Performance comparison between maximum likelihood estimation and variational method for estimating simple linear regression parameter

Yekti Widyaningsih*, *Hakiim Nur* Rizka, and *Titin* Siswantining

Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Depok 16424, Indonesia

Abstract. Variational estimation method is a deterministic approximation technique which involves Bayesian framework while giving a point estimate instead of the usual Bayesian interval estimation. The linear regression model, which has always been a popular model, can benefit from the implementation of variational estimation method. In this paper, the theoretical basis on why variational method can reduce overfitting in linear regression is reviewed. Based on the review, in theory, variational method is more robust to overfitting than MLE. This paper also performed a simulation study. The simulation is done in a manner such that the simulation represents the situation of predicting for new or hidden data. The simulation starts from generating random explanatory data and generates the appropriate response data based on linear regression equation. Then, the randomly generated data is used to estimate the linear regression parameters. The simulation is performed to compare the parameters estimation results from variational method with the method of MLE. The comparison is done using the estimation values and the squared differences between true parameters value and the estimates. Empirical findings show that both methods have relatively close estimate values. It can be seen as the simulation study concludes that both variational and ML yield rather close parameters estimates for simple linear regression case. The estimates closeness gets more obvious as the sample size grows. The study also found that Variational method has performs better in terms of parameters estimation in linear regression when the sample size is small or the data has large variance.

Keywords. Bayesian regression, maximum likelihood, overfitting, simulation

1 Introduction

Linear regression analysis is a statistical analysis method that can be used to represent the relationship between 2 types of variables, independent and dependent [1]. In general, linear

* Corresponding author: yekti@sci.ui.ac.id

regression can be used to represent the relationship of one dependent variable with one or more independent variables [2]. In its application, regression analysis can be used as a tool for predicting the value of the dependent variable, analyzing the relationship between the dependent and independent variables, and predicting changes in the value of the dependent variable based on the relationship between the dependent and independent variables [3].

When applying linear regression analysis to data, it is necessary to build an equation that represents the relationship between the dependent and independent variables. In building this equation, the value of the regression equation coefficient needs to be known. However, the actual value of the coefficient of the regression equation in its application cannot be known, so an estimate is made for that value [4]. One method that is widely used to estimate the coefficients of the regression equation is the maximum likelihood estimation method.

The maximum likelihood method is a method used in estimating the parameters of an assumed probability distribution, given some observed data by searching for the maximum value of the likelihood function [4]. In estimating the coefficients of the regression equation using the maximum likelihood method, distribution assumptions are needed in the regression model. The most used assumption is that errors are normally distributed, homogeneous, identical, and mutually independent [2]. The ease of using the maximum likelihood method is one of the reasons for using this method in linear regression model applications. However, it is not without any drawbacks. One of the problems that is starting to become a concern but arguably less popular is related to overfitting.

Overfitting can be explained as a problem where the model obtained from the data implementation fails to explain the relationship between the dependent and independent variables at unobserved values [5]. Because of this overfitting problem, often modeling that is built based on data will fail to generalize the conclusions of the analysis obtained [6, 7]. In other words, linear regression models built using the maximum likelihood method based on some observed data often fail to fulfill their function as a method for performing statistical inference [7].

The idea behind overfitting of MLE was noted by Bishop [8] as an “unfortunate” property. To understand this, one can recall or look up the general idea of method of MLE. Mathematically, MLE tries to find values that “fit” the equation. In other words, MLE is always fitting the data to the model, no matter how many parameters are added. Thus, in its basic form, MLE does not have the capacity to regulate the parameters of the model. Quoting Bishop [8], the usage of MLE may lead to severe over-fitting if complex models are built based on limited data.

Although seen as a serious problem, overfitting in linear regression is still rarely discussed [9]. To overcome this problem, researchers formulated many solutions. To name a few instances: improving the data used to build the model [6], Bayesian statistics approach [10, 11], and regularization [12]. These solutions have also been used to reduce overfitting in applied statistics. However, it is also known that these solutions have their disadvantages which render them to be unusable in some situations.

For data improvement, one with enough experience will notice that it is not always possible to increase the size or even remove parts of data that has been extracted. Either limited by time or some other reasons. Some data manipulations to increase the sample size such as bootstrapping are also utilized to improve the data. Yet, it is not a magical technique. Such a technique does not necessarily increase the amount of information in the data, which is the main point of trying to improve the data to reduce overfitting. Meanwhile, Bayesian approach and regularization, these are the more sophisticated solutions.

The Bayesian approach incorporates past information with newly extracted data to hopefully gain a better inference [13]. In theory, the Bayesian approach is always a “superior” method in statistics. Obviously, this does not come without price. The Bayesian approach is well known for being hard-to-implement [14]. In addition, those who have experience in

using Bayesian technique should be aware of how computationally expensive this technique is. In other words, Bayesian technique in general is hard to use and requires large amounts of computation. Similarly, regularization also has problems with implementation. Regularization does not incorporate information outside of the data like Bayesian. Thus, this technique tries to penalize irrelevant information in the data. This process is known as bias-variance trade off. By reducing the variance in the model, it also reduces the fit of the model to the data with the expectation of increasing out-of-sample fit of the model. The problem is that this trade off really depends on the data. In other words, the problem of this technique is similar to the problem of data improvement. Therefore, new solution is proposed for a parameter estimation method that can reduce the overfitting problem in linear regression. This parameter estimation method is known as the variational estimation method.

The variational estimation method is an iterative parameter estimation method that is widely used in machine learning to estimate unknown values based on the distribution function through an optimization process [15]. Some examples of the application of this method are the estimation of the logistic regression coefficient [16], the estimation of the spatial logistic regression coefficient [17], and the estimation of the probability equation parameter equation [18]. While the application has been used by many, a basic and easy-to-use algorithm of variational method for linear regression has not been explored deeply. This is because those applications are mostly applied to answer more complex problems such as the non-existence of analytical solutions or hard-to-compute mathematical equations.

