Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2025, 2025:81

https://doi.org/10.28919/cmbn/9177

ISSN: 2052-2541

BINARY LOGISTIC REGRESSION WITH RANDOM EFFECTS FOR PRODUCTIVITY ANALYSIS IN MINING OPERATIONS

AINUN UTARI, ANNA ISLAMIYATI*, ERNA TRI HERDIANI

Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Makassar 90245, Indonesia

Abstract: Binary logistic regression is a statistical method used to examine the relationship between a binary response variable and predictor variables, measured on a categorical or continuous scale. Binary logistic regression modeling is typically applied to cross-sectional data; however, this study extends its application to a binary logistic regression model for longitudinal data, which combines cross-sectional and time-series data. In longitudinal data, time and individual variables are critical components associated with random effects, as fixed predictor variables often do not fully explain temporal variations and inter-individual differences. Furthermore, the response variable is influenced by predictor variables (fixed effects) and random factors (random effects) such as sample selection, time, and area. A model that incorporates fixed and random effects components is called a mixed-effects model. This study develops a mixed-effects logistic regression model by considering the longitudinal data structure and predictor variables as random effects. The analysis is applied to loading equipment productivity data at PT Kaltim Prima Coal, with variables such as spotting time and loading time as fixed effects and environmental factors beyond control as random effects. This approach improves parameter estimation accuracy to 74.63% and provides a deeper interpretation of the factors affecting productivity.

Keywords: binary logistic regression; longitudinal data; mixed effects; loading equipment productivity.

2020 AMS Subject Classification: 62J12.

*Corresponding author

E-mail address: annaislamiyati701@gmail.com

Received February 06, 2025

1. INTRODUCTION

Binary logistic regression is a statistical method used to examine the relationship between a binary response variable and predictor variables, measured on a categorical or continuous scale [1]. Logistic regression has been extensively applied across diverse disciplines, such as education [2], health [3], and social sciences [4], to identify factors influencing specific parameters. The binary logistic approach has also been developed in nonparametric regression, such as spline binary logistic regression [5]. However, most of these studies rely on cross-sectional data. In fields such as economics, epidemiology, and health, time is often crucial for analyzing changes in research subjects over time. Therefore, this study develops a binary logistic regression model using longitudinal data.

Longitudinal data combines cross-sectional and time-series data from repeated observations over different periods on multiple subjects. In longitudinal data analysis, models are categorized into fixed-effects and random-effects models [7]. A model that incorporates both components is referred to as a mixed-effects model. Time and individual variables are crucial in longitudinal data, where time captures changes in subjects over the observation period, and individual variables reflect differences between subjects that may influence the observed response [8], [6], [9]. These variables are closely related to random effects, as fixed predictors often do not fully explain temporal fluctuations and inter-individual differences [10]. Random effects are thus utilized to accommodate unpredictable variations arising from individual differences and temporal changes [11]. This study aims to develop a mixed-effects logistic regression model by considering longitudinal data structure and treating predictor variables as random effects.

Researchers observe cases involving fixed and random variables in data on the productivity targets of loading equipment during overburden removal activities at PT Kaltim Prima Coal (KPC) in 2022. Factors influencing the productivity of loading equipment include fixed variables, such as spotting time and loading time, which can be optimized through production processes. At the same time, bench height, fragmentation (bucket), and excavation & material methods remain random factors that cannot be controlled as they depend on company standards and are influenced by natural conditions, introducing considerable variability [12]. Previous studies [10], [13] have confirmed that controllable operational factors, such as working time and equipment speed, are essential to productivity. Controllable external factors, such as geological conditions and weather, increase randomness in production outcomes. Based on this, the mixed-effects binary logistic regression model for longitudinal data is expected to describe the factors affecting the productivity targets of loading equipment during overburden removal activities at PT Kaltim Prima Coal Sangatta, enabling efforts to improve productivity.

