International Seminar on Student Research in Education, Science, and Technology
ISSN 3047-1532 | Volume 2 April 2025, pp. 144-150

https://www.issrestec.id/
http://journal.ummat.ac.id/index.php/issrestec
Universitas Muhammadiyah Mataram, Indonesia

Modeling Infant Mortality Rate In North Sumatra Using Robust Regression With Generalized Scale (Gs) Estimations

Isti Marfu'ah[1], Yuliana Susanti[2], Etik Zukhronah[3]
[123]Statistics Departement, Sebelas Maret University, istimrfh2@gmail.com

Abstract: Infant Mortality Rate (IMR) is a key indicator in assessing the success of national health development programs. In recent decades, Indonesia's IMR has declined, but consistent efforts are still needed, especially in provinces with high contributions. In 2023, North Sumatra ranked 3rd among provinces with the highest IMR. Therefore, an analysis is needed to identify factors that may contribute to infant mortality in this region. This study uses several variables, including the number of health centers, the percentage of births in health facilities, the number of low-birth-weight babies, and the number of pregnant women detected as HBsAg reactive. The aim is to develop an IMR model in North Sumatra that can serve as an evaluation reference. Regression analysis is a suitable method to understand the influence of these variables on IMR. However, unmet normality assumptions in classical regression may lead to inaccurate estimates. To address this, robust regression can be applied to obtain models that are more resistant to outliers. This study uses the Generalized Scale (GS) estimation method in robust regression. The resulting GS model produces an adjusted R^2 value of 86.76% and an AIC value of 108.2717.

Keywords: Robust Regression, GS Estimation, Infant Mortality.

Article History:
Received: 01-04-2025
Online : 24-04-2025

A. INTRODUCTION

Infant mortality is the number of deaths that occur in infants from birth until they reach the age of 12 months or 1 year. Infants less than 1 year of age are in a very vulnerable phase so that their condition can be one of the determining factors for the health and survival of infants in the future. IMR is one of the important indicators in determining the success of national health development programs, especially those related to the target of achieving Sustainable Development Goals (SDGs) which emphasize the importance of reducing infant and under-five mortality. North Sumatra itself is ranked 3rd as the province with the highest infant mortality in Indonesia in 2023. Based on the 2010 Population Census data, infant mortality decreased significantly from 26 deaths per 1,000 live births to 18.28 deaths per 1,000 live births from the SP2020 Long Form results (BPS Sumatera Utara, 2023). This figure shows a significant improvement, but the target of lower infant mortality must still be pursued so that the health of infants and children in North Sumatra is getting better.

Based on the description above, analysis and research related to infant mortality in North Sumatra are needed to obtain factors that have a significant effect on infant mortality so that in the future infant mortality can still be controlled and solutions can be found so that the downward trend in infant mortality remains consistent and improves for the next few years. One analysis that can be done is regression analysis with Ordinary Least Squares (OLS). Regression analysis is a technique used to examine and model the relationship between two variables (Montgomery et al., 2015). In regression analysis, parameter estimation is required which can be calculated using OLS. There are several assumption tests that are usually referred to as classical assumption tests (Nugraha, 2022). This assumption test must be met so that OLS can produce an accurate model (Sholihah et al., 2023). OLS often does not produce accurate model parameters if the residuals are not normally distributed (Nurdin et al., 2014).

Robust regression is a regression method that can be used when the residual data is not normally distributed which can be caused by outliers. Outliers are observations that deviate far from the data center or other observations. Data containing outliers can not only cause the data to be not normally distributed but can also have an effect on drawing conclusions or decisions in research (Sihombing et al., 2023). Robust regression is intended to accommodate the oddity of the data, while eliminating the identification of outlier data and is also automatic in dealing with outlier data. Robust regression analysis does not normalize the model, but the model produced by the robust method has a high level of accuracy (Susanti et al., 2021). One of the robust regression estimation methods is GS estimation, which is a minimization solution of the M estimate based on the difference of paired scale residuals. M estimation itself works by minimizing the residual function ρ.

Research on IMR modeling has been conducted by Husain and Jamaluddin (2024) using robust regression M estimation shows that the variable that has a significant effect is LBW with an R^2_{adj} of 69.8% (Husain & Jamaluddin, 2024). Research by Hamidah et al. (2023) on robust regression comparing the S estimate and found that the GS estimate is better in modeling the number of stunted toddlers by producing an R^2_{adj} of 99.6%. Other studies also show a significant influence between place of birth and infant mortality (Rukmono et al., 2021).

