UNP JOURNAL MANAGEMENT CENTER

math.ppj.unp.ac.id
p-ISSN 2716-0726
e- ISSN 2716-0734

Indexed

Article History
Vol. 4, No. 1, 2025

Subject Areas:
Statistics, Applied Statistics

Keywords:
Linear_regression, polynomial_local, regression_analysis, stunting

Author for correspondence:
Fadhilah Fitri
e-mail: fadhilahfitri@fmipa.unp.ac.id

RANGKIANG MATHEMATICS JOURNAL
Journal of Mathematics and Mathematics Education

Comparison of Linear Regression and Polynomial Local Regression in Modeling Prevalence of Stunting

Fadhilah Fitri[1] and Mawanda Almuhayar[2]

[1] Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Padang

[2] Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Andalas

Abstract- Stunting is one of the main focuses of the government in Indonesia. This is because nutritional status is one of the benchmarks of community welfare. Stunting can be influenced by various societal aspects such as health, economy, social status, and education. One factor that is thought to be closely related to stunting is the level of education. Therefore, the prevalence of stunting and the level of education will be modeled; in this case, the mean years of schooling is used. Modeling uses two approaches: parametric through linear regression and nonparametric through local polynomial regression. This study compares both models to see which method better explains the stunting phenomenon. The comparison is made through the determination coefficient value or R^2, Root Mean Square Error or RMSE, and the fitted curve plot. The results of R^2 and RMSE for both models were obtained. The linear regression model has an R^2 of 32.94% and an RMSE of 4.84. Meanwhile, for the local polynomial model, it is R^2 43.44% and RMSE 4.32. Based on these results, it can be concluded that local polynomial regression is better at modeling the relationship between the prevalence of stunting and mean years of schooling in Indonesia. This finding confirms that the polynomial local regression method can capture phenomena that occur for data that do not follow a particular pattern.

1. Introduction

Stunting is a significant public health issue characterized by toddlers having below-average height due to inadequate nutritional intake over an extended period. This condition can cause various future risks that can potentially reduce the quality of Human Resources (Huey & Mehta, 2016; Montenegro et al., 2022; Raiten & Bremer, 2020). Stunting has long been a serious challenge for Indonesia, with its peak

occurring in 2013, where the prevalence reached 37% (Aryastami, 2017). The government's efforts have begun to show positive results, where, based on the 2023 Indonesian Health Survey, the prevalence of stunting has dropped to 21.5%, although it only dropped 0.1% from the previous year (Badan Kebijakan Pembangunan Kesehatan, Kementerian Kesehatan, 2023). However, this figure is still below the WHO threshold of less than 20% (De Onis et al., 2019). Therefore, more efforts are needed to reduce stunting rates, including finding the factors that cause stunting.

The factors influencing stunting are very complex, including health, economic, social, and educational aspects. One factor that is thought to be closely related to stunting is the level of education. Several studies that have been conducted show that maternal education has a significant impact on the prevalence of stunting (Casale et al., 2018; Laksono et al., 2022; Makoka & Masibo, 2015; Tahangnacca et al., 2020). In this study, the Mean Year of Schooling will be used as a whole for both women and men. Higher education can increase public awareness of nutrition, sanitation, and childcare patterns, thus potentially reducing the risk of stunting. One method that can be used is regression analysis. This analysis is used to determine whether a variable affects another variable. The regression approaches are parametric, nonparametric, and semi-parametric. This study will use parametric and non-parametric approaches. In the parametric regression method, linear regression is used, while the nonparametric method is polynomial local regression.

The Linear Regression approach relies on a predefined linear relationship between independent and dependent variables, which may not adequately represent complex data structures. As a parametric regression approach, it is also strict with the assumptions (Suparti et al., 2019). Its reliance on the assumption of linearity can lead to biased results if the true relationship is nonlinear. Furthermore, Linear Regression can also be sensitive to outliers and may not perform well when the data does not meet the assumptions of normality and homoscedasticity (Schmidt & Finan, 2018; Yang et al., 2019). However, Linear Regression is often simpler to implement and interpret, making it a popular choice for many applications.

