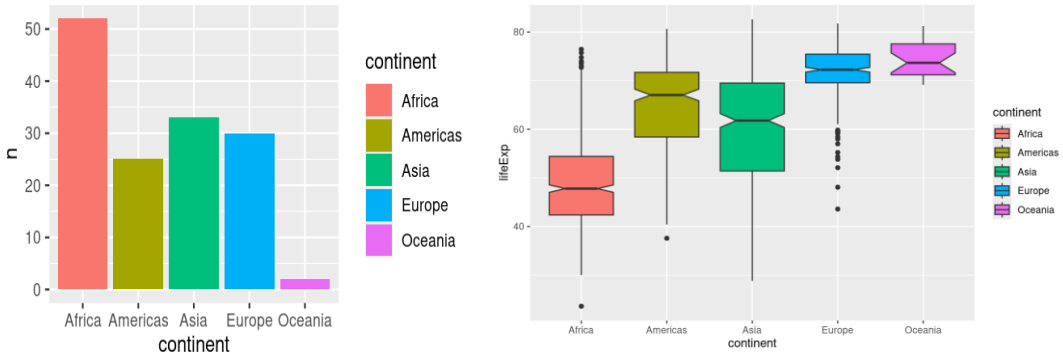
   a. Conducting error normality test using Kolmogorov-Smirnov test. error normality hypothesis test is the null hypothesis (H_0) claims that the alternative hypothesis and the error data are regularly distributed.(H_1) is the error data does not normally distributed.
   b. Conduct a partial test using the *t-test statistic* to determine which variables have a significant influence on life expectancy. The test examines the null hypothesis that a specific regression coefficient equals zero.

$$H_0: \beta_j = 0,$$
$$H_1: \beta_j \neq 0.$$

6. After modeling the life expectancy data, parameter estimation was carried out using Restricted Maximum Likelihood (REML), BootStrapped REML, and Bayesian MCMC.

3.3 Overview of Life Expectancy in Different Continents

Based on the data in Gapminder Data obtained in Table 1, each country was surveyed 12 times every 5 years and in the data the way the numbers are distributed of surveys for each country was the same. The distribution of the number of countries on each continent can be done with a bar graph presented in Fig. 1 below. It can be seen in Fig. 1 that the distribution of the number of countries surveyed or recorded is not balanced on each continent where the African continent has more than 50 countries surveyed, then the Asian continent is in second place with more than 30 countries, then followed by the Americas Continent, the European Continent, and the Oceania Continent. Based on Fig. 1, it can be seen that the Boxplot above displays the distribution of life expectancy by continent. Each boxplot provides information about the distribution of data such as median, quartile and outlier values. The African continent has the lowest median life expectancy among all continents. The distribution of data is quite wide, which shows large variations between countries.



(a) (b)

Figure 1. (a) Bar chart of the Number of Continent, (b) Boxplot of Life Expectancy by Continent

It can be seen in the African continent some low outliers, namely countries with very low life expectancies around 30 and above. Life expectancy generally ranges from 40 to 55 years. The Americas continent has a higher median life expectancy than the African and Asian continents. The Americas continent has a narrower distribution than Africa which indicates less variability between countries. There is one low outlier around 38 years. Most countries in the Americas continent have life expectancies between 60 and 75 years. The Asian continent has a lower median life expectancy than the European and Oceanian

continents, but higher than the African continent. The variation in the data in the Asian continent is large with some low outliers. The life expectancy in the Asian continent is widespread around 40 to 80 years.