2. PRELIMINARIES

Binary logistic regression with mixed effects combines fixed and random effects in the longitudinal data analysis, where multiple observations are collected from each individual (y_{it}, x_{it}, z_{it}). The binary logistic regression model for longitudinal data is as follows:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \boldsymbol{z}_{it}\boldsymbol{\gamma} + \varepsilon_{it}; i = 1, \dots, n; t = 1, \dots, T \quad (1)$$

where:

$\boldsymbol{\beta}$ = $[\beta_1 \quad \cdots \quad \beta_p]^T$ a vector of fixed effect coefficient parameters with dimensions $p \times 1$.

$\boldsymbol{x}_{it}$ = $[x_{1it} \quad \cdots \quad x_{pit}]$ a vector of fixed effect predictor variables with dimensions $1 \times p$.

$\boldsymbol{\gamma}$ = $[\gamma_1 \quad \cdots \quad \gamma_q]^T$ a vecto r of random effect parameters with dimensions $q \times 1$

$\boldsymbol{z}_{it}$ = $[z_{1it} \quad \cdots \quad z_{qit}]$ a vector of random effect predictor variables with dimensions $1 \times q$.

ε_{it} = a skalar of random errors which follows a logistic distribution with a mean of 0 and variance $\pi^2/3$.

The parameters of the mixed-effects binary logistic regression model for longitudinal data are estimated using the Maximum Likelihood Estimation (MLE) approach. The principle of MLE is to find parameter estimation by maximizing the likelihood function, which requires the response variable to follow a specific distribution [14]. If the response y_{it} is biner $y_{it} \in \{0,1\}$ it follows a binomial distribution with a probability density function (pdf) expressed as follows:

$$f(y_{it}) = \pi(x_{it}, z_{it})^{y_{it}}\left(1 - \pi(x_{it}, z_{it})\right)^{1-y_{it}} \quad (2)$$

The MLE method is a parameter function that maximizes the likelihood function to estimate parameters. Given the probability of y_{it} with model parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and predictor variables (x_{it}, z_{it}) the joint likelihood for all time observations for all time observations $t = 1, \dots, T$ for the$-i$ individua as:

$$L_i(\boldsymbol{\beta}, \boldsymbol{\gamma_i}) = \prod_{t=1}^{T} f(y_{it}|x_{it}, z_{it}, \boldsymbol{\beta}, \boldsymbol{\gamma_i}) \quad (3)$$

And the total likelihood for all individuals $i = 1, \dots, n$ is then given by:

$$L(\boldsymbol{\beta},\boldsymbol{\gamma_i}) = \prod_{i=1}^{n} L_i(\boldsymbol{\beta},\boldsymbol{\gamma_i}) \tag{4}$$

Since $\boldsymbol{\gamma_i}$ is an unobserved random effect and follows $\boldsymbol{\gamma_i} \sim N(0,\sigma_u^2)$, the likelihood function in Equation (4) is integrated concerning the distribution $f(\boldsymbol{\gamma_i})$ obtain the marginal likelihood in Equation (5).

$$L(\boldsymbol{\beta},\boldsymbol{\gamma_i}) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} L_i(\boldsymbol{\beta},\boldsymbol{\gamma_i}) \,.\, f(\boldsymbol{\gamma_i}) d\boldsymbol{\gamma_i}$$

$$= \int_{-\infty}^{\infty} \left[\prod_{i=1}^{n} f(y_{it}|\boldsymbol{x_{it}},\boldsymbol{z_{it}},\boldsymbol{\beta},\boldsymbol{\gamma_i})\right] f(\boldsymbol{\gamma_i}) d\boldsymbol{\gamma_i} \tag{5}$$

Where $L_i(\boldsymbol{\beta},\boldsymbol{\gamma_i})$ is the joint likelihood for individual i, $f(\gamma_i)$ is the probability density function (pdf) of the random effect, which follows a multivariate normal distribution:

$$f(\gamma_i) = \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left(-\frac{\gamma_i'\gamma_i}{2\sigma_u^2}\right) \tag{6}$$

