Journal of English Language and Education
ISSN 2597- 6850 (Online), 2502-4132 (Print)
Journal Homepage: https://jele.or.id/index.php/jele/index

OPEN ACCESS Article

Modeling Expected Years of Schooling Based on Socioeconomic Factors in Banten Province Using the Gamma Regression Model

https://doi.org/10.31004/jele.v11i2.2020

*Resti Oktaviani[a]

[1]Military Mathematics Study Program, Faculty of Mathematics and Natural Sciences, Military Defense University of the Republic of Indonesia, Indonesia
Corresponding Author: restioktaviani729@gmail.com

ABSTRACT

Despite rapid economic growth in Banten Province, the disparity in Expected Years of Schooling (HLS) between districts/cities remains significant due to socioeconomic factors. This study aims to analyze the effect of Poverty Percentage, Gender Development Index (IPG), and Beneficiary Families (KPM) on HLS using Gamma Regression. A quantitative approach using secondary data from BPS Banten (2020-2024), the population of all districts/cities (n=42 total sampling observations), was analyzed through GLM R Studio with the link log function. The results show that IPG has a significant positive effect on HLS (p=0.007), while Poverty Percentage and KPM have a non-significant negative effect. The model has a high fit (AIC=79.05, MAPE=3.62%, RMSE=0.581) without heteroscedasticity, autocorrelation, or multicollinearity. Gender equality has been shown to be crucial for increasing access to education, although poverty remains a structural barrier.

Keywords: *Banten Province, Gamma Regression, Gender Development Index, Expected Years of Schooling, Socioeconomic Factors*

Article History:
Received 20th February 2026
Accepted 09th March 2026
Published 10th March 2026

INTRODUCTION

Education is a key pillar in human development that improves the quality of human resources and the nation's competitiveness, with Expected Years of Schooling (HLS) as a key indicator reflecting the average length of formal education expected by an individual (Mahya & Widowati, 2021; Arofah & Rohimah, 2019). Banten Province shows rapid economic growth on the island of Java, with the provincial HDI reaching 76.35 in 2024, but is accompanied by significant socioeconomic disparities between districts/cities (Mahya & Widowati, 2021; Rahmawati & Hidayah, 2024).

Despite this, HLS variation across regions in Banten remains high, with impoverished areas like Lebak and Pandeglang having lower scores than urban areas like South Tangerang (Dewi et al., 2025; Rahmawati & Hidayah, 2024). BPS data shows that relatively high poverty correlates with low HLS, while better-off areas tend to have superior access to education (Mahya & Widowati, 2021; Dewi et al., 2025).

The problem is further complicated by poverty, gender disparities, and social assistance such as Beneficiary Families (KPM), which hinder educational equality, with poor families prioritizing basic needs over their children's schooling (Dewi et al., 2025; Rahmawati & Hidayah, 2024). Previous research confirms that women's education reduces poverty, but the GPI in Banten is still not optimal in supporting equitable HLS (Rahmawati & Hidayah, 2024; Dewi et al., 2025).

HLS data in Banten is continuous, positive, and right-skewed, so conventional linear regression is not appropriate due to heteroscedasticity, requiring a GLM approach such as Gamma Regression for accurate analysis (Mahya & Widowati, 2021; Rahmawati & Hidayah, 2024).

This study aims to analyze the influence of Poverty Percentage, GDI, and KPM on HLS in Banten Province using Gamma Regression with log link, which is urgent because it provides an empirical basis for inclusive education policies amidst Banten's inequality in 2020-2024 (Mahya & Widowati, 2021; Dewi et al., 2025). Its novelty lies in the application of Gamma Regression specifically to right-skewed HLS data at the Banten district/city level, going beyond previous general studies that focused on national HDI or linear regression, thus offering an accurate predictive model for gender and poverty interventions (Rahmawati & Hidayah, 2024; Mahya & Widowati, 2021).

METHOD

Types and Methods of Research

This study uses a quantitative approach with a descriptive and analytical design to analyze the relationship between Expected Years of Schooling (HLS) as a response variable and socioeconomic factors such as Poverty Percentage, Gender Development Index (IPG), and Beneficiary Families (KPM) in Banten Province for the 2020-2024 period. The Gamma Regression method as part of the Generalized Linear Model (GLM) was chosen because it suits the characteristics of HLS data which is continuous, positive, and right-skewed distribution, thus being able to overcome heteroscedasticity and non-constant variance that often appear in socioeconomic data (Mahya & Widowati, 2021; Sugiyono, 2021; Creswell & Creswell, 2023). This approach allows for accurate parameter estimation through Maximum Likelihood Estimation (MLE) with a log link function, as recommended for predictive models on regional education indicators (Rahmawati & Hidayah, 2024; Sudaryono, 2022).

