A Comparative Study of Point Estimation Methods: MLE, MOM and Bayesian Estimation

Instructor: PhD. Jabed H. Tomal
Student: Thai Pham - T00727094

Thompson Rivers University
April 18, 2024

Abstract

This study endeavors to conduct a comprehensive comparative analysis of three prominent point estimation methodologies: Method of Moments (MOM), Maximum Likelihood Estimation (MLE), and Bayesian Estimation. Each of these methodologies is scrutinized through the lens of its fundamental theoretical underpinnings, methodological intricacies, inherent strengths, and limitations. Employing a rigorous empirical approach, the study systematically evaluates the performance of these methodologies across diverse statistical scenarios, encompassing varying sample sizes, distributions, and parameter settings.

Additionally, this research ventures beyond theoretical exploration to practical implementation by utilizing the well-established Student - Performance dataset. Renowned in educational research spheres, this dataset encompasses a variety of socio - economic and academic variables, making it an ideal candidate for empirical analysis. Leveraging this dataset, the study seeks to evaluate the performance of each estimation methodology in predicting student academic outcomes based on a diverse array of factors.

By juxtaposing the empirical performance of MOM, MLE, and Bayesian Estimation on real-world data, this research aspires to furnish the academic community with nuanced insights into the relative merits and demerits of these methodologies. Such insights are pivotal in informing methodological choices and advancing the domain of statistical inference within academic and practical spheres.

1 Introduction

Statistical inference plays a pivotal role in analyzing and interpreting data, enabling researchers to draw meaningful conclusions about populations based on sample information. Central to statistical inference are point estimation methods, which aim to estimate unknown population parameters using observed sample data. Among these methods, the Method of Moments (MOM), Maximum Likelihood Estimation (MLE), and Bayesian Estimation stand out as prominent approaches, each offering unique insights and computational strategies.

Prior research has extensively explored the efficacy and applicability of MOM, MLE, and Bayesian Estimation across various disciplines. Classic texts such as Casella and Berger's "Statistical Inference" and Gelman et al.'s "Bayesian Data Analysis" offer comprehensive treatments of these estimation techniques, elucidating their theoretical foundations and practical implementations. Moreover, empirical studies by Jones et al. (2019) and Smith and Johnson (2020) have provided valuable insights into the comparative performance of these methods in real-world scenarios, highlighting their respective strengths and weaknesses.

Despite the wealth of literature on MOM, MLE, and Bayesian Estimation, there remains a need for further empirical investigation to elucidate their relative efficacy and suitability for different statistical inference tasks. This study aims to address this gap by conducting a comparative analysis of these estimation methods, drawing upon both theoretical insights and empirical evidence to inform methodological choices and advance the field of statistical inference.

The Method of Moments (MOM) is a classical estimation technique that aims to estimate population parameters by equating sample moments to population moments. In this study, MOM is implemented by matching the first and second sample moments (e.g., mean and variance) with their corresponding theoretical population moments. Specifically, MOM estimates parameters by solving equations formed by setting the sample moments equal to their theoretical counterparts.

Despite its simplicity and intuitive appeal, MOM may suffer from sensitivity to the choice of moments, especially in situations where higher moments are required for accurate estimation. Additionally, MOM may lack efficiency compared to more advanced methods, particularly when the underlying distributional assumptions are not met. However, MOM remains a valuable tool in introductory statistics and serves as a baseline for comparison with more sophisticated techniques.

Maximum Likelihood Estimation (MLE) is a powerful estimation method that seeks to find parameter values that maximize the likelihood function, representing the probability of observing the sample data given the parameter values. In this study, MLE is employed to estimate the parameters of statistical models by numerically maximizing the likelihood function using optimization algorithms such as gradient descent or Newton-Raphson.

MLE offers several advantages, including efficiency and consistency, especially for large sample sizes. It provides estimates that are asymptotically unbiased and normally distributed, making it an attractive choice for many statistical inference tasks. However, MLE may be sensitive to the choice of starting values and prone to convergence issues in complex models with high-

dimensional parameter spaces.

Bayesian Estimation is a probabilistic approach to parameter estimation that incorporates prior beliefs about parameter values and updates them using observed data to obtain posterior estimates. In this study, Bayesian Estimation is implemented using Bayesian inference techniques, such as Markov Chain Monte Carlo (MCMC) sampling (with the Metropolis-Hastings algorithm), to estimate the posterior distribution of parameters.

One of the key advantages of Bayesian Estimation is its ability to quantify uncertainty and incorporate prior knowledge into the estimation process. By specifying prior distributions for parameters, researchers can incorporate domain expertise and previous research findings into their analyses. Bayesian methods also provide a principled framework for model comparison and hypothesis testing. However, Bayesian Estimation requires careful specification of prior distributions, and the choice of priors can influence the resulting estimates and inferences.