Thus, the likelihood function in Equation (5) becomes as in Equation (7):

$$L(\boldsymbol{\beta},\boldsymbol{\gamma_i}) = \int_{-\infty}^{\infty} \left[\prod_{i=1}^{n} f(y_{it}|\boldsymbol{x_{it}},\boldsymbol{z_{it}},\boldsymbol{\beta},\boldsymbol{\gamma_i})\right] \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left(-\frac{\gamma_i'\gamma_i}{2\sigma_u^2}\right) d\boldsymbol{\gamma_i} \tag{7}$$

The likelihood function in Equation (7) poses analytical challenges, requiring numerical methods for its solution. The Adaptive Gauss Hermite Quadrature (AGHQ) method [15] is applied by multiplying and dividing the likelihood function by the normal density. The Gauss Hermite integration formula is expressed in Equation (8) as follows:

$$\int_{-\infty}^{\infty} f(x)dx \approx \sum_{m=1}^{M} w_m\, f(a_m) e^{a_m^2} \tag{8}$$

Where x_m are the quadrature points and w_m are the Gauss-Hermite wights. Thus, the likelihood function in Equation (8) can be written as Equation (10).

$$L_i = \left(\sigma_i^* \sqrt{2}\right) \int_{-\infty}^{\infty} f(w_i) dw_i \tag{9}$$

$$= \left(\sigma_i^* \sqrt{2}\right) \sum_{m=1}^{M} w_m^* \exp\left(a_m^2\right) \left(\frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left(-\frac{\left(a_m\sigma_i^*\sqrt{2}+\mu_i^*\right)^2}{2\sigma_u^2}\right)\right) \prod_{t=1}^{T} f\left(y_{it}|\boldsymbol{x}_{it}^T\boldsymbol{\beta}, a_m\sigma_i^*\sqrt{2}+\mu_i^*\right) \tag{10}$$

It is known that $\sigma_u^2 = \frac{\rho}{1-\rho}$, so the likelihood function for n observations in Equation (10), after substituting σ_u^2 becomes Equation (11):

$$L=\prod_{i=1}^{n}\left(\sigma_i^*\sqrt{2}\right)\left(\frac{2\pi\rho}{1-\rho}\right)^{\frac{1}{2}}\sum_{m-1}^{M}w_m^*\exp\left(a_m^2\right)\exp\left(-\frac{\left(a_m\sigma_i^*\sqrt{2}+\mu_i^*\right)^2(1-\rho)}{2\rho}\right)\prod_{t=1}^{T}\left(\pi\left(\boldsymbol{x}_{it}^T\boldsymbol{\beta}+a_m\sigma_i^*\sqrt{2}+\mu_i^*\right)\right)^{y_{it}}\left(1-\pi\left(\boldsymbol{x}_{it}^T\boldsymbol{\beta}+a_m\sigma_i^*\sqrt{2}+\mu_i^*\right)\right)^{1-y_{it}} \tag{11}$$

The next step is to maximize the likelihood function by transforming it into the form of the log-likelihood function as follows:

$$\log(L)=\sum_{i=1}^{n}log\left[\left(\sigma_i^*\sqrt{2}\right)\left(\frac{2\pi\rho}{1-\rho}\right)^{\frac{1}{2}}\sum_{m=1}^{M}w_m^*\exp\left(a_m^2\right)\exp\left(-\frac{\left(a_m\sigma_i^*\sqrt{2}+\mu_i^*\right)^2(1-\rho)}{2\rho}\right)\prod_{t=1}^{T}\left(\pi\left(\boldsymbol{x}_{it}^T\boldsymbol{\beta}+a_m\sigma_i^*\sqrt{2}+\mu_i^*\right)\right)^{y_{it}}\left(1-\pi\left(\boldsymbol{x}_{it}^T\boldsymbol{\beta}+a_m\sigma_i^*\sqrt{2}+\mu_i^*\right)\right)^{1-y_{it}}\right] \tag{12}$$

The log-likelihood function is maximized by differentiating the log-likelihood function concerning each parameter, and then the first derivative is equal to zero. Based on the first derivative results, the obtained estimates do not have a closed-form solution and cannot be solved analytically. Therefore, the estimation process continues using the Newton-Raphson iteration method. The iteration process stops when the results converge $\left\|\boldsymbol{\theta}_{(k+1)}-\boldsymbol{\theta}_{(k)}\right\|<\varepsilon$, where ε represents the smallest positive number is 10^{-16}.

