Penalized Multivariate Adaptive Regression Splines with Generalized Cross Validation for Modeling Health Insurance Ownership in East Java

Deby Victoria[1], Addina Nurkamila[1], Yanuar Ibnu Ridho[1], Rafly Tawekal[1], and Dita Amelia[1*]
[1]Departement of Mathemathics, University of Airlangga, Indonesia
*Corresponding author: dita.amelia@fst.unair.ac.id

Received: 8 January 2026
Revised: 20 May 2026
Accepted: 25 May 2026

ABSTRACT – **The average programmatic health insurance coverage is a key indicator of public welfare and the effectiveness of healthcare policies. This study proposes the use of Penalized Multivariate Adaptive Regression Splines (PMARS) with Generalized Cross Validation (GCV) to model the determinants of this average coverage rate across 38 regencies and cities in East Java Province, using secondary data from the 2025 BPS East Java publication. Several PMARS specifications are evaluated by varying polynomial degrees, numbers of basis functions, and penalty values to identify the optimal model structure. The PMARS approach effectively captures nonlinear relationships, interaction effects, and variable importance within a flexible regression framework. The optimal model is a quadratic specification consisting of 16 basis functions, a maximum interaction of two, and an optimal penalty parameter, explaining 94.46% of the variability in the average health insurance coverage with a minimum GCV value of 5.49654. Access to improved sanitation, clean water, and health complaints are identified as the most influential determinants, all exhibiting positive associations that drive the need for formal financial protection. These results demonstrate the effectiveness of the GCV-PMARS methodology for modeling complex socio-health data and provide empirical insights to support policies aimed at strengthening universal health coverage, ensuring equitable programmatic utilization, and advancing Sustainable Development Goal 3 on Good Health and Well-Being.**

Keywords– **Health Insurance Ownership; GCV-PMARS Method; East Java Province; Nonparametric Regression.**

I. INTRODUCTION

Health is a fundamental pillar of human development, shaping quality of life, productivity, and societal well-being [1]. In Indonesia, equitable healthcare access is primarily facilitated through the National Health Insurance program (Jaminan Kesehatan Nasional/JKN), which aims to provide financial protection and reduce catastrophic out-of-pocket medical expenses [2]. As of June 30, 2024, JKN coverage reached 273.5 million individuals (96.8% of the national population), while East Java recorded 95.83% coverage as of March 31, 2025, approaching the Universal Health Coverage (UHC) target of 98% [3],[4],[5]. Despite this achievement, disparities persist in the distribution of insurance participation across the 38 regencies and cities in East Java, indicating that aggregate coverage may mask structural inequalities in healthcare access. This study therefore shifts the focus from total coverage to the average percentage across three BPS-reported coverage categories: BPJS Kesehatan PBI, BPJS Kesehatan Non-PBI, and Jamkesda. While BPJS Kesehatan is classified into PBI and Non-PBI, Jamkesda represents a complementary local government program, enabling a more nuanced subregional analysis of equity in healthcare access [6],[7].

Environmental conditions, health behaviors, and healthcare infrastructure display substantial regional heterogeneity and play a crucial role in shaping health insurance ownership patterns [8]. Variations in health complaint prevalence, smoking behavior, sanitation coverage, and access to clean water remain pronounced across districts in East Java [4]. Similarly, healthcare service availability is unevenly distributed, with marked differences in the number of health facilities between regencies [4]. These factors interact in complex and potentially nonlinear ways, particularly in areas facing high health risks and limited infrastructure, where barriers to JKN enrollment often persist despite elevated healthcare needs. Conventional linear regression approaches struggle to capture such interaction complexity, frequently leading to biased estimates and limited policy relevance.

To overcome these methodological limitations, this study employs Penalized Multivariate Adaptive Regression Splines (PMARS) with Generalized Cross Validation (GCV) to examine the determinants of average programmatic JKN coverage across East Java's regencies and cities. PMARS effectively captures nonlinear relationships and interaction effects while controlling model complexity through internal penalization, ensuring robust and interpretable results [9]. Unlike LASSO-penalized MARS approaches that impose external $\ell 1$ penalties [10], this study utilizes the native complexity penalty within the earth package optimized via GCV. This approach represents a methodological advancement, as no prior studies have applied PMARS-GCV to analyze environmental, behavioral, and healthcare service determinants of JKN ownership disparities using this averaged metric in East Java.

This research contributes empirically by identifying the dominant nonlinear drivers of health insurance disparities at the regency/city level, providing evidence-based guidance for targeted policy interventions. The findings support Indonesia's efforts to accelerate UHC achievement while aligning with Sustainable Development Goal 3, which emphasizes universal health coverage and equitable access to healthcare by 2030 [5]. Beyond East Java, this study offers

a replicable analytical framework for subnational health policy analysis in diverse developing country contexts.