Data Analysis Instruments and Techniques

The main research instrument is secondary data from the Central Statistics Agency (BPS) of Banten Province, including the dependent variable HLS and the independent variables Poverty Percentage, GDI, and KPM, which were processed using R Studio software with the stats, MASS, and car packages for GLM modeling. Data analysis techniques include descriptive exploration (mean, min-max, distribution), Gamma distribution assumption tests, regression model estimation, and diagnostics such as the Wald test, Likelihood Ratio, Deviance, Breusch-Pagan, Durbin-Watson, VIF, MAPE, and RMSE to validate model suitability (Mahya & Widowati, 2021; Emzir, 2021; Creswell & Creswell, 2023). This process is supported by visualizations such as residual plots, QQ-plots, and partial effects to interpret the partial influence of variables, ensuring the model's robustness to right-skewed data typical of education indicators (Rahmawati & Hidayah, 2024; Sugiyono, 2021).

Population and Sample

The study population comprised all regencies/cities in Banten Province (Serang, Lebak, Pandeglang, Tangerang, South Tangerang, Tangerang City, and Serang City) with annual data from BPS for 2020-2024, resulting in a panel analysis unit of 42 observations (7 regions x 6 years). The sample was total sampling because it covered the entire population without external generalization, in accordance with the characteristics of representative administrative secondary data for regional analysis (Mahya & Widowati, 2021; Sudaryono, 2022; Emzir, 2021). This approach allows for spatial mapping of HLS inequality between regions, with southern areas such as Lebak showing lower values than northern areas (Rahmawati & Hidayah, 2024; Creswell & Creswell, 2023).

Research Procedures

The research procedure was carried out in stages: first, collection and exploration of secondary data from BPS; second, testing the HLS distribution to confirm right-skewed; third, building a Gamma Regression model via the glm() function with a log link; fourth, parameter significance testing (Wald, Likelihood Ratio) and diagnostic assumptions; and fifth, interpretation of partial effects and accuracy validation (AIC, MAPE, RMSE). Processing in R Studio ensured reproducibility, with a focus on interpreting exponential coefficients for the proportional impact of variables on HLS (Mahya & Widowati, 2021; Sugiyono, 2021; Sudaryono, 2022). All steps followed standard quantitative analysis protocols for GLM on

social data, resulting in a reliable predictive model for Banten education policy (Rahmawati & Hidayah, 2024; Emzir, 2021; Creswell & Creswell, 2023).

FINDINGS AND DISCUSSIONS

Result

Descriptive Analysis of Research Data

This study uses secondary data from the Central Statistics Agency (BPS) of Banten Province for the 2020–2024 period. The variables analyzed include Expected Years of Schooling (HLS) as the dependent variable, and Poverty Percentage, Beneficiary Families (KPM), and the Gender Development Index (IPG) as the independent variables.

Table 1. Descriptive Statistical Results

Variables	Minimum	Maximum	Average
Expected Years of Schooling (Years)	11.97	14.70	13.22
Poverty Percentage (%)	2.29	10.72	6.07
Beneficiary Family (Soul)	11,008	823,986	234,291
Gender Development Index	79.81	95,079,816	90.00

The residual values of the model range between− 0,071to indicate that the model predictions are relatively close to the actual data.0,076

Table 2. Correlation Results between Variables

No	Relationship between variables	Types of Correlation	r value
1	Poverty Percentage and KPM	Positive	0.435
2	Poverty Percentage and GDI	Negative	-0.642
3	KPM and IPG	Negative	-0.335

The correlation between variables shows that the Poverty Percentage is positively correlated with the KPM (r = 0.435), while the Poverty Percentage is negatively correlated with the GDI (r = –0.642). This indicates that areas with high poverty rates tend to have lower gender equality.