On the other hand, polynomial local regression can capture non-linear patterns that may occur. Polynomial local regression is a non-parametric method that estimates the relationship between dependent and independent variables by performing local regression around each data point. This method is adaptive to data fluctuations to capture different relationships across various independent variable values. This method also does not rely on the assumption of the relationship between the dependent and independent variables, so it is more flexible in detecting non-linear patterns (Fan & Gijbels, 1996; Miller & Hall, 2010). However, like other non-parametric regression methods, the resulting model is difficult to interpret, so it is more often displayed in a scatterplot. Polynomial regression produces equations that describe curves through linear slopes, making them difficult to interpret directly (Stimson et al., 1978).

This study aims to compare the performance of classical linear regression models and polynomial local regression in modeling the relationship between mean years of schooling and prevalence of stunting. This approach is expected to identify whether the relationship between the two variables is linear or has a more complex pattern, so that the results can provide a better understanding of this relationship.

2. Methods

The methods used in this study are parametric linear regression and nonparametric polynomial local regression.

(a) Data and Variables

The data used in this study are data from 38 provinces in Indonesia in 2023. Data are retrieved from BPS-Statistics (Central Bureau of Statistics) Indonesia for Mean Years of Schooling. Meanwhile, stunting prevalence data results from the Indonesian Health Survey (Badan Kebijakan Pembangunan Kesehatan, Kementerian Kesehatan, 2023). The stunting data used is data on the prevalence of stunting in children aged 0-23 months. The variables used can be seen in Table 1:

Table 1. Variables and Definitions

Variable	Definition	Unit

Y	Prevalence of Stunting in Children Aged 0-23 Months	Year
X	Mean Years of Schooling	Percentage

(b) The Steps of Analysis

Data analysis was carried out with the help of R software and using the KernSmooth package. The steps taken in this study are:

1. Data description
   At this stage, the condition of the data will be examined through descriptive analysis.
2. Model prespesification
   Model prespecification aims to see the relationship between dependent and independent variables. One way to do this is by creating a scatterplot.
3. Modeling using Linear Regression
   The regression curve describes the relationship between dependent and independent variables. Regression analysis can be divided into 3 based on the regression function: parametric, nonparametric, and semiparametric. The regression relationship can be modeled. The regression relationship can be modeled as:

$$Y_i = m(X_i) + \varepsilon_i \quad (1)$$

   where m It is a regression function.
   Parametric Regression has a known function and has assumptions to be met. In Linear Regression, the form of $m(.)$:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (2)$$

   Linear regression relies on several assumptions, including linearity, homoscedasticity, and absence of multicollinearity and autocorrelation (Gujarati & Porter, 2009). Violations of these assumptions can significantly impact model performance and interpretation. Meanwhile, Nonparametric Regression merely assumes that $m(.)$ It is a smooth function. The Nonparametric Regression function is unknown; hence, we cannot perform pre-specification to determine the form of the function (Härdle et al., 2004)

4. Modeling the data using Polynomial Local Regression
   Approximation for the regression function $m(x)$ locally by a polynomial of order p using Taylor expansion (Fan & Gijbels, 1996):

$$m(x) \approx m(x_0) + m'(x_0)(x - x_0) + \frac{m''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p$$

   Local polynomial regression works by performing a polynomial approximation around each point x_0 Using weighted least squares, minimize:

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \right\}^2 K_h(X_i - x_0)$$

   The weights are assigned using a kernel function (K) such as Gaussian, so the bandwidth (h) must be optimized. One method to obtain the optimal h is to use the Direct Plug-in (DPI), which is defined as follows:

$$\hat{h}_{DPI} = C_1(K) \left[\frac{\hat{\sigma}_1{}^2(\hat{\lambda}_{AMSE})(b-a)}{\hat{\theta}_{22}{}^{0.05}(\hat{g}_{AMSE})n} \right]^{\frac{1}{5}}$$

$$\hat{\lambda}_{AMSE} = C_3(K) \left[\frac{\hat{\sigma}_Q{}^4(\widehat{N})(b-a)}{\hat{\theta}_{22}{}^{0.05}(\hat{g}_{AMSE})^2 n^2} \right]^{\frac{1}{9}}$$

$$\hat{g}_{AMSE} = C_2(K) \left[\frac{\hat{\sigma}_Q{}^2(\widehat{N})(b-a)}{\left|\hat{\theta}_{24}{}^{Q}(\widehat{N})\right| n} \right]^{\frac{1}{7}}$$

   Further information is available in the following article (Ruppert et al., 1995). It is called local because each weight is assigned to each point x. The local linear estimator is a special case of the local polynomial estimator for $p = 1$. Besides bandwidth, the order polynomial will also be optimized.
5. Comparing the Linear Regression model with the Polynomial Local Regression model
   Comparisons are made based on the coefficient of determination and RMSE values. Comparisons are

math.ppj.unp.ac.id Rangkiang Mathematics Journal Vol 4, No. 1, 2025

also made in scatterplots that display the data and the models formed.