II. LITERATURE REVIEW

A. Nonparametric Regression Framework

Regression analysis helps model the relationship between a response variable and one or more predictors. When the true form of this relationship is unclear and doesn't fit straightforward linear or polynomial shapes, nonparametric methods work best. These approaches assume only that the regression function is smooth and fits within a specific function class, letting the data dictate the curve's form without rigid parametric constraints [11].

In essence, the nonparametric regression model takes the form

$$y_i = f(x_i) + \varepsilon_i, i = 1,2,\dots,n \tag{1}$$

In this model, y_i denotes the response variable, x_i represents the predictor (or vector of predictors), $f(x_i)$ signifies the unknown smooth regression function, and ε_i is the random error term, typically assumed to follow a normal distribution $N(0,\sigma_\varepsilon^2)$.

B. Spline Regression and Penalized Splines

Spline regression is a nonparametric regression approach used to model nonlinear relationships between predictor and response variables. The general spline regression model can be expressed as [12]:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_q x^q + \sum_{i=1}^{K} \gamma_i (x - k_i)_+^q \tag{2}$$

where q is the polynomial order, k_i are knot locations, K represents the number of knots, while $(x-k_i)_+^q$ signifies the truncated power basis function. To avoid overfitting when many knots are used, penalized splines introduce a penalty term on the spline coefficients through the Penalized Least Squares (PLS) criterion [13],[12]. In matrix form, PLS is defined as

$$PLS = (\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{Y} - \boldsymbol{X\beta}) + \lambda^2 \boldsymbol{\beta}^T \boldsymbol{D\beta} \tag{3}$$

In this context, λ indicates the smoothing parameter, while $\mathbf{D}$ serves as the penalty matrix.

The penalty matrix $\mathbf{D}$ is constructed such that only the spline coefficients are penalized, while the low-order polynomial part (intercept and global trend) is left unpenalized. Suppose $\boldsymbol{\beta}$ consists of p unpenalized polynomial coefficients and r penalized spline coefficients; then $\mathbf{D}$ has the block structure

$$\boldsymbol{D} = \begin{bmatrix} \mathbf{0}_{pxp} & \mathbf{0}_{pxr} \\ \mathbf{0}_{rxp} & \boldsymbol{I}_{rxr} \end{bmatrix}$$

where $\mathbf{0}_{.\times.}$ is a zero matrix and $\mathbf{I}_{r\times r}$ is the identity matrix of size $r \times r$.

The zero blocks imply that the polynomial coefficients are not penalized, whereas the identity block means that each spline coefficient receives a quadratic (ridge-type) penalty, so that the penalty term $\boldsymbol{\beta}^T\mathbf{D}\boldsymbol{\beta}$ effectively sums only the squared spline coefficients and controls the roughness of the fitted curve [12].

Minimizing the PLS criterion with respect to $\boldsymbol{\beta}$ yields the penalized spline estimator

$$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda^2 \boldsymbol{D})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{4}$$

which shows that $\boldsymbol{\lambda^2 D}$ acts as a penalty term augmenting $\mathbf{X^TX}$ and shrinking the spline coefficients toward zero as λ increases.

C. Multivariate Adaptive Regression Splines (MARS)

Multivariate Adaptive Regression Splines (MARS), [14] offers a versatile nonparametric regression technique for capturing nonlinear relationships and interactions between predictor variables. MARS constructs the regression function through a linear combination of basis functions (BF), typically hinge functions expressed as $\max(0, x - t)$or $\max(0, t - x)$ at different knot locations.

Generally, the MARS model takes the form [14]:

$$\hat{f}(x) = \alpha_0 + \sum_{m=1}^{M} \alpha_m \prod_{k=1}^{K_m} [s_{km}.(x_{v(k,m)} - t_{km})] \tag{5}$$

where:

$\alpha_0$is the intercept,
$\alpha_m$are coefficients of the m-th basis function,
$M$is the number of non-constant basis functions,
$K_m$is the interaction degree of the m-th term,
$s_{km} \in \{-1,1\}$determines the side of the knot,
$x_{v(k,m)}$is the predictor involved, and
$t_{km}$is the knot location.

MARS models are built in two stages: forward selection (adding basis functions to improve fit) and backward pruning (removing non-contributing basis functions to control model complexity).

D. Penalized Multivariate Adaptive Regression Splines (PMARS)

While MARS excels at modeling nonlinear patterns, too many basis functions can cause overfitting. Penalized Multivariate Adaptive Regression Splines (PMARS) enhances the traditional MARS method by adding a specific complexity penalty to the basis functions [14], [15].

The criterion minimized in PMARS is the Penalized Residual Sum of Squares (PRSS):

$$PRSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \sum_{m=1}^{M_{max}} \lambda_m P(\psi_m) \tag{6}$$

with $\hat{y}_i = \boldsymbol{\theta}^\top \boldsymbol{\psi}(\bar{\mathbf{d}}_i)$, where:

y_i is the observed response at observation i,

$\hat{y}_i$ is the fitted value,

$\bar{\mathbf{d}}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ is the predictor vector,

$\boldsymbol{\psi}(\bar{\mathbf{d}}_i)$ is the vector of MARS basis functions evaluated at $\bar{\mathbf{d}}_i$,

$\boldsymbol{\theta} = (\theta_0, \dots, \theta_{M_{\max}})^\top$ is the vector of coefficients,

$M_{\max}$ is the maximum number of basis functions,

λ_m is the penalty parameter for the m-th basis function, and

$P(\psi_m)$ measures the complexity of basis function ψ_m.