To use the variational estimation method as an estimate of the linear regression coefficient, Drugowitsch [19] built a hierarchical modeling for linear regression through the Bayes principle. The modeling is built by setting prior assumptions that the linear regression parameters have a normal distribution which is explained through hyperparameters with a gamma distribution. The variational method is then used to find the approximate values of the parameters of the posterior distribution for the Bayesian linear regression [19]. Using this Bayesian concept, the estimation results of the variational method for the linear regression coefficients obtained will be more robust to overfitting if the appropriate conjugate prior distribution is used [10].

These theoretical findings of reducing the overfitting problem have great significances in the field of statistics and anything related. As noted by Dalicandro et al. [9], the prevalence of overfitting is something that needs to be cautioned when one wants to model a data using MLE linear regression. Overfitting will render the results of the linear regression model prone to error. Obviously, reducing the overfitting by getting a more accurate estimate to the linear regression model parameters helps in formulating a better analysis' results using linear regression. This paper is an exploration of a study to improve the performance of linear regression model by reducing the overfitting problems.

To see whether variational method reduce the overfitting in linear regression, an experimental study is performed to compare the performance of maximum likelihood and variational methods in linear regression model. This comparison study is performed by comparing the parameters estimates and the MSE between variational methods and MLE in a simulation. Using the two measures, the accuracy for both estimation methods can be compared. The simulation includes randomly generating the explanatory data and the corresponding response data based on linear regression equation. The generated data is then used to do the estimation of the linear regression model's parameters. As an addition, the study also does a consistency analysis by replicating the simulations several times and computes the squared difference between the true value of the parameters and the estimates. This procedure is parallel to that of checking the unbiasedness of an estimator. Obviously, since variational method does not have closed form for its estimate, this study used the average values of the estimates instead of expected value. The idea behind this performance

comparison is derived from the performance comparison study for different model by Hardouin [17].

The linear regression model used in the simulation is limited to simple linear regression to make it easier to write the algorithms. Estimating the coefficients of multiple linear regression can be done by simply changing the function parameters in the algorithm. Simulation studies were carried out using RStudio software with the statistical programming language R version 4.2.1.

2 Theoretical review

The Variational estimation method is an approximation method which utilizes a Bayesian principle and framework. That is, variational method assumes that the parameters of the linear regression have uncertainty on their values. Then, variational method approximates the posterior distribution over the regression parameters using variational distribution. By optimizing this variational distribution relative to its parameters, one can obtain the estimated value of the regression parameters. Regarding the overfitting problem and variational estimation method as its solution in linear regression, the following are brief theoretical reviews which helps understanding the 2 ideas.

2.1 Linear regression

Linear regression assumes that a regression function can be used to represent how each change in the independent variables affect the dependent variable. Linear regression with one dependent and one independent variable is called simple linear regression, while multiple linear regression refers to linear regression with one dependent and more than one independent variable [4].

The term linear in linear regression is used to describe the relationship between the regression parameter and the dependent variable which can be represented using the regression function as written below.

$$y_i = \boldsymbol{\beta}' \boldsymbol{x_i} + \varepsilon_i , \qquad i = 1, 2, \dots, n \tag{1}$$

where yi represent the dependent variable and xi is the vector of size k for the independent variables for the i-th observation. The **β** in the Equation (1) represent the so-called regression coefficient/parameter vector with size k. While the ϵ_i serves to function as the "error" or random chance.

One can also write regression function collectively using the following equation.

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{2}$$

X is referred to as the design matrix with size n × k. In the application of linear regression, by having data $\boldsymbol{D} = \{\boldsymbol{X}, \boldsymbol{y}\}$, one needs to estimate the value of parameters in β such that a response point y* of a new data point x* can be predicted using the mean response model below:

$$E(y^*) = \boldsymbol{\beta}' \boldsymbol{x}^* \tag{3}$$

the Equation (3) is obtained by utilizing the Equation (1) while applying the assumption of normality to the error term. In a mathematical notation, this is usually written as,

$$\varepsilon_i \sim N(\mu, \sigma^2), \qquad i = 1, 2, \dots, n \tag{4}$$

To do the parameter estimation when assumption of normality, that is Equation (4) holds for the linear regression model, maximum likelihood estimation (MLE) can be utilized to estimate the regression coefficients in Equation (1) [2]. To do this, first establish the likelihood function for the regression function which can be written as,

$$p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})\right) \tag{5}$$

Equation (5) is built using the collection of n normally distributed random samples with mean 0 and variance σ^2. For simplicity of notation for the following discussion, the likelihood in Equation (5) is changed into $p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta})$ and σ^2 is assumed to be known.

2.2 Overfitting in linear regression

In statistics, one will most likely try to achieve inference. This concept is that one needs to simply learn from some examples to understand a population. This is also applied in linear regression. Linear regression studies the relationship between variables by building a regression equation using observed data. Then, an inference about the relationship between variables can be formulated for both observed and unobserved data [4]. However, a problem arises when one is constructing a regression equation.

Following [4] to estimate the regression coefficients in regression equation, when the assumption of normality is fulfilled, MLE can be used to estimate the regression coefficients. When using MLE in linear regression, the problem of "only considering what is known" is of concern. Applying MLE to regression likelihood function is equivalent to solving maximization problem for the likelihood with respect to the unknown coefficients. Thus, overfitting happens. Quoting Bishop [8], method of MLE has an unfortunate property that is it has no mechanism within its framework to prevent overfitting.

To illustrate Bishop's point, let's say there are 1000 data points and a model of linear regression with 1000 coefficients. In this scenario, the method of MLE will give a perfect fit for the sample. However, this obviously leads to high error if such model is generalized into the population. Method of MLE works just like fitting each of these data points into the model until a perfect fit occurs [8]. In another perspective with similar understanding, MLE works by building an estimation based on current data [7]. One needs to consider the population data and not limiting estimation for model's parameters based on some samples. Thus, alternatives for method of MLE to reduce the overfitting problem are necessary.