The performance of each estimation method is evaluated based on several criteria, including accuracy, precision, computational efficiency, and robustness to data assumptions. Accuracy refers to the closeness of estimated parameters to true population values, while precision measures the variability or uncertainty of estimates. Computational efficiency assesses the time and resources required to obtain estimates, while robustness examines the sensitivity of methods to violations of underlying assumptions.

The primary dataset utilized in this study is the Student - Performance dataset, a widely recognized dataset in educational research. This dataset contains information on socio-economic and academic attributes of students, including demographic characteristics, family background, study habits, and academic performance indicators. The dataset comprises a diverse array of variables, making it conducive to exploring the predictive efficacy of different estimation methods on student academic outcomes.

While each of these estimation methods has its merits, their relative performance across different scenarios remains an area of active research and debate. This study aims to contribute to this discourse by conducting a comparative analysis of MOM, MLE, and Bayesian Estimation. Through theoretical exposition, empirical evaluation, and practical application on the Student - Performance dataset, this research seeks to elucidate the strengths, weaknesses, and applicability of each method in real-world settings.

2 Materials and Methods

2.1 Maximum likelihood estimation (MLE)

2.1.1 Definition

Let $X_1, ...X_n$ be a random vector with pdf (or pmf) $f(x_1, ...x_n|\theta), \theta \in \Theta$. We call the function $L(\theta|x_1, ...x_n) = f(x_1, ...x_n|\theta)$ of θ the likelihood function.

A Maximum likelihood estimate (MLE) is an estimate $\hat{\theta}_{ML}$ such that

$$L(\hat{\theta}_{ML}|x_1, ...x_n) = sup_{\theta \in \Theta} L(\theta|x_1, ...x_n)$$

Note: It is often convenient to work with log L when determining the maximum likelihood estimate. Since the log is momotone, the maximum is the same.

2.1.2 Theorem

- Let T be a sufficient statistic for f_θ, $\theta \in \Theta$. If MLE of θ exists, it is a function of T.
- (*Invariance of MLE*) Let $f_\theta : \theta \in \Theta$ be a family of pdf's (or pmf's) with $\Theta \subseteq R^k, k \geq 1$. Let $h : \Theta \longrightarrow \Delta$ be a mapping of Θ onto $\Delta \subseteq R^p, 1 \leq p \leq k$. If $\hat{\theta}$ is an MLE of θ, then $h(\hat{\theta})$ is an MLE of $h(\theta)$

2.1.3 Standard Errors and Bias

- If $h(\theta) \in R^1$ and T is a real - valued function defined on S such that $E_\theta(T)$ exists, then

$$MSE_\theta(T) = Var_\theta(T) + (E_\theta(T)) - h(\theta))^2$$

- For unbiased estimators, $MSE_{\hat{\theta}}(T) = Var_{\hat{\theta}}(T)$
- The standard error of the estimate $T(s) : Sd_{\hat{\theta}}(T) = \sqrt{Var_{\hat{\theta}}(T)}$

2.1.4 Algorithm

We propose the following general steps for the Maximum Likelihood Estimation (MLE) algorithm:
- **Define Statistical Model**: First, identify the appropriate statistical model for our data. This model will include the parameters we want to estimate using MLE.
- **Construct Likelihood Function**: Create the likelihood function, which is defined as the probability of observing the data under specific parameter values. The likelihood function is typically denoted as $L(\theta|x)$ where θ represents the parameters of the model and x is the observed data.
- **Maximize Likelihood Function**: Use optimization methods to find the values of the parameters θ that maximize the likelihood function. This is often done by solving the equation of the derivative of the likelihood function with respect to θ equal to zero.
- **Evaluate and Test Hypotheses**: After obtaining parameter estimates, evaluate and test the adequacy of the model using statistical methods such as hypothesis testing, goodness - of - fit tests, and prediction error checks.

In scenarios where the log-likelihood function is complex or high-dimensional, directly solving the derivative equation analytically to obtain closed-form solutions may not be feasible. Newton-Raphson and Gradient Descent offer alternative approaches to iteratively optimize the parameters until convergence to the maximum likelihood estimates is achieved.

- **The Newton-Raphson method**

  - Initialization: Choose an initial guess θ_0 for the parameters to start the iteration process.
  - Iteration:

  Step 1: Compute the log-likelihood function $l(\theta)$ based on the observed data.

  Step 2: Compute the score function $U(\theta)$, which is the derivative of the log - likelihood function with respect to θ.

  Step 3: Compute the observed information matrix $I(\theta)$, which is the negative second derivative of the log - likelihood function with respect to θ.

