BAREKENG: Journal of Mathematics and Its Applications
March 2023 Volume 17 Issue 1 Page 0487–0494
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss1pp0487-0494

ESTIMATION OF MAXIMUM LIKELIHOOD WEIGHTED LOGISTIC REGRESSION USING GENETIC ALGORITHM (CASE STUDY: INDIVIDUAL WORK STATUS IN MALANG CITY)

Dahlia Gladiola Rurina Menufandu[1*], Rahma Fitriani[2], Eni Sumarminingsih[3]

[1,2,3]Departemen of Statistics, Faculty of Mathematics and Natural Sciences, Brawijaya University
Jl. Veteran Ketawanggede Kecamatan Lowokwaru, Kota Malang, 65145, Indonesia

*Corresponding author's e-mail: * **menufandu.dahlia@gmail.com***

Article History:

Received: 21st November 2022
Revised: 4th February 2023
Accepted: 15th February 2023

Keywords:

Genetic Algorithm;
Weighted Logistic Regression

ABSTRACT

Weighted Logistic Regression (WLR) is a method used to overcome imbalanced data or rare events by using weighting and is part of the development of a simple logistic regression model. Parameter estimation of the WLR model uses Maximum Likelihood estimation. The maximum likelihood parameter estimator value is obtained using an optimization approach. The Genetic algorithm is a computational optimization algorithm that is used to optimize the estimation of model parameters. This study aims to estimate the Maximum Likelihood Weighted Logistic Regression with the applied genetic algorithm and determine the significant variables that affect the working status of individuals in Malang City. The data used results from data collection from the National Labor Force Survey of Malang City in 2020. The results of the analysis show that the variable education completed and the number of household members have a significant effect on individual work status in Malang City.

How to cite this article:

D. G. R. Menufandu, R. Fitriani and E. Sumarminingsih., "ESTIMATION OF MAXIMUM LIKELIHOOD WEIGHTED LOGISTIC REGRESSION USING GENETIC ALGORITHM (CASE STUDY: INDIVIDUAL WORK STATUS IN MALANG CITY)," *BAREKENG: J. Math. & App.*, vol. 17, iss. 1, pp. 0487-0494, March 2023.

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

Research Article • Open Access

1. INTRODUCTION

Logistic regression is a data analysis technique in statistics that is used to determine the effect of several variables where the response variable is categorical, both nominal and ordinal with predictor variables that can be categorical or continuous [1]. Logistic regression is one of the most commonly used classical classification methods. The problem that is often faced in the classification model with logistic regression is that the data is distributed unequally between different classes, which can be referred to as data imbalance.

Weighted Logistic Regression (WLR) is a development of logistic regression for imbalanced data or rare events using the weighted method. This method considers two stages of correction. The first is to add weights to the log-likelihood of binary logistic regression and the second is to add an element of bias corrections [2]. Parameter estimation for the WLR model uses the maximum likelihood estimation method.

The Maximum Likelihood Estimation (MLE) is a method to get the optimum parameter estimator value by maximizing the likelihood function (objective function). This method has several properties including unbiased, efficient, and consistent to achieve good parameter estimation. If the parameter estimation step using the maximum likelihood will produce a non-linear equation, then solving the equation to obtain the parameter estimation value uses a numerical optimization approach [3].

Newton-Raphson (NR) is a numerical optimization method commonly used in estimating Maximum Likelihood WLR. This method requires a gradient vector and a Hessian matrix. The gradient vector is obtained from the first derivative of the log-likelihood and the Hessian matrix is the second derivative [4]. However, the NR method has a problem that requires an initial value (the first derivative of the gradient vector), and obtaining the inverse matrix of the Hessian matrix is large which results in processing being long to converge due to the large matrix size. Therefore, this study will use a genetic algorithm approach as an alternative method to estimate the Maximum Likelihood of WLR.

Genetic algorithm (GA) is a computational optimization algorithm that can be used to optimize the estimation of model parameters based on random search methods inspired by the principles of biological systems, such as evolution, mutation, and the like, to overcome problems encountered in solving probability equations [5]. The Genetic algorithm is one of the most powerful heuristic methods for solving optimization problems. The main reason for choosing GA in obtaining parameter estimates is that the use of AG guarantees convergence to a globally optimal solution in large-scale and complex nonlinear optimization problems [6].

