Available online at http://iorajournal.org/indx.php/orics /index

Operations Research: International Conference Series

e-ISSN: 2722-0974
p-ISSN: 2723-1739

Vol. 5, No. 1, pp. 20-26, 2024

Enhancing Risk Management Strategies: GAM Analysis of Health Insurance Claim Determinants

Azkanul Wahyu[1*], Muhammad Dhafin Qinthar Ramdhani[2]

[1,2]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Padjadjaran, Sumedang, Indonesia 45363*

**Corresponding author email: azkanul20001@mail.unpad.ac.id*

Abstract

Health insurance plays a crucial role in providing financial protection and ensuring access to necessary healthcare services. The awareness of Indonesian society regarding the importance of health insurance continues to grow, as evidenced by a 22% increase in premium income according to AAJI data as of March 2023. Despite the benefits of health insurance, an increasing number of insurance participants raises risks for insurance companies. The Generalized Additive Models (GAM) P-Spline can overcome these problems. The non-linear relationship between claim amount with age, body mass index, and blood pressure can be modelled with GAM P-Spline. The formed GAM model with PIRLS unable to give a clear information of relationship between variables explicitly, but can be seen by the shape of the function of each predictor associated with the link function used.

Keywords: Health Insurance, GAM P-Spline, PIRLS, link function.

1. Introduction

Health insurance plays a crucial role in providing financial protection and ensuring access to necessary healthcare services for policyholders. According to research conducted by Inventure Indonesia in November 2020, 78.7% of respondents agreed that, as a result of the pandemic, having insurance, whether life insurance or health insurance, has become a necessity (Perdana et al., 2022). In Indonesia, public awareness of the importance of health insurance continues to grow. Based on data from the Indonesian Life Insurance Association (AAJI), as of March 2023, there has been a 22% increase in premium income compared to the 2022 period.

The more participants in health insurance, the greater the risk faced by insurance companies. This is because with an increasing number of health insurance participants, the likelihood of health insurance claims also increases. Therefore, insurance companies must set the right premiums to manage the risks they face and ensure they have sufficient reserves to cover potential health insurance claims. In determining premiums, insurance companies must consider factors such as age, gender, health history, and the lifestyle of health insurance participants. By setting the right premiums, insurance companies can manage risks more effectively and provide better protection to their customers.

In this study, the relationship between participant data such as BMI, age, gender, lifestyle, and others will be analyzed in relation to the amount of health insurance claims. A better understanding of this relationship can help insurance companies identify and manage health risks more effectively. Additionally, this research has the potential to improve the accuracy of risk assessment and premium determination, thereby enhancing the financial sustainability of health insurance companies.

This research chooses to use a generalized additive model (GAM) as the most suitable statistical analysis method to formulate the complex relationship between participant data variables and the amount of health insurance claims. GAM has advantages in handling nonlinear patterns, complex interactions between variables, and flexibility in modeling heteroskedasticity. In this case, the use of GAM allows for a deeper understanding of the information from the data and facilitates a more accurate understanding of patterns that may be overlooked by more traditional linear approaches.

By understanding these relationships, insurance companies can adjust premiums accurately, optimize marketing strategies for more targeted market segments, and provide better protection for policyholders.

2. Literature Review

2.1. Monte Carlo

The Monte Carlo method is an algorithm for solving problems that utilizes random processes or randomization in the form of probability simulation. This approach employs probability by generating random numbers for estimation. The Monte Carlo simulation process is carried out through a series of iterations, and the number of iterations is determined by generating as many random numbers as needed for the simulation (Halton et al., 2005).

The following are the steps of the Monte Carlo simulation (Zakhary et al., 2011):

1. Data group
2. Calculation of relative frequency distribution
3. Calculation of cumulative relative frequency distribution
4. Determination of random number intervals
5. Calculation and generation of random numbers
6. Monte Carlo simulation

The Monte Carlo method utilizes existing sampled data (historical data) with known data distributions. This approach relies on the strong law of large numbers, indicating that the more random variables used, the closer the approach will be to the exact value (Seifoddini et al., 2021).