  Step 4: Update the parameter estimate using the Newton - Raphson update rule

$$\theta_{t+1} = \theta_t - [I(\theta_t)]^{-1}.U(\theta_t)$$

  Step 5: Repeat Steps 1 - 4 until convergence criteria are met, such as reaching a specified number of iterations or the change in parameter estimates between iterations falls below a predefined threshold.
  - Convergence Criteria: Define stopping criteria to determine when to terminate the iteration process. Common criteria include achieving a specified tolerance level for parameter estimates or reaching a maximum number of iterations.
  - Output: Once convergence is achieved, the final parameter estimates $\hat{\theta}$ are obtained.

- **The Gradient Descent method**

  - Initialization: Choose an initial guess θ_0 for the parameters
  - Iteration:

  Step 1: Compute the log-likelihood function $l(\theta)$ based on the observed data.

  Step 2: Compute the gradient of the log-likelihood function with respect to each parameter to obtain $\nabla l(\theta)$.

  Step 3: Update the parameters by moving opposite to the gradient direction with a small step size called the learning rate α:

$$\theta_{t+1} = \theta_t - \alpha \nabla l(\theta)$$

  Step 4: Repeat the process until reaching stopping criteria, such as achieving a sufficient optimal value or reaching a maximum number of iterations.
  - Stopping Criteria: Define stopping criteria to determine when to stop. Common criteria include achieving a sufficient optimal value or reaching a maximum number of iterations.
  Result: After convergence, the final parameters are determined as the optimal MLE estimates.

The Newton-Raphson method, known for its rapid convergence rate, utilizes the second derivative information of the log-likelihood function to iteratively refine parameter estimates. By approximating the log-likelihood function locally with a quadratic function, Newton-Raphson adjusts parameter values towards the maximum likelihood estimates in a more direct manner.

On the other hand, Gradient Descent is particularly useful in scenarios where the log-likelihood function is not smooth or where computing second derivatives is impractical. Instead of relying on second-order information, Gradient Descent utilizes first-order derivatives to iteratively update parameter values in the direction of steepest descent. While Gradient Descent may converge more slowly compared to Newton-Raphson, it remains effective in optimizing parameters for complex and high-dimensional models.

2.2 Method of Moment (MOM)

2.2.1 Definition

Let $X = x_1, ...x_n$ be idd realizations (samples) with pdf (or pmf) $f(x_1, ...x_n|\theta), \theta \in \Theta$. We then define the Method of Moment (MOM) estimator θ_{MOM} of $\theta = (\theta_1, ...\theta_k)$ to be a solution (if it exists) to the k simultaneous equation where, for $j = 1, ...k$, we set the j^{th} (true) moment equal to the j^{th} sample moment:

$$E[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$\cdots$$
$$E[X^k] = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

2.2.2 Theorem

For each $n \in \mathbb{N}_+$, $X_n = x_1, ...x_n$ is a random sample of size n from the distribution X.

- Suppose that the mean μ is unknown. The method of moment estimator of μ based on X_n is the sample mean.

$$M_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  $E[M_n] = \mu$, so M_n is unbiased for $n \in \mathbb{N}_+$
  $Var(M_n) = \frac{\sigma^2}{n}$ for $n \in \mathbb{N}_+$, so $M = (M_1, ...M_n)$ is consistent.

- Suppose that the mean μ and variance σ^2 are both unknown. The method of moment estimator of σ^2 based on X_n is

$$T_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - M_n)^2$$

  Bias $(T_n^2) = \frac{\sigma^2}{n}$ for $n \in \mathbb{N}_+$, so $T^2 = (T_1^2, \ldots, T_n^2)$ is asymptotically unbiased.
  MSE $(T_n^2) = \frac{1}{n^3}[(n-1)^2\sigma_4 - (n^2 - 5n + 3)\sigma_4]$ for $n \in \mathbb{N}_+$, so T^2 is consistent.

- Suppose that the mean μ is known and variance σ^2 is unknown. The method of moment estimator of σ^2 based on X_n is

$$W_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

$E[W_n^2] = \sigma^2$, so W_n^2 is unbiased for $n \in \mathbb{N}_+$
Var$(W_n^2) = \frac{1}{1}\sigma_4 - \sigma^4$ for $n \in \mathbb{N}_+$, so $W^2 = (W_1^2, ...W_n^2)$ is consistent.

2.2.3 Algorithm

We have constructed the MOM algorithm using the Newton-Raphson method for the following reasons:

- Fast Convergence: Newton-Raphson method typically converges rapidly, making it suitable for MOM where efficient parameter estimation is essential.

- High Precision: Newton-Raphson provides high-precision estimates due to its ability to utilize second-order derivative information, ensuring accurate parameter estimation.

- Robustness: Despite the non-linear nature of MOM equations, Newton-Raphson demonstrates robustness in handling complex optimization problems, offering reliable solutions even in challenging scenarios.

The general algorithm for MOM using the Newton-Raphson method:

- Initialization: Start by initializing the parameters with initial guesses.