3. MAIN RESULTS

PT KPC is one of the largest open-pit coal mining companies globally, playing a significant role in meeting domestic and international market demands. Achieving optimal production outcomes requires continuous efforts to improve productivity through the efficient performance of mechanical equipment, particularly the productivity of loading machinery. The loading equipment consists of three units, with repeated observations recorded daily, making the dataset suitable for longitudinal analysis. Predictor variables include Spotting Time (x_1) and Loading Time (x_2), measured on a ratio scale and treated as fixed effects. Fragmentation and Bucket Capacity (x_3) and Excavation and Material Method (x_4) are also measured on an ordinal scale and modeled as random effects. The response variable (Y) is binary categorical, making the mixed-effects binary logistic regression model an appropriate analytical method. This approach accounts for longitudinal data by analyzing repeated observations from the same unit over different periods while considering variability between units (random effects) and within units (fixed effects). The mixed-effects binary logistic regression model will be applied to the longitudinal dataset to evaluate the achievement of productivity targets. The influence of predictor variables on the response variable will be assessed using a simultaneous test to identify significant predictors, followed by partial tests to refine the model [14].

The test statistics in the parameter testing was conducted simultaneously using the G-test or LRT (Likelihood Ratio Test) with $\alpha = 5\%$ where $G = 62.854 > \chi^2_{(0.05;2)} = 5.991$, thus the null hypothesis (H_0) is rejected. This indicates that predictor variables (x, z) have a significant effect on the response variable (y), making the model suitable for further analysis. In the next phase, partial tests are conducted to identify which predictor variables significantly influence the response variable. The results of the partial test for the β parameters are presented in Table 1.

Table 1 Results of Partial Testing for Parameter $\boldsymbol{\beta}$

Parameter	Estimated Coefficient	Std. error	$\lvert W_j \rvert$	*p*-value	Information
β_1	-5.246	0.876	-5.987	2 x 10^{-9}	significant
β_2	-2.088	0.416	-5.013	5 x 10^{-7}	significant

Table 1 shows that with a significance level of $\alpha = 5\%$ or 0.05 and $p - value < \alpha$, the null hypothesis H_0 is rejected. This indicates that all predictor variables, namely spotting time (x_1) and loading time (x_2) in the binary logistic regression model with random effects on longitudinal data have a significant effect on the binary response variable (y). Meanwhile, the random effects or γ parameter are presented in Table 2.

Table 2 Results of Partial Testing for Parameter $\boldsymbol{\gamma}$

Parameter	$LRT(G^2)$	$\chi^2_{(0.05;2)}$	*p-value*	Information
$subject$	125.5	5.991	2.2 x 10^{-16}	Significant
$time$	0.092	3.841	0.761	not significant
γ_3	2.126	3.841	0.144	not significant
γ_4	1.865	3.841	0.172	not significant

Table 2 shows that the criteria for the G-test or LRT (Likelihood Ratio Test) with a significance level of $\alpha = 5\%$ where $G > \chi^2_{(\alpha;df)}$ or $p - value < \alpha$ indicate that only the subject variable significantly affects the response variable (y). Conversely, the time, γ_3, and γ_4 parameters do not significantly affect the response variable. Consequently, the probability function for the significant variables in the mixed-effects binary logistic regression model for longitudinal data is formulated as follows:

$$\pi_{it}(x_{it}, z_{it}) = \frac{\exp\left(10.0703 - 5.246\mathrm{x}_1 - 2.088\mathrm{x}_2 - \gamma_{subject}\right)}{1 + \exp\left(10.0703 - 5.246\mathrm{x}_1 - 2.088\mathrm{x}_2 - \gamma_{subject}\right)}$$