In contrast to Newton-Raphson for maximum likelihood estimation, the genetic algorithm approach has the advantage that in the process it does not require differential initial value requirements, does not calculate the gradient matrix, is easy to converge, and has more flexible assumptions [7],[8]. This method gives good results in estimating the ML estimator when compared to optimization using traditional optimizations such as Newton-Raphson (NR), Nelder-Mead (NM), and iterative re-weighting algorithm (IRA) [9].

This study uses data from the National Labor Force Survey data collected by the Central Statistics Agency of Malang City in 2020. The unit of observation that will be examined in this study is the status of individuals working in the workforce. The Labor Force Participation Rate in 2020 is 66.41% while the Open Unemployment Rate in 2020 is 9.61% [10]. This is for the absence of data imbalance. For handling unbalanced data, the WLR method will be used.

Based on this study will utilize the estimation of Maximum Likelihood Weighted Logistic Regression with genetic algorithms and determine the significant variables that affect the working status of individuals in Malang City.

2. RESEARCH METHODS

2.1 Data Source

The data used in this study is secondary data obtained from the results of the August National Labor Force Survey data collection conducted by the Central Statistics Agency of Malang City in 2020. The unit of observation that will be studied in this study is the individual working in the sample household.

2.2 Research Variable

The research variables used are described in **Table 1**.

Table 1. Research Variables

Symbol	Variable	Category	Scale
Y	Individual working status classification	0 : Work 1 : Doesn't work	Nominal
X_1	Gender	1 : Man 2 : Woman	Nominal
X_2	Age	1 : 15-30 years old 2 : 31-55 years old 3: $>$ 55 years old	Ordinal
X_3	Last Education	1 : No School 2 : Primary School 3 : Junior high school 4 : Senior high school 5 : College	Ordinal
X_4	Marital status	1 : Not married 2 : Married 3 : Divorced	Nominal
X_5	Health status	1 : Sick 2 : Health	Nominal
X_6	Number of household members	1 : 1-2 people 2 : 3-5 people 3: $>$ 5 people	Nominal

2.3 Data Analysis Method

Data analysis method to get the maximum likelihood estimation weighted logistic regression solution using genetic algorithm. The specific steps of the algorithm used include:

1. Identify the fitness function and the initial parameters of the algorithm:
   - Fitness function is log-likelihood weighted logistic regression.
   - Initial parameters of the genetic algorithm include population size (N) = 500, number of generations = 200, probability of crossover (Pc) = 0,95, probability of mutation (Pm) = 0,1
2. Generating the initial population of chromosomes (N) chromosomes generated from the search space through initialization. The initial population is denoted by $[{\theta_1}^{(0)}, {\theta_2}^{(0)}, \dots, {\theta_N}^{(0)}]$ where $\theta = [\beta_0, \beta_1, \beta_2, \dots, \beta_6]$ in this study. Each solution is represented through a chromosome and a fitness function, chromosomes in the population are represented by $\theta_j, j = 1, \dots N$.
3. Evaluate the fitness value of m iteration, ln $L_W(\boldsymbol{\beta})^{(m)}$, for each chromosome in the population.
4. Carrying out the selection process at a predetermined selection rate, the solutions (individuals) having the worst fitness function values, in light of the evaluation of individuals having executed the previous step, are replaced by new individuals generated randomly. Moreover, a certain number (EN) of individuals, having the best fitness values, are accepted as elite individuals and they are directly transferred without any modification to the new generation.
5. Using the roulette wheel (based on the principle that there is a greater chance of being selected if there is better fitness) as a proportional selection method, two parent candidate individuals are selected from the individuals, other than the elite individuals.
6. Perform crossover and mutation operators, as a perturbation mechanism, to candidate individuals according to Pc and Pm probabilities. Crossover of parents is conducted to obtain new offspring individuals and mutate new individuals. A new $(m+1)$nd generation $[{\theta_1}^{(m+1)}, {\theta_2}^{(m+1)}, \dots, {\theta_N}^{(m+1)}]$ is obtained.
7. Finally, set m = m + 1 and continue the iteration with the fitness evaluation step until the convergence criteria is satisfied. When evolution stops, the solution with the best fitness value in the last population is the best solution.
8. Getting the estimator $(\tilde{\beta})$.
9. Perform a parameter significance test.
10. Conclusion.