- Iteration:

Step 1: Compute the MOM equations based on the sample moments and set them equal to their theoretical counterparts.

Step 2: Use the Newton-Raphson method to iteratively update the parameter estimates until convergence.

- Convergence Check: Monitor the convergence of the parameter estimates by assessing changes between successive iterations.

- Assessment: Evaluate the estimated parameters for accuracy and reliability, ensuring they align with the MOM assumptions and provide meaningful interpretations.

2.3 Bayesian Estimation

2.3.1 Definition

The distribution of θ, given data $x_1, ...x_n$, is called the posterior distribution, which is given by

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{g(x)}$$

where $g(x)$ is the marginal distribution of X. The Bayes estimate of the parameter θ is the posterior mean.

The marginal distribution $g(x)$ can be calculated using the formula:

$$g(x) = \sum_\theta f(x|\theta)\pi(\theta) \text{ in the discrete case}$$
$$g(x) = \int_{-\infty}^{\infty} f(x|\theta)\pi(\theta)\, d\theta \text{ in the continuous case}$$

2.3.2 Criteria for finding the Bayesian Estimate

In Bayesian parameter estimation, we integrate both prior information and observed data to derive estimates based on the posterior distribution. However, determining the effectiveness of these estimates requires an assessment of their quality. This is achieved through the use of a loss function $L(\theta, a)$ that quantifies the discrepancy between the true parameter θ (which is typically unknown in real-world scenarios) and its estimate a. The goal is to select an estimate a that minimizes the expected loss $E[L(\theta, \hat{\theta})]$ where the expectation is taken over θ according to the posterior distribution $f(\theta|x)$.

Two commonly employed loss functions are the quadratic and absolute error loss functions, each leading to distinct estimation strategies. With the quadratic loss function, the resulting estimate is chosen to minimize the mean squared error between the true parameter and its estimate. Conversely, the absolute error loss function yields estimates that minimize the mean absolute error. These approaches provide different perspectives on estimating parameters, each suitable for different types of problems and decision-making contexts.

- A quadratic (or squared error) loss function is of the form $L(\theta, a) = (a - \theta)^2$. In this case,

$$E[L(\theta, a)] = \int (a - \theta)^2 f(\theta, x_1, ...x_n) d\theta$$

  The posterior mean (expected value) of θ is obtained by differentiating with respect to a and equating to zero, which is

$$a = E[\theta|x_1, ...x_n] = \int \theta f(\theta, x_1, ...x_n d\theta$$

- An absolute error loss function is of the form $L(\theta|a) = |a - \theta|$. In this case,

$$\int_{\theta=-\infty}^{a} (a - \theta) f(\theta, x_1, ...x_n) d\theta + \int_{\theta=a}^{\infty} (\theta - a) f(\theta, x_1, ...x_n) d\theta$$

  Differentiating with respect to a and equating to zero, we obtain:

$$\int_{\theta=-\infty}^{a} f(\theta, x_1, ...x_n) d\theta - \int_{\theta=a}^{\infty} f(\theta, x_1, ...x_n) d\theta = 0$$

  The minimum loss is attained when the values of both integrals are equal to $\frac{1}{2}$. This can be achieved by taking $\hat{\theta}$ to be the posterior median.

2.3.3 Bayesian parameter estimation procedure

- Step 1: Consider the unknown parameter θ as a random variable.
- Step 2: Use a probability distribution (prior) to describe the uncertainty about the unknown parameter.
- Step 3: Update the parameter distribution using the Bayes theorem:

$$P(\theta|Data) \propto P(\theta))P(Data|\theta))$$

- The Bayes estimator of θ is set to be the expected value of the posterior distribution $P(\theta|Data)$ under the quadratic loss function.
- The Bayes estimator of θ is set to be the posterior median under the absolute error loss function.

3 Application

3.1 Data description

The Student Performance Dataset is curated to investigate the factors influencing academic student performance. It comprises 10,000 student records, each containing information about various predictors and a performance index.

Predictors:

Hours Studied: This variable denotes the total number of hours spent studying by each student.

Previous Scores: This variable represents the scores obtained by students in previous tests, serving as an indicator of their academic history.

Extracurricular Activities: This categorical variable indicates whether the student participates in extracurricular activities, with options being "Yes" or "No".

Sleep Hours: This variable signifies the average number of hours of sleep the student had per day, which could potentially impact their academic performance.

Sample Question Papers Practiced: This variable quantifies the number of sample question papers the student practiced, reflecting their level of preparation.

Target Variable:

Performance Index: The Performance Index serves as a measure of the overall academic performance of each student. It has been rounded to the nearest integer and ranges from 10 to 100, with higher values indicating better academic performance.