The interpretation of coefficient values in the logistic regression model is conducted in the form of odds ratios. The odds ratio values for predictor variables deemed significant about the productivity of loading equipment during overburden removal activities at PT Kaltim Prima Coal in 2022 can be shown in Table 3. The odds ratio for the variable Spotting Time (x_1) is $\exp(-5.246) = 0.0053$, which means that each one-unit increase in spotting time decreases the probability of achieving the productivity of the loading equipment by 0.0053, assuming other variables remain constant. The negative coefficient indicates that an increase in spotting time (the time required for the equipment to position itself and start loading material) reduces the productivity of achieving the productivity of the loading equipment at PT KPC. Therefore, the longer the spotting time, the lower the probability of achieving the productivity target. The odds ratio for the variable Loading Time (x_2) is $\exp(-2.088) = 0.1241$, which means that each one-unit increase in loading time decreases the probability of achieving the productivity of the loading equipment by 0.1241, assuming other variables remain constant. The negative coefficient indicates that an increase in loading time reduces the probability of achieving the productivity of the loading equipment at PT KPC. Therefore, the longer the loading time, the lower the probability of achieving the productivity target. The subject variable (identified as the type of excavator) exhibits very high variance, with a value of 5.69968 and a standard deviation of 2.3874. The subject variable is a random effect that significantly affects productivity achievement. The high variance indicates substantial differences among types of excavators, which consistently affect the probability of achieving productivity. This may reflect that certain excavators have a baseline productivity level that differs significantly from others, suggesting that selecting excavator types considerably impacts productivity outcomes.

Table 3 Odds Ratio Values of the Mixed-Effects Binary Logistic Regression Model for Longitudinal Data

Parameter	Estimated Coefficient	Odds Ratio
Intercept	10.0703	23672.5
β_1	-5.246	0.0053
β_2	-2.088	0.1241
$\gamma_{subject}$	5.699	

The relationships between factors influencing the productivity of loading equipment can be seen in Figures 1, 2, and 3.

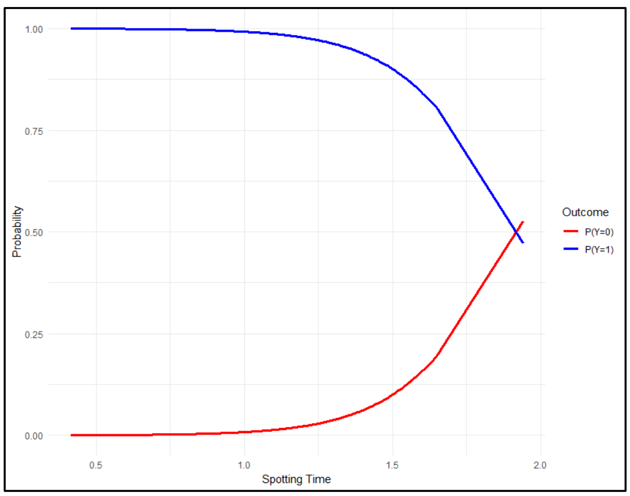

Figure 1. Graph of the Relationship between Probabilities P(Y=1), P(Y=0) and Spotting Time x_1

Figure 1 shows the probability distribution of two binary outcomes, P(Y=1) and P(Y=0), with the predictor variable Spotting Time (x1). The probability P(Y=1) increases exponentially, approaching one when the spotting time is low, and decreases significantly as the spotting time increases. Conversely, the probability P(Y=0) decreases exponentially, approaching zero when the spotting time is low and increases significantly as the spotting time increases. This indicates that the longer the spotting time, the higher the probability that the productivity of the loading equipment fails to be achieved.