Reducing overfitting of a model is not anything new in scientific study. To address overfitting in linear regression, one can use the regularization method. Regularization is a popular technique used in statistics and its derivatives. In principle, the idea of regularization is to introduce a mechanism to regulate the model fit to the sample data [20]. This idea is in line with the problem of overfitting in the method of MLE. A regularized linear regression tries to reduce the model fit to the sample data in hope of achieving better accuracy in the population. While it is almost impossible in many cases to fit the population, regularization technique measures the out-of-sample fit using parts of the data which is not used to build the model. Thus, one of the first steps in the implementation of regularization is to separate the data into training and test data. The regression equation is built using the training data

and then test data is used to tune the regression model such that its fit is balanced for both training and the test data. This process is known as bias-variance tradeoff [20]. A rather recent example of the regularization method is done by Kolluri et al. [12] by using a more subtle regularization technique which helps in reducing the overfitting problem.

Another way to reduce overfitting problem is by utilizing Bayesian approach [13]. Bayesian linear regression works by incorporating a "prior". This prior is a representation of probability density of the regression parameters based on the past knowledge. In other words, the Bayesian approach assumes that the regression parameters have many different possible values [13]. This prior combined with the data is used to build the posterior which is the probability density of the regression parameters after considering the data. While it sounds unintuitive for many, it does help in reducing the problem of overfitting by not yielding just a point of estimates for the regression parameters estimation. In other perspective, the Bayesian approach has the advantage of incorporating information outside of the data. Jones et al. [11] explore the Bayesian approach and improve the overall performance by using the posterior.

The two solutions for overfitting above have been used by many. However, they are not without any drawbacks. In the case of regularization, the bias-variance tradeoff itself is flawed. There is no guarantee that this tradeoff leads to improving accuracy towards the actual population. That is because regularization in essence only works using data that is available. This problem is solved by the Bayesian approach. The prior in Bayesian acts to accommodate the randomness in the model. Incorporating relevant information outside of the data to build the model improves the accuracy of Bayesian regression in modeling out-of-sample data. However, the Bayesian approach is neither an easy-to-use method nor is it easy to compute. In light of these problems, the Variational estimation method is proposed as a solution.

2.3 Bayesian principle for linear regression

Proceeding using the setup from linear regression section, building the Bayesian framework to do linear regression coefficients estimation can be done by utilizing the basic Bayesian principle. Suppose we wish to do comparison for a set of n probability distributions with parameter: $\{\theta i\}, i = 1, 2, \dots, n$. Given a sample data $\boldsymbol{D}$ we then evaluate the posterior distribution.

$$p(\theta_i|\boldsymbol{D}) \propto p(\theta_i)\, p(\boldsymbol{D}|\theta_i)$$

p(θ_i) is referred as the prior information or prior distribution and $p(\boldsymbol{D}|\boldsymbol{\theta_i})$ as the "evidence" or also known as likelihood. To apply this principle in regards of parameter estimation in linear regression, first assume that the regression parameters $\boldsymbol{\beta}$ has uncertainty. This uncertainty leads to inference about the true value of the parameters is not singular but rather some interval with a defined probability density function over it. This concept is usually termed as "assigning prior distribution" [21]. Thus, following from the Equation (2), the likelihood and prior for the vector of parameters $\boldsymbol{\beta}$ can defined by:

$$p(y \mid \boldsymbol{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} N(\beta' x_i, \sigma^2) \tag{6}$$

$$p(\boldsymbol{\beta}|\tau) = N(m_0, \tau^{-1} I) \tag{7}$$

In this setup, the distribution for the prior is multivariate normal with the choice of mean parameter is m_0=0 and hyper parameter τ is introduced to represent the randomness of the regression parameters $\boldsymbol{\beta}$. The prior for the hyper parameter τ is also defined using the following equation.

$$p(\tau) = Gamma(a_0, b_0) \tag{8}$$

where a_0 and b_0 is the parameter for the distribution of τ. Combining the setup from Equation (6) to (8) and using the conditional probability rule, Bolstad [13] writes the joint distribution over all variables using the following equation.

$$p(y, \boldsymbol{\beta}, \tau \mid \boldsymbol{X}) = p(y \mid \boldsymbol{X}, \boldsymbol{\beta}) p(\boldsymbol{\beta} \mid \tau) p(\tau) \tag{9}$$

The left side of the Equation (9) is the posterior distribution of the linear regression in Equation (2).

2.4 Variational linear regression

As mentioned in the beginning, variational method works by first setting up the framework in a Bayesian approach. Then, using the defined posterior distribution in Equation (9), variational method can be implemented. In short, variational method looks for an approximation for the posterior by means of optimization. Thus, the step after defining the posterior in Equation (9) is to look for suiTABLE approximation followed by the appropriate optimization method. Based on Cawley and Talbot [22], a Bayesian approach by way defining a conjugate prior helps in reducing the overfitting. This combined with the optimization for the approximate distribution, variational method should in theory not only help in reducing the overfitting but also has lower computational burden than that of Bayesian approach [15].

To do regression coefficients using variational, we need to approximate the posterior in Equation (9) using variational distribution. Derivation for this has been done by Drugowitsch [19]. A brief review and discussion for the procedure is explained below.

First define the variational distribution to approximate the posterior distribution from Equation (9) as:

$$q(\boldsymbol{\beta}, \tau) = q(\boldsymbol{\beta}) q(\tau) \equiv p(\boldsymbol{\beta}, \tau \mid \boldsymbol{X}, \boldsymbol{y}) \tag{10}$$

The assumption here is that independence is hold for $\boldsymbol{\beta}$ and τ. The partition into 2 marginal distributions as noted in Equation (10) is based on the mean-field variational theory [15]. This concept essentially helps the optimization of the variational distribution. Using Equation (10) one needs to evaluate simpler problems, that is the 2 marginal variational distributions, to find the solution for the more complex problem, the joint variational distribution. Thus, to find the optimal variational distribution, evaluation of q(β) and q(τ) is performed. Using the previously defined posterior (10), evaluation of the distribution form for q(β) and q(τ) can be done by taking the expected logarithm of the posterior distribution relative to the parameter that is not being optimized. In other words, the logarithm of a marginal variational distribution is evaluated by computing expected logarithm of posterior distribution (9) regarding the parameters of the other marginal variational distribution [19].