6. Conclusion
   The model that best explains the relationship between these two variables will be concluded. The conclusion is drawn based on the value of the coefficient of determination, RMSE, and the fitting curve.

3. Results and Discussion

(a) Descriptive Statistics

Descriptive statistics of this data are presented in Table 2:

Table 2. Descriptive Statistics of Data

Variable	N	Minimum	Quartile 1	Mean	Quartile 3	Maximum
Y	38	7.00	17.48	20.80	23.32	36.60
X	38	3.720	8.115	8.706	9.422	11.450

Based on the statistical information displayed in Table 2, the data tends to be homogeneous because the distribution is not too far apart.

(b) Model Prespesification

The respecification can be done by plotting the independent and dependent variables in a scatter plot. The result is in Figure 1.

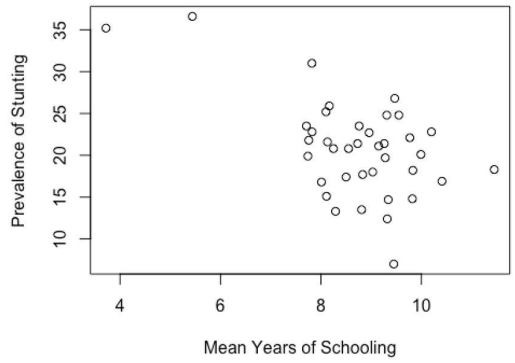

Figure 1. Scatter plot between Prevalence of Stunting and Mean Years of Schooling

At the model specification stage, Figure 1 shows whether the relationship pattern between the dependent and independent variables follows a certain distribution pattern. The data does not follow a certain distribution.

(c) Equation Linear Regression Model

The estimated model of linear regression is as follows:

$$Y = 42.913 - 2.54X$$

With R^2 32.94% and RMSE 4.84. This means that 32.94% of the variation in the prevalence of stunting can be explained by the mean years of schooling in the model. Other variables outside the model or random error explain the rest. A classical assumption test was carried out on the model, and the results showed that the model met the assumptions of normality, heteroscedasticity, and autocorrelation. The multicollinearity test is not needed because there is only one independent variable. Figure 2 shows a regression line plot that illustrates the relationship between the prevalence of stunting and mean years of schooling.

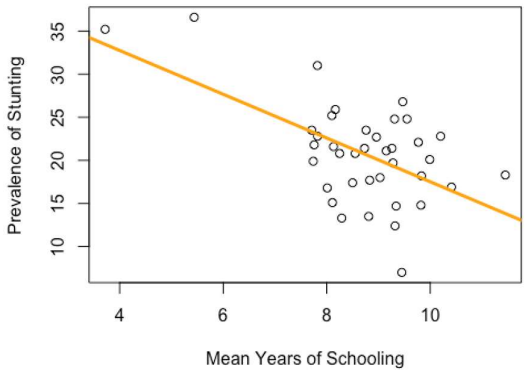

Figure 2. Scatter plot of Linear Regression Model

Figure 2 shows that the regression line does not follow the data pattern but is at the average.

(d) Polynomial Local Regression Model

Optimum bandwidth obtained using the dpill function in the KernSmooth package. This function uses direct plug-in methodology. The obtained result is 0.6861906. The degree of the polynomial used is selected based on the following plot in Figure 3:

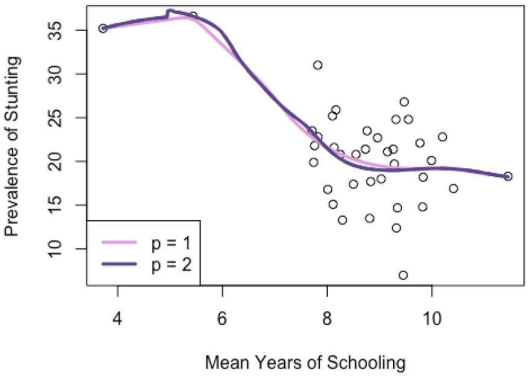

Figure 3. Fitted Curve of Polynomial Local Regression with Different Order of p

Based on Figure 2, the degree of the polynomial chosen is 1 because it looks smoother. Therefore, the shape of the fitted curve for local polynomial regression is in Figure 4:

