JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 9, No. 3, July 2025, pp. 946-960

Multilevel Semiparametric Modeling with Overdispersion and Excess Zeros on School Dropout Rates in Indonesia

Arna Ristiyanti Tarida[1*], Anik Djuraidah[1], Agus Mohamad Soleh[1]
[1]School of Data Science, Mathematics and Informatics, IPB University, Indonesia
arnaristiyanti@apps.ipb.ac.id

Article History:
Received : 11-03-2025
Revised : 05-05-2025
Accepted : 07-05-2025
Online : 03-07-2025

Keywords:
Semiparametric multilevel;
Overdispersion;
Excess zeros;
B-Spline;
School dropout rates.

ABSTRACT

This study aims to identify key factors influencing high school dropout rates in Indonesia by applying advanced statistical modeling that accounts for complex data characteristics. Dropout data often display overdispersion (variability greater than expected) and excess zeros (many students not dropping out), which, if ignored, can bias conclusions. To address this, we compare parametric models, Zero-Inflated Poisson Mixed Model (ZIPMM), Zero-Inflated Generalized Poisson Mixed Model (ZIGPMM), and Zero-Inflated Negative Binomial Mixed Model (ZINBMM), with their semiparametric counterparts (SZIPMM, SZIGPMM, SZINBMM). The semiparametric models use B-spline functions to capture nonlinear relationships between predictors and dropout rates, with flexibility. Model performance was evaluated using Akaike Information Criterion (AIC) and Root Mean Square Error (RMSE) across 100 simulation repetitions to ensure robustness. Results show that the semiparametric ZIGPMM (SZIGPMM) outperformed other models, achieving the lowest average AIC (18969.62), suggesting the best trade-off between model fit and complexity. The optimal spline configuration used knot point 2 and order 3, with a Generalized Cross-Validation (GCV) score of 9.4107. Key predictors of dropout include school status (public or private), student-teacher ratio, distance from home to school, parental education level, parental employment status, and number of siblings. These findings provide actionable insights for education policymakers, emphasizing the need to address structural and socioeconomic barriers to reduce dropout rates effectively.

https://doi.org/10.31764/jtam.v9i3.30102

A. INTRODUCTION

Regression analysis is a fundamental statistical method for modelling the relationship a response variable and one or more explanatory variables. It can be categorized into three main approaches: parametric, nonparametric, and semiparametric regression. Parametric regression specifies a predefined functional form, whereas nonparametric regression makes minimal assumptions and captures patterns directly from the data (Mahmoud, 2021). Semiparametric regression combines the strengths of both by incorporating flexible smoothing techniques such as splines along with fixed parametric components. Spline regression, particularly basis splines (B-splines) and penalized splines, is widely used in semiparametric modelling due to its flexibility and numerical stability. B-spline functions are advantageous for managing high spline orders and dense knot placements, which often cause numerical instability (Beccari & Casciola, 2021; Chudy & Woźny, 2022).

Semiparametric methods, such as kernel, spline regression, local polynomials, have been applied in various studies. For example, local polynomial semiparametric for longitudinal data

(Utami et al., 2024) and truncated spline model to analyze factors influencing the growth of a cashless society (Pramaningrum et al., 2024). Furthermore, mixed-effects models can be integrated into semiparametric frameworks to account for hierarchical or multilevel structures within data, such as regional or institutional clusters. Belloc et al. (2011) employed a semiparametric mixed-effects model to assess individual-level dropout determinants in Italian universities. Masci et al. (2022) proposed a semiparametric multinomial mixed-effects model, allowing for flexible modeling of hierarchical educational data with unobserved heterogeneity across academic programs.

Count data, such as school dropouts often exhibit overdispersion and excess zeros. Two characteristics that violating Poisson model assumpsions of equidispersion (the variance equals the mean) (Agresti, 2015). However, this assumption often does not hold in real-world settings. Overdispersion occurs when the empirical variance in the data is greater than that predicted by the model (Dean & Lundy, 2016). It may result from unobserved heterogeneity, outliers, data clustering, or an excessive number of zero counts (Hardin & Hilbe, 2018). While Negative Binomial (NB) models address overdispersion, and Zero-Inflated models like Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) handle excess zeros. However, ZINB models are known to suffer from convergence issues when misapplied to equidispersed data, leading to biased regression coefficient estimates and overestimation of predictor significance (Fernandez & Vatcheva, 2022). The Zero-Inflated Generalized Poisson (ZIGP) model offers greater flexibility in handling both issues. Recent studies highlight that semiparametric extensions of zero-inflated models further improve performance, especially with hierarchical and nonlinear data structures (Almasi et al., 2016; Aráujo et al., 2023; Mahmoodi et al., 2016). Despite their relevance, few studies have applied these advanced modeling frameworks to social indicators such as education.

Dropout rates are critical indicators of educational access and quality, influenced by socio-economic, regional, and policy-related factors. According to Sustainable Development Goal Target 4.1, every child should have access to free, equitable, and quality education (UNESCO, 2025). However, dropout rates remain a challenge, especially in developing countries. Studies have shown that despite the implementation of tuition-free education policies, dropout remains prevalent due to factors such as absenteeism, household economic hardship, and policy implementation gaps (Gbaguidi & Adetou, 2024; Ole Kinisa, 2019). In Indonesia, despite the implementation of 12-year compulsory education, dropout rates persist at concerning levels, particularly at the senior secondary level. In 2022, dropout rates in provinces like West Nusa Tenggara (0.88%), Papua (0.59%), and Gorontalo (0.54%) were notably higher than the national average (MoECRT, 2022).