Through this penalized framework, only relevant basis functions are retained, resulting in a model that is simpler, more stable, and has better generalization ability.

E. Model Selection Criteria: Generalized Cross Validation (GCV) and Coefficient of Determination

1) Generalized Cross Validation (GCV)

During model building, the MARS/PMARS procedure consists of two stages: forward selection, which sequentially adds basis functions, and backward pruning, which removes non–contributory basis functions. To select the optimal model, this research utilizes the Generalized Cross Validation (GCV) criterion, which balances goodness of fit and model complexity [14], [16]. The GCV is defined as

$$PGCV = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\left(1 - \frac{C(M)}{N}\right)^2} \tag{7}$$

here, N represents the total number of observations, while y_i denotes the observed response for the i -th observation, $\hat{y}_i$ represents the predicted value for the i -th observation, and $C(M)$ serves as a measure of model complexity based on the effective number of basis functions. The preferred model is the one with the smallest GCV value, as it provides the best trade-off between predictive accuracy and parsimony [17].

2) Coefficient of Determination (R^2)

The coefficient of determination (R^2) serves as an additional goodness-of-fit metric, measuring the percentage of variation in the response variable accounted for by the model [18]. It is defined as one subtracting the ratio of sum of squared errors (SSE) to total sum of squares (SST).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \; x \; 100\% \tag{8}$$

where y_i represents the observed response for the i -th observation, and $\hat{y}_i$ is denotes the predicted response for the i-th unit, and $\bar{y}$ is the sample mean of the observed responses. Values of R^2 closer to 1 indicate that a larger proportion of the variance in the response is explained by the model, corresponding to a better fit, whereas values near 0 indicate limited explanatory power.

F. Model Selection Criteria: Generalized Cross Validation (GCV) and Coefficient of Determination

The simultaneous test is performed to assess the overall significance of the regression model in explaining the variability of the dependent variable Y [19]. This evaluation is carried out using the F-test, as formulated by Wicaksono [20], with the test statistic expressed as

$$F_{\text{calc}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2/M}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(N - M - 1)}. \tag{9}$$

The null hypothesis H_0 posits that all regression coefficients are simultaneously equal to zero ($\alpha_1 = \alpha_2 = \cdots = \alpha_M = 0$), implying that the independent variables do not jointly affect the dependent variable. In contrast, the alternative hypothesis H_1 states that at least one regression coefficient α_m differs from zero. The null hypothesis is rejected when the computed F-statistic exceeds the critical value $F_{\alpha(M, N-M-1)}$ or when the associated *p-value* is lower than the significance level α, indicating that the independent variables collectively contribute significantly to the model.

In addition to the simultaneous test, a partial test is conducted to determine the significance of each independent variable individually in influencing the dependent variable [21]. This analysis employs the t-test statistic, following Wicaksono [20], which is calculated as

$$t_{\text{calc}} = \frac{\hat{\alpha}_m}{\text{se}(\hat{\alpha}_m)}, \text{where se}(\hat{\alpha}_m) = \sqrt{\text{Var}(\hat{\alpha}_m)}. \quad (10)$$

The hypotheses tested in the partial analysis consist of $H_0: \alpha_m = 0$, indicating that the m-th independent variable has no significant effect on the dependent variable, and $H_1: \alpha_m \neq 0$. The null hypothesis is rejected if the absolute value of the calculated t-statistic exceeds the critical value $t_{\alpha/2,\, N-M-1}$ or if the corresponding *p-value* is smaller than the specified significance level α. This result suggests that the independent variable under consideration has a statistically significant influence on the dependent variable.

G. Diagnostic Checks: Normality, Multicollinearity, and Heterogeneity

The normality test is applied to evaluate whether the research data satisfy the assumption of normal distribution required in regression analysis. In this study, data normality is assessed using the Shapiro–Wilk test, which is particularly suitable for small to moderate sample sizes. The hypotheses tested are H_0, indicating normally distributed data, and H_1, indicating non-normal data. The Shapiro–Wilk test statistic is defined as

$$W = \frac{\left(\sum_{i=1}^{n} a_i\, x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (11)$$

where $x_{(i)}$ denotes the i-th order statistic, $\bar{x}$ represents the sample mean, and a_i are constants derived from the expected values and covariance structure of order statistics from a normal distribution and are automatically determined by statistical software. The null hypothesis is accepted when the *p-value* is greater than or equal to the significance level α, indicating that the normality assumption is met.