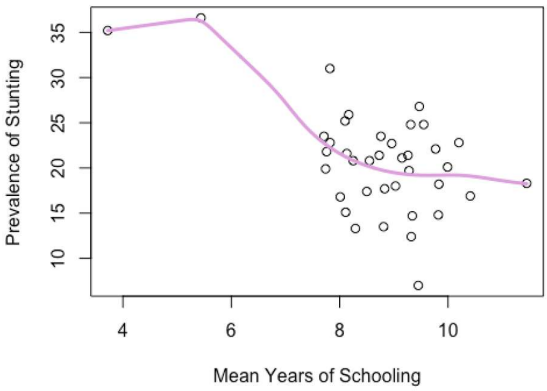

Figure 4. Fitted Curve of Polynomial Local Regression

The model has R^2 43.44% and RMSE 4.32. The Coefficient of Determination value of 43.44% means that 43.44% of the variation in the dependent variable (Y) can be explained by the model's independent variable (X). Other variables outside the model or random error explain the rest.

(e) Comparison

The linear regression model and local polynomial regression model produced from data processing will be compared based on the values of the determination coefficient, RMSE, and the fitted curve obtained. This comparison is presented in Table 3:

Table 3. Comparison Between the Two Models

Method	R^2	RMSE
Linear Regression	32.94%	4.84
Plynomial Local Regression	43.44%	4.32

Based on the Coefficient of Determination, it can be seen that the local polynomial model has a higher value. It means that the local polynomial model is better at explaining the effect of mean years of schooling on the prevalence of stunting. This is in line with the RMSE value, where the local polynomial model has a smaller RMSE value, which means this model is more appropriate. Furthermore, we compare the curves formed from each model in a plot in Figure 5:

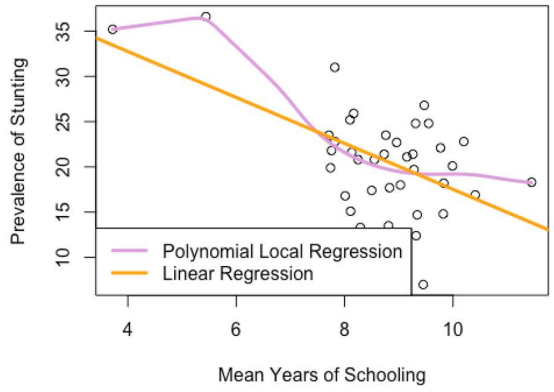

Figure 5. Fitting Curve of Polynomial Local Regression vs Linear Regression

In Figure 5, the curve for the local polynomial regression model better fits the existing data than the curve for the linear regression model.

(f) Discussion

This comparison has also been done in previous studies. In a study by Ajona et al. (2022), they use multiple linear regression (MLR) and polynomial local regression (LPR). The conclusion: 1) the LPR model had a high R^2 of 0.863, whereas the MLR model had a low value of 0.495; 2) The LPR model provided the best match, and it also explains the impact of conditioning factors on the biodegradation rate. The next study shows that LPR outperformed linear regression in predicting stock prices, with LPR achieving a slightly lower MAPE of 6.54% compared to 6.55% for linear regression (Satriyo et al., 2023). The next paper compares LPR and linear regression, finding that LPR with a Gaussian kernel is the best model for predicting Indonesia's non-oil and gas export values (Fauzi & Sofia Yanti, 2023).. Then, both models will be compared for sales forecasting in a snack food company. The result is that LPR has the lowest MAPE (Heni, Roberta et al., 2023). LPR captures nonlinearities and provides a flexible, data-driven approach, unlike linear regression, which assumes a constant relationship across the dataset. The local method also minimizes the influence of outliers, enhancing forecasting accuracy for rainfall predictions (George et al., 2016). Based on these studies, we can conclude that LPR outperforms linear regression in various fields.