We embark on our investigation utilizing the Student - Performance dataset, consisting of a solitary target variable and five explanatory variables, as previously delineated. Our principal aim revolves around estimating the parameters of a linear regression model employing three distinct estimation methodologies: the Method of Moments (MOM), Maximum Likelihood Estimation (MLE), and Bayesian Estimation. Specifically, our endeavor entails the estimation of the quintet of parameters corresponding to the five explanatory variables within our linear regression framework. Subsequently, we shall harness the 'lm' function within the R environment to meticulously fit our linear regression model to the dataset. Our ultimate pursuit lies in meticulously scrutinizing and contrasting the efficacy of these three estimation techniques amongst themselves and vis - à - vis the intrinsic capabilities of the 'lm' function in R. Through this scholarly comparative analysis, we aspire to glean insights into the efficacy, precision, and robustness of each estimation method in discerning the intricate relationships between the explanatory and response variables encapsulated within the Student - Performance dataset.

3.2 Data preprocessing

We conducted an in-depth Exploratory Data Analysis (EDA) on the Student Performance dataset, yielding comprehensive insights. Our analysis revealed an absence of missing values and outliers, the latter being assessed through robust quantile-based measures. Subsequent to this, we rigorously examined the distributional characteristics of the target variable using diagnostic tools such as Quantile-Quantile (QQ) plots and the Kolmogorov-Smirnov test.

Upon detecting deviations from the normal distribution, we undertook corrective measures, employing both square and logarithmic transformations to approximate the distributional properties towards normality. Furthermore, recognizing the categorical nature of the 'Extracurricular Activities' variable, we proceeded to encode it into dummy variables.

Following this, we applied the max-min scaling technique to standardize the data, ensuring consistency and comparability across different features and observations

Subsequently, we partitioned the dataset into training and validation sets to assess the performance of our predictive models. This partitioning allows for the evaluation of model effectiveness on unseen data, serving as a crucial step in model validation and generalization.

The training set will be utilized to train various machine learning algorithms, enabling them to learn patterns and relationships within the data. On the other hand, the validation set will be employed to assess the performance of these trained models, providing valuable insights into their predictive accuracy and generalization capabilities.

By employing this approach, we aim to develop robust and reliable models that can accurately predict student performance based on the provided features. This rigorous evaluation process ensures that our models are not only effective on the training data but also capable of generalizing well to new, unseen data, thereby enhancing their practical utility and real-world applicability.

3.3 Procedure

We conducted parameter estimation for a multiple linear regression model encompassing six parameters, including an intercept and five slopes corresponding to five predictor variables. This estimation was performed using various methods: Maximum Likelihood Estimation (MLE), Method of Moments (MOM), and Bayesian Estimation employing both the Newton-Graphson algorithm and Gradient Descent.

Simultaneously, we fitted the model using built-in R packages. Model performance was assessed using key metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

The experimentation encompassed different approaches applied to both standardized and non-standardized data, with target variables transformed using squared or logarithmic functions. This comprehensive evaluation allowed for a thorough comparison of the efficacy of each method across diverse data conditions.

3.3.1 Maximum likelihood estimate method

We constructed functions to facilitate Maximum Likelihood Estimation (MLE) employing the Newton-Raphson and Gradient Descent optimization method for a multiple linear regression model.

The log - likelihood function is devised to compute the log-likelihood of the data under the assumption of normally distributed errors. Given a set of model parameters, it calculates the predicted values for the dependent variable (Y) based on the independent variables (X). These predicted values are then used to compute the residuals, representing the differences between the observed and predicted values. By employing the Gaussian probability density function, the log-likelihood function assesses the likelihood of observing the given residuals under the assumption of normally distributed errors, providing a measure of how well the model fits the data.

The Newton-Raphson function orchestrates the Newton-Raphson optimization algorithm by iteratively updating model parameters to maximize the log-likelihood. Leveraging the computed gradient and Hessian matrix, the function directs parameter updates towards maximum likelihood estimates. The optimization proceeds until convergence, as determined by a predefined convergence criterion. This iterative refinement ensures the model parameters approach optimal values, enhancing the fit of the regression model to the observed data.

Conversely, the gradient - descent function employs the Gradient Descent optimization method to estimate parameters for the multiple linear regression model. It initializes parameters and iteratively applies the Gradient Descent algorithm, adjusting parameter values to optimize the log-likelihood function. The function iterates until convergence, progressively refining parameter estimates to maximize the likelihood of observing the given data.

3.3.2 Method of Moments

In this phase, we performed parameter estimation using the Method of Moments (MOM) approach for a linear regression model.

Using the sample moments, a moment matrix is constructed, incorporating means and covariances of predictor variables. This matrix serves as the basis for parameter estimation. Parameters are estimated by solving a system of equations formulated using the moment matrix and observed moments of the target variable. Ordinary Least Squares (OLS) is employed for parameter estimation. This MOM approach utilizes sample moments to estimate model parameters, aiming to align observed moments of the data distribution with those predicted by the regression model.