In addition, the multicollinearity test is performed to identify the presence of high linear relationships among independent variables within the regression model. This test utilizes the Variance Inflation Factor (VIF) as an indicator of multicollinearity, which is calculated using the formula

$$VIF = \frac{1}{1 - R_j^2}, j = 1,2, \ldots, p \quad (12)$$

where R_j^2 is obtained by regressing the j-th explanatory variable on the remaining predictors. Multicollinearity is indicated when VIF > 10 (or tolerance <0.1) ; otherwise, the model is considered free from serious multicollinearity.

Finally, potential heteroskedasticity (non-constant error variance across observations) was examined using the Breusch-Pagan (BP) test. The BP test evaluates whether the variance of the regression residuals depends on the explanatory variables, which is particularly relevant in cross-sectional data where variability across observational units may occur.

The hypotheses are:

$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2 = \sigma^2$ (homoskedasticity)

$H_1 : i \text{ such that } \sigma_i^2 \neq \sigma^2$ (heteroskedasticity)

The BP test is based on an auxiliary regression of squared residuals on the original explanatory variables. The test statistic is computed as:

$$BP = nR^2 \quad (13)$$

where n is the number of observations and R^2 is the coefficient of determination obtained from regressing the squared residuals on the independent variables. Under the null hypothesis, the BP statistic follows a chi-square distribution with degrees of freedom equal to the number of predictors.

The decision is based on the p-value (or chi-square critical value): H_0 is rejected when $p < \alpha$, indicating the presence of heteroskedasticity in the model residuals.

H. Implementation Procedure for GCV-PMARS in This Study

The Penalized MARS procedure begins with data preparation in R by importing the dataset, defining the response and predictors, and checking missing values and outliers. The model is built through a forward pass that adds basis functions to capture nonlinear patterns under various BF, MI, and MO settings, then controlled using penalty values and backward pruning to remove insignificant functions. The best model is chosen based on the smallest GCV and largest R^2. Model significance is assessed with F- and t-tests, residual assumptions are verified via diagnostics, and the final model is interpreted by identifying significant predictors and explaining nonlinear and interaction effects.

III. METHODOLOGY

A. Data Source

This study employs secondary data obtained from the publication *East Java Province in Figures 2025*, issued by the Statistics Indonesia (Badan Pusat Statistik, BPS) of East Java Province. The dataset comprises regency- and city-level information across East Java Province for the year 2024.

B. Research Variables

The variables employed in this study are presented in Table 1.

Table 1 Research Variable

Variable	Variable Description	Unit	Scale
Y	Average Percentage of Coverage per Health Insurance Program	Percentage	Ratio
X_1	Percentage of Households with Access to Improved Sanitation	Percentage	Ratio
X_2	Percentage of Households with Access to Improved Drinking Water	Percentage	Ratio
X_3	Percentage of Population Aged 15 Years and Above Who Smoked in the Past Month	Percentage	Ratio
X_4	Percentage of Population Experiencing Health Complaints	Percentage	Ratio
X_5	Number of Healthcare Facilities	Unit (Number of Facilities)	Ratio

C. Data Analysis

The data analysis in this study was conducted through the following sequential steps:

1. Compiling descriptive statistics to summarize the initial characteristics of the response and predictor variables.
2. Preparing the dataset for Penalized Multivariate Adaptive Regression Splines (PMARS) modeling.
3. Constructing the initial PMARS model using the forward selection procedure.
4. Applying penalty parameters and performing backward pruning to control model complexity.
5. Selecting the optimal model based on the minimum Generalized Cross Validation (GCV) value and the maximum R^2 values.
6. Conducting overall and partial significance tests for the selected Penalized MARS model.
7. Performing residual diagnostics to evaluate model adequacy.
8. Interpreting the final Penalized MARS model and identifying significant predictor variables.
9. rawing conclusions and discussing the implications of the findings.

IV. RESULTS AND DISCUSSIONS

A. Statistics Descriptive

This study presents descriptive statistical summaries produced for the response variable (Y), the percentage of health insurance coverage, and five associated predictor variables (X).

Table 2 Statistics Descriptive

Variable	N	Mean	Standard Deviation	Minimum	Median	Maximum
Y	38	24.58	5.77	11.76	25.72	33
X_1	38	86.22	11.49	50.31	87.84	98.56
X_2	38	97.74	3.57	82.87	98.74	100
X_3	38	27.28	4.09	19.19	27.37	38.68
X_4	38	29.18	5.19	15.88	28.75	39.36
X_5	38	2761.4	8432.6	209	1422	53848

Based on Table 2, decent sanitation averages 86.22% (50.31%–98.56%) and suitable drinking water is very high at 97.74% (82.87%–100%). Smoking prevalence (15+) averages 27.28% (19.19%–38.68%), while health complaints average 29.18% (15.88%–39.36%). Total health services average 2,761.4 units with a wide range (209–53,848), indicating strong regional disparities. Health insurance coverage averages 24.58% (11.76%–33%), showing relatively low and uneven coverage across regions.

