Journal of Research in Mathematics Trends and Technology (JoRMTT) Vol. 3, No. 1, 2021 | 34-39

Journal of Research in Mathematics Trends and Technology

Application of Expectation Maximization Algorithm in Estimating Parameter Values of Maximum Likelihood Model

J. E. Simarmata[1*], H. Bete[1], and S. A. Purba[2]

[1]Mathematics Education Study Program, Universitas Timor, Kefamenanu, 85613, Indonesia
[2]Department of Mathematics, Universitas HKBP Nomensen, Pematangsiantar, 21132, Indonesia

Abstract. Parameter estimation is an estimation of the population parameter values based on data or samples of a population. Parameter estimation can be solved by several methods, one of which is the Maximum Likelihood method. The focus of this research is to estimate the parameter value of normal distribution data with the Maximum Likelihood based on the iteration algorithm. The iteration algorithm that will be used is the Expectation Maximation Algorithm with help of the Matlab 2016a program. Based on the results obtained that the estimation value of the parameter μ and σ^2 for an accident data in Indonesia based on age group with using Expectation-Maximization algorithm is $\mu = 1725{,}176$ and $\sigma^2 = 1445795{,}2$ with 2 iterations.

Keyword: Parameter Estimation, Maximum Likelihood, EM Algorithm.

Abstrak. *Estimasi parameter adalah penaksiran terhadap nilai-nilai parameter populasi berdasarkan data atau sampel yang diambil dari populasi. Estimasi parameter dapat dilakukan dengan beberapa metode, salah satu diantaranya dengan metode Maximum Likelihood. Fokus dari penelitian ini adalah mengestimasi nilai parameter suatu data berdistribusi normal dengan Maximum Likelihood berdasarkan algoritma iterasi. Algoritma iterasi yang akan digunakan adalah algoritma Algoritma Expectation Maximation dengan bantuan program Matlab 2016a. Berdasarkan hasil yang diperoleh bahwa nilai estimasi parameter μ dan σ^2 untuk sebuah data kecelakaan yang terjadi di Indonesia berdasarkan kelompok umur dengan menggunakan algoritma Expecation Maximization adalah $\mu = 1725{,}176$ dan $\sigma^2 = 1445795{,}2$ dengan jumlah iterasi sebanyak 2.*

Kata Kunci: *Estimasi Parameter, Maximum Likelihood, Algoritma EM*

Received 12 February 2021 | Revised 23 February 2021 | Accepted 26 February 2021

1. Introduction

Parameter estimation is an estimation of the population parameter values based on data or samples of a population. There are several methods for estimating the parameters of a generalized linear model, including Maximum Likelihood Estimation (MLE) using a distribution approach by maximizing the likelihood function [1].

Maximum Likelihood Estimator (MLE) is a method for estimating parameters from a data set that follows a certain distribution [2], [3]. Generally, the maximum of a function cannot be

*Corresponding author at: Mathematics Education Study Program, Universitas Timor, Kefamenanu, 85613, Indonesia

E-mail address: justinesimarmata@unimor.ac.id

 e-ISSN: 2656-1514, DOI: 10.32734/jormtt.v3i1.8331
Journal Homepage: https://talenta.usu.ac.id/jormtt

solved analytically if there is an implicit and nonlinear form, so it can be solved using the Expectation-Maximization Algorithm [4].

The Expectation-Maximization (EM) algorithm is a commonly used algorithm to calculate maximum likelihood estimates used for situations that include missing observations. According to [5], [6] EM algorithm is one of the algorithms that used for data classification or grouping. The EM algorithm can be interpreted as an algorithm that functions to find the estimated value of the Maximum Likelihood of the parameters in a probabilistic model [7]–[9].

Research on estimating parameter values using the Maximum Likelihood method has indeed received great attention from researchers [10]. The focus of this research is to estimate the parameter values of Maximum Likelihood model based on the EM algorithm with the aim to know the estimated value of the algorithm.

2. Research Methods

The steps of this research that related to the purpose is to follow:

1. Collecting data from the Indonesian National Police Korlantas regarding the number of accidents in Indonesia.
2. Checking using the SPSS program to find out the type of data distribution.
3. Find the likelihood function
4. Find the log-likelihood function
5. Estimating parameters using the EM algorithm using the help of the Matlab 2016a program

3. Results and Discussion

3.1. Table Accident Data in Indonesia by Age Group

In this paper, parameter estimation will be carried out on accident data obtained from the Indonesian Police Korlantas. In estimating parameters using the Maximum Likelihood method, it is necessary to know the type of distribution of the accident data. This paper checks the type of population distribution of the accident data using the SPSS program. So that it is obtained that the data is normally distributed, this is indicated by obtaining of significance value is 0,124 (0,124 > 0,05) it can be said that the data is normally distributed.