In addition to comparing methods, many developments have been made to the LPR method, including combining it with a model to improve output gap estimates and forecasts (Fritz, 2021). LPR is sensitive to outliers, which can skew results. A novel framework introduces similarity kernels incorporating independent and dependent variables, enhancing robustness and accuracy in noisy environments (Shulman, 2025). The next paper introduces a novel local bandwidth estimation procedure for LPR. The method enhances accuracy and computational speed for large datasets (Samarov, 2015). LPR was also

developed for spatial data use (Kurisu & Matsuda, 2022). Besides the paper discussed here, many developments have been carried out to improve and enhance the performance of LPR. Therefore, it can be seen that this method still has much room to develop and become more powerful. In addition, this method can also be used in various fields.

4. Conclusion

It can be concluded that Polynomial Local Regression is better at modeling the Prevalence of Stunting and Mean Years of Schooling. The Polynomial Local Regression model can explain 43.44% of the variation in the Prevalence of Stunting, which is accounted for by Mean Years of Schooling. The fitting curve of the model is in Figure 6:

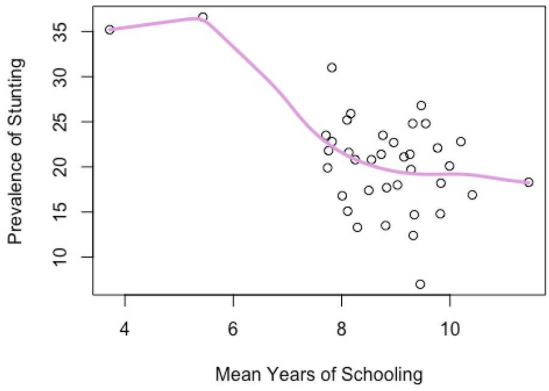

Figure 6. Fitting Curve of Polynomial Local Regression

For datasets characterized by non-linear relationships, Polynomial Local Regression will likely yield more accurate insights. This study's limitation is that it only uses one independent variable. For further research, it is recommended that other dependent variables be added so that the stunting phenomenon in Indonesia can be better explained.

References

1. Ajona, M., Vasanthi, P., & Vijayan, D. S. (2022). Application of multiple linear and polynomial regression in the sustainable biodegradation process of crude oil. Sustainable Energy Technologies and Assessments, 54, 102797. https://doi.org/10.1016/j.seta.2022.102797
2. Aryastami, N. K. (2017). Kajian Kebijakan dan Penanggulangan Masalah Gizi Stunting di Indonesia. Buletin Penelitian Kesehatan, 45(4), 233–240. https://doi.org/10.22435/bpk.v45i4.7465.233-240
3. Badan Kebijakan Pembangunan Kesehatan, Kemeterian Kesehatan. (2023). Survei Kesehatan Indonesia (SKI) 2023 Dalam Angka.
4. Casale, D., Espi, G., & Norris, S. A. (2018). Estimating the pathways through which maternal education affects stunting: Evidence from an urban cohort in South Africa. Public Health Nutrition, 21(10), 1810–1818. https://doi.org/10.1017/S1368980018000125
5. De Onis, M., Borghi, E., Arimond, M., Webb, P., Croft, T., Saha, K., De-Regil, L. M., Thuita, F., Heidkamp, R., Krasevec, J., Hayashi, C., & Flores-Ayala, R. (2019). Prevalence thresholds for wasting, overweight, and stunting in children under 5 years. Public Health Nutrition, 22(1), 175–179. https://doi.org/10.1017/S1368980018002434
6. Fan, J., & Gijbels, I. (1996). Local polynomial modelling and its applications (First CRC Press reprint 2003). CRC Press.
7. Fauzi, F., & Sofia Yanti, T. (2023). Pemodelan Regresi Polinomial Lokal pada Nilai Ekspor Non-Migas di Indonesia. Bandung Conference Series: Statistics, 3(2), 531–537. https://doi.org/10.29313/bcss.v3i2.8512
8. Fritz, M. (2021). Improved Output Gap Estimates and Forecasts Using a Local Linear Regression. The 7th International Conference on Time Series and Forecasting, 32. https://doi.org/10.3390/engproc2021005032
9. George, J., Janaki, L., & Parameswaran Gomathy, J. (2016). Statistical Downscaling Using Local