In the case of the previous setup, the optimal variational distribution of q(τ) is,

$$\log q^*(\tau) = E_\beta \left[\log p(y, \boldsymbol{\beta}, \tau \mid \boldsymbol{X})\right] + constant \tag{11}$$

One can examine the Equation (11) yields the equation which leads to the logarithmic form of a gamma(α_i, β_i) distribution [19] with the parameters:

$$\alpha_i = \alpha_0 + \frac{c}{2}; \beta_i = \beta_0 + \frac{1}{2} E_{\boldsymbol{\beta}}[\boldsymbol{\beta}^t \boldsymbol{\beta}] \tag{12}$$

with c as the number of independent variables. The index i for the Equation (12) is used for the iterative procedure and to differentiate with the model parameters vector. Then, for the optimal of q($\boldsymbol{\beta}$) is:

$$\log q^*(\boldsymbol{\beta}) = E_\tau \left[\log p(\boldsymbol{y}, \boldsymbol{\beta}, \tau \mid \boldsymbol{X})\right] + constant$$

This leads to the fact that the optimal distribution for q(β) is in the form of multivariate normal distribution with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{S}$ is given by:

$$\boldsymbol{S} = \left(\frac{\alpha_i}{\beta_i}\boldsymbol{I} + \frac{1}{\sigma^2}\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}; \boldsymbol{m} = \frac{1}{\sigma^2}\boldsymbol{S}\boldsymbol{X}^t\boldsymbol{y} \tag{13}$$

in which $\frac{\alpha_i}{\beta_i}$ came from the parameters of $q^*(\tau)$ or the Equation (12). This also leads to the following equation to hold.

$$E_{\boldsymbol{\beta}}[\boldsymbol{\beta}^t\boldsymbol{\beta}] = \boldsymbol{m}^t\boldsymbol{m} + tr\,(\boldsymbol{S}) \tag{14}$$

To evaluate the variational distribution, start by initializing the value of parameters in either of the two marginal distribution q($\boldsymbol{\beta}$) or q(τ). Then, estimate the other one based on the initialization of the first distribution. Repeat the process using the result for each step until convergence criterion is satisfied.

Regarding convergence criterion, variational method used a function known as "lower bound" [23] to decide if an approximation is close enough. In the case of linear regression, Drugowitsch [19] derived the lower bound which can be written as the following equation:

$$\begin{aligned} L[q] = {} & E_{\boldsymbol{\beta}}[\log\, p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta})\,] + E_{\boldsymbol{\beta},\tau}[\log p(\boldsymbol{\beta}|\tau)\,] + E_\tau[\log p(\tau)\,] - E_\tau[\log q(\tau)] \\ & - E_{\boldsymbol{\beta}}[\log q(\boldsymbol{\beta})\,]. \end{aligned} \tag{15}$$

To compute the lower bound in Equation (15), the following results from Drugowitsch [19] can be used:

$$E_{\boldsymbol{\beta}}[\log\, p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta})\;\;] = \frac{n}{2}\log\left(\frac{\lambda}{2\pi}\right) - \frac{\lambda}{2}\boldsymbol{y}^t\boldsymbol{y} + \lambda\boldsymbol{m}^t\boldsymbol{X}^t\boldsymbol{y} - \frac{\lambda}{2}\,tr\left(\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{m}\boldsymbol{m}^t + \boldsymbol{S})\right) \tag{16}$$

$$E_{\boldsymbol{\beta},\tau}[\log p(\boldsymbol{\beta}|\tau)\,] = -\frac{c}{2}\log 2\pi + \frac{c}{2}(\psi(\alpha_i) - \log\,\beta_i\,) - \frac{\alpha_i}{2\,\beta_i}\left(\boldsymbol{m}^t\boldsymbol{m} + tr\,(\boldsymbol{S})\right) \tag{17}$$

$$E_\tau[\log p(\tau)] = \alpha_0 \log\beta_0 + (\alpha_0 - 1)(\psi(\alpha_i) - \log\beta_i\,) - \beta_0\frac{\alpha_i}{\beta_i} - \log\Gamma(\alpha_0) \tag{18}$$

$$E_\tau[\log q(\tau)\,] = -\log\Gamma(\alpha_i) + (\alpha_i - 1)\psi(\alpha_i) + \log\beta_i - \alpha_i \tag{19}$$

$$E_{\boldsymbol{\beta}}[\log q(\boldsymbol{\beta})\,] = -\frac{1}{2}\log|\boldsymbol{S}| - \frac{c}{2}\,(1+\log 2\pi\,) \tag{20}$$

n in Equation (16) represents the number of observations. In Equations (16) to (20) λ is used as the precision parameter which is equal to the inversed value of variance of error term. This definition is chosen for mathematical purpose based on Drugowitsch [19]. $\psi(\_)$ represents digamma distribution while $\Gamma(\_)$ Represents gamma distribution. Utilizing Equations (16) to (20), the lower bound in Equation (15) can be computed.

The lower bound of Equation (15) is the equivalence of the general lower bound for variational mean-field approach in [15]. This lower bound essentially measures how close our variational method approximation to the posterior distribution of the linear regression is. Thus, assigning a value B to this lower bound such that $L[q] < \epsilon$ equivalent to stating the minimum "distance" of our approximation to the posterior distribution [15] of linear regression.

In summary, the procedure of the Variational method in estimating the parameters of linear regression is as follow:

- Initiates the value for α_0 and β_0.
- Compute the variational parameters using Equations (12), (13), and (14).
- Compute lower bound function of Equation (15) using the Equations (16) to (20).
- The lower bound of Equation (15) helps in determining the closeness of the variational approximation to the posterior distribution in Equation (9). Thus, set a value ϵ such that if $L[q] < \epsilon$, the iteration stops.
- If the iteration stops, then the regression parameters estimates can be extracted from the variational distribution.
- Else, the procedure restarts from computing the variational parameters in Equations (12), (13), and (14) until $L[q] < \epsilon$.

3 Results and discussion

First, the simulation in this paper is done using the assistance of R studio with R language version 4.2.1. Most of the functions used in the simulation are available in default installment except for a better plotting function that used ggplot2 package, tidying up and data transformation package with tidyr and dplyr, lastly matrix calculation utilized matrixcalc package.