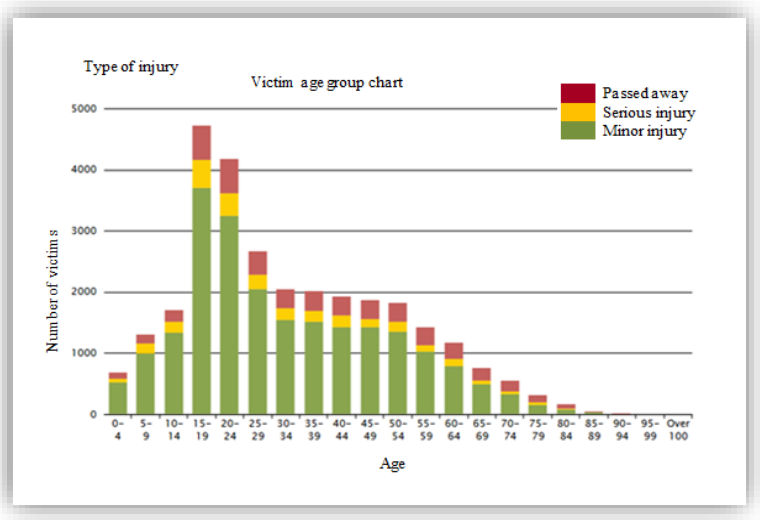

Figure 1. Graph of accident data in Indonesia by age group

3.2. Graphic Parameter Estimation in Accident Data Using Maximum Likelihood Method

Parameter estimation using the maximum likelihood required the value of the population distribution of the accident data. By using SPSS it is known that the data is normally distributed. The following is a joint probability density function (pdf) of a normal distribution with parameters μ and σ^2 is

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

So that the likelihood and log-likelihood functions from the normal distribution can be obtained as follows:

The likelihood function of the normal distribution is

$$\begin{aligned} L(\theta|y_1, y_2, \dots, y_N) &= \prod_{i=1}^{N} L(\theta|y_i) = \prod_{i=1}^{N} f(y_i) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}} \\ &= (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\sum_{i=1}^{N} \frac{(y_i-\mu)^2}{2\sigma^2}\right) \end{aligned} \tag{2}$$

The log-likelihood function of the normal distribution is

$$\begin{aligned} l(\theta|y_1, y_2, \dots, y_N) &= \log L(\theta|y_1, y_2, \dots, y_N) = \log\left[\prod_{i=1}^{N} L(\theta|y_i)\right] \\ &= \ln\left[(2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\sum_{i=1}^{N} \frac{(y_i-\mu)^2}{2\sigma^2}\right)\right] \\ &= -\frac{N}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{N} \frac{(y_i-\mu)^2}{2\sigma^2} \end{aligned} \tag{3}$$

In this research, parameter estimation will be carried out μ and σ^2 based on the EM Algorithm.

3.3. Expectation-Maximization Algorithm (EM Algorithm)

Parameter estimation with EM algorithm has two stages, namely the expectation stage (E-step) then the maximization stage (M-step). Parameter estimation using the EM algorithm uses the maximum likelihood. Therefore, the likelihood function and the log-likelihood distribution of a population are needed. Which in this case obtained the likelihood function and the log-likelihood of the observed distribution. The difference is that the EM algorithm is an estimation process for missing data.

- Expectation Stage (E-Step)

At this stage, the Q function is needed as follows:

$$\begin{aligned} Q\left(\mu,\sigma^2\middle|\mu^{(t)},\sigma^{2(t)}\right) &= E_{x|y,\mu^{(t)},\sigma^{2(t)}}[\log f(X|\theta)] \\ &= E_{x|y,\mu^{(t)},\sigma^{2(t)}}\left[-\frac{N}{2}\log(2\pi\sigma^2)-\sum_{i=1}^{M}\frac{(y_i-\mu)^2}{2\sigma^2}-\sum_{i=M+1}^{N}\frac{(y_i-\mu)^2}{2\sigma^2}\right] \\ &= -\frac{N}{2}\log(2\pi\sigma^2)-\sum_{i=1}^{M}\frac{(y_i-\mu)^2}{2\sigma^2}-\sum_{i=M+1}^{N}\frac{E_{x|y,\mu^{(t)},\sigma^{2(t)}}(y_i-\mu)^2}{2\sigma^2} \end{aligned} \quad (4)$$

In this case

$$\begin{aligned} E_{x|y,\mu^{(t)},\sigma^{2(t)}}(y_i-\mu)^2 &= E_{x|y,\mu^{(t)},\sigma^{2(t)}}({y_i}^2-2y_i\mu+\mu^2) \\ &= \left(\mu^{(t)}\right)^2+\sigma^2-2\mu^{(t)}\mu+\mu^2 \end{aligned} \quad (5)$$

So

$$Q\left(\mu,\sigma^2\middle|\mu^{(t)},\sigma^{2(t)}\right) = -\frac{N}{2}\log(2\pi\sigma^2)-\sum_{i=1}^{M}\frac{(y_i-\mu)^2}{2\sigma^2}-\frac{n-m}{2\sigma^2}\left[\left(\mu^{(t)}\right)^2+\sigma^2-2\mu^{(t)}\mu+\mu^2\right] \quad (6)$$