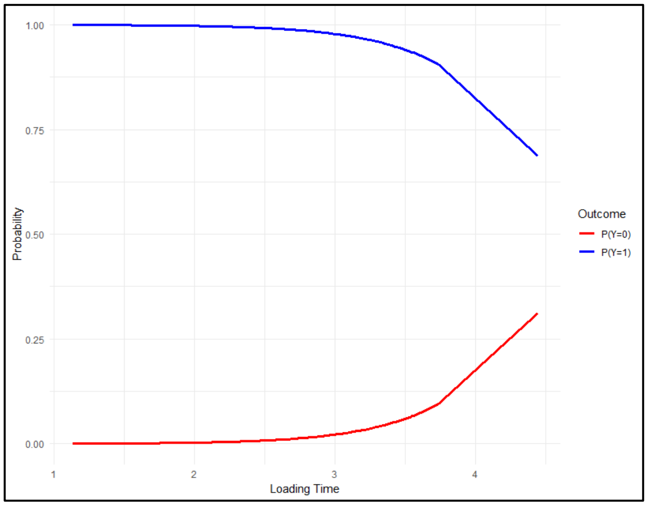

Figure 2. Graph of the Relationship between Probabilities P(Y=1), P(Y=0) and Loading Time x_2

Figure 2 illustrates the probability distribution of two binary outcomes, P(Y=1) and P(Y=0), with the predictor variable Loading Time (x2). The probability P(Y=1) increases, approaching one when the loading time decreases and significantly declines as the loading time increases. Conversely, the probability P(Y=0) decreases, approaching zero when the loading time decreases but increases significantly as the loading time increases. This indicates that the longer the loading time, the higher the probability that the productivity of the loading equipment fails to be achieved.

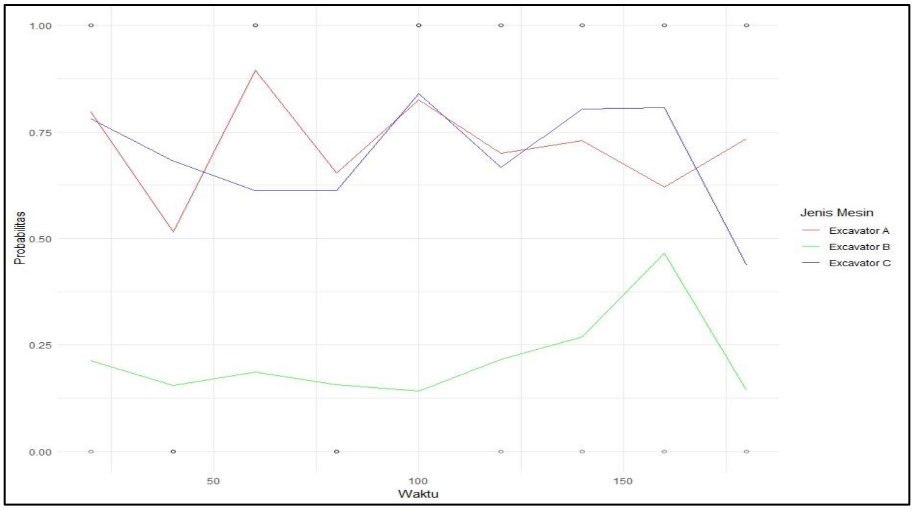

Figure 3. Graph of the Relationship between Probabilities P(Y=1), P(Y=0) and Subject as Random Effect

Figure 3 shows that the blue curve illustrates the differences in the probability of achieving productivity across different types of excavators over time, reflecting the influence of random effects within the model. Excavator A demonstrates a fluctuating probability pattern but maintains higher probabilities at most points, suggesting that this type of excavator is more likely to achieve the desired productivity. This indicates that Excavator A may be more reliable or perform better than the other excavators. Excavator B shows lower and more stable probabilities over time, with a slight increase at specific time points before declining. This suggests that Excavator B may have limitations in achieving consistent productivity, with some instances showing better opportunities but remaining low overall. Excavator C exhibits a significant decline in probabilities at the final time point. This indicates that the excavator's performance decreases drastically, or other factors may substantially reduce the productivity of achieving productivity at specific times. Overall, Excavator A demonstrates the highest performance with consistently high probabilities of achieving productivity, whereas Excavator C shows a significant decline in performance during the later period.