B. METHOD

This study uses secondary data taken from the official publication of the North Sumatra Provincial Health Office entitled North Sumatra Health Profile in 2023. The object used as the dependent variable is IMR data per district in North Sumatra in 2023, while the independent variables are the number of puskesmas (X_1), the percentage of births in health facilities (X_2), the number of low-birth-weight babies (X_3), and the number of pregnant women detected with HbsAG reactive (X_4).

This study used the GS estimation robust regression analysis method. The analysis was conducted with the help of R Studio software. The following are the steps taken.

1. Collect data including infant mortality rate per district in North Sumatra, number of health centers, percentage of births in health facilities, number of low-birth-weight infants, and number of pregnant women detected with HbsAG reactive.
2. Estimating the regression coefficient model using the OLS model.
3. Testing the classical assumptions of the regression model
   a. Conducting a normality test using Shapiro Wilk test
   b. Conducting a homoscedasticity test using Breusch Pagan test
   c. Conducting a non-autocorrelation test between residuals using the Durbin Watson test
   d. Conducting a non-multicollinearity test between independent variables using VIF.
4. Identifying outliers using the DFFITS method, if there are outliers then the data can be analyzed using robust regression.
5. Estimating robust regression coefficients using M estimation. The steps of M estimation are as follows
   a. Performing initial estimation with MKT to get the initial value of $\hat{\beta}_0$.
   b. Calculating the value of the residuals
   $$e_i = y_i - \hat{y}_i \tag{1}$$
   c. Calculating the standard deviation value
   $$\hat{\sigma}_M = \frac{MAD}{0.6745} = \frac{\text{median}|e_i - \text{median}(e_i)|}{0.6745} \tag{2}$$

d. Calculating the value of u_i

$$u_i = \frac{e_i}{\hat{\sigma}_M} \qquad (3)$$

e. Calculating weights using Tukey Bisquare weights

$$w_i = \begin{cases} \left(1-\left(\frac{u_i}{c}\right)^2\right)^2 & , |u_i| \leq c \\ 0 & , |u_i| > c \end{cases}$$

(4)

The value of c is the Tukey Bisquare weighting constant in the M estimation of 4,685.

f. Calculating the parameter $\hat{\beta}_M$ with the weight w_i.

g. Repeating steps b-f until a converged value of $\hat{\beta}_M$ is obtained.

6. Estimating the robust regression coefficient of GS estimation with the initial error used is the pairwise error obtained from the convergent M estimation error. The following are the GS estimation steps.

a. Calculating the pairwise residuals of the converged M-estimated residuals.

$$\Delta e_{ii}(\beta) = (\Delta e_{12}(\beta), \dots, e_{n-1}(\beta))\,' \qquad (5)$$

b. Calculating the standard deviation value.

$$\hat{\sigma}_{GS} = \frac{\text{median}|e_{ii} - \text{median}(e_{ii})|}{0{,}6745} \qquad (6)$$

c. Calculating the value of u_i.

$$u_i = \frac{e_i}{\hat{\sigma}_{GS}} \qquad (7)$$

d. Calculating the weight value with the weighting function w_{iGS}.

$$w_i = \begin{cases} \begin{cases} \left(1-\left(\frac{u_i}{c}\right)^2\right)^2 & , |u_i| \leq c \text{ , iteration} = 1 \\ 0 \text{ , } |u_i| > c & \end{cases} \\ \frac{\rho(u_i)}{u_i^2} \quad \text{, iteration} > 1 \end{cases} \qquad (8)$$

where $\rho(u_i)$ is defined as the Tukey Bisquare objective function.

$$\rho(u_i) = \begin{cases} \frac{u_i^2}{2} - \frac{u_i^2}{2c^2} + \frac{u_i^6}{6c^2}, & |u_i| \leq c \\ \frac{c^2}{6} & , |u_i| > c \end{cases} \qquad (9)$$

The value of c is the Tukey Bisquare weighting constant in the GS estimation of 0,9958.

e. Calculating the estimator $\hat{\beta}_{GS}$ by the IRLS method with w_i weights so that a new error is obtained.

f. Repeating iterations until a convergent $\hat{\beta}_{GS}$ is obtained.

7. Determine the value of R^2_{adj} and AIC in the GS estimation robust regression model.

C. RESULTS AND DISCUSSION

1. Regression Model with the Ordinary Least Squares (OLS)

The regression model using OLS on IMR data in North Sumatra obtained the following regression equation.

$$\hat{Y} = 15.240025 - 0.184481\,X_1 - 0.117475\;X_2 + 0.044703\;X_3 + 0.026202\;X_4$$

The R^2_{adj} value is 73.58%, which means that the independent variables used in this study are the number of puskesmas, the percentage of births in health facilities, the number of lowbirth weight babies, and the number of pregnant women detected with reactive HbsAG can explain the dependent variable infant mortality rate by 73.58% and the remaining 26.42% is explained by other factors outside the model.