2.4 Weighted Logistic Regression

Weighted Logistic Regression (WLR) is an extension of the simple logistic regression model introduced by King and Zeng [2]. WLR is a bias correction method that consists of two correction steps. First add the weight w_i to the log-likelihood of binary logistic regression, to complete the difference in the proportion of events in the sample and population

$$\ln \mathrm{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i \left[y_i \ln \left(\frac{e^{x_i{}^T \boldsymbol{\beta}}}{1+e^{x_i{}^T \boldsymbol{\beta}}} \right) + (1-y_i) \ln \left(\frac{1}{1+e^{x_i{}^T \boldsymbol{\beta}}} \right) \right] \quad (1)$$

Where, if $y_i = 1$ than $w_i = \left(\frac{\tau}{\bar{y}}\right) = w_1$ and if $y_i = 0$ than $w_i = \left(\frac{1-\tau}{1-\bar{y}}\right) = w_0$. $\bar{y}$ representing the proportions in the sample, τ representing the proportions in the sample, and w_i representing the weight.

Second, adding an element of bias correction aims to reduce bias and variance. The estimated bias corrected for WLR is [11]:

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \boldsymbol{Bias}\left(\hat{\boldsymbol{\beta}}\right) \quad (2)$$

where $\boldsymbol{Bias}\left(\hat{\boldsymbol{\beta}}\right) = (\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{D}\xi$, $\xi_i = 0{,}5Q_{ii}\left((1+w_1)\hat{\pi}_i - w_i\right)$, Q_{ii} representing the diagonal element of $\boldsymbol{Q} = (\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X})^{-1}\boldsymbol{X}^T$ and $\boldsymbol{D} = diag\,\{\hat{\pi}_i(1-\hat{\pi}_i)w_i\}$.

To test the parameter estimates generated by the WLR model, we use a partial test with the Wald statistic test [12]:

$\mathrm{H}_0 : \beta_j = 0$
$\mathrm{H}_1 : \beta_j \neq 0,\ j = 1,2,\ldots,p$
Test statistics:

$$W_j = \left(\frac{\tilde{\beta}_j}{SE(\tilde{\beta}_j)}\right)^2;\ j = 1,2,..,p \quad (3)$$

where $\tilde{\beta}_j$ = estimator value β_j dan $Se\left(\tilde{\beta}_j\right)$ = *standard error* β_j
Wald's test follows a *Chi-Squared* distribution with degree of freedom one. Formula $SE\left(\tilde{\beta}\right)$ WLR [13]:

$$SE\left(\tilde{\beta}\right) = \sqrt{\left(\frac{n}{n+p}\right)^2}\left(\boldsymbol{X^T D X}\right)^{-1} \quad (4)$$

Wald test rejection region (H_0) if the value of $W_j > \chi^2_{(\alpha,1)}$ or $p - value < \alpha$.

2.5 Genetic Algorithm

Genetic Algorithm (GA), an iteration-based search technique proposed by Holland, is a very popular heuristic algorithm for finding the optimal solution of the objective function [14]. This algorithm adopts the genetic system of living things which includes reproduction, crossover, and mutation operators using the principle of natural selection, namely the individual who can survive is a strong individual. The concept of genetic algorithms which is based on genetic science causes the terms used in genetic algorithms to be widely adapted from that science [15]. The basic components that must be met in the genetic algorithm are [16]:

1) Coding Scheme
   The first step in the genetic algorithm is to represent the solution variables sought as a chromosome arrangement. One chromosome must represent one solution. The coding scheme used in this study is a real number scheme.
2) Fitness Function
   The fitness function is used to measure the level of goodness of the solution being sought from a set of solutions. A number of solutions generated in the population are evaluated using the fitness value. The chromosome that has the highest fitness value is the best solution. In each iteration, the algorithm will find and maintain the chromosome that has the highest fitness value.
3) Selection
   Selection aims to provide reproductive opportunities for individuals who have high fitness. The method that is often used for parent selection is roulette wheel selection (RWS).

4) Crossover
   Crossover is a method of mating two parent chromosomes (parents) to get a new individual (child). Individuals were randomly selected to perform with a crossover probability between 80% - 95%. The method used for crossbreeding is a one-point crossover.
5) Mutation
   A Mutation is a process to change the arrangement of genes in a chromosome to make a difference in the population by avoiding a condition where the chromosomes converge when the fitness value has not reached optimally. A good mutation probability value ranges from 0.5%-1%.