The European continent has a high median life expectancy with a value close to the life expectancy of the Oceanian continent. Small variations in the European continent indicate that countries in Europe have relatively uniform life expectancies. Some low outliers such as countries with life expectancy values lower than the European average, this is likely from Eastern Europe. The European continent is generally in the range of 70 to 80 years. The Oceanian continent has the highest median life expectancy among all continents. It can be seen that there are almost no outliers and the boxplot is very narrow because it only consists of two countries, namely Australia and New Zealand. The life expectancy of the Oceanian continent is very stable at over 75 years. Thus, the Oceanian continent and the European continent have the highest and most stable life expectancy levels. The African continent has the lowest life expectancy with high variation and many lagging countries. The Americas and Asia are in the middle, with the Americas slightly ahead of Asia in median and stability values European continent has a high median life expectancy with a value close to the life expectancy of

Figure 2. GDP Per Capita based on Life Expectancy using Rstudio Software

Based on Fig. 2 above, the scatter plot illustrates the relationship between GDP per capita and life expectancy. It can be observed that across all years, there is a positive trend, indicating that countries with higher GDP per capita tend to have higher life expectancy. The large black bubbles represent population size — the larger the bubble, the larger the population of the country. For example, China and India have large bubbles, suggesting that changes in these countries have a global impact. Colors represent the distribution of countries across continents. It is evident that Europe and Oceania have many countries with high GDP and high life expectancy. Asia and the Americas have a moderate range, including both wealthy and developing countries. African countries are mostly represented in the bottom-left area, reflecting low income and low life expectancy. Looking at the panel year by year, we can see that the cluster of bubbles on the right has grown from 1952 to 2007. Africa has shown gradual progress but remains significantly behind other continents. Asia has shifted upward and to the right. When GDP per capita reaches around 10,000 to 20,000, the increase in life expectancy begins to slow down. Thus, while GDP per capita and life expectancy are strongly correlated, their relationship is not linear. Over time, many countries have moved toward the upper-right quadrant, indicating an overall improvement in social and economic indicators. Therefore, economic and demographic factors significantly influence life expectancy.

3.4 Mixed-Effects Models with Restricted Maximum Likelihood

In Mixed Effects model, normality test on the error terms is necessary to determine whether they follow a normal distribution. The normalcy test's findings are presented in Table 2 below. The normality test was conducted using the Shapiro-Wilk test, which shows that the Mixed Effects model with both Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML) estimation methods follows a normal distribution. Furthermore, parameter estimation in Mixed Effects Models can be carried out using Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML). The parameter estimation results for this model, obtained using RStudio software, are presented in Table 2. The random effect used in this model is the year, indicating that there is variation across years that affects life expectancy in each continent. As we can see Table 2, that the variables significantly influencing life expectancy in each country under both ML and REML estimation methods are GDP per Capita, and the continents of Americas, Asia, Europe, and Oceania. However, the population variable does not significantly affect life expectancy in each country. Countries in the Oceania continent have the highest estimated life expectancy compared to the Americas, Asia, and Europe. This indicates that people in Oceania tend to live longer than those in other continents.

Table 2. Estimation Parameter of Mixed-Effect Model with ML and REML

	Maximum Likelihood (ML)		**Restricted Maximum Likelihood (REML)**	
Fixed Effects:				
Variables	**Coefficient (Std. Error)**	**t-value**	**Coefficient (Std. Error)**	**t-value**
(Intercept)	48.18204 (1.4713703)	32.74637*	48.18246 (1.5324175)	31.44212*
GDP Per Capita	0.00030 (0.0000199)	15.22125*	0.00030 (0.0000199)	15.21650*
Population	0.00000 (0.0000000)	1.18191	0.00000 (0.0000000)	1.17896
Continent Americas	14.26886 (0.4910482)	29.05797*	14.26977 (0.4909103)	29.06798*
Continent Asia	9.34234 (0.4685091)	19.94057*	9.34365 (0.4683787)	19.94893*
Continent Europe	19.30730 (0.5146027)	37.51885*	19.30940 (0.5144621)	37.53319*
Continent Oceania	20.48845 (1.4586392)	14.04627*	20.49120 (1.4582297)	14.05211*
Random Effects:				
Groups	**Standard Deviation**		**Standard Deviation**	
Area	4.995075		5.220602	
Residual	6.820087		6.832206	
Normality Test:	*p-value*		*p-value*	
Shapiro-Wilk	4.518e-07		4.636e-07	