2. **Classical Assumption Test**

   a. Normality Test

   The normality test aims to determine whether the residuals from the model are normally distributed or not. The statistical test used is the Shapiro Wilk test. The residuals from the model is said to be normally distributed if $W \geq W_{\alpha(n)}$ or $p - value >$ α (Permana & Ikasari, 2023). The $p - value = 0.01375 < \alpha = 0.05$ and $W = 0.91546 \leq W_{0.05(33)} = 0.931$. Based on the results of the Saphiro Wilk test, it can be concluded that the residual model is not normally distributed.

   b. Homoscedasticity Test

   Homoscedasticity test is conducted to test whether in a regression there is an inequality of variance of the residuals from one observation to another. One test that can be used is the Breusch-Pagan test (Andriani, 2017). The residual variance is said to be inhomogeneous if $\phi_{hitung} > X^2_{(\alpha,k)}$ or $p - value < \alpha$. The value of $\phi_{hitung} = 2.9439 < X^2_{(0.05,4)} = 9.488$ and $-value = 0.5673 > \alpha = 0.05$ are obtained. Based on the results of the Breusch-Pagan test, it is concluded that the residuals variance is homogeneous.

   c. Non-autocorrelation test

   The non-autocorrelation test aims to test whether there is a correlation between errors in period t and errors in $t - 1$ or the previous period. The method that can be used to detect autocorrelation is the Durbin Watson test. Residues are declared to have autocorrelation if the value of $0 < \mathrm{d} < \mathrm{d_L}$ or $4 - \mathrm{d_L} < \mathrm{d} < d_L$ based on the Durbin Watson table with a value of n = 33 and k = 4 obtained a value of $\mathrm{d_U} = 1.7298$ and $\mathrm{d_L} = 1.1927$ or $p - value < \alpha = 0.05$. Obtained a $p - value = 0.1084$ and a value of $\mathrm{d} = 1.6108$. Based on the results of the Durbin Watson test, it is concluded that there is no autocorrelation between residuals.

   d. Non-multicollinearity test

   The non-multicollinearity test aims to determine whether there is a correlation between the independent variables in the regression model. The presence or absence of multicollinearity can be detected from the VIF (Variance Inflation Factors) value (Montgomery et al., 2015). Variables can be said to have multicollinearity if the VIF value is > 10. The VIF value generated using R-studio software for each independent variable is as shown in Table 1.

Table 1. Non-multicollinearity Test Results

Independent Variable	VIF
X_1	1.267964
X_2	1.027708
X_3	1.049568
X_4	1.257933

   The VIF value for all independent variables X < 10 so it can be concluded that there is no multicollinearity between the independent variables.

3. **Outlier Detection**

The method that can be used to detect the presence or absence of outliers in the data is by using Difference in Fitted Value (DFFITS) (Ihsan et al., 2019). A data can be said to be an outlier if.

$$|DFFITS| > 2\sqrt{\frac{k+1}{n}} = 2\sqrt{\frac{5}{33}} = 0.7786$$

Table 2 shows the observations that are considered as outliers in the IMR data of North Sumatra Province.

Table 2. DFFITS Value on Data

Data	DFFITS	|DFFITS|
7	1.168837	1.168837
8	1.379508	1.379508
13	-1.559372	1.559372
26	-2.769208	2.769208
33	2.063095	2.063095

In the outlier detection results, the 7th, 8th, 13th, 26th and 33rd data are outliers because they have a $|DFFITS|$ value > 0.7786.

4. GS Estimation Robust Regression Model

Calculation of the GS estimate is done by calculating the M estimate until $\hat{\beta}_M$ converges. The calculation results using of R-studio software obtained iteration results as in Table 3.