3. RESULTS AND DISCUSSION

3.1. Data Description

The first step to analyzing the working status of individuals in Malang City is to do a descriptive analysis of the data associated with the research variables. This overview can help provide preliminary information regarding the working status. This work activity includes both those who are working and those who have worked but in the past week while they were not actively working for example, due to leave, illness, and the like. This descriptive analysis uses cross-tabulation between the dependent variable (working status) and the independent variable. Factors that affect individual working status in Malang City.

Table 2. Factors Affecting Individual Work Status in Malang City

Variable	Category	Work	Doesn't work
X_1	Man	91,08%	8,92%
	Woman	88,33%	11,67%
X_2	15-30 years old	75,25%	24,75%
	31-55 years old	99,41%	0,59%
	$>$ 55 years old	87,39	12,61%
X_3	No school	83,04%	16,96%
	Primary school	87,19%	12,81%
	Junior school	73,83%	26,17%
	Senior school	97,03%	2,97%
	College	99,31%	0,69%
X_4	Not married	72,25%	27,75%
	Married	98,51%	1,49%
	Divorced	86,26%	13,74%
X_5	Sick	76,19%	23,81%
	Health	91,02%	8,98%
X_6	1-2 people	95,22%	4,78%
	3-5 people	88,06%	11,94%
	$>$ 5 people	81,25%	18,75%

Table 2, shows the percentage of working status on the variables of gender, age, education, marital status, health status, and the number of household members. In terms of gender, the percentage of men is more than women, which is 91.08% of the total male workforce compared to the female workforce which is only around 88.33%. This can be interpreted that compared to women, men tend to have easy access to jobs. Judging from the age group, it is known that the largest percentage of the working-age population is the 31-55 age group of 99.41%. Based on the education completed, the higher the level of education, the higher the percentage of those who work. This may be related to the expertise possessed. There are 99.31% of the workforce with working status with a university education background. Marital status greatly affects a person to carry out economic activities, because, in addition to aiming to support himself, the person must support the family for which he is responsible, it can be seen that the largest percentage of those with working status are those who are married, which is 98.51%. Next is the number of household members concerning for to working status. Workers are generally supported by health conditions. However, the reality is that it is possible to find workers who are not healthy but still work, as shown in Table 2, there are 76.31% of the working-age population are sick but still working. The number of household members who work in one family is 1-2 people, which is 95.22%. However, it does not require the possibility that a large household will tend to household members who do not have jobs to become dependents of the Head of the Household.

3.2. Estimation of Likelihood Weighted Logistic Regression using Genetic Algorithm

This study has determined population size = 500, crossover probability value (Pc) = 0.95, mutation probability value (Pc) = 0.01, and generation size = 200. The iteration algorithm parameter used is the maximum likelihood value obtained from all stages of the genetic algorithm process against log-likelihood Weighted Logistic Regression as a fitness function. The maximum likelihood estimation value of WLR with the genetic algorithm is presented in **Table 3**.

Table 3. The Result of Estimating the Maximum Likelihood of WLR with a Genetic Algorithm

Parameter	Estimation	Standard Error	Wald	P-value
$\tilde{\beta}_0$	3,459	2,804	1,528	0,369
$\tilde{\beta}_1$	0,354	0,210	2,858	0,214
$\tilde{\beta}_2$	-0,381	0,209	3,347	0,185
$\tilde{\beta}_3$	-0,379	0,091	17,329	0,037
$\tilde{\beta}_4$	-0,518	0,257	4,072	0,153
$\tilde{\beta}_5$	-2,213	1,363	2,648	0,230
$\tilde{\beta}_6$	0,726	0,192	14,358	0,044
Maximum value log-likelihood	**156.48315**			

Table 3, results Based on the maximum likelihood estimator value with the algorithm, the value of the log-likelihood weighted logistic regression function is 156, 4831 and the value of each estimator parameter is the value $\tilde{\beta}_0$= 3,459 , $\tilde{\beta}_1$ = 0,354, $\tilde{\beta}_2$ = -0,381, $\tilde{\beta}_3$= -0,379 , $\tilde{\beta}_4$= -0,518, $\tilde{\beta}_5$= -2,213,and $\tilde{\beta}_6$ = 0,726.

Furthermore, the test statistics using the Wald Test. The decision to reject H_0 is made if the value of Wald $> \chi^2_{(1)}$or if the $\mathrm{p} - \mathrm{value} < \alpha$. Table 3 shows that using the statistical value of the Wald test with$\chi^2_{(0,05,1)} = 3{,}841$ where the decision to reject H_0 is made if Wald $> \chi^2_{(0,05,1)}$so it can be concluded that the education completed (x_3) and the number of household members (x_6) affect the working status of individuals in Malang City.