Maximization Stage (M-Step)

At this stage the Q function is maximized which in this case is done by performing the first derivative of each parameter as follows:

$$\frac{\partial}{\partial\mu}Q\left(\mu\middle|\mu^{(t)}\right)=0 \text{ and } \frac{\partial}{\partial\sigma^2}Q\left(\sigma^2\middle|\sigma^{2(t)}\right)=0 \quad (7)$$

So

$$\begin{aligned} \mu^{(t+1)} &= \frac{\sum_{i=1}^{m}y_i+(n-m)\mu^{(t)}}{n} \\ \sigma^{2(t+1)} &= -\frac{N}{2\sigma^2}+\frac{\sum_{i=1}^{m}(y_i-\mu)^2}{2\sigma^4} \end{aligned} \quad (8)$$

The EM algorithm will be carried out using the help of the Matlab Program. The following are the results of the estimated parameter values from an accident data which are normally distributed based on the EM Algorithm assisted by Matlab:

Table 1. Expectation Maximization Iteration Results

Iteration	μ	σ^2
1	1725.1764706	3669402.4705882
2	1725.1764706	1445795.2041522

Based on the results, the parameter estimation value obtained μ and σ^2 for accident data that occurred in Indonesia based on age group using the EM algorithm is $\mu = 1725{,}176$ and $\sigma^2 = 1445795{,}2$ with the number of iterations is 2.

4. Conclusions

From the results and discussion, it can be said that the parameter estimation results μ and σ^2 of an accident data that occurred in Indonesia based on age groups using the EM algorithm is $\mu = 1725{,}176$ and $\sigma^2 = 1445795{,}2$ with 2 iterations.

REFERENCES

[1] L. J. Bain and M. Engelhardt, Statistical Analysis of Reliability and Life-Testing Models. Routledge, 2017.

[2] S. A. Purba, "Estimasi Parameter Data Berdistribusi Normal Menggunakan Maksimum Likelihood Berdasarkan Newton Raphson," J. Sains Dasar, vol. 9, no. 1, pp. 16–18, 2020.

[3] H. Retnawati, "PERBANDINGAN ESTIMASI KEMAMPUAN LATEN ANTARA METODE MAKSIMUM LIKELIHOOD DAN METODE BAYES," J. Penelit. dan Eval. Pendidik., vol. 19, no. 2, pp. 145–155, Oct. 2015, doi: 10.21831/pep.v19i2.5575.

[4] A. I. Yunita, A. K. Jaya, and G. M. Tinungki, "Model Regresi Bivariate Zero-Inflated Poisson Pada Kematian Ibu dan Bayi," ESTIMASI J. Stat. Its Appl., vol. 3, no. 1, pp. 33–40, 2022.

[5] S. Suhada, G. L. Ginting, and Hondro, "Penerapan Data Mining Untuk Memprediksi Besarnya Pembayaran Pajak Kendaraan Pada: (Badan Pengelolaan Pajak Dan Retribusi Daerah Upt Samsat Medan Selatan) Menggunakan Algoritma Expectation Maximization," Bull. Inf. Technol., vol. 2, no. 2, pp. 69–75, 2021.

[6] D. Mitra, D. Kundu, and N. Balakrishnan, "Likelihood analysis and stochastic EM algorithm for left truncated right censored data and associated model selection from the Lehmann family of life distributions," Japanese J. Stat. Data Sci., vol. 4, no. 2, pp. 1019–1048, Dec. 2021, doi: 10.1007/s42081-021-00115-1.

[7] N. Atikah, S. Rahardjo, D. L. Afifah, and N. Kholifia, "Modelling Spatial Spillovers of regional economic growth in East Java: an empirical analysis based on Spatial Durbin Model," J. Phys. Conf. Ser., vol. 1872, no. 1, p. 012029, May 2021, doi: 10.1088/1742-6596/1872/1/012029.

[8] F. Mufidah, I. Susanto, and E. Zukhronah, "Pengelompokan Negara Berdasarkan Populasi Urban Dengan Algoritma Expectation Maximization," Pros. SNAST, pp. 94–100, 2021.

[9] M. M. Lucini, P. J. van Leeuwen, and M. Pulido, "Model Error Estimation Using the Expectation Maximization Algorithm and a Particle Flow Filter," SIAM/ASA J.

Uncertain. Quantif., vol. 9, no. 2, pp. 681–707, Jan. 2021, doi: 10.1137/19M1297300.

[10] L. Novais and S. Faria, "Comparison of the EM, CEM and SEM algorithms in the estimation of finite mixtures of linear mixed models: a simulation study," Comput. Stat., vol. 36, no. 4, pp. 2507–2533, Dec. 2021, doi: 10.1007/s00180-021-01088-1.