Dropout data are typically non-negative integers, right-skewed, zero-inflated, overdispersed, and hierarchical structured by region and school level. Additionally, the relationship between predictors such as economic status or geographical location and dropout likelihood may be nonlinear. Therefore, this study proposes the use of parametric and semiparametric zero-inflated mixed models, particularly the ZIP, ZIGP, and ZINB frameworks, to model dropout rates in Indonesian high schools. This approach is expected to provide more accurate estimates and policy-relevant insights into the determinants of school dropouts.

B. METHODS

1. Overdispersion

Overdispersion is a condition in which the variance of the response variable exceeds its mean (Agresti, 2013). This violates the assumption of equidispersion in Poisson regression. According to Hilbe (2011), applying Poisson regression in the presence of overdispersion may lead to underestimated standard errors. Consequently, this increases the risk of incorrectly rejecting the null hypothesis. Overdispersion can be detected by calculating the deviance divided by the model's degrees of freedom (Myers et al., 2010). A ratio greater than one indicates the presence of overdispersion.

2. ZIP, ZIGP, and ZINB Distribution

This study focuses on three count data distributions: ZIP, ZIGP, and ZINB. The probability mass functions for each distribution are presented below within the context of a two-level hierarchical framework.

$$\boldsymbol{P}(\boldsymbol{y_{ij}}|\boldsymbol{\mu_{ij}},\boldsymbol{p_{ij}}) = \begin{cases} \boldsymbol{p_{ij}} + (\boldsymbol{1}-\boldsymbol{p_{ij}})\boldsymbol{e}^{-\boldsymbol{\mu_{ij}}} & , \boldsymbol{y_{ij}} = \boldsymbol{0} \\ (\boldsymbol{1}-\boldsymbol{p_{ij}})\dfrac{\boldsymbol{e}^{-\boldsymbol{\mu_{ij}}}\boldsymbol{\mu_{ij}}^{\boldsymbol{y_{ij}}}}{\boldsymbol{y_{ij}}!} & , \boldsymbol{y_{ij}} > \boldsymbol{0} \end{cases} \tag{1}$$

$$P(y_{ij}|\mu_{ij},\alpha,p_{ij}) = \begin{cases} p_{ij} + (1-p_{ij})\exp\left(-\dfrac{\mu_{ij}}{1+\alpha\mu_{ij}}\right) & , y_{ij} = 0 \\ (1-p_{ij})\left(\dfrac{\mu_{ij}}{1+\alpha\mu_{ij}}\right)^{y_{ij}}\dfrac{(1+\alpha y_{ij})^{y_{ij}-1}}{y_{ij}!}\exp\left[\dfrac{-\mu_{ij}(1+\alpha y_{ij})}{1+\alpha\mu_{ij}}\right] & , y_{ij} > 0 \end{cases} \tag{2}$$

$$P(y_{ij}|\mu_{ij},\alpha,p_{ij}) = \begin{cases} p_{ij} + (1-p_{ij})\left(\dfrac{1}{1+\alpha\mu_{ij}}\right)^{\frac{1}{\alpha}} & , y_{ij} = 0 \\ (1-p_{ij})\dfrac{\Gamma\left(y_{ij}+\frac{1}{\alpha}\right)}{\Gamma(y_{ij}+1)\Gamma\left(\frac{1}{\alpha}\right)}\left(\dfrac{1}{1+\alpha\mu_{ij}}\right)^{\frac{1}{\alpha}}\left(1-\dfrac{1}{1+\alpha\mu_{ij}}\right)^{y_{ij}} & , y_{ij} > 0 \end{cases} \tag{3}$$

Here, p_{ij} denotes the probability of an excess zero, where $i = 1, \dots, n_j$ and $j = 1, \dots, m$. The parameter α represents the overdispersion.

3. Linear Mixed Model (LMM)

LMM extends the linear model by incorporating random effects to account for variability across groups. The general form of the mixed model can be expressed as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon} \tag{4}$$

Let $\boldsymbol{y}$ be an $n \times 1$ vector of observed responses. The matrix $\boldsymbol{X}$ is the $n \times p$ design matrix for fixed effects, where p is the number of fixed predictors, and $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed effect parameters. The matrix $\boldsymbol{Z}$ is the $n \times k$ design matrix of random effects, where k is the number of random effects, $\boldsymbol{u}$ is the corresponding $n \times k$ vector of random effect parameter. The residual error is denoted by ε, an $n \times 1$ vector. It is assumed that $\boldsymbol{u} \sim N(0, \boldsymbol{G})$ and $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{R})$, with $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ being independent. Here, $\boldsymbol{G} = Var(\boldsymbol{u})$ and $\boldsymbol{R} = Var(\boldsymbol{\varepsilon})$ are variance-covariance matrices that

involve unknown dispersion or variance component (σ^2). Under this model, the expected value and variance-covariance structure of $\boldsymbol{y}$ are given by:

$$E(\boldsymbol{y}) = \boldsymbol{X\beta}, \quad Var(\boldsymbol{y}) = \boldsymbol{R} + \boldsymbol{ZGZ'}$$

Parameter estimation in the Linear Mixed Model is typically carried out using the Maximum Likelihood Estimation (MLE) method. The log-likelihood function for the model is given by:

$$\log(L) = (\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})'\boldsymbol{R^{-1}}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu}) + \boldsymbol{u'G} \tag{5}$$

The estimators for the fixed and random effects are obtained as follows:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\boldsymbol{X'V^{-1}X})^{-1}\boldsymbol{X'V^{-1}y} \\ \widehat{\boldsymbol{u}} &= \boldsymbol{GZ'V^{-1}}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \end{aligned} \tag{6}$$