*) Significant at $\alpha = 5\%$

The parameter estimation results using 1.000 simulations with bootstrapped REML are presented in Table 3 below. In Table 3, each row shows the estimated fixed effects along with the 95% confidence intervals from 1,000 bootstrap samples. It can be seen that the intercept lies within a confidence interval between 47.63 and 48.71. This means that the average life expectancy is within this range when all predictors are set to zero. The GDP per Capita variable shows a significant positive effect, as its confidence interval does not cross zero. This suggests that life expectancy is generally positively correlated with GDP per capita. For the population variable, the confidence interval crosses zero, suggesting that it does not have a significant effect on life expectancy. The Americas show higher life expectancy compared to Asia based on the confidence interval in Table 3. Europe has a significantly higher life expectancy than both the Americas and Asia. Oceania shows the highest life expectancy compared to the Americas, Asia, and Europe. Thus, it can be concluded that all continents—except for Africa—exhibit significantly higher life expectancy.

Table 3. Estimation Parameter of Mixed-Effect Model with bootstrapped REML

Variable	**2.5%**	**97.5%**
Intercept	47.627	48.714
GDP Per Capita	0.000	0.000
Population	-0.000	0.000

Variable	2.5%	97.5%
Continent Americas	13.331	15.270
Continent Asia	8.405	1.021
Continent Europe	18.268	20.028
Continent Oceania	17.655	23.346

3.5 Mixed-Effects Models with Bayesian Inference

Mixed Effects Model using the Bayesian inference can be implemented through Bayesian MCMC. In the Bayesian framework, we use the *rstanarm* package in RStudio. In the mixed-effects model, we nested GDP per Capita, Population, and Country. The Bayesian-adjusted mixed-effects model is used to predict life expectancy based on the year. The first step in Bayesian modeling is specifying the likelihood and prior distributions. Before performing mixed-effects modeling using the Bayesian approach, we need to check for convergence across several model chains. Convergence can be assessed in two ways: numerically and visually. Based on Table 4, it can be concluded that the MCSE values for all parameters are close to zero (average ≈ 0.0), indicating that the uncertainty due to the Monte Carlo simulation is very low.

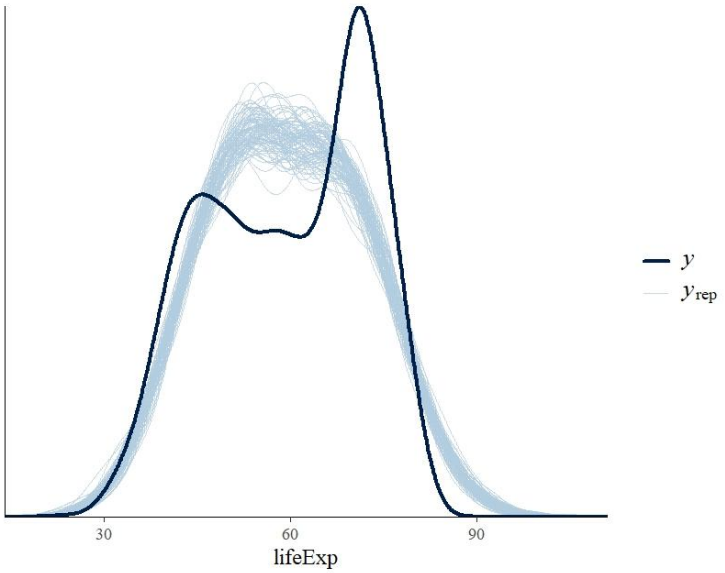
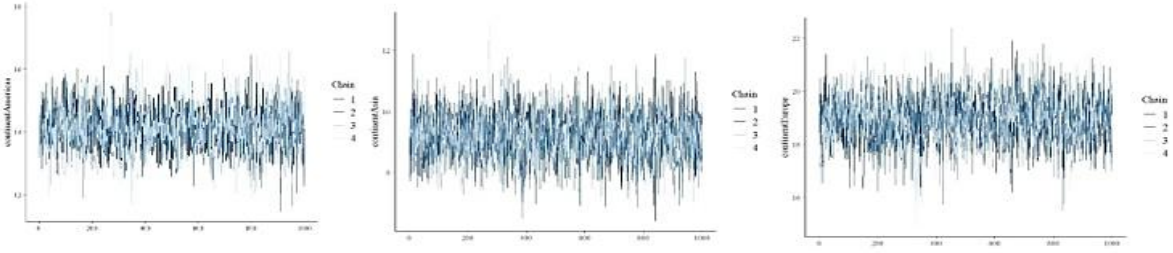
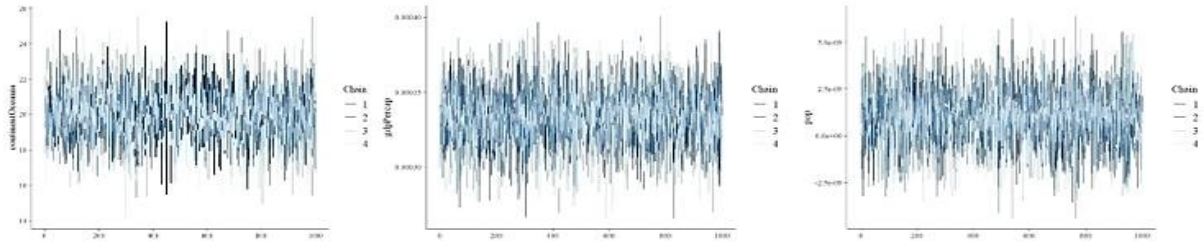