B. Model Estimation Penalized MARS

Prior to estimating the model, multicollinearity among the explanatory variables was examined, with the results summarized in Table 3.

Table 3 Multicolinearity Test Results

Variable	VIF
X_1	1.73
X_2	1.18
X_3	1.68
X_4	1.20
X_5	1.09

The multicollinearity test indicates that all independent variables have VIF values below 10, suggesting no serious multicollinearity and confirming that all variables can be reliably included in the model. Model selection was carried out in RStudio through a systematic combination of Basis Function (BF), Maximum Interaction (MI), and Minimum Observation between knots (MO). The BF values evaluated were 10, 15, 20, and 25; MI values were 1, 2, and 3; MO values were 0, 1, 2, and 3; and penalty values were 0, 1, and 2. The optimal model was selected based on the minimum GCV value, with parsimony used as a tie-breaking criterion when multiple models yielded identical GCV values. The model combinations for BF = 25 are summarized in Table 4.

Table 4 Best Model Based on Penalized MARS Combinations

Minimum Interaction Level	2
Number of Basis Functions	25
Penalty Value	0 (terms only)
Minimum Observations per Knot	1
GCV	5.49654
R-Square	0.9446
Number of Basis Functions After Penalization	16

The basis functions can be estimated once the best model is selected. Initially, the model produced 25 basis functions. However, because penalization in this study was applied only to the basis functions (terms) and not to the knots, the penalized MARS procedure controlled model complexity without restricting the number of knots formed. After penalization to prevent overfitting, the final model retained 16 basis functions. The estimates of these selected basis functions are presented in Table 5.

Table 5 Estimated Significant Basis Functions of the Best Model

BF	Basis Function	Parameter Estimate
Intercept	1	27.1603
BF1	$max(0;\ X_1 - 82.99)$	-1.05583
BF2	$max(0;\ 89.99 - X_1)$	-11.9256
BF3	$max(0;\ X_1 - 89.99)$	6.364742
BF4	$max(0;\ X_1 - 91.47)$	-5.56412
BF5	$max(0;\ X_2 - 98.22)$	-37.4167
BF6	$max(0;\ 98.34 - X_2)$	-2.34258
BF7	$max(0;\ X_2 - 98.34)$	42.89687
BF8	$max(0;\ 29.46 - X_3)$	-3.90363
BF9	$max(0;\ 25.67 - X_4)$	1.569166
BF10	$\max(0; X_4 - 25.67)$	-0.50712
BF11	$max(0;\ 1398 - X_5)$	0.509242
BF12	$max(0; 89.99 - X_1) * X_2$	0.103701
BF13	$max(0; 89.99 - X_1) * X_4$	0.061056
BF14	$X_2 * max(0; 1398 - X_5)$	-0.00509
BF15	$max(0; 29.46 - X_3) * X_4$	0.144429

Based on Table 5, the penalized MARS model for estimation is obtained as follows.

$$\hat{y} = 27.1603 - 1.05583BF1 - 11.9256BF2 + 6.364742BF3 - 5.56412BF4 - 37.4167BF5 - 2.34258BF6 + 42.89687BF7 - 3.90363BF8 + 1.569166BF9 - 0.50712BF10 + 0.509242BF11 + 0.103701BF12 + 0.061056BF13 - 0.00509BF14 + 0.144429BF15$$

Suppose the conditions are $X_1 > 91.47$, $X_2 > 98.34$, $X_3 < 29.46$, $X_4 > 25.67$, and $X_5 < 1398$. Then, the estimated model $\hat{y}$ becomes:

$$\hat{y} = 27.1603 - 1.05583(X_1 - 82.99) + 6.364742(X_1 - 89.99) - 5.56412(X_1 - 91.47) - 37.4167(X_2 - 98.22) + 42.89687(X_2 - 98.34) - 3.90363(29.46 - X_3) - 0.50712(X_4 - 25.67) + 0.509242(1398 - X_5) - 0.00509(1398 - X_5) * X_2 + 0.144429(29.46 - X_3) * X_4$$

$$\hat{y} = 117.49393668 - 0.255216X_1 - 1.628613X_2 + 3.903631X_3 + 3.74775434X_4 - 0.509242X_5 + 0.00509X_2X_5 - 0.144429X_3X_4$$

C. Significance Testing of the Penalized MARS Model

Significance testing within the PMARS framework is conducted to examine both parameter relevance and overall model adequacy. The joint test determines whether the basis function coefficients collectively affect the response variable. The corresponding test statistics were obtained using R software and are presented below.

Table 6 Simultaneous Test of Basis Coefficients in the Penalized MARS Model

Test Statistic	Value
F-Statistic	475.58
p-value	2.2×10^{-16}

As shown in Table 6, the computed F-statistic of 475.58 is greater than the critical value $F_{0.05;15;22}$= 2.172, while the *p-value* of 2.2×10⁻¹⁶ is smaller than α = 0.05. Accordingly, the null hypothesis is rejected, indicating that the basis function coefficients collectively exhibit a significant association with the response variable. Subsequently, an individual significance test was performed to assess the effect of each basis function coefficient in the PMARS model. The outcomes of this partial testing are reported in Table 7.