Table 3. Coefficient Value of $\hat{\beta}_M$ each Iteration at Estimation M

Iteration	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
1	14.772646	-0.171053	-0.113798	0.039722	0.022076
2	14.428985	-0.161395	-0.110934	0.037519	0.020520
3	14.308973	-0.157414	-0.110128	0.037128	0.020147
4	14.268531	-0.155834	-0.109932	0.037041	0.020026
5	14.262762	-0.155386	-0.110010	0.037019	0.020019
6	14.267465	-0.155380	-0.110129	0.037016	0.020045
7	14.273433	-0.155498	-0.110219	0.037020	0.020072
8	14.277778	-0.155613	-0.110272	0.037024	0.020089
9	14.280208	-0.155688	-0.110295	0.037026	0.020099
10	14.281273	-0.155727	-0.110303	0.037028	0.020103
⋮	⋮	⋮	⋮	⋮	⋮
20	14.281295	-0.155739	-0.110298	0.037028	0.020102
21	14.281296	-0.155739	-0.110298	0.037028	0.020102
22	14.281297	-0.155739	-0.110298	0.037028	0.020102
23	14.281297	-0.155739	-0.110298	0.037028	0.020102

Table 3 shows that the value of $\hat{\beta}_M$ converges and the calculation process stops at the 23rd iteration with the M estimation robust regression model, namely.

$$\hat{Y}_M = 14.281297 - 0.155739\, X_1 - 0.110298\, X_2 + 0.037028\, X_3 + 0.020102\, X_4$$

The M estimation model produces an AIC value of 114.6261 and an R^2_{adj} value of 76.37, which means that the independent variables used in this study, namely the number of puskesmas, the percentage of births in health facilities, the number of low birth weight babies, and the number of pregnant women detected with reactive HbsAG, can explain the dependent variable infant mortality rate by 76.37% and the remaining 23.63% is explained by other factors outside the model.

The next step is to perform calculations using GS estimation using paired residuals from M estimation until $\hat{\beta}_{GS}$ converges with iteration results as in Table 4.

Table 4. Coefficient Value of $\hat{\beta}_{GS}$ each Iteration at Estimation GS

Iteration	$\widehat{\boldsymbol{\beta}}_0$	$\widehat{\boldsymbol{\beta}}_1$	$\widehat{\boldsymbol{\beta}}_2$	$\widehat{\boldsymbol{\beta}}_3$	$\widehat{\boldsymbol{\beta}}_4$
1	14.772646	-0.171053	-0.113798	0.039722	0.022076
2	14.428985	-0.161395	-0.110934	0.037519	0.020520
3	14.308973	-0.157414	-0.110128	0.037128	0.020147
4	14.268531	-0.155834	-0.109932	0.037041	0.020026
5	14.262762	-0.155386	-0.110010	0.037019	0.020019
6	14.267465	-0.155380	-0.110129	0.037016	0.020045
7	14.273433	-0.155498	-0.110219	0.037020	0.020072
8	14.277778	-0.155613	-0.110272	0.037024	0.020089
9	14.280208	-0.155688	-0.110295	0.037026	0.020099
10	14.281273	-0.155727	-0.110303	0.037028	0.020103
⋮	⋮	⋮	⋮	⋮	⋮
26	15.476677	-0.202797	-0.120719	0.039440	0.026330
27	15.476680	-0.202797	-0.120719	0.039440	0.026330
28	15.476682	-0.202797	-0.120719	0.039440	0.026330
29	15.476684	-0.202797	-0.120719	0.039440	0.026330
30	15.476684	-0.202797	-0.120719	0.039440	0.026330

Table 4 shows that the $\hat{\beta}_{GS}$ value converges and the calculation process stops at the 30th iteration with the robust GS estimation model, namely.

$$\hat{Y}_{GS} = 15.476684 - 0.202797X_1 - 0.120719\,X_2 + 0.039440\,X_3 + 0.026330\,X_4$$

The GS estimation model produces an AIC value of 108.2717 and an R^2_{adj} value of 86.76%.

D. CONCLUSIONS AND SUGGESTIONS

The robust regression model of GS estimation with Tukey Bisquare weights on infant mortality data is $\hat{Y}_{GS} = 15.476684 - 0.202797X_1 - 0.120719\,X_2 + 0.039440\,X_3 + 0.026330\,X_4$. The model shows that every one increase in the number of health centers will decrease the infant mortality rate by 0.202797, every one percent increase in births in health facilities will decrease the infant mortality rate by 0.120719, every one increase in low-weight babies will increase the infant mortality rate by 0.039440, and every one increase in pregnant women detected with HbsAG reactive, the infant mortality rate will increase by 0.026330.

The GS estimation model produces an AIC value of 108.2717 and an R^2_{adj} value of 86.76%, which means that the independent variables used in the study can explain the dependent variable by 86.76% and the remaining 13.24% is explained by other factors outside the model. Other researchers who are interested in continuing this research can consider trying other robust regression estimation methods. The variables used can also be studied more deeply to find other variables that may have more influence on infant mortality rates in North Sumatra.