This simulation study consists of 3 main parts: single estimation, training-and-test simulation, and replicated estimation. The first simulation aims to study the comparison of estimation from the 2 methods. This part will do comparison in 3 different models with each model the coefficient is estimated on 3 different sample sizes in which the choice follows Morris et al. [24]. The second simulation aims to do comparison of the models' prediction accuracy for new data built from the 2 methods. The third simulation studies the pattern consistency of both estimation which can viewed as an extension of the first simulation.

To do the comparison, the measurements to be used are the estimates value and MSE. These measures are also used in a similar comparison study for variational method by Hardouin [17]. The two measures are the easiest indication of how good the estimates are [2]. In addition, the average of squared differences between the true parameters value and the estimates is used following [17] and [19] in order to analyze the consistency of the estimate's quality. Lastly, to study the closeness of the parameter estimates, graphical representations of the value estimates for the two methods are also used. These graphics are to be used to illustrate the distance between each of the estimate's value from both methods.

To begin the simulation, the true model is set up such that generating random sample for both dependent and independent variables is possible. For the first until third simulation, the model is as stated in Equation (1) with the limitation of one independent variable and one intercept term. The independent variable is generated randomly from standard normal distribution while the dependent variable is generated using the regression model as in Equation (1) with the error term is assigned to follow normal distribution with variance characterized using precision parameter λ. The precision parameter is the inversed value of variance. The formal relationship between precision and the variance of error term is $\frac{1}{\lambda} = \sigma^2$. This relationship is chosen for computational purpose based on Drugowitsch [19]. Then, for each simulation, both methods are used to estimate the regression coefficients for the corresponding model. The first simulation compares the estimates of the regression coefficient from the 2 methods. Similarly, the third simulation also compares the estimates of the regression coefficient from the 2 methods. The difference is that the third simulation compares the estimates in a cumulative way that comes from the replication. Meanwhile, the second simulation compares the accuracy of prediction value for the dependent variable based on new input of independent variable.

To understand the behavior of the estimation, this simulation does some combinations of parameters which characterize 3 different simple linear regression models. First, which in the following section referred as model 1, has the true value the regression parameters set as $\beta_0 = \beta_1 = \lambda = 1$. Thus, model 1 represents the simple linear regression model with rather small variance. For model 2, the true values of the regression parameters are set as $\beta_0 = 0$ and $\beta_1 = \lambda = 1$. Thus, model 2 represents simple linear regression model without intercept. Lastly, model 3 has the following value for the true regression parameters: $\beta_0 = \beta_1 = 1$ and $\lambda = \frac{1}{9}$. Thus, model 3 represents a simple linear regression model with a relatively large variance in the data. To summarize the model specification, Table 1 shows the true value of each model's parameters.

Using the estimation procedure from previous study, a function is built in the R language to estimate the regression coefficients using variational method. The initialization used for the variational parameters are $\alpha_0 = 0.1$ and $\beta_0 = 0.01$, precision parameter λ is assigned according to the model, and tolerance value for the lower bound is set at 0.000001. Meanwhile, to estimate the regression coefficients using MLE method, a function from default package in the R Studio software namely lm() function is used. Then, using these 2 functions, fitting into regression model and parameter estimation procedure can be done.

3.1 First simulation: Single estimation comparison

For this simulation comparison between the two methods is done using the result of the estimation value themselves and using the mean of squared error (MSE). The evaluation of performance is done by looking at the difference between estimation and the true value of the parameters and the value of MSE.

This simulation study does 3 different numbers of sample sizes n that is 50, 100, and 1000 for each of the models. This is done to see if there is any difference in accuracy of the estimation as the sample size grows. The data is generated independently for each model in each sample size. For example, simulation for model 1 with sample size n = 50, data used for this is different and independent from the data used in simulation for model 1 with sample size $n = 100$ or model 2 with sample size $n = 50$.

For model 1 with the set of parameters is $\beta_0 = \beta_1 = \lambda = 1$, the simulation and estimation procedures yield the results seen in Table 2 and Table 3.

Table 1. Estimation comparison for β_0 for Model 1.

Model	β_0	β_1	λ
1	1	1	1
2	0	1	1
3	1	1	1/9

Table 2. Estimation comparison for β_0 for Model 1.

N	True value	MLE	Variational
50	1	1.3111	1.2936
100	1	0.9591	0.9483
1000	1	0.9552	0.9542

Table 3. Estimation comparison for β_1 for Model 1.

N	True value	MLE	Variational
50	1	0.9982	0.9886
100	1	0.8832	0.8735
1000	1	1.0307	1.0296

In the case of accuracy of estimation for linear regression parameters in the first simulation, both MLE and variational do not have any apparent difference between each other. However, there is an interesting finding. Even though the accuracy of estimation gets better as the sample size grows, variational method performs slightly better than MLE in all the sample size except for $n = 50$. In this finding, one can relate this to the previously mentioned fact about the problem of overfitting in MLE. The variational method is more likely to sacrifice its fit to the observed data but compensate it in a way to avoid overfitting. This can also be observed in Table 4 of MSE comparison.

Referring to the result of Table 4, it is very tempting to conclude that MLE has a better fit than variational method. Thus, one may also conclude that MLE has better accuracy than variational. To check whether this is true in the case of predicting new data point, the later simulation will do a training-to-test simulation which represents a situation for predicting new data.

Since the first simulation for model 1 is done, the following are results from the simulation for model 2 and 3. For simplicity, the results are shown in Table 5 to Table 7.

A slightly different pattern from the first simulation. In terms of accuracy of parameters estimation, variational method has a slight upper hand compared to MLE. On the other hand, the measure of fit compared using MSE shows a similar result from the first simulation. So far, there is no apparent advantage in terms of parameter estimates of Variational method relative to method of MLE. However, note that Variational tends to have slightly better

estimates in the case of sample size $n = 50$ and $n = 100$ compared to MLE. This advantage will be discussed in the latter part of this section.

Table 4. MSE comparison for Model 1.

