



Recovery Rate of Pulmonary Tuberculosis Patients Using 2-Parameter Gamma Regression Model with Weighted Least Square Approach

Hendra H. Dukalang^{1,2}, Joko Purwadi^{1,3,*}, Sukma Adi Perdana^{1,4}, and Setia Ningsih⁵

¹ Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia

² Syariah Banking, IAIN Sultan AMAI Gorontalo, Indonesia, Gorontalo 96181, Indonesia

³ Mathematics, Faculty of Applied Science and Technology, Universitas Ahmad Dahlan, 55166, Yogyakarta, Indonesia

⁴ Management of Islamic Education, STAIN Sultan Abdurrahman Kepri, 29157, Indonesia

⁵ Department of Mathematics, Universitas Negeri Gorontalo, Bone Bolango 96554, Indonesia

*Corresponding author. joko@math.uad.ac.id

Abstract. This paper discuss about the recovery rate of Pulmonary Tuberculosis (TB) Patient. data analysis is carried out using regression methods. Regression models are generally built on assumptions following the Normal distribution, but in practice empirically, this assumption is not always correct because it is possible that the data distribution is asymmetric and may even be thicker or thinner-tailed than the normal distribution. The gamma regression model is used when the values of the response variables under the study are positively skewed following the gamma distribution. Based on the results of the estimated parameters of the regression model with a weight least square and the interpretation that has been carried out, it can be concluded that if the TB patient-free variables such as Age X1, Indication of shortness of breath X3, Indication of Cough X4, and previous history of pulmonary TB X6 are of low value, it is estimated that the treatment period of the TB patient will also be low or the patient will recover faster.

Keywords: pulmonary tuberculosis, recovery rate, gamma regression model, Weighted least square

1 Introduction

Pulmonary Tuberculosis is a disease caused by Mycobacterium Tuberculosis and is contagious. Transmission occurs when Pulmonary Tuberculosis patients are coughing or sneezing so that Mycobacterium Tuberculosis bacteria spread into the air in the form of sputum splashes (nuclei droplets) because once coughing

can produce 3000 sputum splashes [1] Generally, sputum splashes can last a long time when in the room. Tuberculosis bacteria can stay in the air for hours even when the room is dark and damp until it is finally inhaled by others.

The National Tuberculosis Program in Indonesia provides essential reporting, which allows for early discovery and treatment, minimizes complications, prevents transmission, and lowers mortality [2]. Tuberculosis (TB) is the most common cause of mortality caused by infectious germs globally. TB affected 9.9 million individuals in 2020, and 1.5 million TB patients died [3]. Antimicrobial drug resistance in *Mycobacterium tuberculosis* jeopardizes the success of who's anti-TB policy. Diagnostic advancements and the availability of new anti-TB medications have resulted in significant changes in the care of individuals with drug-resistant TB [4]. New anti-TB medications (bed aquiline, delamanid, and pretomanid) have been licensed for the treatment of drug-resistant tuberculosis, as have significant improvements in treatment recommendations and regimens [5].

TB in Indonesia is still the third leading cause of death in the world. The World Health Organization annually began publishing the Global Tuberculosis Report in 1997. Noted in the Global Tuberculosis Report released in 2021, Indonesia is the second country to contribute to the reduction of TB cases below India by arounds 13% - 14%. then on 2022, it becomes the third country under India and China with a reduced rate of 18% - 19% [6]. The high number of pulmonary tuberculosis patients in Indonesia, which is accompanied by a large reduction every year, encourages researchers to find out what factors are most influential on the cure rate of pulmonary Tuberculosis. So that it can reduce the number of cases of TB sufferers.

To find out this, data analysis is carried out using regression methods. Regression models are generally built on assumptions following the Normal distribution, but in practice empirically, this assumption is not always correct because it is possible that the data distribution is asymmetric and may even be thicker or thinner-tailed than the normal distribution.

GLMs, which were initially introduced by Wedderburn [7] and are an extension of conventional linear models, may be used to describe discrete and continuous dependent variables without making any assumptions about general normality or constant variance [8]. GLM has three components: (i) the response variable, which must be part of the exponential distribution family, (ii) the linear predictor, and (iii) the connecting function, which must be a monotonous differentiated function and connects the linear predictor with the average response on the observation [9].