Figure 3. Plot Posterior Predictive Check the Observed Distribution

Based on Fig. 3, it can be shown that the results of the posterior predictive distribution (thin light blue line) and the observed data (thick dark blue line) do not have similar distributions. However, the replication model is generally quite good with most of the simulation lines (y_rep) following the original distribution pattern (y), indicating that the model is quite successful in replicating the general shape of the distribution. There are similar shapes and peaks in the lifeExp range of around 55-75. In Fig. 3 it can be seen the trace plots also exhibit good mixing, as indicated by random fluctuations without signs of high autocorrelation, which results in a large effective sample size (n_eff) and efficient parameter estimates. Moreover, no stuck chains or extreme values deviating substantially from the posterior mean were observed, indicating that the sampling process proceeded stably without numerical issues. Overall, the trace plots provide visual validation of sampling quality and support the conclusion that the posterior estimates can be used with high confidence for subsequent inference and prediction.



(a) (b) (c)



(d) (e) (f)

Figure 4. Trace Plots for Monitoring Model Parameter Convergence
(a) Continent Americas, (b) Continent Asia, (c) Continent Europe, (d) Continent Oceania,
(e) GDP Per Capita, (f) Population (pop)

This is important because it suggests that the posterior estimates are stable across simulation replications and are not substantially influenced by random sampling noise. All parameters exhibit an $\hat{R}$ value of 1.0, indicating that the Markov chains have reached convergence. In this context, an $\hat{R}$ value close to 1 implies that the between-chain variance is nearly equal to the within-chain variance, supporting the reliability of the posterior estimates. The effective sample size (n_eff) for all parameters ranges from 542 to 6,121, which is considered high and indicates that the posterior distribution is adequately sampled to produce precise estimates. MCMC diagnostic evaluation was conducted to assess the quality of the posterior estimates for the parameters in the Bayesian logistic regression model. The trace plot for each predictor variable is presented in the following Figure 4.

Table 4. Posterior Distributions of Model Parameters and MCMC Diagnostic for Posterior Quality Assesment

Variable	Mean	SD	10%	50%	90%	MCSE	$\hat{R}$	n_eff
(Intercept)	48.1	1.2	46.7	48.1	49.6	0.1	1.0	542
GDP Per Capita	0.0	0.0	0.0	0.0	0.0	0.0	1.0	5453
Population	0.0	0.0	0.0	0.0	0.0	0.0	1.0	6121
Continent Americas	14.1	0.7	13.3	14.1	15.0	0.0	1.0	2397
Continent Asia	9.2	0.8	8.2	9.2	10.2	0.0	1.0	1670
Continent Europe	18.9	0.8	17.9	19.0	20.0	0.0	1.0	1011
Continent Oceania	20.0	1.5	18.0	20.0	22.0	0.0	1.0	5244
Mean_PPD	59.5	0.2	59.2	59.5	59.8	0.0	1.0	3959
Log_posterior						0.2	1.0	1225