N	MLE	Variational
50	1.3264	1.3269
100	0.8640	0.8643
1000	1.0198	1.0198

Table 5. Estimation comparison for β_0 for Model 2.

N	True value	MLE	Variational
50	0	-0.0268	-0.0281
100	0	0.1219	0.1201
1000	0	0.0343	0.0341

Table 6. Estimation comparison for β_1 for Model 2.

N	True value	MLE	Variational
50	1	1.2070	1.1508
100	1	0.9328	0.9078
1000	1	1.0023	1.0001

Table 7. MSE comparison for Model 2.

N	MLE	Variational
50	0.7535	0.7555
100	0.9524	0.9529
1000	1.0384	1.0384

The simulation for model 3 seen in Table 8 to Table 10 yields a slight difference of performance pattern than two previous simulations. A clear difference between MLE and variational estimation shows at the sample size of $n = 100$. In this sample size, variational method appears to underestimate the value of slope or the β_1. Meanwhile, in the case of $n = 50$ and $n = 1000$, variational method yields a slightly better estimation than MLE.

Measure of fit is still the same as the two previous simulations, that is showing superiority of MLE.

From the simulation for the 3 models above there is another finding that is a pattern where both MLE and variational method yield estimates that get closer in value as the sample size grows larger. Thus, a study to see whether this pattern is only coincidence or not, simulation for several trials would be conducted. This study is done later in the third simulation.

Table 8. Estimation comparison for β_0 for Model 3.

N	True value	MLE	Variational
50	1	1.3890	1.2200
100	1	0.9634	0.8500
1000	1	1.0310	1.0216

Table 9. Estimation comparison for β_1 for Model 3.

N	True value	MLE	Variational
50	1	1.1551	1.0454
100	1	0.8026	0.6850
1000	1	1.0030	0.9939

Table 10. MSE comparison for Model 3.

N	MLE	Variational
50	8.3730	8.4152
100	10.8014	10.8250
1000	9.4205	9.4207

3.2 Second simulation: Training-and-test simulation comparison

From the first simulation, there is one question that needs to be asked. That is whether variational method in simple linear regression reduces the overfitting tendency which was mentioned in the previous section. The next simulation procedure uses the simulation type that utilizes training-to-test dataset. The procedure is similar to the first simulation. However, instead of generating data then fit it into the linear regression, this simulation generates 2 types of data. The first data is data used to estimate the regression coefficients, while the second data is used to do the prediction accuracy comparison. The first data is called training dataset and the second data is test dataset. For each data, the sample size is set at $n = 1000$.

To generate independent variable X for both training and test data, randomization is done using the standard normal distribution. Meanwhile, for dependent variable y, differentiation for generating the data is done. For training data, y is generated using the regression equation such that it has the form of Equation (2) with error term randomized from normal distribution

characterized by the precision parameter corresponding to the model. On the other hand, y in test data is generated using the regression equation but the value of the standard deviation is 1.2. This differentiation is done to mimic the situation in the application of linear regression such as [10]. That is, building model data using observed sample data and using the model to predict target variable using new input value.

The comparison is done for the previous 3 different settings of the linear regression parameter. Then, for each model, prediction accuracy comparison is done in 3 trials using the measure of squared difference between the true value and the predicted value. Obviously, for each of these trials, it always begins with data generation.

For the prediction comparison, Table 11 to Table 13 show the comparison for the squared residuals of test dataset for the 3 models. As shown in Table 10, residual for variational method estimation is rather close to MLE. This finding is obviously not a good indication to support the argument of using variational method in simple linear regression to overcome overfitting problem. However, this is not in any way saying that variational method is not an alternative solution to overfitting. This finding happens as the consequence of this simulation study limitation, which is by nature of simple linear regression: can hardly spark overfitting problem. In other words, simple linear regression is not a model that "forces" its fit into the data since there are only so many parameters in the model. Thus, answering the question in the first line of this sub section, Variational method does not help overfitting in simple linear regression. That is because reducing what is already simplistic does not help improving the model itself.

Table 11. Average of squared residuals comparison in a single estimation for Model 1.

Trial	MLE	Variational
1	1.4083	1.4083
2	1.4215	1.4137
3	1.4213	1.4285

Table 12. Average of squared residuals comparison in a single estimation for Model 2.

Trial	MLE	Variational
1	1.4456	1.4456
2	1.5263	1.5259
3	1.5201	1.5199

Table 13. Average of squared residuals comparison in a single estimation for Model 3.

Trial	MLE	Variational
1	1.4482	1.4508
2	1.4513	1.4518
3	1.4130	1.4095

From the first and second simulation however, a finding that is found similar to the previous simulation is that MLE and variational estimates for regression coefficients tend to have similar estimates values as the sample size grows. The third simulation focuses on visualizing this finding from some replicated trials.

3.3 Third simulation: Replicated estimation simulation comparison

The performance consistency of variational and ML method so far is that both methods yield estimates which values get closer as the sample size grows. This can be seen in the result of the first simulation. For model 1 and 2, this pattern is very apparent. However, it does not look as apparent for model 3. Thus, for this sub section, the simulation focuses on performance comparison of variational and ML method for model 3.

The setting of the true value for the regression parameters of this simulation is the same as the previously mentioned model 3. In other words, this last simulation uses the setting of $\beta_0 = \beta_1 = 1$ and $\lambda = \frac{1}{9}$. as the true value of the regression parameters. This simulation compares the behavior of maximum likelihood and variational method estimates in 3 different sample sizes that is $n = 50, 100,$ and 1000. For each of these settings, 100 replications of estimation comparison similar to the first simulation is done. This simulation compares the relative distance of the estimates from the 2 estimation methods and the true value of the regression parameters.

First, look at the following graphic of parameter estimates for β_0 from the 2 parameter estimation methods in 100 replications with sample size n = 50.

Figure 1 shows that for the sample size of 50, MLE and variational methods do not have clear differences in term of how close the estimates to the true value. This is of course only true visually from the difference between the 2 estimates that can be seen in Table 14. From this computation, maximum likelihood method appears superior compared to variational method.