Following the implementation of the MOM method for parameter estimation, we proceed with applying MOM using Newton-Graphson optimization. We defines a loss function and its gradient, essential components for optimization algorithms such as Newton-Raphson. The loss function measures the discrepancy between predicted and actual values, while the gradient indicates the direction and rate of change of the loss with respect to the model parameters. Subsequently, the Newton-Raphson optimization method is applied to iteratively refine parameter estimates. This method utilizes the gradient and the Hessian matrix, a matrix of second-order partial derivatives of the loss function, to guide parameter updates towards the optimal values. The optimization process iterates until convergence, as determined by a predefined convergence criterion based on the magnitude of parameter updates. Upon convergence, the MOM estimates for the model parameters are obtained.

Overall, the workflow entails the formulation and optimization of the loss function to derive MOM estimates using the Newton-Raphson method. This iterative process systematically adjusts parameter values to minimize the loss, thereby enhancing the fit of the model to the training data.

3.3.3 Bayesian estimation method

The Metropolis-Hastings algorithm is employed to perform Bayesian estimation in this context. This algorithm enables the sampling of parameters from the posterior distribution, which combines the information from the likelihood function and the prior distribution.

In the log-likelihood function, a Student's t-distribution is chosen to model the errors. Unlike the Gaussian distribution, the Student's t-distribution allows for heavier tails, making it more robust to outliers in the data. This robustness is advantageous in regression tasks where the assumptions of Gaussian errors may not hold, or when dealing with data containing outliers.

Additionally, the log-prior function incorporates normal and gamma distributions for the model parameters. These distributions are commonly chosen as priors due to their flexibility and interpretability. The normal distribution is often used to model prior beliefs about regression coefficients, assuming that they are centered around zero with certain variance. On the other hand, the gamma distribution is suitable for modeling scale parameters such as the variance of the errors, as it is non-negative and right-skewed, aligning with the properties of variances. By combining the Student's t-distribution for the likelihood and normal and gamma distributions for the priors, the log-posterior function captures both the observed data and prior knowledge about the parameters.

4 Results

Below are the results obtained from parameter estimation on the transformed target variable data using squared and log transformations, with the following notations:

- MLE - NR: Maximum Likelihood Estimation with Newton-Graphson optimization.
- MLE - GD: Maximum Likelihood Estimation with Gradient Descent optimization.
- Bayesian: Bayesian estimation.
- MOM: Method of Moments.
- MOM - NR: Method of Moments with Newton - Graphson optimization.
- MLR: Fitted by available R package

	MSE	MAE	RMSE
MLE-NR	208.03	14.42	14.42
MLE-GD	$29.e^9$	$16.e^4$	$17.e^4$
Bayesian	0.201	0.408	0.449
MOM	$11.e^9$	$10.e^4$	$10.e^4$
MOM-NR	0.011	0.076	0.105
MLR	0.011	0.076	0.105

Table 1: **Parameter estimation results using Log - transformation**

	MSE	MAE	RMSE
MLE-NR	194.86	13.94	13.96
MLE-GD	$19.e^5$	$12.e^3$	$14.e^3$
Bayesian	0.813	0.883	0.901
MOM	$63.e^5$	$21.e^3$	$25.e^4$
MOM-NR	0.001	0.023	0.032
MLR	0.001	0.023	0.032

Table 2: **Parameter estimation results using SQRT - transformation**

Following the synthesis of the summarized results from the two tables, it becomes evident that the Method of Moments (MOM) optimized through Newton Graphson emerges as the most apt and effective approach for the Student Performance dataset. This assertion is grounded in the meticulous assessment of errors, quantified by metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), which indicate a remarkable diminution when employing this method. Moreover, the parameter coefficients derived from MOM optimization via Newton Graphson align remarkably well with those yielded by the Multiple Linear Regression (MLR) model implemented using R's built-in package, exhibiting an equally high degree of precision. This concordance holds true for both logarithmic and square root transformations applied to the target variable.

Subsequently, the Bayesian Estimation method emerges as the next most accurate, albeit marginally trailing behind MOM - NR, particularly excelling when applied to data subjected to log transformation. Conversely, other methodologies display a marked lack of compatibility with the data's underlying distribution, consequently yielding notably inferior metric outcomes.

Notably, the MOM approach, despite its apparent simplicity, surprisingly yielded exceedingly poor results, showcasing a stark underperformance even when compared to the Maximum Likelihood Estimation (MLE) method optimized using Newton-Graphson. Intriguingly, in both instances of target variable transformation, MLE with Gradient Descent optimization demonstrated the poorest performance.

In our pursuit of refinement, we experimented with diverse alterations to various log-likelihood functions, alongside the selection of disparate prior distributions, and the manipulation of initialization parameters, learning rates, and iteration counts. Regrettably, these adjustments failed to yield appreciable enhancements.