Three key diagnostic indicators were examined: Monte Carlo Standard Error (MCSE), Gelman–Rubin convergence diagnostic ($\hat{R}$), and Effective Sample Size (n_eff). The results are presented in Table 4. Before assessing the results of the estimation, we assess sampling quality. Mixed-Effect model uses *rstanarm* default settings for MCMC and runs without warnings. For numerical checks of sampling quality, we describe the summary function. The parameter estimation results using the Bayesian approach are presented in Table 4 above. The log-posterior parameter has an MCSE of 0.2, and $\hat{R}$ of 1.0, and an effective sample size (n_eff) of 1,225, indicating that the posterior target function has also achieved stable estimation. Overall, these results suggest that the constructed Bayesian model satisfies key convergence diagnostics, thereby supporting its interpretability and further use for decision-making purposes. The continent variable has a significant effect on life expectancy.

4. CONCLUSION

Mixed-Effect Model in this study uses several inferences. Frequentist inference is done using Restricted Maximum Likelihood and Bootstrapped REML, and Bayesian Inference. Although these three methods use different stages, the estimation results of this method are the same. Based on the results of the REML estimation, it can be concluded that the population does not have a significant effect on the number bootstrapped REML life expectancy. estimation results semiparametric of likelihood which uses a certain confidence interval. From the results of the REML bootstrapped estimation, it can be concluded that the population also has no effect on life expectancy. Furthermore, Bayesian inference with the MCMC approach

can be concluded that all predictor variables have converged as seen in the Trace plot. The results of the Bayesian MCMC estimation show that the variables that have a significant effect on life expectancy are GDP Per Capita and continents. Therefore, all of these models it can be concluded that life expectancy captures random effects as years are influenced by GDP Per Capita, Country, and Continent. However, this model still uses simple predictor variables so that for further research it is recommended to use more complex predictor variables.

Author Contributions

Asysta Amalia Pasaribu: Conceptualization, Methodology, Writing-Original Draft, Software, Validation. Kusman Sadik: Data Curation, Resources, Draft Preparation. Anang Kurnia: Formal Analysis, Validation. All authors discussed the results and contributed to the final manuscript.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors

Acknowledgment

The authors express their sincere gratitude to the Statistics and Data Science Program of the School of Data Science, Mathematics, and Informatics at IPB University, and to Bina Nusantara University for their valuable support in this work.

Declarations

The authors declare no competing interest.

Declaration of Generative AI and AI-assisted Technologies

AI-assisted technology (ChatGPT) was used to support sentence restructuring and clarity improvements. The authors confirm that the underlying ideas, arguments, data analyses, and conclusions are original and were not generated by AI. All AI-assisted edits were critically reviewed and validated by the authors

REFERENCES

[1] S. W. Raudenbush, *HLM 6: HIERARCHICAL LINEAR AND NONLINEAR MODELING*. SCIENTIFIC SOFTWARE INTERNATIONAL, 2004.

[2] K. Lee, "ASSOCIATIONS BETWEEN CUMULATIVE POVERTY AND CHILDREN'S ACADEMIC OUTCOMES: ABSOLUTE VERSUS RELATIVE POVERTY," *The British Journal of Social Work*, p. bcaf001, 2025.doi : https://doi.org/10.1093/bjsw/bcaf001

[3] A. M. Eyasu, T. Zewotir, and Z. G. Dessie, "IMPACT OF CROP COMMERCIALIZATION ON MULTIDIMENSIONAL POVERTY IN RURAL ETHIOPIA: PROPENSITY SCORE APPROACH," *Frontiers in Public Health*, vol. 12, p. 1412670, 2025.doi : https://doi.org/10.3389/fpubh.2024.1412670