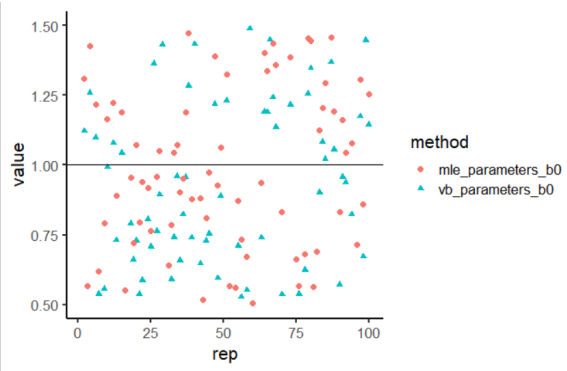

Fig. 1. Plot of estimated β_0 in 100 replications from MLE and variational methods for n = 50 (Horizontal black line shows the true value of β_0).

Table 14. Average of squared difference of estimates and true value of β_0 for n = 50 for MLE and variational method in 100 replications.

MLE	Variational
0.2097	0.2401

However, once again the variational method shows a significant improvement on the accuracy of parameter estimation as the sample size grows larger. This can be seen from Fig. 2 and Fig. 3.

This improvement in terms of parameter estimation accuracy can be seen clearly by computing the average squared difference between the estimates and the true value of the parameter in the 100 replications seen in Table 15. This finding shows that variational method estimates will be close to the ML estimates as the sample size grows larger.

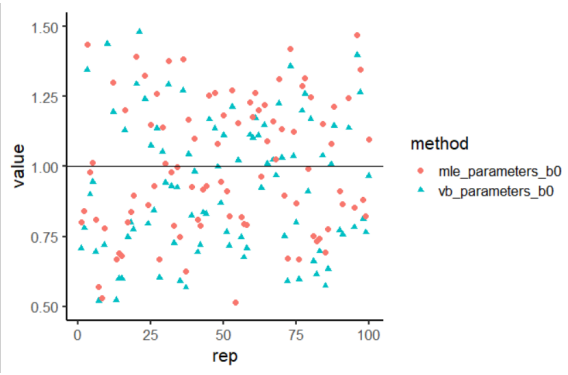

Fig. 2. Plot of estimated β_0 in 100 replications from MLE and variational methods for n = 100 (Horizontal black line shows the true value of β_0).

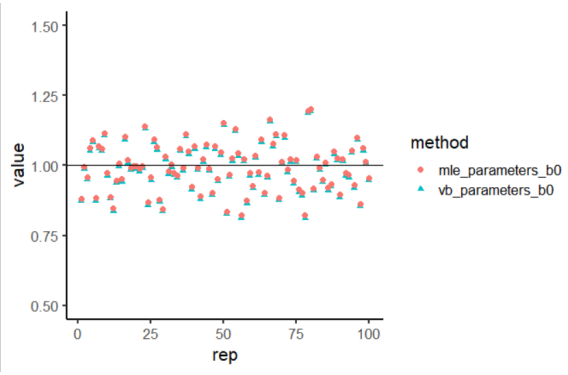

Fig. 3. Plot of estimated β_0 in 100 replications from MLE and variational methods for n = 1000 (Horizontal black line shows the true value of β_0).

Table 15. Average of squared difference of estimates and true value of β_0 for n = 100 and 1000 for MLE and variational method in 100 replications.

N	MLE	Variational
100	0.0853	0.0902
1000	0.0071	0.0072

A similar pattern is also found in the case of estimates for β_1 which can be seen in Fig. 4 to Fig. 6 and Table 16.

From Fig. 4 to Fig. 6, the pattern of closeness as the sample size grows is obvious. This is also supported by the Table 15. While this sounds like a very intuitive finding, the

behavior of variational method in small sample size can be of consideration. Recall that variational method uses semi-Bayesian approach which in nature considers both the data and prior. When the data is big enough, the Bayesian approach will consider the data more than prior. However, if the data is small, Bayesian maintains its flexibility in estimating the regression parameters. Thus, in the case of a small sample, the Bayesian approach does not overfit the model to the data.

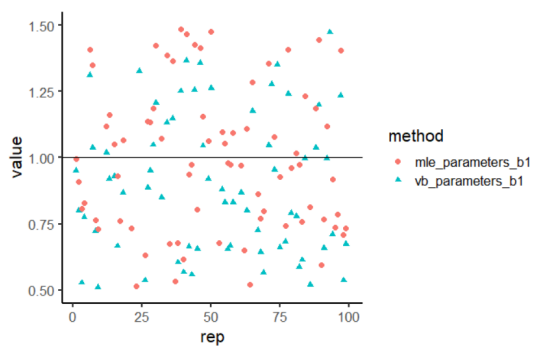

Fig. 4. Plot of estimated β_1 in 100 replications from MLE and variational methods for n = 50 (Horizontal black line shows the true value of β_1).

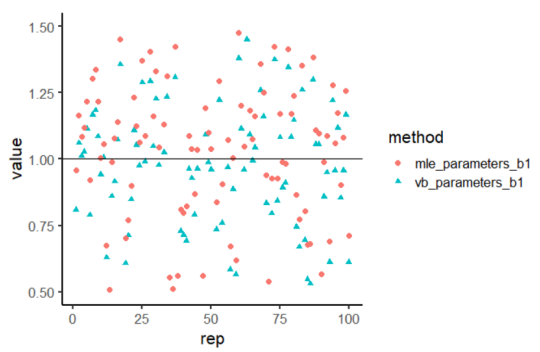

Fig. 5. Plot of estimated β_1 in 100 replications from MLE and variational methods for n = 100 (Horizontal black line shows the true value of β_1).

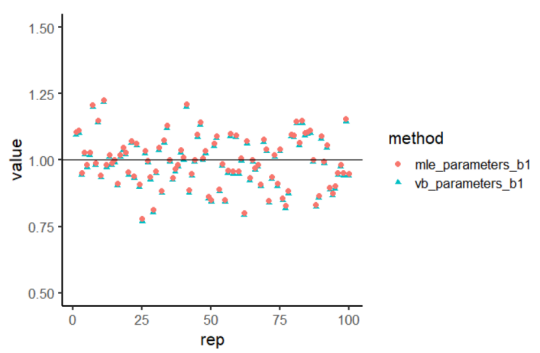

Fig. 6. Plot of estimated β_1 in 100 replications from MLE and variational methods for n = 1000 (Horizontal black line shows the true value of β_1).