ACKNOWLEDGMENTS

All praise and gratitude are sincerely offered to Allah SWT for His abundant mercy and blessings, which have enabled the successful completion of this research. The authors also extend their sincere appreciation to Universitas Sebelas Maret for its support and assistance throughout this study.

REFERENCES

Andriani, S. (2017). Uji Park Dan Uji Breusch Pagan Godfrey Dalam Pendeteksian Heteroskedastisitas Pada Analisis Regresi. *Al-Jabar : Jurnal Pendidikan Matematika*, *8*(1), 63–72. https://doi.org/10.24042/ajpm.v8i1.1014

Badan Pusat Statistik Sumatera Utara. (2023). *Sensus Penduduk 2020 - Sumatera Utara*. *09*, 1–44. http://sp2010.bps.go.id/

Hamidah, I., Susanti, Y., & Sugiyanto. (2023). *Seminar Nasional LPPM UMMAT Perbandingan Estimasi Scale Dan Estimasi Generalized Scale Estimation Dalam Pemodelan Balita Stunting Indonesia*. *2*(April), 539–546.

Husain, A., & Jamaluddin, S. R. W. (2024). Pemodelan Data Angka Kematian Bayi Menggunakan Regresi Robust. *SAINTEK: Jurnal Sains, Teknologi & Komputer*, *1*(1), 1–7. https://jurnal.larisma.or.id/index.php/SAINTEK/article/view/326

Ihsan, H., Sanusi, W., & Nurfadillah, N. (2019). Estimasi Parameter Regresi Linear Pada Kasus Data Outlier Menggunakan Metode Estimasi Method Of Moment. *Journal of Mathematics, Computations, and Statistics*, *1*(1), 38. https://doi.org/10.35580/jmathcos.v1i1.9176

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2015). *Introduction to Linear Regression Analysis-4th edition*. 6.

Nugraha, B. (2022). *Pengembangan Uji Statistik: Implementasi Metode Regresi Linier Berganda dengan Pertimbangan Uji Asumsi Klasik*. Applied Mathematics and Computation (Issue 9), Pradina Pustaka. https://books.google.co.id/books?hl=en&lr=&id=PzZZEAAAQBAJ&oi=fnd&pg=PR5&dq=+Pengembangan+uji+statistik:+Implementasi+metode+regresi+linier+berganda+dengan+pertimbangan+uji+asumsi+klasik.+In+Applied+Mathematics+and+Computation+(Issue+9).+&ots=KxyX_3ubqj&s

Nurdin, N., Raupong, & Islamiyati, A. (2014). Penggunaan Regresi Robust Pada Data Yang Mengandung Pencilan Dengan Metode Momen. *Matematika, Statistika Dan Komputasi*, *10*(2), 115.

Permana, R. A., & Ikasari, D. (2023). Uji Normalitas Data Menggunakan Metode Empirical Distribution Function Dengan Memanfaatkan Matlab Dan Minitab 19. *Semnas Ristek (Seminar Nasional Riset Dan Inovasi Teknologi)*, *7*(1), 7–12. https://doi.org/10.30998/semnasristek.v7i1.6238

Rukmono, P., Anggunan, Pinilih, A., & Yuliawati, Si. Sh. (2021). Hubungan Antara Tempat Melahirkan Dengan Angka Kematian Neonatal Di Rsud Dr. H. Abdoel Moeloek Provinsi Lampung. *Student Journal, Prevalensi Hbsag Positif Antara Donor Darah Sukarela Dengan Donor Darah Pengganti Di Utd Pmi Provinsi Lampung Tahun 2019-2020*, *1*, 435–444.

Sholihah, S. M., Aditiya, N. Y., Evani, E. S., & Maghfiroh, S. (2023). Konsep Uji Asumsi Klasik Pada Regresi Linier Berganda. *Jurnal Riset Akuntansi Soedirman*, *2*(2), 102–110. https://doi.org/10.32424/1.jras.2023.2.2.10792

Sihombing, P. R., Suryadiningrat, Sunarjo, D. A., & Yuda, Y. P. A. C. (2023). Identifikasi Data Outlier (Pencilan) dan Kenormalan Data Pada Data Univariat serta Alternatif Penyelesaiannya. *Jurnal Ekonomi Dan Statistik Indonesia*, *2*(3), 307–316. https://doi.org/10.11594/jesi.02.03.07

Susanti, Y., Pratiwi, H., & Qona’ah, I. (2021). *Regresi Robust Teori dan Penerapannya* (I). UNS PRESS.