The variance components are estimated using the maximum likelihood approach with the following log-likelihood function:

$$\log(L)_{ML} = -\tfrac{1}{2}[\log|\boldsymbol{V}| + \boldsymbol{y'V^{-1}}(\boldsymbol{I} - \boldsymbol{X}[\boldsymbol{X'V^{-1}X}]^{-1}\boldsymbol{X'V^{-1}})\boldsymbol{y}] - \tfrac{n}{2}\log(2\pi) \tag{7}$$

However, estimation of variance components using MLE tends to be biased. Therefore, the Restricted Maximum Likelihood (REML) method is preferred. The REML log-likelihood function is given by:

$$\boldsymbol{\log(L)_{REML}} = \boldsymbol{\log(L)_{ML}} - \tfrac{1}{2}\boldsymbol{\log}\left|\boldsymbol{X'V^{-1}X}\right| + \tfrac{\boldsymbol{p}}{\boldsymbol{2}}\boldsymbol{\log}\,(\boldsymbol{2\pi}) \tag{8}$$

(Jiang & Nguyen, 2021; Ruíz et al., 2023).

4. B-Spline

B-Spline functions are constructed from a set of smooth and flexible basis functions, making them suitable for modeling nonlinear relationships. A regression model incorporating a B-Spline of order m with k knots can be expressed as follows:

$$y_i = \textstyle\sum_{j=1}^{m+k} \beta_j\, B_{(j-m),m}(x_i) + \varepsilon_i\,, \quad i = 1,2,\dots,n \tag{9}$$

where y_i is the response variable; β_j denotes the regression coefficient associated with the B-Spline basis; $B_{(j-m),m}(x_i)$ is the B-Spline basis function of order m; $t_1, \dots, t_k$the knot points; and ε is the random error term. To construct a B-Spline function of order m with k internal knot points $t_1, \dots, t_k$, where $a < t_1 < \cdots < t_k < b$, an additional $2m$ boundary knots are introduced. These are defined by repeating the boundary values a and b, each m times, to ensure that the basis functions are well-defined over the entire domain.

$$t_{-(m-1)}, \dots, t_{-1}, t_0, t_{k+1}, \dots, t_{k+m}$$

Specifically, the additional boundary knots are defined as $t_{-(m-1)} = \cdots = t_0 = a$ and $t_{k+1}, \dots, t_{k+m} = b$, where a and b represent the lower and upper bounds of the domain, respectively. This extension ensures that the B-spline basis functions are properly defined and continuous across the entire interval $[a, b]$. According to Perperoglou et al. (2019), the basis of a B-Spline function of order m with k knot points $t_1, \dots, t_k$ can be defined recursively as follows:

$$B_{j,m}(x) = \frac{x - t_1}{t_{j+m-1} - t_j} B_{j,m-1}(x) + \frac{t_{j+m} - x}{t_{j+m} - t_{j+1}} B_{j+1,m-1}(x) \tag{10}$$

for $j = -(m-1), \dots, k$, the zeroth-order (piecewise constant) B-spline basis functions are defined as:

$$B_{j,1}(x) = \begin{cases} 1, & \text{if } t_j \le x < t_{j+1} \\ 0, & \text{other} \end{cases}$$

The B-Spline function of order m with k knot points, where $\lambda = \{t_1, t_2, \dots, t_k\}$ then be expressed as:

$$f_\lambda = \sum_{j=1}^{m+k} \beta_{\lambda j} B_{\lambda j - m, m}(x_i)$$

The B-Spline regression model can be expressed as follows:

$$y_i = \sum_{j=1}^{m+k} \beta_{\lambda j} B_{\lambda j - m, m}(x_i) + \varepsilon_i \tag{11}$$

Equation (11), when expressed in matrix form, becomes:

$$\boldsymbol{y} = \boldsymbol{B_\lambda \beta_\lambda} + \boldsymbol{\varepsilon}$$

The parameter $\boldsymbol{\beta_\lambda}$ is estimated using the least squares method. The estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ is obtained by minimizing the Residual Sum of Squares (RSS), defined as:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} = \left(\boldsymbol{B_\lambda^T B_\lambda}\right)^{-1} \boldsymbol{B_\lambda^T y}$$

In nonparametric regression, the model estimation using B-Spline basis functions can be written as:

$$\hat{y} = \boldsymbol{\beta_\lambda} \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} = \boldsymbol{\beta_\lambda} \left(\boldsymbol{B_\lambda^T B_\lambda}\right)^{-1} \boldsymbol{B_\lambda^T} = \boldsymbol{S_\lambda y} \tag{12}$$

with $\boldsymbol{S_\lambda} = \boldsymbol{\beta_\lambda} \left(\boldsymbol{B_\lambda^T B_\lambda}\right)^{-1} \boldsymbol{B_\lambda^T}$ is a symmetric and positive definite matrix.

5. SZIPMM, SZIGPMM, SZINBMM

The SZIPMM, SZIGPMM, and SZINBMM are semiparametric multilevel models designed for analizing count data, particularly in the presence of zero-inflation. The log function for the mean of these three models can be formulated as follows:

$$\log(\boldsymbol{\mu_{ij}}) = \boldsymbol{\eta_{ij}} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_{ij1}} + \cdots + \boldsymbol{\beta_k x_{ijk}} + \boldsymbol{f(x_{ij})} + \boldsymbol{u_j} \tag{13}$$

where $\boldsymbol{\mu_{ij}}$ is the mean response, $\boldsymbol{x_{ij}}$ is the explanatory variable for the $\boldsymbol{i}$-th observation within the $\boldsymbol{j}$-th group, with corresponding parameter $\boldsymbol{\beta}$, $\boldsymbol{f}(\boldsymbol{x_{ij}})$ is the nonparametric function (spline) that models the nonlinear relationship between $\boldsymbol{x_{ij}}$ and $\boldsymbol{\mu_{ij}}$, and $\boldsymbol{u_j}$ is the random effect for group $\boldsymbol{j}$, assumed to follow anormal distribution: $\boldsymbol{u_j} \sim \boldsymbol{N(0, \sigma_u^2)}$. The model for zero-inflated probabilities is:

$$\boldsymbol{logit(p_{ij})} = \log\left(\frac{\boldsymbol{p_{ij}}}{\boldsymbol{1-p_{ij}}}\right) = \boldsymbol{\gamma_0} + \boldsymbol{\gamma_1 z_{ij1}} + \cdots + \boldsymbol{\gamma_k z_{ijk}} + \boldsymbol{g(z_{ij})} + \boldsymbol{v_j} \tag{14}$$