Polynomial Regression for Rainfall Predictions – A Case Study. Water Resources Management, 30(1), 183–193. https://doi.org/10.1007/s11269-015-1154-0
10. Gujarati, D. N., & Porter, D. C. (2009). Basic econometrics (5th ed). McGraw-Hill Irwin.
11. Härdle, W., Werwatz, A., Müller, M., & Sperlich, S. (2004). Nonparametric and Semiparametric Models. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-17146-8
12. Heni, Roberta, Solihin, Jasan Supratman, & Muhendra, R. (2023). Pengembangan model peramalan penjualan menggunakan metode regresi linier dan polinomial pada industri makanan ringan (Studi Kasus: CV. Stanley Mandiri Snack). TEKNOSAINS : Jurnal Sains, Teknologi Dan Informatika, 10(2), 185–192. https://doi.org/10.37373/tekno.v10i2.456
13. Huey, S. L., & Mehta, S. (2016). Stunting: The Need for Application of Advances in Technology to Understand a Complex Health Problem. EBioMedicine, 6, 26–27. https://doi.org/10.1016/j.ebiom.2016.03.013
14. Kurisu, D., & Matsuda, Y. (2022). Local polynomial trend regression for spatial data on $\mathbb{R}^d$ (Version 7). arXiv. https://doi.org/10.48550/ARXIV.2211.13467
15. Laksono, A. D., Wulandari, R. D., Amaliah, N., & Wisnuwardani, R. W. (2022). Stunting among children under two years in Indonesia: Does maternal education matter? PLOS ONE, 17(7), e0271509. https://doi.org/10.1371/journal.pone.0271509
16. Makoka, D., & Masibo, P. K. (2015). Is there a threshold level of maternal education sufficient to reduce child undernutrition? Evidence from Malawi, Tanzania, and Zimbabwe. BMC Pediatrics, 15(1), 96. https://doi.org/10.1186/s12887-015-0406-8
17. Miller, H., & Hall, P. (2010). Local polynomial regression and variable selection. In Institute of Mathematical Statistics Collections (pp. 216–233). Institute of Mathematical Statistics. https://doi.org/10.1214/10-IMSCOLL615
18. Montenegro, C. R., Gomez, G., Hincapie, O., Dvoretskiy, S., DeWitt, T., Gracia, D., & Misas, J. D. (2022). The pediatric global burden of stunting: Focus on Latin America. Lifestyle Medicine, 3(3), e67. https://doi.org/10.1002/lim2.67
19. Raiten, D. J., & Bremer, A. A. (2020). Exploring the Nutritional Ecology of Stunting: New Approaches to an Old Problem. Nutrients, 12(2), 371. https://doi.org/10.3390/nu12020371
20. Ruppert, D., Sheather, S. J., & Wand, M. P. (1995). An Effective Bandwidth Selector for Local Least Squares Regression. Journal of the American Statistical Association, 90(432), 1257–1270. https://doi.org/10.1080/01621459.1995.10476630
21. Samarov, D. V. (2015). The Fast RODEO for Local Polynomial Regression. Journal of Computational and Graphical Statistics, 24(4), 1034–1052. https://doi.org/10.1080/10618600.2014.949724
22. Satriyo, S. A. L., Adi Rizky Pratama, & Rahmat. (2023). Perbandingan metode linear regresi dan polynomial regresi untuk memprediksi harga saham studi kasus Bank BCA. INFOTECH : Jurnal Informatika & Teknologi, 4(1), 59–70. https://doi.org/10.37373/infotech.v4i1.602
23. Schmidt, A. F., & Finan, C. (2018). Linear regression and the normality assumption. Journal of Clinical Epidemiology, 98, 146–151. https://doi.org/10.1016/j.jclinepi.2017.12.006
24. Shulman, Y. (2025). Robust Local Polynomial Regression with Similarity Kernels (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2501.10729
25. Stimson, J. A., Carmines, E. G., & Zeller, R. A. (1978). Interpreting Polynomial Regression. Sociological Methods & Research, 6(4), 515–524. https://doi.org/10.1177/004912417800600405
26. Suparti, Santoso, R., Prahutama, A., Devi, A. R., & Sudargo. (2019). Modeling longitudinal data based on Fourier regression. Journal of Physics: Conference Series, 1217(1), 012105. https://doi.org/10.1088/1742-6596/1217/1/012105
27. Tahangnacca, M., Amiruddin, R., Ansariadi, & Syam, A. (2020). Model of stunting determinants: A systematic review. Enfermería Clínica, 30, 241–245. https://doi.org/10.1016/j.enfcli.2019.10.076
28. Yang, K., Tu, J., & Chen, T. (2019). Homoscedasticity: An overlooked critical assumption for linear regression. General Psychiatry, 32(5), e100148. https://doi.org/10.1136/gpsych-2019-100148