Pak. J. Statist.
2025 Vol. 41(1), 17-34

GENERALIZED ADDITIVE MODEL FOR VEHICLE INSURANCE PREMIUM CALCULATION BASED ON MILEAGE AND CONTRACT DURATION

Fevi Novkaniza[§], **Alfina Wijaya** and **Sindy Devila**
Department of Mathematics, Faculty of Mathematics and Natural Sciences
University of Indonesia, Kampus FMIPA UI, Depok 16424, Indonesia
[§] Corresponding author

ABSTRACT

Premiums are sums of money stipulated by insurance or reinsurance companies and agreed upon by policyholders; they are payable according to the terms of insurance or reinsurance agreements. When calculating premium rates, insurance companies typically consider the risk exposure of the insured vehicle, which is crucial in estimating the number of claims. While the duration of the insurance contract is a primary factor in assessing risk exposure, other elements, such as the distance traveled, also significantly impact accident risk. This study aims to enhance the computation of risk exposure for vehicles by incorporating both the distance traveled and the duration of the insurance contract. The objective is to evaluate the combined effect of mileage and insurance contract duration on the number of claims via a generalized additive model (GAM). The GAM is chosen for its ability to capture potential nonlinear relationships between covariates and response variables. In this research, the GAM is constructed via cubic splines, and model coefficients are estimated via the penalized iteratively reweighted least squares (PIRLS) method. Upon estimating the model coefficients, the GAM is used to predict claim frequency, which can subsequently inform the relativity of premium rates compared with a reference premium. This methodology is then applied to vehicle insurance claim data to establish more accurate premium rates, considering both distance traveled and contract duration.

KEYWORDS

Cubic spline; generalized cross validation; penalized iteratively reweighted least squares; pay-as-you drive insurance; reference premium.

2020 AMS Subject Classification: 62P05

1. INTRODUCTION

A premium is a sum of money stipulated by an insurance company or reinsurance company and agreed upon by the policyholder to be payable based on the insurance agreement or reinsurance agreement or a sum of money stipulated based on the provisions of laws and regulations that govern compulsory insurance programs for benefits [1]. In general, vehicle insurance companies assume that the risk exposure of an insured individual is proportional. For example, consider two cars, car a and car b, with the same brand, type,

year of production, and engine capacity. Car a is insured for a duration of 6 months, whereas car b is insured for 1 year. The insurance company perceives the risk exposure of car b to be greater than that of car a. However, in practice, the risk exposure associated with vehicle usage is influenced not only by the duration of the insurance policy but also by other factors, such as the distance traveled. Therefore, this paper calculates the risk exposure of vehicles by considering the distance traveled and the duration of the insurance contract.

Several studies have indicated a significant relationship between the distance traveled in kilometers and the risk of car accidents [2,3,4,5]. Litman concluded that there is a positive and nonlinear relationship between the number of accidents and the distance traveled within one year. The curve representing this relationship shows a positive but non-constant association, increasing during the first 35,000 kilometers but decreasing in the interval between 35,000 and 37,500 kilometers. The curve subsequently rises again but not as sharply as it initially increases. Moreover, the relationship between the distance traveled and the number of accidents in a year is not proportional. For example, a vehicle covering an annual distance between 25,000- and 30,000-kilometers experiences six times more journeys than a vehicle covering less than 5,000 kilometers. However, the accident rate is only approximately 2.4 times higher [2].

Boucher, Côté, and Guillen undertook a study to examine the joint effects of distance traveled and duration as covariates on claim frequency via generalized additive models (GAMs) [6]. GAMs are a class of generalized linear models (GLMs) [7] that incorporate smoothing functions of covariates into linear predictors, allowing for more flexible modeling of nonlinear relationships between the response variable and covariates [8,9]. In the GAM, smoothing functions are constructed via various techniques, such as kernel smoothing, smoothing splines, and locally estimated scatterplot smoothing (Loess). One common smoothing technique is splines, which employ continuous piecewise polynomial functions at specified knot points [10]. Cubic splines, used by Boucher, Côté, and Guillen, are cubic polynomial basis functions that are continuous up to the second derivative at the knots, making them suitable for capturing nonlinear relationships in GAMs [6,8].