The log-likelihood function for a semiparametric multilevel model for zero-inflated data, which accounts for both fixed and random effects as well as nonparametric (spline) components, can be formulated as follows:

$$\log L(\beta, \gamma, u, v) = \sum_{j=1}^{m}\sum_{i=1}^{n_j} \log P\left(Y_{ij} = y_{ij} \middle| X_{ij}, Z_{ij}, u_j, v_j\right) - \frac{\lambda}{2} B^T D\beta - \frac{\lambda}{2}\gamma^T D\gamma$$
$$-\frac{1}{2}\sum_{j=1}^{m}\left(\frac{u_j^2}{\sigma_u^2} + \frac{v_j^2}{\sigma_v^2}\right) \tag{15}$$

with $P\left(Y_{ij} = y_{ij} \middle| X_{ij}, Z_{ij}, u_j, v_j\right) = \begin{cases} p_{ij} + (1-p_{ij}) f\left(0 \middle| \mu_{ij}, \alpha\right), & jika\ y_{ij} = 0 \\ (1-p_{ij}) f\left(y_{ij} \middle| \mu_{ij}, \alpha\right), & jika\ y_{ij} > 0 \end{cases}$

Equation (15) estimates the parameters using the Penalized Maximum Likelihood Estimation (PMLE) and the Expectation-Maximization (EM) algorithm.

6. Data

The data utilized in this study are secondary data obtained from the Ministry of Education, Culture, Research, and Technology (MoECRT) in 2022. The data were sourced sourced from the Educational Data System (EDS), an official platform for managing educational data and statistics in Indonesia. The hierarchical structure of the dataset comprises two levels: level 1 units represent senior high schools, which are nested within level 2 units corresponding to the 34 provinces of Indonesia. In this study, random effects are specified at the provincial level, which delegates the responsibility for managing secondary education to provincial governments rather than to district authorities (Law Number 23 of 2014 on Regional Government), as shown in Table 1.

Table 1. List of Variables Used in The Study

Variable	Description	Measurement	Data Source
Y	The number of school dropouts (per 100 students)	Count	EDS, MoECRT
X_1	School status (1: Public; 2: Private)	Categorical	EDS, MoECRT
X_2	Student-teacher ratio	Ratio	EDS, MoECRT
X_3	Percentage of students whose home-to-school distance exceeds 5 km	Percentage	EDS, MoECRT
X_4	Percentage of students receiving KIP (government financial aid)	Percentage	EDS, MoECRT
X_5	Percentage of students whose father's education level is below high school	Percentage	EDS, MoECRT
X_6	Percentage of students whose fathers are unemployed	Percentage	EDS, MoECRT
X_7	Percentage of students with more than three siblings	Percentage	EDS, MoECRT
u_j	Province (Level 2)	Categorical	EDS, MoECRT
u_{ij}	School (Level 1)	Categorical	EDS, MoECRT

7. Research Procedures

a. Data exploration and data testing.

b. Determining the best number of knots and order of B-Spline (nonparametric component). Optimization based on the minimum Generalized Cross Validation (GCV) score.

$$GCV(\lambda) = \frac{n^{-1}RSS(\lambda)}{(n^{-1}trace[\boldsymbol{I} - \boldsymbol{S}_{\boldsymbol{\lambda}}])^2}$$

c. Splitting data into training (90%) and testing (10%).

d. Modeling the data using training data with parametric models: ZIPMM, ZIGPMM, ZINBMM, and semiparametric models: SZIPMM, SZIGPMM, SZINBMM.

1) Calculating Akaike Information Criterion (AIC) for each model.

$$AIC = -2\ln L + 2k$$

with L is the likelihood of the model and k is the number of the parameters.

2) Predicting the response variable for both training and testing data.

3) Computing the Root Mean Square Error (RMSE)

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

e. Repeating steps 3-4, 100 times. Evaluate models based on average AIC and RMSE. A model with the lowest average AIC and RMSE is the best model.

f. Estimating the parameters and testing the significance of the model parameters based on the best model.

C. RESULT AND DISCUSSION

1. Data Exploration

The descriptive statistics of high school dropout data in 2022 are presented in Table 2. A total of 14,210 high schools across 34 provinces and 514 districts were analized. The table indicates that the average number of dropouts is low, close to zero, suggesting the presence of excess zeros in the data. Additional, the relatively low variance implies that most data points are concentrated around zero, with a few notable outliers. Several variables, such as X_2, X_3, X_6, and X_7, display a right-skewed distribution, as evidenced by their mean values exceeding their medians. Variables X_5, X_6, and X_7 exhibit high variance, suggesting a broad range of values or the presence of significant outliers, as shown in Table 2.