Table 7 Partial Test of Basis Function Coefficients in the Penalized MARS Model

Parameter	S.E.	T-*Ratio*	*P-Value*
Intercept	2.08	13.045	7.87×10^{-12}
BF1	14.67	2.924	0.007852
BF2	0.57	-4.081	0.000495
BF3	0.09	5.487	1.63×10^{-5}
BF4	0.94	-4.154	0.000414
BF5	1.55	4.115	0.000455
BF6	3.44	-3.471	0.002168
BF7	0.01	9.362	3.95×10^{-9}
BF8	0.13	-4.049	0.000535
BF9	0.33	4.814	8.27×10^{-5}
BF10	0.04	2.959	0.007252
BF11	1.43	-3.901	0.000767
BF12	0.00	-5.46	1.74×10^{-5}
BF13	0.04	4.121	0.000449
BF14	13.80	-2.71	0.012775
BF15	0.36	-2.901	0.008282

The critical region for the partial test is to reject H_0 if $|t| > t_{0,025;22}$ = 2.074 or if the p-value is below α = 0.05. Based on Table 7, all basis functions have *p-values* < 0.05; therefore, H_0 is rejected for all coefficients. This indicates that every basis function in the model is significantly related to the response variable.

D. Variable Importance

Variable importance serves to order the predictor variables based on their relative impact on the response variable. The corresponding importance values within the model are presented below.

Table 8 Variable Importance

Variable	Variable Name	Importance Level
X_2	Percentage of Households with Safe Drinking Water	100%
X_1	Percentage of Households with Adequate Sanitation	68.7%
X_4	Percentage of Health Complaints	68.7%
X_5	Total Health Services	63%
X_3	Percentage of Smokers Aged 15 and Above	28.8%

The results indicate that basic environmental quality, specifically the percentage of households with safe drinking water and adequate sanitation, plays a dominant role in shaping the average participation rate across health insurance programs. Regions with better WASH (Water, Sanitation, and Hygiene) facilities tend to exhibit higher and more balanced programmatic coverage, likely because healthier environments are associated with greater community awareness and readiness to engage with formal, structured health protection schemes [22], [23]. Furthermore, indicators of health needs and behavioral risks, reflected by health complaints and smoking prevalence among productive-age individuals, also significantly drive this average coverage by elevating health risks and necessitating structured financial protection. The total availability of health services provides essential structural support but acts as a secondary driver compared to fundamental environmental conditions [24]. Overall, policy efforts aiming to ensure equitable and balanced utilization across JKN programs should integrate healthcare service provision with foundational environmental improvements and targeted behavioral interventions.

E. Residual Assumption Testing

Two assumption tests were applied in this study, namely the normality test and the constant variance assumption test. The first step was the normality test, the results of which are presented in Table 9.

Table 9 Results of the Shapiro–Wilk Test

Test Statistics	Value
W	0.9746
p-value	0.5295

The test rejects H_0 when the *p-value* is below $\alpha = 0.05$. Based on Table 9, the *p-value* is greater than 0.05, so H_0 is accepted; the W value of 0.9746 (close to 1) also indicates normality, meaning the residuals are normally distributed. Next, a constant variance (heteroskedasticity) check was performed using the Breusch–Pagan test, with results shown in Table 10.

Table 10 Result of Breusch-Pagan Test

Test Statistics	Value
Breusch-Pagan	0.0027825
p-value	0.9579

The null hypothesis is rejected if the *p-value* is less than $\alpha = 0.05$. Based on Table 10, the *p-value* exceeds 0.05, indicating acceptance of $\boldsymbol{H_0}$. The very small Breusch-Pagan statistic (0.0027825) further confirms the absence of heteroskedasticity, indicating that the residuals are homoscedastic.

F. Interpretation of the Penalized MARS Model Results

The best Penalized MARS model comprises 16 basis functions, a maximum interaction degree of 2, and a minimum of one observation between knots. It shows strong performance with $R^2 = 0.9446$(explaining 94.46% of the variation) and a low GCV of 5.49654. After confirming significant variables and residual assumptions, Figure 1 presents the comparison between observed and predicted values.1.