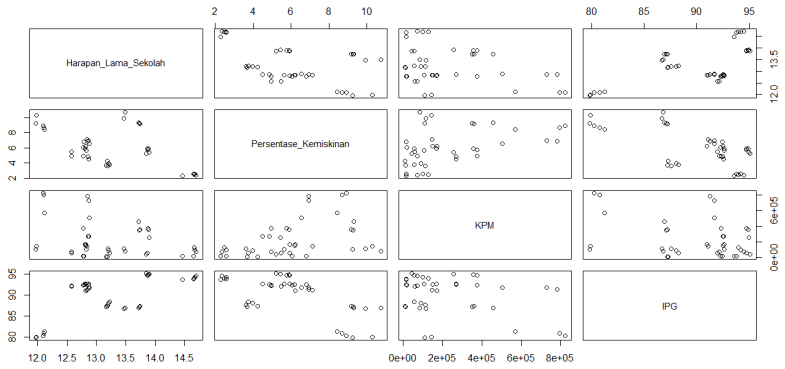

Figure 1. Correlation between Variables

Figure 1 shows the relationships between the variables used in the study. It can be seen that Expected Years of Schooling has a negative relationship with the Poverty Percentage and the KPM, and a positive relationship with the GPI. This pattern indicates that increasing poverty tends to decrease expected years of schooling, while increasing GPI is associated with an increase in expected years of schooling.

Table 3. The Results of the Correlation Analysis between Independent Variables Show the Following Values

Variables	Poverty Percentage	KPM	IPG
Poverty Percentage	1.00	0.43	-0.64
KPM	0.43	1.00	-0.33
IPG	-0.64	-.33	1.00

This correlation value indicates the absence of high multicollinearity between variables because all correlation coefficients are <0.8. Thus, all variables can be used in the Gamma regression model.

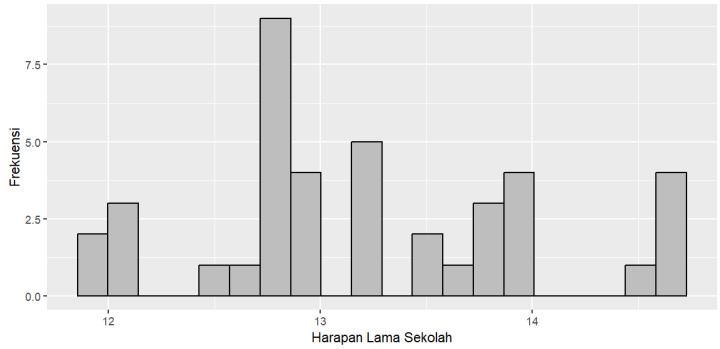

Figure 2. Distribution of Expected Years of Schooling

The distribution of the HLS variable appears positively skewed, with most values falling in the 12–14 year range. This pattern aligns with the assumptions of a Gamma distribution, as the data are positive and continuous.

Gamma Regression Modeling Results

Modeling was carried out using Gamma Regression with a log link function, because the response variable (HLS) is positive and continuous.

The regression model formed is as follows:

$$\log(\mu) = 2.066 - 0.00376(\text{Kemiskinan}) - 2.983 \times 10^{-8}(\text{KPM}) + 0.00605(\text{IPG})$$

Table 4. Gamma Regression Model Estimation Results

Variables	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	2,066	0.207	9,959	<0.001	***
Poverty Percentage	-0.00376	0.00427	-0.879	0.385	ns
KPM	-2.98e-08	3.48e-08	-0.857	0.397	ns
IPG	0.00605	0.00212	2,857	0.007	**

The model has an AIC value of 79.05, BIC value of 87.49, and residual deviance of 0.0755 (df = 36). The Likelihood Ratio test yields a p-value of 8.32×10^{-5}, which means that the model with predictor variables is much better than the model without variables.

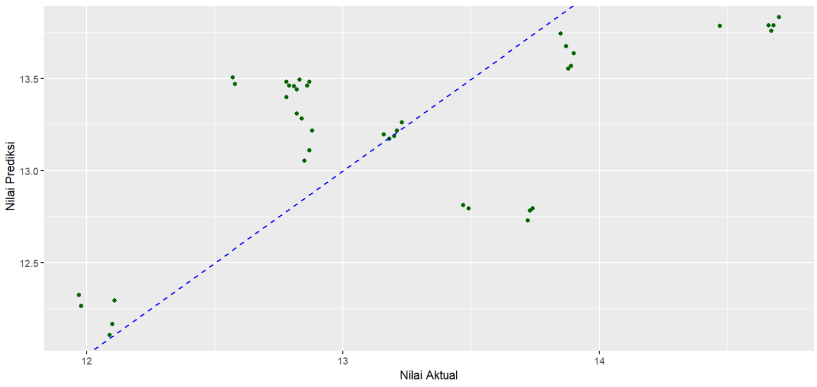

Figure 3. Plot of Actual Vs Predicted Values

Figure 3 shows that the predicted values of the Gamma model are close to the actual Expected Years of Schooling values, with a distribution pattern that relatively follows the diagonal line. This indicates that the model has a good level of accuracy.