Table 2. Descriptive Statistics of All Variables

Variable	Min	Median	Mean	Max	Var
Y	0.00	0.00	0.51	100.00	9.45
X_2	0.00	14.00	15.35	324.00	120.42
X_3	0.00	1.07	1.84	46.48	7.97
X_4	0.00	0.00	0.25	100.00	2.87
X_5	0.00	52.31	54.10	100.00	824.12
X_6	0.00	13.43	16.67	100.00	210.32
X_7	0.00	2.73	9.40	100.00	219.21

Excess zeros in the response variable are assessed descriptively through the histogram shown in Figure 1. The distribution reveals that 81.69% of cases report no dropout incidents, indicating the presence of excess zeros. This suggests that excess zeros should be accounted for in the analysis.

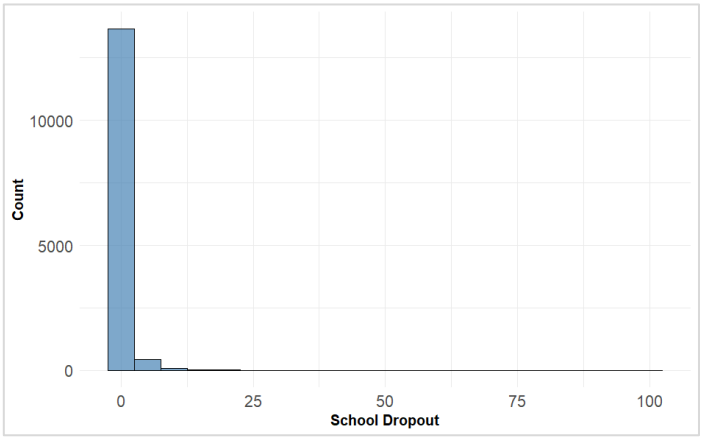

Figure 1. Histogram of the Number of High School Dropouts in Indonesia

Thematic mapping is used to visualize school dropout rates by province. Figure 2 illustrates the distribution of high school dropouts per 10,000 students across 34 provinces in Indonesia. The highest dropout rates are observed in West Nusa Tenggara (83 students), Papua (56 students), and Gorontalo (54 students). In contrast, D.I. Yogyakarta (1 student), Bali (2 students), and Central Java (4 students). This variation in dropout rates highlights regional disparities in educational outcomes, which may be influenced by differences in socio-economic conditions, access to education, infrastructure, and local government policies. Provinces with

higher dropout rates often face challenges such as limited educational resources, higher poverty rates, or geographical barriers. Understanding these spatial patterns is essential for developing targeted interventions aimed at reducing school dropout rates and promoting educational equity across regions.

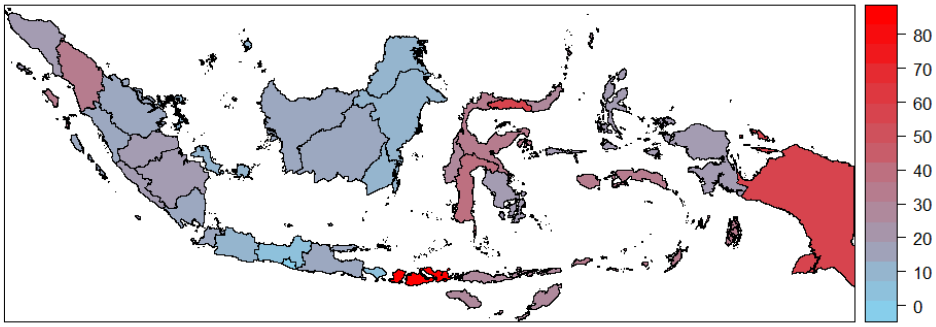

Figure 2. Thematic Map of the Number of High School Dropouts per 10,000 Students in Each Province

Figure 3 presents the scatter plot between the explanatory variables and the response variable. The variable X_1 is a categorical, indicating school status, with public and private schools comprising 49.2% and 50.8%, respectively. The variables X_2, X_3, X_4, X_5, X_6, and X_7 are numeric. The plots reveal that the relationships between these numeric variables and the response variable do not follow a clear linear trend. Due to the absence of consistent parametric patterns, a semiparametric approach is applied for further analysis.

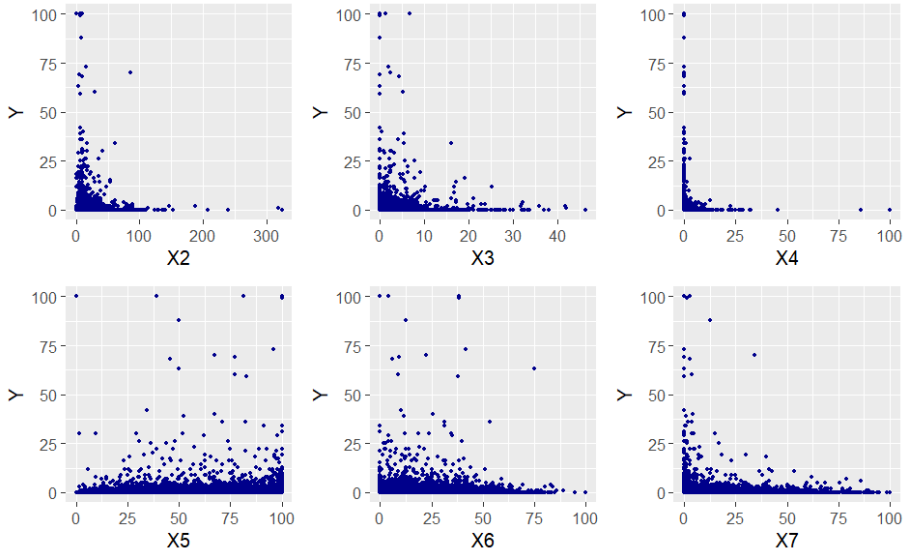

Figure 3. Scatter Plot of Response Variables and Explanatory Variables

As shown in Table 3, all variables have low Variance Inflation Factor (VIF) values (none exceeding 10), indicating the absence of multicollinearity among the explanatory variables. Therefore, the variables are suitable for further analysis.