Subsequently, the model was evaluated using a classification table and accuracy values to demonstrate its overall performance, as presented in Table 4.

Table 4 Classification and Accuracy Value of the Best Model

Actual	Prediction		Total	Accuracy
	0	1		
0	158	51	209	74.63%
1	86	245	331	
Total	244	296	540	

Table 4 shows the correct classifications for achieving productivity totaled 245. The model's accuracy of 74.63% indicates that it performs reasonably well in predicting the data overall.

CONFLICT OF INTERESTS

The authors declared that there is no conflict of interest.

REFERENCES

[1] A. Agresti, An Introduction to Categorical Data Analysis, Wiley, 2007. https://doi.org/10.1002/0470114754.

[2] Nurhasanah, A. Rusyana, A.R. Fitriana, Binary Logistic Regression for Identification of High School Student

Interest in Banda Aceh City in Continuing Study at Universitas Syiah Kuala, J. Phys.: Conf. Ser. 1882 (2021), 012034. https://doi.org/10.1088/1742-6596/1882/1/012034.

[3] R.D. Joshi, C.K. Dhakal, Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches, Int. J. Environ. Res. Public Health 18 (2021), 7346. https://doi.org/10.3390/ijerph18147346.

[4] K.M. Kim, J.H. Kim, H.S. Rhee, B.Y. Youn, Development of a Prediction Model for the Depression Level of the Elderly in Low-Income Households: Using Decision Trees, Logistic Regression, Neural Networks, and Random Forest, Sci. Rep. 13 (2023), 11473. https://doi.org/10.1038/s41598-023-38742-1.

[5] A. Islamiyati, Anisa, M. Zakir, et al. The Use of The Binary Spline Logistic Regression Model on the Nutritional Status Data of Children, Commun. Math. Biol. Neurosci. 2023 (2023), 37. https://doi.org/10.28919/cmbn/7935.

[6] N.M. Laird, J.H. Ware, Random-Effects Models for Longitudinal Data, Biometrics 38 (1982), 963–974. https://doi.org/10.2307/2529876.

[7] M. Mouchart, The Econometrics of Panel Data, Institut de Statistique Universite Catholiqe de Lauvain, Paris, 2009. https://perso.uclouvain.be/michel.mouchart/source/DataPanelLink090511.pdf.

[8] G.M. Fitzmaurice, N.M. Laird, J.H. Ware, Applied Longitudinal Analysis, Wiley, Hoboken, 2012.

[9] L.C. Liu, D. Hedeker, A Mixed-Effects Regression Model for Longitudinal Multivariate Ordinal Data, Biometrics 62 (2006), 261–268. https://doi.org/10.1111/j.1541-0420.2005.00408.x.

[10] H. Wu, J.T. Zhang, Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches, John Wiley & Sons, Inc., New York, (2006).

[11] W. Hulin, Z. Jin-Ting, Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches, Wiley, 2006. https://doi.org/10.1002/0470009675.

[12] H. Ichwanto and Team MOD KPC, Technical Handbook, PT Kaltim Prima Coal, Kalimantan Timur, Indonesia, 2017.

[13] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, John Wiley & Sons, Inc., New York, (2013).

[14] D.W. Hosmer, S. Lemeshow, R.X. Sturdivant, Applied Logistic Regression, Wiley, 2013. https://doi.org/10.1002/9781118548387.

[15] J.S. Butler, R. Moffitt, A Computationally Efficient Quadrature Procedure for the One-Factor Multinomial Probit Model, Econometrica 50 (1982), 761-764. https://doi.org/10.2307/1912613.

[16] R. Ehlers, Maximum Likelihood Estimation Procedures for Categorial Data, Thesis, University of Pretoria, South Africa, 2002.