[4] N. Diz-Rosales, M. J. Lombardía, and D. Morales, "POVERTY MAPPING UNDER AREA-LEVEL RANDOM REGRESSION COEFFICIENT POISSON MODELS," *Journal of Survey Statistics and Methodology*, vol. 12, no. 2, pp. 404–434, 2024. doi : https://doi.org/10.1093/jssam/smad036

[5] A. E. N. Mouteyica and N. Ngepah, "EXPLORING HEALTH OUTCOME DISPARITIES IN AFRICAN REGIONAL ECONOMICS COMMUNITIES: A MULTILEVEL LINEAR MIXED-EFFECT ANALYSIS," *BMC Public Health*, vol. 25, no. 1, p. 175, 2025.doi : https://doi.org/10.1186/s12889-025-21306-5

[6] Q. Liu *et al.*, "DEVELOPMENT AND VALIDATION OF PREDICTION MODELS FOR INCIDENT REVERSIBLE COGNITIVE FRAILTY BASED ON SOCIAL-ECOLOGICAL PREDICTORS USING GENERALIZED LINEAR MIXED MODEL AND MACHINE LEARNING ALGORITHMS: A PROSPECTIVE COHORT STUDY," *Journal of Applied Gerontology*, vol. 44, no. 2, pp. 255–266, 2025.doi : https://doi.org/10.1177/07334648241270052

[7] M. Scandola and E. Tidoni, "RELIABILITY AND FEASIBILITY OF LINEAR MIXED MODELS IN FULLY CROSSED EXPERIMENTAL DESIGNS," *Advances in Methods and Practices in Psychological Science*, vol. 7, no. 1, p. 25152459231214454, 2024.doi : https://doi.org/10.1177/25152459231214454

[8] S. Smout, K. Champion, S. O'Dean, M. Teesson, L. Gardner, and N. Newton, "ANXIETY, DEPRESSION AND DISTRESS OUTCOMES FROM THE HEALTH4LIFE INTERVENTION FOR ADOLESCENT MENTAL HEALTH: A CLUSTER-RANDOMIZED CONTROLLED TRIAL," *Nature Mental Health*, vol. 2, no. 7, pp. 818–827, 2024.doi : https://doi.org/10.1038/s44220-024-00246-w