Table 3. VIF Value of Explanatory Variables

Variable	VIF
X_1	1.075
X_2	1.005
X_3	1.011
X_4	1.005
X_5	1.283
X_6	1.219
X_7	1.142

2. Zero Inflation and Overdispersion Test

The zero inflation was assessed using a score test. As presented in Table 4, the test yielded a p-value of <2.22e-16, leading to the rejects the null hypothesis at a 5% significance level. This result provides strong evidence of excess zeros in the response variable.

Table 4. Zero Inflation Test

Chi-square	p-value
11947.999	< 2.22e-16

Furthermore, overdispersion was tested to determine whether the data exhibits variability greater than expected. As shown in Table 5, the test produced a p-value of 0, which is below the 5% significance level. This indicates the presence of overdispersion in the data.

Table 5. Overdispersion Test

Obs.Var/Theor.Var	Statistic	p-value
18.469	262429.8	0

3. Optimization of the Number of B-Spline Knots

The variables selected as the nonparametric component is X_5. The optimal number and order of knots were determined based on the minimum Generalized Cross-Validation (GCV) score. Knot combinations were evaluated from one to ten knots and orders ranging from one to three. As shown in Table 6, the optimal configuration is two knots with order three, yielding the lowest GCV score of 9.4107. The optimal knot locations are at 15.92 and 95.24, corresponding to the 0.1 and 0.9 quantiles.

Table 6. Knot, Order, and GCV Optimal

Knot	Order	GCV	Knot Position	Knot Quantiles
2	2	9.4126	15.92, 95.24	0.1, 0.9
2	3	9.4107	15.92, 95.24	0.1, 0.9
3	3	9.4116	15.92, 52.31, 95.24	0.1, 0.5, 0.9
4	3	9.4128	15.92, 40, 66.67, 95.24	0.1, 0.37, 0.64, 0.9

4. Model Comparison

The next stage involves modeling using both parametric and semiparametric approaches. To account for variability due to random data partitioning, each model is repeated 100 times. This repetition ensures that the results are more stable and not influenced by a single instance of data splitting. The significance of the explanatory variable parameters is evaluated using the Wald Test. Table 7 presents the parameter estimates and the percentage of significance for the parametric models, including ZIPMM, ZIGPMM, and ZINBMM. Meanwhile, Table 8 summarizes the results for the semiparametric models: SZIPMM, SZIGPMM, and SZINBMM. The result from ZIPMM and SZIPMM indicate that all parameters are statistically significant. This is due to the use of both models when dealing with overdispersed data, tends to result in the significance test rejecting the null hypothesis.

Table 7. Parameters Estimation for Parametric Models

Parameter	ZIPMM	% Sig	ZIGPMM	% Sig	ZINBMM	% Sig
β_0	0.295*	88%	-1.039*	100%	-1.786*	100%
β_1 $(Private)$	0.974*	100%	-0.367*	100%	0.430*	100%
β_2	-0.025*	100%	0.008*	100%	0.001	13%
β_3	0.019*	100%	0.025*	100%	0.049*	100%
β_4	-0.060*	89%	-0.022	0%	-0.064*	91%
β_5	0.006*	100%	0.003*	100%	0.009*	100%
β_6	0.004*	88%	0.004*	92%	0.010*	98%
β_7	-0.007*	100%	0.003	46%	-0.001	1%
Random Effect						
Province $(Intercept)$	0.187		0.157		0.240	
Zero-Inflation $(Intercept)$	1.275*	100%	-18.517	0%	-18.003	0%

* significant at the 5% level

Table 8. Parameters Estimation for Semiparametric Models

Parameter	SZIPMM	% Sig	SZIGPMM	% Sig	SZINBMM	% Sig
β_0	2.819*	100%	-1.950*	100%	-1.093*	100%
β_1 $(Private)$	0.999*	100%	-0.344*	100%	0.458*	100%
β_2	-0.019*	100%	0.008*	100%	0.001	10%
β_3	0.020*	100%	0.023*	100%	0.045*	100%
β_4	-0.066*	99%	-0.021	0%	-0.066*	97%
β_{51}	-3.386*	100%	0.587	2%	-1.446*	94%
β_{52}	-1.814*	100%	1.554*	100%	0.509	13%
β_{53}	-2.310*	100%	0.949*	82%	-0.357	11%
β_{54}	2.241*	100%	1.221*	100%	-0.058	7%
β_{55}	-1.074*	100%	1.198*	100%	0.860*	87%
β_6	0.003*	76%	0.004*	97%	0.010*	100%
β_7	-0.005*	99%	0.003*	85%	0.001	0%
Random Effect Province $(Intercept)$	0.147		0.153		0.217	
Zero-Inflation $(Intercept)$	1.261*	100%	-18.534	0%	-17.988	0%

* significant at the 5% level

Each model generates a different spline curve (Figure 4), ilustrating how each model captures the relationship between X_5 (the percentage of students whose fathers' education level is below high school) and the model's output (school dropout rates). The SZINBMM and SZIPMM models exhibit similar patterns, although SZIPMM displays a sharper nonlinear effect. The value of $f(X_5)$ decreases significantly until around $X_5 \approx 10$ and then gradually increases. The value of $f(X_5)$ remains negative across most of the X_5 range, indicating that this variable primarily contributes to a decrease in the response variable. In contrast, the SZIGPMM model shows an increasing trend, with $f(X_5)$ remaining relatively stable after $X_5 \approx 25$. The changes are more gradual compared to the other models, suggesting that the effect of the variable X_5 in this model is more stable. The variable X_5 tends to have a positive relationship with school dropout rates. This means that as the percentage of students whose fathers have less than a high school education increases, the likelihood of dropping out tends to increase.