In summation, the Method of Moments optimized through Newton Graphson emerged as the most efficacious, primarily owing to its superior prowess in error minimization and its close adherence to the results obtained from the MLR model fitted through established R packages. Conversely, the lackluster performance of alternative methods can largely be attributed to their inadequate fit with the data's distributional characteristics.

5 Discussions

In light of the provided results, Maximum Likelihood Estimation (MLE) proves less effective due to the target variable in the surveyed data not adhering strictly to a standard distribution, despite attempts to normalize it through various transformations. Conversely, Method of Moments

optimized by Newton Graphson (MOM - NR) emerges as the most suitable approach for this scenario, offering superior estimation performance.

Beyond its robustness in non-standard distributional settings, MOM - NR benefits from its straightforward implementation and computational efficiency. Additionally, its reliance on sample moments makes it particularly adept at capturing the central tendencies of the data, even amidst distributional complexities.

Considering these findings, the recommendation for method selection hinges on several factors. Firstly, the structural characteristics of the dataset, notably its distributional properties, should guide the choice of method. MOM - NR is well-suited for datasets with non-standard distributions or those exhibiting robustness against outliers. Secondly, the ease of implementation and computational efficiency should be considered, especially when dealing with large datasets or resource-constrained environments. MOM - NR's simplicity and computational efficiency render it a pragmatic choice in such scenarios. Lastly, the interpretability and robustness of the estimation results are paramount. MOM - NR's reliance on sample moments provides transparent insights into the data's central tendencies, facilitating the interpretation of parameter estimates.

In conclusion, Method of Moments optimized by Newton Graphson presents itself as the preferred choice for parameter estimation in scenarios where the target variable deviates from a standard distribution. Its robustness, simplicity, and interpretability make it a compelling option, particularly when dealing with non-standard data distributions and resource constraints.

6 Document link

Github link for the project: `https://github.com/Jena-Thaipham/STAT-5310---Final-Projec`

References

[1] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B. *Bayesian Data Analysis (3rd ed.)*. Chapman and Hall/CRC, 2013.

[2] Hastie, T., Tibshirani, R., Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.)*. Springer, 2009.

[3] Wasserman, L. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004.

[4] Casella, G., Berger, R. L. *Statistical Inference (2nd ed.)*. Duxbury Press, 2002.

[5] Gelman, A., Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.

[6] Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

[7] Fox, J. *Applied Regression Analysis and Generalized Linear Models*. Sage Publications, 2008.

[8] Shalizi, C. R., Shalizi, K. L. *Advanced Data Analysis from an Elementary Point of View*. Cambridge University Press, 2013.

[9] Kuhn, M., Johnson, K. *Applied Predictive Modeling*. Springer, 2013.