2.2. *Generalized Additive Models* (GAM)

Generalized Additive Models (GAM) eliminate the assumption of some function to obtain a model that shows the relationship between the response variable ant its predictor variables (Sibhatu et al., 2022). The general form of the GAM model can be formulated as follows:

$$Y = \beta_0 + \sum_{j=1}^{p} f_j(X_{ij}) \tag{1}$$

The equation can be rewritten as model below:

$$Y = \beta_0 + f_1(X_{i1}) + f_2(X_{i2}) + \ldots + f_j(X_{ij});\ i = 1, 2, \ldots, n$$

with

Y : Response variable
β_0 : Intercept coefficient
X_j : Predictor variables for j
p : Number of predictor variables
f_j : Smoothing function for predictor variables j

Parameter estimation in GAM can be done with Penalized Iteratively Re-weighted Least Squares (PIRLS). In addition, smoothing function estimation based on cross validation criteria, such as Generalized Cross Validation (GCV) (Wood, 2006).

2.3. Penalized Spline B-Spline Basis

Splines are segmented polynomials in which each segment is joined together with knots. This segmented nature makes splines have good flexibility (Hidayat, 2019). B-Spline is polynomial functions that are segmented on the interval x formed by the knots and estimated for each segment at a certain polynomial degree (de Boor, 2001). The basis of k-th B-Spline with degree v is as many as u knots $t_0, t_1, \ldots, t_u$ for $k = 1, \ldots, v + u$. The basis can be defined recursively as follows:

$$B_k(x; v) = \frac{x - t_k}{t_{k+v-1} - t_k} B_k(x; v-1) - \frac{x - t_{k+v}}{t_{k+v} - t_{k+1}} B_{k+1}(x; v-1)$$

with

$$B_k(x; v = 0) = \begin{cases} 1, & \text{if } t_k \le x \le t_{k+1} \\ 0, & \text{for other } x \end{cases}$$

The B-Spline function uses too many knots that tend to overfit and require a penalty on the adjacent coefficients of the B-Spline (Eilers & Marx, 1996). P-Spline is a smoother consisting of several basis functions of B-Splines that do not exceed the number of observations and whose regression coefficients are penalized.

2.4. Generalized Additive Models P-Splines

Equation (1) can be rewritten as follows:

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} f_j(X_{ij}) \tag{2}$$

with

$g(\mu_i)$: Link function
Y : Response variable
β_0 : Intercept coefficient
X_j : Predictor variables for j
p : Number of predictor variables
f_j : Smoothing function for predictor variables j

Generalized Additive Models P-Splines are GAMs that contain additive functions using P-Spline as the smoothing function. Model (2) can be written as follows:

$$g(\mu) = \eta = \beta_0 + \sum_{l=1}^{p_1^*} a_{1l} B_{1l}(x_1; v_1) + \sum_{l=1}^{p_2^*} a_{2l} B_{2l}(x_2; v_2) + \cdots + \sum_{l=1}^{p_j^*} a_{jl} B_{jl}(x_j; v_j) \tag{3}$$

The model can be rewrite as follows:

$$g(\mu) = \boldsymbol{B}\boldsymbol{a} = \eta$$

with

$\boldsymbol{B}$: Regressor matrix for the B-Spline basis
$\boldsymbol{a}$: Coefficient vector
$\boldsymbol{B}_j$: Regressor matrix for the j-th predictor B-Spline basis

$$\boldsymbol{B}_j = \begin{bmatrix} b_1(x_{1j}) & \cdots & b_l(x_{1j}) \\ \vdots & \ddots & \vdots \\ b_1(x_{nj}) & \cdots & b_l(x_{nj}) \end{bmatrix}$$

$\boldsymbol{a}_j$: Vektor koefisien prediktor ke-j

$$\boldsymbol{a}_j = \begin{pmatrix} a_{j1} \\ a_{j_2} \\ \vdots \\ a_{jl} \end{pmatrix}$$

The unknown parameter estimates in GAM can be obtained by maximizing the likelihood function as follows:

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^T \boldsymbol{S}\boldsymbol{\beta}$$

with $\boldsymbol{S} = \sum_j \lambda_j \boldsymbol{S}_j$ and λ_j is the smoothing parameter whose value is known (Wood, 2006). According to Marx & Eilers (1998), the estimation form can be obtained as follows:

$$\boldsymbol{\beta}^{[k+1]} = \left(\boldsymbol{B}^T \boldsymbol{W}^{[k]} \boldsymbol{B} + \boldsymbol{S}\right)^{-1} \boldsymbol{B}^T \boldsymbol{W}^{[k]} \boldsymbol{z}^{[k]}$$