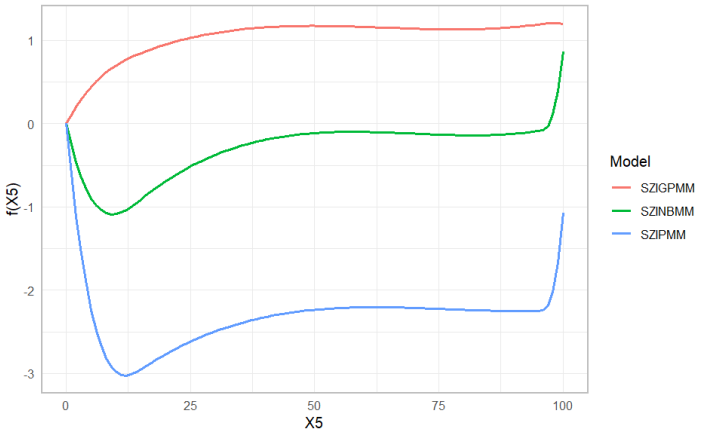

Figure 4. Spline Functions for X_5

Model comparison is conducted using the average of AIC, the average of RMSE for training and testing data. The best model is selected based on the lowest AIC and RMSE values. As shown in Table 9, the SZIGPMM model achieves the lowest AIC, while the RMSE values across all models are relatively close, indicating comparable predictive accuracy. This suggests that all models are suitable for prediction. However, SZIGPMM is considered the best model due to its optimal trade-off between model fit and complexity, as evidenced by the lowest AIC and a stable spline curve.

Table 9. Model Comparison

Model	AIC	RMSE Train	RMSE Test
ZIPMM	26009.95	3.073	2.841
ZIGPMM	18988.38	3.078	2.829
ZINBMM	19382.18	3.067	2.825
SZIPMM	25349.87	3,107	2,920
SZIGPMM	18969.62	3,079	2,829
SZINBMM	19327.25	3,061	2,823

Based on the significant variables, modeling using SZIGPMM can be expressed as follows:

$$\hat{y} = \exp(-1.950 - 0.344(X_1 = private) + 0.008X_2 + 0.023X_3 + 1.554B_{-1,2}(X_5) + 0.949B_{0,2}(X_5) + 1.221B_{1,2}(X_5) + 1.198B_{2,2}(X_5) + 0.004X_6 + 0.003X_7)$$

The negative coefficients for X_1 (private schools) indicate that, on average private schools have lower dropout rates than public schools, holding other variables constant. X_2 (student-teacher ratio), X_3 (distance from home to school), X_6 (parental employment status) and X_7 (number of siblings), all have positive coefficients, indicating that as these variables increase, the expected dropout rate also increases. The basis spline coefficients for X_5 (parental education level) are all positive, indicating a nonlinear but overall positive relationship between X_5 and dropout rates. Suggesting that the higher the percentage of parents with education below high school, the higher the tendency for school dropout rates. Overall, these findings imply that socioeconomic and school-level factors significantly influence dropout rates, with both linear and nonlinear effects captured in the model.

The significant factors influencing dropout rates include school status (public or private), student-teacher ratio, distance to school, parental education, employment status, and the number of siblings. This is consistent with research from Mubarokah et al. (2016) that shows that the student-teacher ratio significantly affects dropout rates. Rahma & Arcana (2019) identify several significant factors influencing the risk of school dropout among adolescents in Papua: the employment status of household heads, the education level of household heads, and the number of household members.

D. CONCLUSION AND SUGGESTIONS

The analysis revealed that the SZIGPMM model is the best choice for modeling high school dropout rates in Indonesia, as indicated by the lowest AIC and the stability of the spline components. This model effectively addressed overdispersion and excess zeros in the count data, confirming the advantages of semiparametric multilevel models combined with B-Splines for capturing nonlinear effects. One limitation of this study is the complexity of the resulting model. This may pose challenges for interpretation and practical application by policymakers without statistical expertise. Beyond the technical performance, the findings highlight several significant factors contributing to high school dropout rates. These include school status (public or private), student-teacher ratio, distance from home to school, parental education level, parental employment status, and the number of siblings. These factors suggest that both institutional conditions and family background play a crucial role in influencing students decisions to remain in school. The results emphasize the need for targeted educational policies that address these underlying issues to reduce dropout rates and improve educational attainment across regions.

ACKNOWLEDGEMENT

This research was made possible due to the support from relevant parties. The author expresses gratitude to Beasiswa Unggulan (BU) for funding the master's studies at the Department of Statistics and Data Science. Special thanks to the Data and Information

Technology Center, Ministry of Education, Culture, Research and Technology, for providing the analyzed data, leading to valuable research outcomes.