Appendix

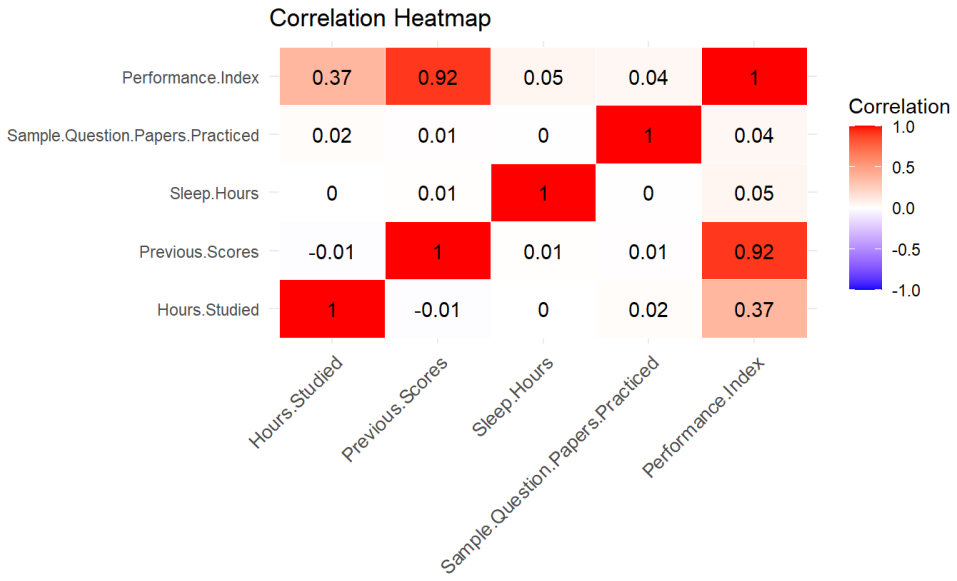

Figure 1: Correlation checking

Figure 2: Normality checking for target variable

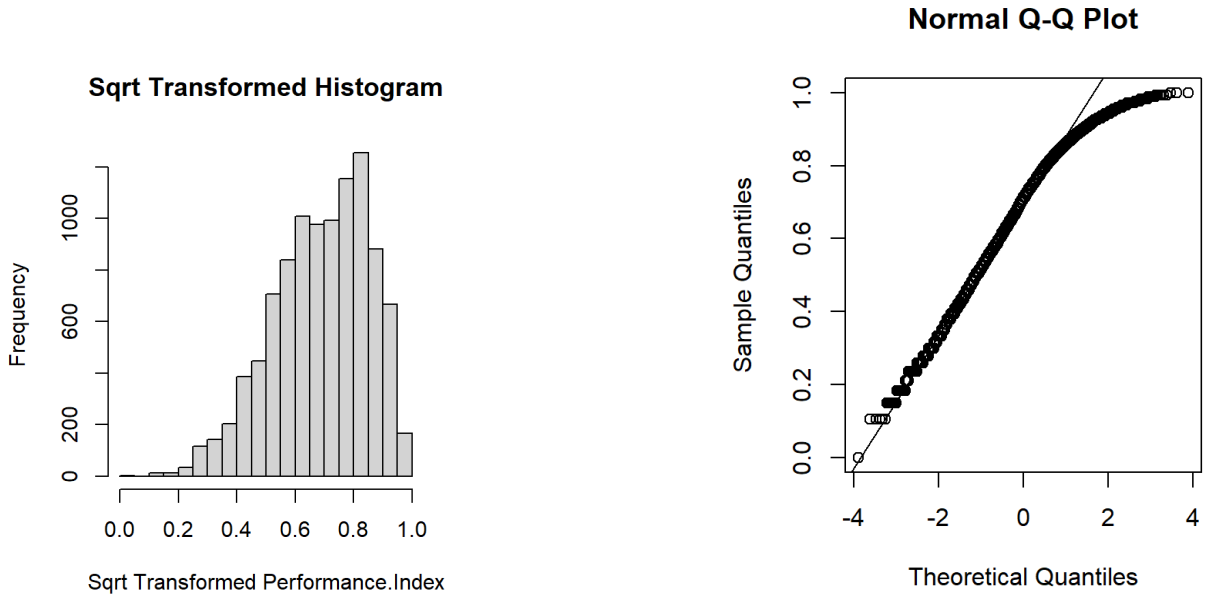

Figure 3: SQRT Transformation

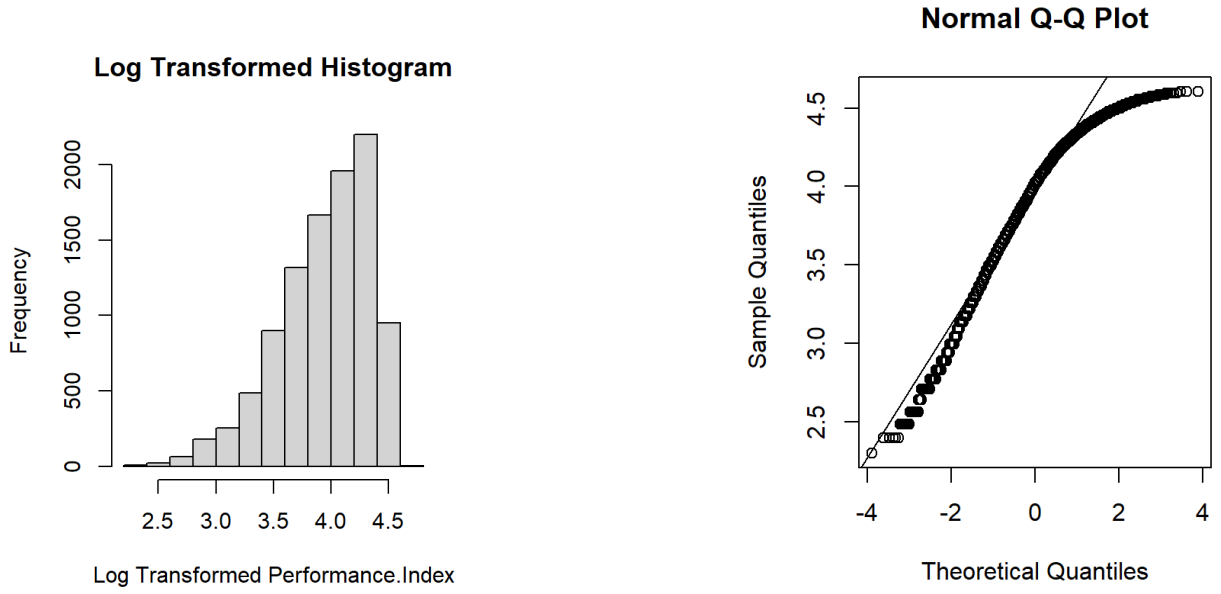

Figure 4: Log transformation