Table 16. Average of squared difference of estimates and true value of β_1 for n = 100 and 1000 for MLE and variational method in 100 replications.

N	MLE	Variational
50	0.2146	0.2242
100	0.0947	0.0952
1000	0.0093	0.0091

4 Conclusion and recommendation

This paper discusses a more recent method of estimation for linear regression parameters known as the variational bayes method. Variational method utilized the framework of Bayesian statistics approach in order to estimate the parameters of the linear regression. By utilizing Bayesian approach, estimation procedure in variational method does not rely solely on the available data but also acknowledging a randomness nature of the value of the parameters themselves. As such, variational method benefitted from the Bayesian approach in a way that it does not overfit to the data.

Theoretical review suggests that Variational method has a unique characteristic. Quoting from Blei et al. [15] Variational method has similar the flexibility of that Bayesian approach in a sense that the estimation procedure does not rely on the data by itself. Combining Bayesian approach and optimizing the approximation based on the posterior distribution, Variational method can reduce the overfitting in the linear regression model while maintaining the easy-to-use of the model.

Variational method also has other uniqueness where it does not need to compute the whole distribution of posterior such as in Bayesian approach. Instead, it only optimizes a family of posterior distribution which is referred by Blei [15] as variational family. By optimizing this variational family, one can estimate the joint posterior of the actual target distribution. In the case of linear regression, the optimized variational distributions, assuming a normal conjugate prior for the regression parameters and gamma distribution for the hyper-parameter, will have the form of gamma and normal distributions.

These two characteristics have great significance in the field of statistics. Jaakkola & Jordan [16] use Variational method as the alternative to estimate the logistic regression parameters. It is found that the advantage of using Variational method as the parameters estimation method in logistic regression is the reduction of the computation required compared to some other iterative approximation method. Hardouin [17] and Rizka [18] implemented the Variational method for spatial model. The findings of the research are that Variational method has more efficient computation and able to yield a more accurate parameters estimation in a data with large variance. The latter finding is similar to what has been found in the simulation done in this paper.

In the simulation study, it is found that variational method yields estimate which value gets closer to MLE method estimate as the sample size grows. This finding is found in the case of simple linear regression model. One thing that needs to be highlighted is that, in the case of data with small sample size or large variance, it is found that Variational method tends to be more accurate than MLE. This finding is closely related to the fact that Variational method in estimating the linear regression parameters utilizes Bayesian Framework [19]. In summary, when the data that is available is small or has large variance, Variational method tends to have better parameters estimate.

As for reduction of overfitting in simple linear regression, variational method is constrained to the form of the model itself. That is, since the model is not overly complex to begin with, variational method does not necessarily reduce overfitting in simple linear regression. However, in the case of a more complex linear regression, different findings may be found. Thus, a recommendation to improve the comparison study is to try another model such as multiple linear regression with more than 3 independent variables.

Acknowledgements

This study was funded by the Directorate of Research and Development, Universitas Indonesia (DRPM UI) as an additional output of the International Indexed Publication Grant (PUTI) Q2 2022—2023 No: NKB-668/UN2.RST/HKP.05.00/2022.

References

1. K. Kumari and S. Yadav, J. Pract. Cardiovasc. Sci. **4**, 33–36 (2018).
2. W. Mendenhall and T. T. Sincich, *A Second Course in Statistics: Regression Analysis* (Pearson Education, 2020).
3. P. Ali and A. Younas, Evid. Based. Nurs. **24**, 116–118 (2021).
4. D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis* (John Wiley & Sons, 2021).
5. M. A. Babyak, Psychosom. Med. **66**, 411–421 (2004).
6. X. Ying, J. Phys. Conf. Ser. **1168**, 022022 (2019).
7. K. Takezawa et al., Open J. Stat. **2**, 309–312 (2012).
8. C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, Vol. 4 (Springer, New York, 2006).
9. L. Dalicandro, J. A. Harder, D. Mazmanian, and B. Weaver, Quant. Meth. Psych. **17**, 1–6 (2021).
10. A. Pacifico, Econom. Rev. **40**, 148–176 (2021).
11. R. Jones, C. Klemenjak, S. Makonin, and I. V. Bajic, preprint arXiv:2009.07756 [cs.AI] (2020).
12. J. Kolluri, V. K. Kotte, M. Phridviraj, and S. Razia, *Proceeding of the 2020 4th International Conference on Trends in Electronics and Informatics* (48184) (IEEE, 2020), pp. 934–938.
13. W. M. Bolstad and J. M. Curran, *Introduction to Bayesian Statistics* (John Wiley & Sons, 2016).
14. D. E. Jonas et al., 13-EHC039-EF (Agency for Healthcare Research and Quality, US, 2013).
15. D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, J. Am. Stat. Assoc. **112**, 859–877 (2017).
16. T. S. Jaakkola and M. I. Jordan, *Proceeding of the Sixth International Workshop on Artificial Intelligence and Statistics* (PMLR, 1997), pp. 283–294.
17. C. Hardouin, Spat. Stat. **31**, 100365 (2019).
18. H. Rizka and Y. Widyaningsih, AIP Conf. Proc. **2374**, 030007 (2021).
19. J. Drugowitsch, preprint arXiv:1310.5438 [stat.ML] (2013).

20. T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York, 2009).
21. M. van Oijen, in *Bayesian Compendium* (Springer Nature, Switzerland, 2020) pp. 17–21.
22. G. C. Cawley and N. L. Talbot, J. Mach. Learn. Res. **8**, 841−861 (2007).
23. T. P. Morris, I. R. White, and M. J. Crowther, Stat. Med. **38**, 2074–2102 (2019).
24. M. J. Beal, Ph.D. Thesis, University College London, United Kingdom, 2003.