2.5. Smoothing Parameter Selection

The selection of smoothing parameters can be done using Generalized Cross Validation (GCV). According to Eubank (1999), the GCV value is obtained from the following formula:

$$GCV = \frac{1}{n}\frac{\sum_{i=1}^{n}\left(y_i - \hat{f}_i\right)^2}{\left[1 - \frac{tr\left(\boldsymbol{A}(\lambda_j)\right)}{n}\right]}$$

with

y_i : The actual value of the i-th observation of the response variable
$\hat{f}$: Predicted value of the i-th observation of the response variable
n : Number of observations
$tr(\boldsymbol{A}_\lambda)$: The sum of the main diagonal element matrix $\boldsymbol{A}_\lambda = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{D})^{-1}\boldsymbol{X}^T$

3. Materials and Methods

3.1. Materials

The data used in this research is simulation data obtained from the Monte Carlo simulation. There are 100 data points consisting of three independent variables: age, body mass index, and blood pressure. The dependent variable used is the claim amount.

3.2. Methods

In additive models, the parameters are estimated by penalized least squares. However, in Generalized Additive Models P-Splines, the parameters can be estimated using Penalized Iteratively Re-weighted Least Squares (PIRLS) to achieve optimal results. In general, the maximum penalized likelihood estimation can be done through the following iteration process:

1. Determine the initial estimated value of $\boldsymbol{\beta}^{[k]}$
2. Determine the initial estimated value of $\boldsymbol{\eta}^{[k]} = \boldsymbol{X}\boldsymbol{\beta}^{[k]}$
3. Determine the initial estimated value of $\boldsymbol{\mu}^{[k]} = g^{-1}(\boldsymbol{\eta}^{[k]})$
4. Calculating the weight matrix $\boldsymbol{W}^{[k]}$
5. Calculating the pseudodata $\boldsymbol{z}^{[k]}$
6. Calculating the value of $\beta^{[k+1]}$ as follows:
$$\boldsymbol{\beta}^{[k+1]} = \left(\boldsymbol{B}^T\boldsymbol{W}^{[k]}\boldsymbol{B} + \boldsymbol{S}\right)^{-1}\boldsymbol{B}^T\boldsymbol{W}^{[k]}\boldsymbol{z}^{[k]}$$
k increment.
7. Repeat the steps starting from $k = 0$ until it converges, i.e.
$$\left\|\boldsymbol{\beta}^{[k+1]} - \boldsymbol{\beta}^{[k]}\right\| < \delta$$
with δ is a very small number (Wood, 2006).

4. Results and Discussion

In this study, there are 100 data points consisting of three independent variables such as age, body mass index, and blood pressure. The dependent variable used is the claim amount. Before modeling the claim amount against each predictor variable, author will first identify the relationship pattern between each variable response and predictor using a scatterplot.

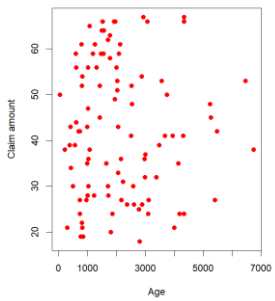

Figure 1: Scatterplot between claim amount and age

Figure 1 shows that the relationship between claim amount and age is randomly distributed. The relationship pattern cannot be specified parametrically.

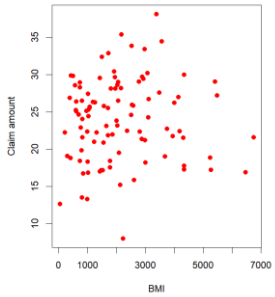

Figure 2: Scatterplot between claim amount and body mass index

Figure 2 shows that the relationship between claim amount and body mass index is randomly distributed. The relationship pattern cannot be specified parametrically.

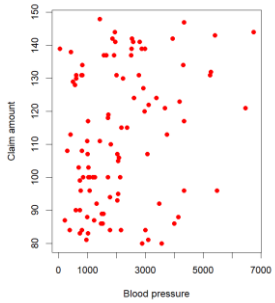

Figure 3: Scatterplot between claim amount and blood pressure

Figure 3 shows that the relationship between claim amount and blood pressure is randomly distributed. The relationship pattern cannot be specified parametrically.