There are several data distributions whose relaxation is able to capture asymmetric patterns and thicknesses on the tail one of the data is the Gamma distribution. Classical analysis will not give superior findings, particularly with its statistical inference of model parameters, therefore Gamma distributions are designed to address asymmetric data patterns because these distributions are designed as flexible and adaptive distributions, thus a more efficient approach and do not require data normalization can be obtained [10]. The gamma distribution

is a kind of exponential distribution that was developed by Swiss mathematician Leonard Euler in the 18th century. Gamma distributions are classified into various categories. A frequent Gamma distribution is the two-parameter gamma distribution. This distribution has both shape and scale parameters. Several studies have examined Gamma regression. Al-Abood [11] examined the Gamma regression model using two estimation methods, namely Maximum Likelihood Estimation and Weighted Least Square. Schutz [12] Generalized Gamma Distribution (GGD) i.e in the case of developing the classification of image textures and modeling the coefficients of the wavelengths obtained from the algorithm for calculating centroids of several parameters.

A gamma regression model is employed more rapidly when a favourably skewed response variable follows the gamma distribution with a specific set of independent variables [13] [14] [15]. Because the condition of independence of explanatory variables is rarely applied in gamma regression models, the multicollinearity problem arises, which implies the maximum probability estimator (MLE) is unstable and produces a high variance [16]. As a result, determining confidence intervals and evaluating model regression parameters becomes problematic. The Gamma regression model is one of the most extensively used models for examining real-world data issues [13] [16], such as medical research, health economics [17], and automobile insurance claims. When the values of the study's response variables are positively skewed, the gamma regression model is applied.

2 Method

This research used secondary data from a study conducted by Wini Aprilia. With the observation unit is a patient with pulmonary TB at the hospital. Prof. Dr. Hi. Aloei Saboe. The research variables that will be used in this study are variables that are suspected to affect the response variables for the recovery time of TB patients as seen in Table 1.

Table 1. Level of Students' Engagement

Variable	Information
Y	The time it takes for the patient to recover
X1	Age
X2	Gender
X3	Shortness of Breath
X4	Cough
X5	Fever
X6	History of tuberculosis disease
X7	Work
X8	Smoke

2.1 Gamma Distribution and Regression

The basic form of gamma distribution characteristics with two parameters according to probability density function is as follows:

$$f(y|\alpha, \theta) = \frac{y^{\alpha-1} e^{-y/\theta}}{\theta^\alpha \Gamma(\alpha)}, \quad y > 0; \alpha > 0; \theta > 0$$

$$\text{with, } \Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt; \text{ Mean} = \alpha\theta; \text{ Variance} = \alpha\theta^2;$$

$$\text{and MGF} = (1 - \theta t)^{-\alpha}.$$

Gamma regression is one of the regressions that can describe the relationship between the variable Y as a Gamma distributed response variable and the variables X_1, X_2, \dots, X_k as the predictor variables [20]. The matrix form of the response variable, predictor variable, and gamma regression parameter are as follows:

$$y = [y_1, y_2, \dots, y_k]_{(n \times 1)}^T;$$

$$X = [x_1, x_2, \dots, x_k]_{(n \times 1)}^T;$$

$$\beta = [\beta_0, \beta_1, \dots, \beta_k]_{(n \times 1)}^T;$$

The following is the univariate gamma regression model for the two parameters and utilizing the log link function.

$$E(Y) = e^{x^T \beta}$$

$$\text{where } x^T \beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad \text{and}$$

X_1, X_2, \dots, X_k are the predictor variable. Based on equation (2), the θ value are obtained as follow:

$$\theta = \frac{e^{x^T \beta}}{\alpha} = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}}{\alpha}$$

The probability density function of gamma univariate regression is derived by substituting equation for its density equation.

2.2 Parameter Estimation of Gamma Regression Model

Maximum likelihood estimation (MLE) and weighted least squares (WLS) estimation methods are used to estimate the parameters of gamma regression mod-

$$f(y) = \frac{y^{\alpha-1} \exp\left(-y/\frac{e^{x^{\gamma\beta}}}{\alpha}\right)}{\left(\frac{e^{x^{\gamma\beta}}}{\alpha}\right) \Gamma(\alpha)}; \alpha > 0, y > 0$$

els. In this study, WLS are used to estimate the regression line by minimizing the sum of the squares of the error of each observation of the line by dividing the ordinary OLS regression equation [17]. From the linear regression equation, parameter estimation is obtained from $\beta_0, \beta_1, \dots, \beta_k$ minimizing the Number of Residual Squares (SSE) is as follows:

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\sum_{i=1}^n \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_k x_{ik}) \right]^2$$

A WLS estimate is an estimate that produces the following equation [13],

$$\hat{\beta}_w = (\mathbf{X}'\mathbf{D}_w\mathbf{X})^{-1} \mathbf{X}'\mathbf{X}_w\mathbf{Z}$$

where $Z_i = \ln Y_i + \ln r_i + \Gamma(r_i), i = 1, \dots, n$.