4. CONCLUSIONS

Genetic algorithm can be an alternative method to get the Maximum Likelihood Estimator Weighted Logistic Regression (WLR). Based on the results of the analysis conducted, the variables that have a significant effect on the status of individuals working in Malang City use the WLR model for imbalanced data or rare events, including education completed and the number of household members.

REFERENCES

[1] X. Zou, Y. Hu, Z. Tian, and K. Shen, "Logistic Regression Model Optimization and Case Analysis," *Proc. IEEE 7th Int. Conf. Comput. Sci. Netw. Technol. ICCSNT 2019*, pp. 135–139, October 19-20 2019.

[2] M. Maalouf, D. Homouz, and T. B. Trafalis, "Logistic regression in large rare events and imbalanced data: A performance comparison of prior correction and weighting methods," *Comput. Intell.*, vol. 34, no. 1, pp. 161–174, February 2017.

[3] R. K. Vinayak, W. Kong, G. Valiant, and S. Kakade, "Maximum likelihood estimation for learning populations of parameters," *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 11217–11226, February 2019.

[4] D. E. A. Sulasih, S. Wulan Purnami, and S. Puteri Rahayu, "the Theoretical Study of Rare Event Weighted Logistic Regression for Classification of Imbalanced Data," *Int. Conf. Sci. Technol. Humanit.*, pp. 159–169, December 07 2015.

[5] A. M. García, I. Santé, M. Boullón, and R. Crecente, "Calibration of an urban cellular automaton model by using statistical techniques and a genetic algorithm. Application to a small urban settlement of NW Spain," *Int. J. Geogr. Inf. Sci.*, vol. 27, no. 8, pp. 1593–1611, November 2013.

[6] Z. Xia, K. Mao, S. Wei, X. Wang, Y. Fang, and S. Yang, "Application of genetic algorithm support vector regression model to predict damping of cantilever beam with particle damper," *J. Low Freq. Noise Vib. Act. Control*, vol. 36, no. 2, pp. 138–147, June 2017.

[7] E. Demir and Ö. Akkuş, "An Introductory Study on ' How the Genetic Algorithm Works in the Parameter Estimation of Binary Logit Model ?,'" *Int. J. Sci. Basic Appl. Res.*, vol. 19, pp. 162–180, January 2015.

[8] A. Salim and M. R. Alfian, "Optimalisasi Regresi Logistik Menggunakan Algoritma Genetika Pada Data Klasifikasi," *J. Teknol. Inf. dan Terap.*, vol. 6, no. 2, pp. 50–55, December 2019.

[9] A. Yalçınkaya, İ. G. Balay, and B. Şenoğlu, "A new approach using the genetic algorithm for parameter estimation in multiple linear regression with long-tailed symmetric distributed error terms: An application to the Covid-19 data," *Chemom. Intell. Lab. Syst.*, vol. 216, June 2021.
[10] Badan Pusat Statistik, "Laporan Eksekutif Ketenagakerjaan Kota Malang 2020," BPS Kota Malang, 2020.
[11] Z. Qiu, H. Li, H. Su, G. Ou, and T. Wang, "Logistic regression bias correction for large scale data with rare events," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8347 LNAI, no. PART 2, pp. 133–144, December 2013.
[12] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression Second Edition*. Canada: John Wiley and Sons,Inc, 2000.
[13] M. Maalouf and M. Siddiqi, "Weighted logistic regression for large-scale imbalanced and rare events data," *IIE Annu. Conf. Expo 2014*, vol. 59, pp. 142–148, March 2014.
[14] E. E. Elmahdy and A. W. Aboutahoun, "A new approach for parameter estimation of finite Weibull mixture distributions for reliability modeling," *Appl. Math. Model.*, vol. 37, no. 4, pp. 1800–1810, February 2013.
[15] T. Alam *et al.*, "Genetic Algorithm : Reviews , Implementations and Applications," *Int. J. Eng. Pedagog.*, vol. 10, no. 6, pp. 57–77, August 2020.
[16] S. Katoch, S. S. Chauhan, and V. Kumar, *A review on genetic algorithm: past, present, and future*, vol. 80, no. 5. Multimedia Tools and Applications, February 2021.