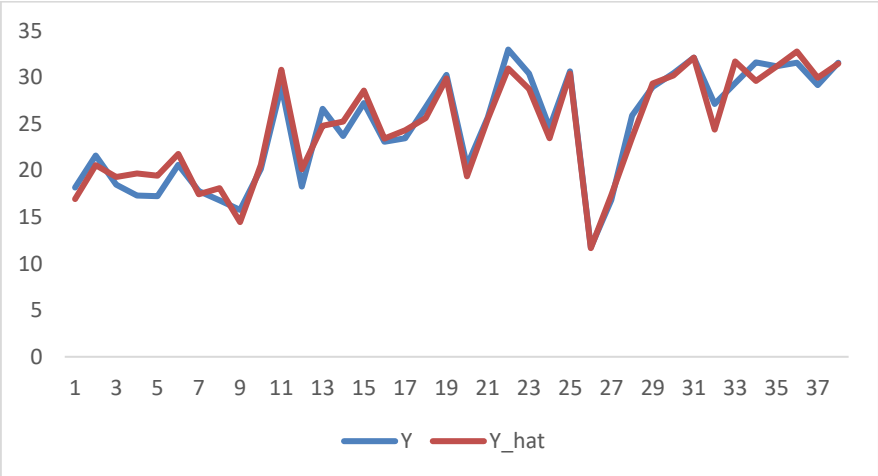

Figure 1 Comparison Plot of Actual and Predicted Values

The Penalized MARS prediction plot shows a strong agreement between predicted and observed values of health insurance ownership (Y), supporting regional classification based on significant basis functions. The model indicates that insurance ownership is associated with environmental conditions, public health status, and service availability: proper sanitation is negatively related to participation due to lower perceived health risks, whereas access to safe drinking water and smoking prevalence are positively associated with ownership. Health complaints exhibit the strongest positive association with insurance utilization, while a higher number of health facilities is linked to lower insurance use, potentially reflecting alternative access routes [25]. Interaction effects suggest that sanitation and safe water jointly reduce health risks, sanitation weakens the influence of health complaints, water access combined with health facilities strengthens utilization, and smokers who experience health complaints tend to underutilize insurance [26]. In East Java, Bojonegoro records the highest observed coverage (Y = 33; $X_1 = 96.43$, $X_2 = 98.74$, $X_3 = 31.21$, $X_4 = 21.28$, $X_5 = 1937$), while the best PMARS model produces an estimated value of $\hat{y} = 24.82$, obtained as follows:

$$\hat{y} = 117{,}49393668 - (0{,}255216 \times 96{,}43) - (1{,}628613 \times 98{,}74) + (3{,}903631 \times 31{,}21) + (3{,}74775434 \times 21{,}28) - (0{,}509242 \times 1937) + (0{,}00509 \times 98{,}74 \times 1937) - (0{,}144429 \times 31{,}21 \times 21{,}28)$$

$$\hat{y} = 24{,}82044921$$

Thus, the model underestimates the observed value by 8.18 points (33 vs. 24.82), although it still reflects the overall pattern captured by the PMARS specification.

V. CONCLUSIONS AND SUGGESTIONS

The analysis shows that GCV-PMARS is able to model the complex, non-linear relationship between the average programmatic health insurance coverage and its determinants across regencies/cities in East Java. The best model is obtained with a quadratic specification using 16 basis functions, a maximum interaction of two, and a minimum observation between knots of one, yielding a GCV of 5.49654 and an $R^2 = 0.9446$. This indicates that 94.46% of the variation in the average coverage rate can be explained by the percentage of improved sanitation, access to safe drinking water, adult smoking prevalence, health complaints, and the number of health service facilities.The results demonstrate that improved sanitation, safe drinking water, and the availability of health facilities are positively associated with the

average programmatic coverage. Furthermore, adult smoking and health complaints also exhibit a positive association; this reflects the logical dynamic that elevated health risks and behavioral vulnerabilities actively drive a higher community need for formal financial protection and structured healthcare access. Future research is recommended to extend the model with additional socio-economic variables, compare GCV-PMARS with alternative flexible methods, and apply the approach to longitudinal data, so that the robustness of the model and its policy implications for achieving more equitable and balanced programmatic utilization can be further validated.