[9] J. Kanyama Busanga and P. Njuho, "LINEAR MIXED MODEL EFFECTS IN META ANALYSIS OF AGRICULTURAL DATA.," *International Journal of Agricultural & Statistical Sciences*, vol. 20, no. 2, 2024.doi : https://doi.org/10.59467/IJASS.2024.20.511
[10] J. Caviedes, J. T. Ibarra, L. Calvet-Mir, S. Álvarez-Fernández, and A. B. Junqueira, "INDIGENOUS AND LOCAL KNOWLEDGE ON SOCIAL-ECOLOGICAL CHANGES IS POSITIVELY ASSOCIATED WITH LIVELIHOOD RESILIENCE IN A GLOBALLY IMPORTANT AGRICULTURAL HERITAGE SYSTEM," *Agricultural Systems*, vol. 216, p. 103885, 2024. doi : https://doi.org/10.1016/j.agsy.2024.103885
[11] J. C. Coltherd *et al.*, "HEALTHY CATS TOLERATE LONG-TERM DAILY FEEDING OF CANNABIDIOL," *Frontiers in veterinary science*, vol. 10, p. 1324622, 2024.doi : https://doi.org/10.3389/fvets.2023.1324622
[12] F. Massa, M. Scavino, and G. Muniz-Terrera, "A BAYESIAN NON-LINEAR MIXED-EFFECTS MODEL FOR ACCURATE DETECTION OF THE ONSET OF COGNITIVE DECLINE IN LONGITUDINAL AGING STUDIES," *arXiv preprint arXiv:2502.08418*, 2025.doi : https://doi.org/10.3390/stats8030074
[13] K. Zhong, F. L. Schumacher, L. M. Castro, and V. H. Lachos, "BAYESIAN ANALYSIS OF CENSORED LINEAR MIXED-EFFECTS MODELS FOR HEAVY-TAILED IRREGULARLY OBSERVED REPEATED MEASURES," *Statistics in Medicine*, vol. 44, no. 3–4, p. e10295, 2025.doi : https://doi.org/10.1002/sim.10295
[14] P.-F. Wang, L.-H. Dong, L.-F. Xie, and Z. Miao, "CONSTRUCTION OF BIOMASS MODELS FOR LARIX OLGENSIS PLANTATION USING HIERARCHICAL BAYESIAN SEEMINGLY UNRELA-TED REGRESSION," *Ying yong sheng tai xue bao= The journal of applied ecology*, vol. 36, no. 5, pp. 1298–1308, 2025.
[15] E. Smenderovac *et al.*, "MIXED MODEL APPROACHES CAN LEVERAGE DATABASE INFORMATION TO IMPROVE THE ESTIMATION OF SIZE-ADJUSTED CONTAMINANT CONCENTRATIONS IN FISH POPULATIONS," *Environmental Science & Technology*, vol. 59, no. 10, pp. 4797–4806, 2025.doi : https://doi.org/10.1021/acs.est.4c10303
[16] N. Laird, N. Lange, and D. Stram, "MAXIMUM LIKELIHOOD COMPUTATIONS WITH REPEATED MEASURES: APPLICATION OF THE EM ALGORITHM," *Journal of the American Statistical Association*, vol. 82, no. 397, pp. 97–105, 1987.doi : https://doi.org/10.1080/01621459.1987.10478395
[17] R. I. Jennrich and M. D. Schluchter, "UNBALANCED REPEATED-MEASURES MODELS WITH STRUCTURED COVARIANCE MATRICES," *Biometrics*, pp. 805–820, 1986. doi : https://doi.org/10.2307/2530695
[18] T. Asparouhov, "GENERAL MULTI-LEVEL MODELING WITH SAMPLING WEIGHTS," *Communications in Statistics—Theory and Methods*, vol. 35, no. 3, pp. 439–460, 2006. doi : https://doi.org/10.1080/03610920500476598
[19] G. Molenberghs and G. Verbeke, *Models for discrete longitudinal data*, vol. 22. Springer, 2005.
[20] A. Veiga, P. W. Smith, and J. J. Brown, "THE USE OF SAMPLE WEIGHTS IN MULTIVARIATE MULTILEVEL MODELS WITH AN APPLICATION TO INCOME DATA COLLECTED BY USING A ROTATING PANEL SURVEY," *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 63, no. 1, pp. 65–84, 2014.doi : https://doi.org/10.1111/rssc.12020
[21] R. Steele, "MODEL SELECTION FOR MULTILEVEL MODELS," *The SAGE handbook of multilevel modeling*, pp. 109–125, 2013.doi : https://doi.org/10.4135/9781446247600.n7
[22] K.-R. Koch and K.-R. Koch, "BAYES' THEOREM," *Bayesian Inference with Geodetic Applications*, pp. 4–8, 1990.doi : https://doi.org/10.1007/BFb0048702
[23] B. Efron, "BAYES' theorem in the 21st century," *Science*, vol. 340, no. 6137, pp. 1177–1178, 2013.doi : https://doi.org/10.1126/science.1236536
[24] A. Ebert, K. Mengersen, F. Ruggeri, and P. Wu, "CURVE REGISTRATION OF FUNCTIONAL DATA FOR APPROXIMATE BAYESIAN COMPUTATION," *Stats*, vol. 4, no. 3, pp. 762–775, 2021.doi : https://doi.org/10.3390/stats4030045
[25] S. S. Qian, C. A. Stow, and M. E. Borsuk, "ON MONTE CARLO METHODS FOR BAYESIAN INFERENCE," *Ecological modelling*, vol. 159, no. 2–3, pp. 269–277, 2003.doi : https://doi.org/10.1016/S0304-3800(02)00299-5
[26] C. M. Crainiceanu and D. Ruppert, "LIKELIHOOD RATIO TESTS IN LINEAR MIXED MODELS WITH ONE VARIANCE COMPONENT," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 66, no. 1, pp. 165–185, 2004.doi : https://doi.org/10.1111/j.1467-9868.2004.00438.x