The WLS estimation is a linear function of the log-gamma random variable because the moment is known for all values from Z_i and r_i , then the covariate matrix is

It can be concluded that the WLS estimate to obtain β from the estimate $\hat{\beta}_w$ the normal and mean multivariate distribution μ , while the covariate matrix is V^{-1} .

2.3 Distribution Testing

Testing the distribution on response variables using Anderson Darling. Anderson Darling is a statistical method used in distribution conformity testing. According to [18], the Anderson darling test is used as a normality test or goodness of fit for quantitative changes and can be used to test the normality of various kinds of data distribution, namely normal data distribution, lognormal gamma, Weibull, and logistic distribution. The advantage of the Anderson darling test is that the test is more sensitive than the K-S test so in this study the test was used [19].

2.4 Data Analysis

The following are the data analysis steps in this study:

$$\text{Cov}(\hat{\beta}_w) = V^{-1} \text{ dan } V = v_{jl},$$

$$\text{where } v_{jl} = \sum_{i=1}^n \frac{x_{ij}x_{il}}{\Gamma^{(1)}r_i}$$

1. Using the Anderson darling test, examine the distribution of data, particularly the distribution of response variables, such as data on the length of hospitalization of pulmonary tuberculosis patients.
2. Parameter estimation and hypothesis testing for two-parameter gamma regression models
3. Interpretation of the model
4. Reaching conclusions

3 Result and Discussion

3.1 Descriptive Statistics

Table 2 shows the information related to research variables. Variable Y is a response variable that represents the period of treatment of the patient. Furthermore, there are 8 predictor variables, namely X_1 (Age), X_2 (Gender), X_3 (Indications of tightness), X_4 (Indications of Cough), X_5 (Indications of fever), X_6 (Previous history of TB), X_7 (Occupation), and X_8 (Smoking).

Table 2 describes the descriptive statistics for the research variables. From the information provided in Table 2, for the response variable Y , the patient's treatment period has an average of 7.045 days with a minimum period of 1 day and a maximum period of 15 days. For age data, the patients treated had an average of 49.66 years, with the age of the youngest patient being 3 years and the age of the oldest patient being 89 years.

Table 2. Descriptive Statistics of Research Variables

Variable	Mean	St. Dev
Time (Y)	7.0450	3.5760
Age (X_1)	49.6600	22.8600
Gender (X_2)	0.6591	0.4795
Shortness of Breath (X_3)	0.3636	0.4866
Cough (X_4)	0.3636	0.4866
Fever (X_5)	0.4545	0.5037
History of TB (X_6)	0.4318	0.5011
Occupation (X_7)	0.4091	0.4974
Smoking (X_8)	0.8636	0.3471

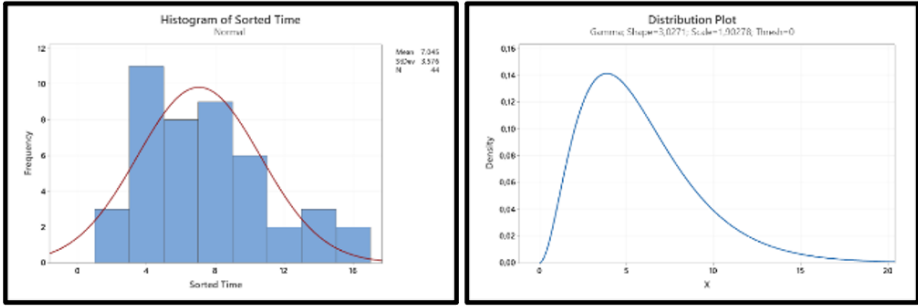


Fig. 1. Distribution of Patient Care Time Data

Figure 2 presents a histogram of the data distribution from the Y response variable during the patient’s treatment period. From the image, it can be observed that the Y data is pressed to the right in accordance with the distribution of gamma distribution. It will then be tested with the Anderson darling test to see the severity of the distribution of Y response data to gamma distribution.