Testing Assumptions and Model Suitability

Table 5. Results of the Gamma Regression Model Diagnostic Test

Types of Diagnostic Tests	Statistics / Test Values	p-value	Interpretation
Deviance Test	–	1,0000	The model fits the data (no significant deviations)
Breusch–Pagan Test	–	0.1377	There is no heteroscedasticity
Durbin–Watson Test	DW = 1.8275	0.291	There is no autocorrelation between residuals
Pregibon Link Test	–	> 0.85	The log link function is used appropriately
Variance Inflation Factor (VIF)	< 2	–	There is no multicollinearity between independent variables
Mean Absolute Percentage Error (MAPE)	3.62%	–	Low prediction error rate
Root Mean Square Error (RMSE)	0.581	–	The model has good accuracy against actual data.

Diagnostic tests showed that the Gamma regression model met all required assumptions—including goodness-of-fit, the absence of heteroscedasticity and autocorrelation, and the appropriate selection of the log link function. Low MAPE and RMSE values indicate that the model has excellent predictive ability for Expected Years of Schooling in Banten Province.

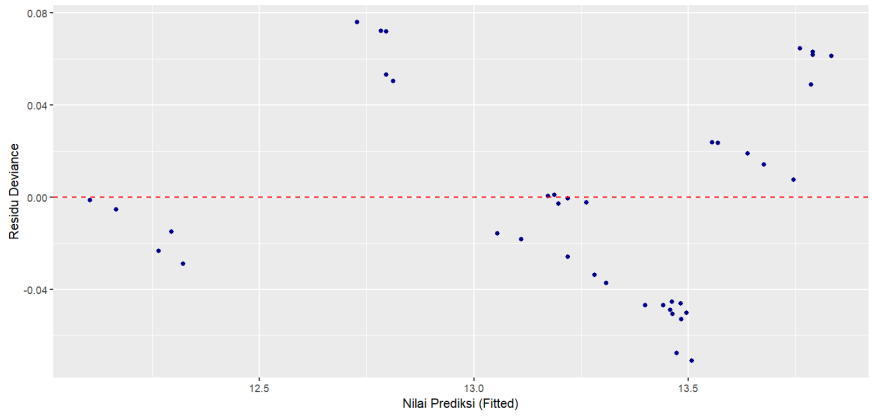

Figure 4. Plot of Residuals vs Predicted Values

This figure shows that the residuals are randomly distributed around zero without any particular pattern, indicating that the assumptions of homoscedasticity and model specifications are met.

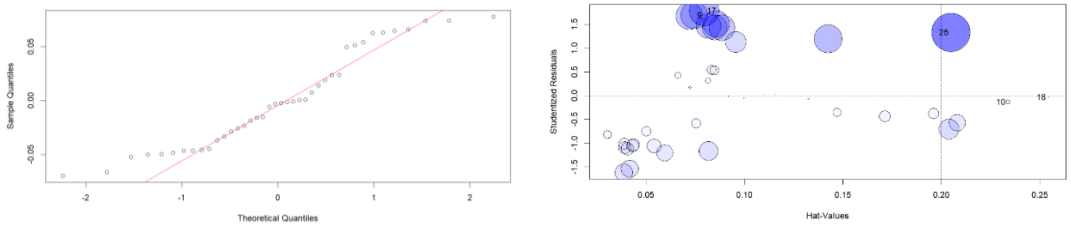

Figure 5. QQplot of Residuals and Outlier/Leverage Graph

The QQ Plot shows that the residuals follow a normal distribution, while the leverage graph shows that no observations are extreme outliers.