REFERENCES

[1] H. Rotinsulu and L. Kawung, "Health as cornerstone of human development: Indonesian perspectives," *Jurnal Pembangunan Manusia*, vol. 12, no. 2, pp. 67–82, 2018.
[2] N. F. Alayda, C. M. Aulia, E. R. Ritonga, and S. H. Purba, "Literature review: Analisis dampak kebijakan Jaminan Kesehatan Nasional (JKN) terhadap akses dan kualitas pelayanan kesehatan," *Jurnal Kolaboratif Sains*, vol. 7, no. 7, pp. 2616–2626, 2024.
[3] BPJS Kesehatan, *Laporan cakupan kepesertaan JKN per 30 Juni 2024*. Jakarta, Indonesia: BPJS Kesehatan, 2024.
[4] Badan Pusat Statistik Provinsi Jawa Timur, *Provinsi Jawa Timur dalam Angka 2025*, vol. 48. Surabaya, Indonesia: BPS Provinsi Jawa Timur, 2025.
[5] Kementerian Koordinator Bidang Pembangunan Manusia dan Kebudayaan, "Universal Health Coverage telah menjangkau 98 persen masyarakat Indonesia," Aug. 9, 2024. [Online]. Available: https://www.kemenkopmk.go.id/universal-health-coverage-telah-menjangkau-98-persenmasyarakat-indonesia
[6] M. S. Hidayat et al., "Decentralisation in Indonesia: The Impact on Local Health Programs," *Kes Mas: Jurnal Fakultas Kesehatan Masyarakat Universitas Ahmad Dahlan*, vol. 12, no. 2, pp. 68–77, 2018, doi: 10.12928/kesmas.v12i2.8906.
[7] D. Fossati, "Is Indonesian Local Government Accountable to the Poor? Evidence from Health Policy Implementation," *Journal of East Asian Studies*, vol. 16, no. 3, pp. 307–330, 2016, doi: 10.1017/jea.2016.17.
[8] I. Ahmad and A. L. Barikha, "Faktor-faktor yang memengaruhi penduduk lanjut usia berobat jalan di Provinsi Jawa Timur," *Jurnal Kependudukan Indonesia*, vol. 17, no. 1, pp. 77–90, 2022.
[9] R. F. Anggraini, "Klasifikasi kabupaten tertinggal di Jawa Timur dengan metode multivariate adaptive regression spline (MARS)," Skripsi, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sunan Ampel, Surabaya, Indonesia, 2021.
[10] M. Ki, J. H. Friedman, and T. Hastie, "LASSO penalized MARS for high dimensional health data analysis," *J. Comput. Graph. Stat.*, vol. 33, no. 1, pp. 89–105, 2024.
[11] R. L. Eubank, *Spline Smoothing and Nonparametric Regression*. New York, NY, USA: Marcel Dekker, 1988.
[12] L. Fahrmeir, T. Kneib, S. Lang, and B. Marx, *Regression: Models, Methods and Applications*, 2nd ed. Berlin, Germany: Springer, 2021.
[13] P. H. C. Eilers and B. D. Marx, *Practical Smoothing: The Joys of P-Splines*. Cambridge, U.K.: Cambridge Univ. Press, 2019.
[14] J. H. Friedman, "Multivariate adaptive regression splines," *Ann. Stat.*, vol. 19, no. 1, pp. 1–67, 1991, doi: 10.1214/aos/1176347963.
[15] P. Taylan and G. W. Weber, "Multivariate adaptive regression splines in mining environmental data," *Optimization*, vol. 56, no. 5–6, pp. 665–681, 2007.
[16] P. Craven and G. Wahba, "Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation," *Numer. Math.*, vol. 31, no. 4, pp. 377–403, 1979, doi: 10.1007/BF01404567.
[17] N. Zakiya, "Model selection in MARS using penalized generalized cross validation," *J. Appl. Stat. Data Sci.*, vol. 6, no. 1, pp. 55–68, 2024.
[18] G. Pratiwi and T. Lubis, "Pengaruh kualitas produk dan harga terhadap kepuasan pelanggan UD Adli di Desa Sukajadi Kecamatan Perbaungan," *Jurnal Bisnis Mahasiswa*, vol. 1, no. 2, pp. 121–134, 2019, doi: 10.60036/jbm.v1i2.
[19] H. Risambessy *et al.*, "Pengujian signifikansi model regresi menggunakan uji F pada penelitian sosial," *Jurnal Penelitian Kuantitatif*, vol. 4, no. 1, pp. 23–32, 2022.
[20] A. Wicaksono *et al.*, "Penggunaan uji F dan uji t dalam analisis regresi: Studi kasus pada data ekonomi," *Jurnal Ekonomi dan Statistika*, vol. 3, no. 2, pp. 89–98, 2014.
[21] S. Mattalunru *et al.*, "Penerapan uji t parsial dalam analisis regresi linier berganda," *Jurnal Statistika Terapan Indonesia*, vol. 6, no. 1, pp. 55–66, 2022.
[22] T. Mulyaningsih, "Pengaruh akses air minum layak dan sanitasi terhadap derajat kesehatan masyarakat di Indonesia," *Jurnal Ilmiah Kesehatan Masyarakat*, vol. 15, no. 2, pp. 120–132, 2023.
[23] A. Jiménez *et al.*, "Water, sanitation and hygiene and the SDGs: A critical review of interlinkages," *Sustainability*, vol. 11, no. 8, pp. 1–18, 2019.
[24] M. Butarbutar *et al.*, "Determinan pemanfaatan pelayanan kesehatan di Indonesia: Peran lingkungan dan perilaku kesehatan," *Jurnal Kesehatan Masyarakat*, vol. 17, no. 2, pp. 101–112, 2022.
[25] H. Maryanto *et al.*, "Akses fasilitas kesehatan dan kepemilikan jaminan kesehatan di Indonesia," *Jurnal Kebijakan Kesehatan Indonesia*, vol. 12, no. 1, pp. 45–58, 2023.
[26] S. N. Agassi, H. S. Kasjono, and M. M. Fauzie, "Hubungan masa kerja, kebiasaan merokok dan olahraga dengan kapasitas vital paru polisi lalu lintas di wilayah kerja Polres Sleman," *Sanitasi: Jurnal Kesehatan Lingkungan*, vol. 9, no. 4, pp. 187–193, 2018.