Therefore, modeling the amount of claims with influencing factors will use a nonparametric approach, such as Generalized Additive Models P-Spline. The number of knots and degrees used is 2, so the number of bases will be as many as $u + v = 2 + 2 = 4$. The equation can be written in the form as follows:

$$E(Y|X) = \frac{1}{\beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3})} \tag{4}$$

The next step is done with PIRLS using the optimal smoothing parameters that have been obtained on each predictor variable to obtain the estimated value of the model parameters. The parameter estimation values of the model are as follows:

Table 1: Results of estimated parameter model

Variable	**Estimated Value of Parameter** (a_{jl})
Intercept	0.0004812269
X_1	$a_{11} = 0.000009817491$
	$a_{12} = 0.00002909947$
	$a_{13} = 0.00003130721$
X_2	$a_{21} = -0.00001125276$
	$a_{22} = -0.00003690625$
	$a_{23} = -0.00003682414$

	$a_{31} = -0.0001282068$
X_3	$a_{32} = -0.0003634469$
	$a_{33} = -0.0003993129$

So, that the Generalized Additive Models P-Spline model (4) is obtained as below:

$$E(Y|X) = 1/0.0004812269 + 0.000009817491\big(B_1(x_1;\ 2)\big) + 0.00002909947\big(B_2(x_1;2)\big) + 0.00003130721\big(B_3(x_1;2)\big) - 0.00001125276\big(B_1(x_2;2)\big) - 0.00003690625\big(B_2(x_2;2)\big) - 0.00003682414\big(B_3(x_2;2)\big) - 0.0001282068\big(B_1(x_3;2)\big) - 0.0003634469\big(B_2(x_3;2)\big) - 0.0003993129\big(B_3(x_3;2)\big)$$

The relationship of each variable from the model above can be seen by the shape of the function of each predictor associated with the link function used.

5. Conclusion

Based on the results of the discussion, it can be concluded that modeling between the amount of claims with age, BMI, and blood pressure can be done with a nonparametric approach, such as GAM. This can be seen from the scatter plot between each predictor variable and the response variable which does not follow the shape of a particular curve function. The Generalized Additive Models P-Spline model formed in the equation cannot be explained explicitly for the increase of each predictor.

References

de Boor, C. (2001). *A Practical Guide to Splines, Revised Edition*. Springer.

Eilers, P. H. C., & Marx, B. D. (1996). Flexible Smoothing with B-splines and Penalties. *Statistical Science*, *11*(2), 89–102. http://www.jstor.org/stable/2246049

Eubank, R. (1999). *Nonparametric Regression and Spline Smoothing* (2nd ed., Vol. 157).

Marx, B. D., & Eilers, P. H. C. (1998). Direct generalized additive modeling with penalized likelihood. In *Computational Statistics & Data Analysis* (Vol. 28).

Halton, J. H. (2005). Quasi-Probability: Why quasi-Monte-Carlo methods are statistically valid and how their errors can be estimated statistically. *Monte Carlo Methods & Applications*, *11*(3).

Hidayat, R., Budiantara, I. N., Otok, B. W., & Ratnasari, V. (2019). An extended model of penalized spline with the addition of kernel functions in nonparametric regression model. *Applied Mathematics and Information Sciences*, *13*(3), 453-460.

Perdana, N. R., Adhasari, G., & Mahadewi, E. P. (2022). Challenges and Implementation of Universal Health Coverage Program in Indonesia. *International Journal of Health and Pharmaceutical (IJHP)*, *2*(3), 589-596.

Seifoddini, J. (2021). Stock option pricing by augmented monte-carlo simulation models. *Advances in Mathematical Finance and Applications*, *6*(4), 1-22.

Sibhatu, K. T., Steinhübel, L., Siregar, H., Qaim, M., & Wollni, M. (2022). Spatial heterogeneity in smallholder oil palm production. *Forest Policy and Economics*, *139*, 102731.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

Zakhary, A., Atiya, A. F., El-Shishiny, H., & Gayar, N. E. (2011). Forecasting hotel arrivals and occupancy using Monte Carlo simulation. *Journal of Revenue and Pricing Management*, *10*, 344-366.