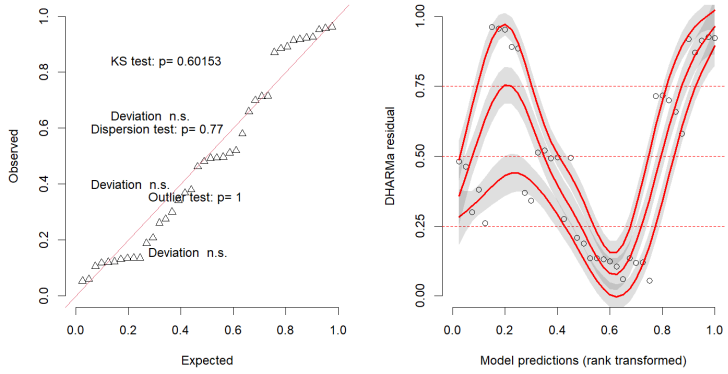

Figure 6. Residue Simulation Test

Simulation tests of the residuals with the DHARMa package showed a uniform distribution of residuals and did not deviate significantly from the expectations of the Gamma model.

Partial Effect Analysis of Predictor Variables

Table 6. The exponential coefficient values show the proportional impact of each variable on HLS:

Variables	Exp(Estimate)	Effects on HLS
Poverty Percentage	0.996	Reduce HLS by 0.37%
KPM	1,000	Almost no effect
IPG	1,006	Increase HLS by 0.61%

To understand the influence of each variable on HLS, partial effect plots from the effects and visreg packages were used.

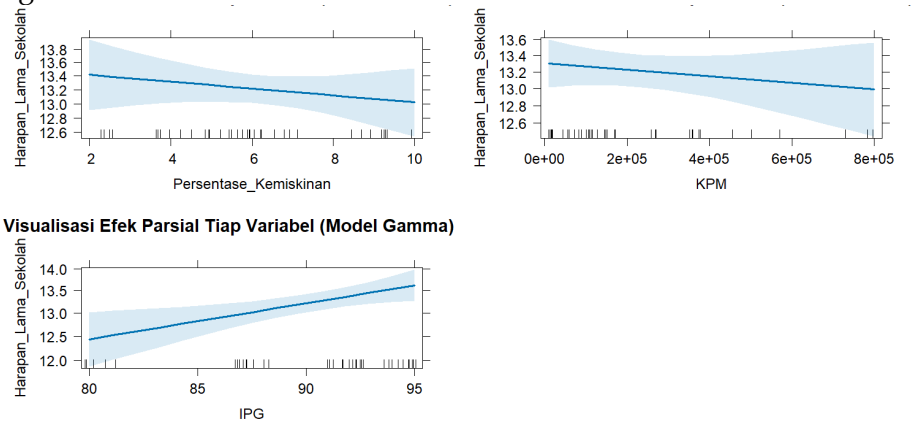

Figure 7. Partial Effect Visualization

The figure shows that an increase in the GPI has a significant positive effect on HLS, while increases in poverty and KPM show no significant changes. The three graphs show that the GPI regression line rises as the GPI value increases, indicating a positive effect on expected years of schooling.

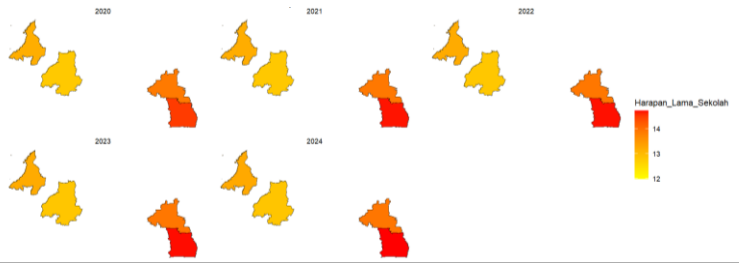

Figure 8. Map of the Distribution of Expected Years of Schooling in Banten Province 2020–2024

The figure above shows the spatial distribution of Expected Years of Schooling (HLS) in Banten Province for the 2020–2024 period. Yellow indicates high HLS values, while red indicates lower HLS. It appears that northern regions, such as Tangerang City and South Tangerang, consistently have higher HLS than southern regions (Lebak and Pandeglang). This

pattern indicates a relatively stable spatial imbalance in educational development in Banten Province from year to year.