3.2 Test Response Variable (Y)

Before estimating parameters for the response variable Y, it is necessary to test the distribution of the response variable Y. The Anderson-Darling (AD) test determines if Y response data fits normal, exponential, Weibull, or gamma distributions. The hypotheses used in such distribution tests are:

$$H_0 : F_Y = F_0(Y)$$

(Data distribution according to a specific distribution)

$$H_1 : F_Y \neq F_0(Y)$$

(The distribution of data does not correspond to a specific distribution)

Table 3. Test Goodness of Fit Anderson Darling

Distribution	AD	P-Value
Normal	1.015	0.010
Exponential	5.153	<0.003
Weibull	0.614	0.104
Gamma	0.539	0.190

Table 3 displays the results of the Anderson Darling test and the response variable distribution parameters are shown in table 4. From the information based on table 3, The Anderson darling test value for the smallest response

variable is 0.539, and the greatest P-value is 0.19. The results of this data show that the response variable is most likely to have a gamma distribution or it can be said that the Y response variable comes from a gamma-distributed population. Table 4 shows that the response variables' gamma distribution parameters are the form parameter of 3.7027 and the scale parameter of 1.9028.

Table 4. Distribution Parameters of Response Variables

Distribution	Location	Shape	Scale	Threshold
Normal	7.0454		3.5761	
Exponential			7.0454	
Weibull		2.1154	7.9737	
Gamma		3.7027	1.9028	

3.3 Gamma Regression Model with Weight Least Square (WLS) In TB Patient Care

Simultaneous testing is carried out on the entire predictor parameters together. The test done to examine whether the parameters in the predictor variables have a simultaneous influence on the form model or not. The hypotheses used are as follows:

$$\begin{aligned}
 H_0 &: \beta_1 = \beta_2 = \dots = \beta_8 = 0 \\
 H_1 &: \exists \beta_j \neq 0, j = 1, 2, \dots, 8
 \end{aligned}$$

concurrent testing on gamma regression model in full is presented in the following table 5.

In Table 5 with a significant degree of 5%, it is obtained that the value of the p-value is less than 5%. It can be said that the conclusion of simultaneous testing is reject H0. Based on this, a conclusion can be drawn that at least one parameter has a simultaneous influence on the Gamma regression model, thus the Gamma regression model is feasible to use in modelling the Duration of treatment of TB patients.

Table 5. Concurrent Test Results of Gamma Regression Models

Sum Square	df	Mean Square	F	p-value
123.46	8	15.432	9.2302	9.91583e-07
58.517	35	1.6719	-	-
181.97	43	-	-	-

Result: Reject H_0 , at least one β parameter is significant.

After simultaneous testing and the result is obtained that there is at least one parameter that has a significant effect on the model, the next step is to test partially. This test is used to determine which predictor variables (factors) have a significant effect on the model. The hypotheses used are as follows.

$$H_0 : \beta_i = 0, \forall_i, i = 1, 2, \dots, 8$$

$$H_1 : \beta_i \neq 0, \forall_i, i = 1, 2, \dots, 8$$

With a significant level of 5%, significant parameters can be obtained in the Gamma regression model. Full partial testing on Gamma regression models is shown in the following table.

Table 6. Gamma Regression Model Concurrent Test Results

Beta	SE	t-score	p-value	Result
2.4845	2.2047	1.1269	0.2675	—
0.0746	0.0114	6.5657	1.3987e-07	***
-0.2112	0.6719	-0.3144	0.75509	—
3.8863	0.8739	4.4471	8.4121e-05	***
-2.3165	1.0721	-2.1607	0.0376	***
1.6192	0.9861	1.6421	0.1095	—
2.4700	0.7241	3.4114	0.0016	***
-1.3834	0.7130	-1.9403	0.0604	—
-0.7771	2.1291	-0.3650	0.7173	—

Significance: *** Significant — Insignificant

It can be seen that at a significance level of 5%, there are four variables that affect the treatment period of TB patients. The influential variables include Age X_1 , Indication of shortness of breath X_3 , Indication of Cough X_4 , and Previous History of Pulmonary TB X_6 .

Based on the results of the calculation of the estimated parameters of the Gamma regression model with the Weight Least Square Method in Table 6, it can be concluded that not all predictor variables have an influence on the length of treatment of TB patients or the response variable Y . From the results in Table 6, the variables X_2 , X_5 , X_7 , and X_8 are not significant at 5%. The resulting model of data processing as a whole by entering all variables will be obtained as follows.