REFERENCES

Agresti, A. (2013). *Categorical Data Analysis* (3rd ed.). John Wiley & Sons, Inc. https://doi.org/10.1002/0470114754

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models Wiley Series in Probability and Statistics* (3rd ed.). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118730034

Almasi, A., Eshraghian, M. R., Moghimbeigi, A., Rahimi, A., Mohammad, K., & Fallahigilan, S. (2016). Multilevel zero-inflated Generalized Poisson regression modeling for dispersed correlated count data. *Statistical Methodology*, *30*, 1–14. https://doi.org/10.1016/j.stamet.2015.11.001

Aráujo, E. G., Vasconcelos, J. C. S., dos Santos, D. P., Ortega, E. M. M., de Souza, D., & Zanetoni, J. P. F. (2023). The Zero-Inflated Negative Binomial Semiparametric Regression Model: Application to Number of Failing Grades Data. *Annals of Data Science*, *10*(4), 991–1006. https://doi.org/10.1007/s40745-021-00350-z

Beccari, C. V., & Casciola, G. (2021). *Stable numerical evaluation of multi-degree B-splines*. http://arxiv.org/abs/2102.03252

Belloc, F., Maruotti, A., & Petrella, L. (2011). How individual characteristics affect university students drop-out: A semiparametric mixed-effects model for an Italian case study. *Journal of Applied Statistics*, *38*(10), 2225–2239. https://doi.org/10.1080/02664763.2010.545373

Chudy, F., & Woźny, P. (2022). *Linear-time algorithm for computing the Bernstein-B\'{e}zier coefficients of B-spline basis functions*. http://arxiv.org/abs/2204.05002

Dean, C. B., & Lundy, E. R. (2016). Overdispersion. *Wiley StatsRef: Statistics Reference Online*, 1–9. https://doi.org/10.1002/9781118445112.stat06788.pub2

Fernandez, G. A., & Vatcheva, K. P. (2022). A comparison of statistical methods for modeling count data with an application to hospital length of stay. *BMC Medical Research Methodology*, *22*(1), 1–21. https://doi.org/10.1186/s12874-022-01685-8

Gbaguidi, V. E., & Adetou, D. (2024). Factors affecting school dropout: Comparative study of rural and urban settings. *International Journal of Educational Management and Development Studies*, *5*(2), 233–256. https://doi.org/10.53378/353073

Hardin, J. W., & Hilbe, J. M. (2018). *Generalized linear models and extensions*. Stata Press. https://www.stata-press.com/books/generalized-linear-models-and-extensions/

Hilbe, J. M. (2011). *Negative Binomial Regression* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511973420

Jiang, J., & Nguyen, T. (2021). *Linear and Generalized Linear Mixed Models and Their Applications* (2nd ed.). Springer. http://www.springer.com/series/692

[MoECRT] *Ministry of Education, Culture, Research and Technology*. Statistik Sekolah Menengah Atas Tahun 2022-2023. https://data.kemdikbud.go.id/publikasi/p/pauddasmen-buku-statistik/statistik-sekolah-menengah-atas-sma-tahun-2022-2023

Ole Kinisa, G. R. (2019). Effectiveness of Educational Policy in Curbing School Dropout in Secondary Schools in Tanzania: A Case of Dodoma City. *International Journal of Scientific and Research Publications (IJSRP)*, *9*(5), 129–159. https://doi.org/10.29322/ijsrp.9.05.2019.p8916

Mahmoodi, M., Moghimbeigi, A., Mohammad, K., & Faradmal, J. (2016). Semiparametric models for multilevel overdispersed count data with extra zeros. *Statistical Methods in Medical Research*, *27*(4), 1–15. https://doi.org/10.1177/0962280216657376

Mahmoud, H. F. F. (2021). Parametric Versus Semi and Nonparametric Regression Models. *International Journal of Statistics and Probability*, *10*(2), 90–109. https://doi.org/10.5539/ijsp.v10n2p90

Masci, C., Ieva, F., & Paganoni, A. M. (2022). Semiparametric Multinomial Mixed-Effects Models: A University Students Profiling Tool. *Annals of Applied Statistics*, *16*(3), 1608–1632. https://doi.org/10.1214/21-AOAS1559

Mubarokah, L., Budiantara, I. N., & Ratna, M. (2016). Pemodelan Angka Putus Sekolah Usia SMP Menggunakan Metode Regresi Nonparametrik Spline di Papua. *Jurnal Sains Dan Seni ITS*, *5*(1), 2337–3520. https://ejurnal.its.ac.id

Myers, R. H., Montgomery, D. C., Vining, G. G., & Robinson, T. J. (2010). *Generalized Linear Models: With Applications in Engineering and the Sciences* (2nd ed.). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470556986

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., & Schmid, M. (2019). A review of spline function procedures in R. In *BMC Medical Research Methodology* (Vol. 19, Issue 1, pp. 1–16). BioMed Central Ltd. https://doi.org/10.1186/s12874-019-0666-3

Pramaningrum, D. S., Fernandes, A. A. R., Iriany, A., & Solimun, S. (2024). The Application of Truncated Spline Semiparametric Path Analysis on Determining Factors Influencing Cashless Society Development. *JTAM (Jurnal Teori Dan Aplikasi Matematika)*, *8*(2), 400–410. https://doi.org/10.31764/jtam.v8i2.19913

Rahma, R. A., & Arcana, I. M. (2019). Risk Level of Dropping Out of School for Adolescent in Papua Province 2018. *Seminar Nasional Official Statistics*, 672–681. https://doi.org/10.34123/semnasoffstat.v2020i1.468

Ruíz, J. S., Montesinos López, O. A., Ramírez, G. H., & Hiriart, J. C. (2023). *Generalized Linear Mixed Models with Applications in Agriculture and Biology*. Springer. https://doi.org/10.1007/978-3-031-32800-8

Law Number 23 of 2014 on Regional Government. (2014). https://peraturan.bpk.go.id/Details/38685/uu-no-23-tahun-2014

UNESCO. (2025). *Sustainable Development Goal 4*. https://www.unesco.org/sdg4education2030/en/sdg4

Utami, T. W., Chamidah, N., & Saifudin, T. (2024). Platelet Modeling in DHF Patients Using Local Polynomial Semiparametric Regression on Longitudinal Data. *JTAM (Jurnal Teori Dan Aplikasi Matematika)*, *8*(1), 231–243. https://doi.org/10.31764/jtam.v8i1.17427