Discussions

The results of this study align with human development theory, which states that gender equality is a crucial factor in improving the quality of education (UNDP, 2023). The increase in the GPI reflects equal access to education between men and women, thus directly impacting the average length of schooling. The negative impact of poverty on education indicates that economic constraints remain a major barrier to participation in formal education, although the effect is not yet statistically significant. The insignificance of the KPM variable indicates that social assistance programs have not had a direct impact on education indicators, as they focus more on meeting basic needs than on investing in education.

CONCLUSION

This study concludes that the Gamma Regression model with a log link function effectively models Expected Years of Schooling (HLS) in Banten Province for the 2020-2024 period, with the Gender Development Index (IPG) having a positive and significant effect on HLS (p=0.007), while the Poverty Percentage and Beneficiary Families (KPM) have a negative but insignificant effect. The model shows a high fit (AIC=79.05, MAPE=3.62%, RMSE=0.581) and meets the assumptions of no heteroscedasticity, autocorrelation, and multicollinearity, thus producing accurate predictions for right-skewed data. These findings confirm the crucial role of gender equality in improving access to education, although poverty remains a structural barrier. However, limitations of this study include the use of secondary data from BPS, which is prone to reporting imperfections, and the focus on only three variables, which does not fully capture spatial dynamics or other factors such as school infrastructure. Future research suggests adding control variables such as per capita education expenditure, a spatial regression approach, or primary data through household surveys to improve generalizability. Practically, these results recommend that the Banten government prioritize programs to increase the GPI through gender-responsive scholarships and the integration of KPM assistance with education incentives, in order to reduce inter-regional HLS disparities and support sustainable human development.

REFERENCES

Arofah, I., & Rohimah, S. (2019). Path analysis for the influence of Life Expectancy, Expected Years of Schooling, Average Years of Schooling on the Human Development Index through Real Expenditure Per Capita in East Nusa Tenggara Province. Jurnal Saintika UNPAM: Jurnal Sains dan Matematika, 2(1), 76-87.

Central Statistics Agency (BPS) Gorontalo. (2024). Human Development Index and Socioeconomic Factors 2024.https://gorontalo.bps.go.id

Creswell, J. W., & Creswell, J. D. (2023). Research design: Qualitative, quantitative, and mixed methods approaches (6th ed.). SAGE Publications.https://doi.org/10.1007/978-3-031-13137-7

Dewi, AK, Septiani, RE, Rahmah, S., & Qurrota Aini, S. (2025). The Impact of Women's Education on Poverty in Indonesia. Economic and Education Journal (Ecoducation), 7(1).https://ejurnal.uibu.ac.id/index.php/ecoducation/article/view/1364

Dewi, NP, Sari, RM, & Pratama, IG (2025). Analysis of the role of women's education in reducing poverty levels in Indonesia. Indonesian Journal of Development Economics, 12(1), 45–56.

Emzir. (2021). Quantitative research methodology (Revised Edition). Prenada Media.

Mahya, AJ, & Widowati, W. (2021). The Effect of Expected Years of Schooling, Average Years of Schooling, and Per Capita Expenditure on the Human Development Index. Prismatika: Journal of Mathematics Education and Research, 3(2).https://ejurnal.uibu.ac.id/index.php/prismatika/article/view/784

Rahmawati, A., & Hidayah, S. (2024). The influence of gender equality on economic growth and education quality in Indonesia. Journal of Social and Development Studies, 15(3), 201–214.

Rahmawati, F., & Miftha'ul Hidayah, Z. (2024). Exploring the Relationship between the Gender Development Index and Economic Growth. EcceS: Economics, Social, and Development Studies, 7(1).https://doi.org/10.24252/ecces.v7i1.13919

One Data Dharmasraya. (2024a). Expected Years of Schooling (HLS) – Conjunto de datos.https://satudata.dharmasrayakab.go.id/pt_BR/dataset/harapan-lama-sekolah

Sudaryono. (2022). Quantitative, qualitative, and mixed methods research methods. Student Library.

Sugiyono. (2021). Quantitative, qualitative, and R&D research methods. Alfabeta.

United Nations Development Programme. (2023). Human development report 2023/2024.http://hdr.undp.org

Widodo, E., Hakim Akbar Alhaqq, F., Putri Ginastuti, A., Alifyah, H., & Nindia, ZP (2025). Panel data regression analysis for modeling the open unemployment rate in West Java Province in 2019-2022. Emerging Statistics and Data Science Journal, 3(3), 650-665.