The equation is a model using both significant and insignificant variables. For the purposes of analysis, then interpretation is carried out only for significant free variables. From the model, it can be interpreted that if the age variable X_1 increases by 1 then the treatment period of patient Y will increase by $\exp(0.0746)$ or increase to 1.0019 times, with the other free variables considered fixed.

$$\hat{\theta} = \frac{\exp \left(\begin{array}{l} 2.845 + 0.0746x_1 + 0.2112x_2 + \\ 3.8863x_3 + 2.3156x_4 + 1.6191x_5 \\ + 2.4700x_6 - 1.3834x_7 - 0,7771x_8 \end{array} \right)}{3.7027}$$

If the free variable Indication of shortness of breath X_3 increases by 1 then the treatment period of patient Y will increase by $\exp(0.0001)$ or increase to 1.0001 times, with the other free variables considered fixed.

If the free variable Indication of cough X_4 increases by 1 then the treatment period of patient Y will increase by $\exp(0.0000)$ or increase to 1.00002 times, with the other free variables considered fixed.

If the free variable Previous pulmonary TB history X_6 increases by 1 then the treatment period of patient Y will increase by $\exp(0.0007)$ or increase to 1.0007 times, with the other free variables considered fixed.

4 Conclusion

According to the research's results of the estimated parameters of the regression model with a weight least square and the interpretation that has been carried out, it can be concluded that if the TB patient-free variables such as Age X_1 , Indication of shortness of breath X_3 , Indication of Cough X_4 , and previous history of pulmonary TB X_6 are of low value, it is estimated that the treatment period of the TB patient will also be low or the patient will recover faster.

References

1. R. Werdhani, *Patofisiologi, Diagnosis, dan Klasifikasi Tuberkulosis*, vol. 1 (Departemen Ilmu Kedokteran Komunitas, Okupasi, dan Keluarga, Fkui, Jakarta, 2019)
2. D. Iskandar, A.A. Suwantika, I.S. Pradipta, M.J. Postma, J.F.M. Van Boven, *The Lancet Global Health* **11**(1), e117 (2022). DOI 10.1016/S2214-109X(22)00455-7
3. W.H. Organization, *Global Tuberculosis Report 2021* (World Health Organization, 2021)
4. W.H. Organization, *Consolidated Guidelines on Tuberculosis Treatment* (World Health Organization, 2020)
5. K. Dheda, T. Gumbo, G. Maartens, K.E. Dooley, M. Murray, J. Furin, E.A. Nardell, R.M. Warren, A. Esmail, et al., *The Lancet Respiratory Medicine* **7**(9), 820 (2019). DOI 10.1016/S2213-2600(19)30263-2
6. W.H. Organization, *Global Tuberculosis Report* (World Health Organization, Geneva, 2022)
7. J.A.N. Wedderburn, R.W. M., *Journal of the Royal Statistical Society. Series A (General)* **135**(3), 370 (1972)
8. A.I. Khuri, *Linear Model Methodology*, 1st edn. (Chapman and Hall/CRC, 2009). DOI 10.1201/9781420010442

9. Y.A.M. Korkmaz, *Journal of Computational and Applied Mathematics* **403** (2022). DOI 10.1016/j.cam.2021.113819
10. E.J. Williams, *Regression Analysis* (John Wiley & Sons, 1959)
11. A.M. Al-Abood, D.H. Young, *IEEE Transactions on Reliability* **35**(2), 216 (1986). DOI 10.1109/TR.1986.4335408
12. A. Schutz, L. Bombrun, Y. Berthoumieu, M. Najim, in *European Signal Processing Conference* (2013)
13. A.M. Al-Abood, D.H. Young, *Communications in Statistics - Theory and Methods* **15**(6), 1865 (1986). DOI 10.1080/03610928608829223
14. V.E. Vinzi, W.W. Chin, J. Henseler, H. Wang, *Springer Handbooks of Computational Statistics* (Springer, 2011). DOI 10.1007/978-3-642-16345-6
15. Z.Y. Algamal, *Journal of Chemometrics* **32**(10) (2018). DOI 10.1002/cem.3054
16. E. Dunder, S. Gumustekin, M.A. Cengiz, *Journal of Applied Statistics* **45**(1), 8 (2018). DOI 10.1080/02664763.2016.1254730
17. A.S. Malehi, F. Pourmotahari, K.A. Angali, *Health Economics Review* **5**(1), 0 (2015). DOI 10.1186/s13561-015-0045-7
18. M.A. Stephens, *Journal of the American Statistical Association* **69**(347), 730 (1974)
19. H. Dukalang, B.W. Otok, I. Zain, H. Yusuf, in *Proceeding of 3rd International Conference on Research, Implementation and Education of Mathematics and Science* (2016), pp. M-37-M-44

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