This paper discusses the development of a GAM on the basis of research by Boucher, Côté, and Guillen [6], who explored the relationships among distance traveled, insurance contract duration, and claim frequency. On the basis of the GAM, it is possible to estimate the smoothing function for each covariate to determine the premium price relativity on the basis of the policyholder's risk profile. Each policyholder's risk profile is represented as a rating variable. The resulting premium rates reflect the relative premium price of policyholders compared with a reference premium. For the analysis in this study, insurance data from research by [12] on automobile insurance are utilized. The dataset comprises information from 10,000 policies, including covariates such as distance traveled, insurance contract duration, claim frequency, and the duration of insurance coverage during observation periods.

2. MATERIALS AND METHODS

2.1 Pay-As-You Drive Pricing System

The pay-as-you drive (PAYD) pricing system is an insurance premium pricing structure for vehicle insurance that is based on the amount of time a vehicle is driven during the insurance contract period [2,12]. The exposure unit is changed from vehicle-year to vehicle-mile, vehicle-kilometer, or vehicle-minute. The system is supported by the advent of new technology, namely, GPS, which can be installed in vehicles to generate accurate data on the distance traveled by the insured. This premium pricing structure is typically offered as an option for policyholders, allowing them to choose between the current insurance premium structure or PAYD. The PAYD pricing structure is designed to enable more accurate insurance premiums, as premiums vary on the basis of the distance traveled. Other rating factors can also be included so that lower-risk drivers pay lower premiums than higher-risk drivers do.

2.2 Reference Premium Rating Factors and Key Ratios

For each policy, the premium is determined by the values of several variables known as rating factors or rating variables, which represent the characteristics or traits of the policyholder. The premium value applied under the assumption that there is no influence from the rating factors is called the base premium or reference premium. The following are examples of characteristics or variables based on categories [13]:

a. Policyholder: Age or gender of an individual, type of business of a company, etc.
b. Insured Object: age or model of the car, type of building, etc.
c. Geographic region: per capita income or population density of the policyholder's residence, etc.

Moreover, the key ratio denoted by Y is the ratio between the response variable (X) and the exposure (w) [13] with the following formula:

$$Y = \frac{X}{w} \tag{1}$$

Suppose that the variable X denotes the number of claims from policyholders in a period (duration) w. The key ratio is the claim frequency, which is the ratio or average number of claims in a period. Table 1 contains several other important and frequently used key ratios.

Table 1
Important key ratios

Exposure (w)	**Response Variable (X)**	**Key Ratio $\left(Y = \frac{X}{w}\right)$**
Duration	Number of claims	Claims frequency
Duration	Claim cost	Pure premium
Number of claims	Claim cost	(Average) Claim severity
Earned premium	Claim cost	Loss ratio
Number of claims	Number of large claims	Proportion of large claims

2.3 Chi-Squared Test

The chi-square test [14] is a statistical test used to evaluate the fit between the observed distribution in the data and the expected distribution. This test is commonly used to test whether data come from a specific distribution, such as the normal, binomial, and Poisson distributions. The hypotheses tested in the chi-square test are as follows:

H_0: The distribution is suitable for modeling the data (there is no significant difference between the observed and expected distributions).

H_1: The distribution is not suitable for modeling the data (there is a significant difference between the observed and expected distributions).

In the goodness-of-fit test via the chi-square test, the chi-square test statistic is calculated via the following equation:

$$X^2_{tested} = \sum_{i=1}^{k} \frac{(e_i - o_i)^2}{e_i} \tag{2}$$

where o_i represents the number of observations or values in the i-th category, e_i represents the expected count for each value in the i-th category, and k represents the number of categories of observation values. The expected count or frequency can be calculated via the following formula:

$$e_i = nPr(Y = j) \tag{3}$$

where n is the number of observations and where $Pr(Y = j)$ is the probability density function at j from the model distribution, with $j = 0,1, \dots p$. The decision rule for this test depends on the critical value derived from the chi-square distribution with the corresponding degrees of freedom and significance level α. The degrees of freedom in the goodness-of-fit test are determined by the number of categories (k) and the number of estimated parameters from the data, i.e., $df = k - l - 1$. The decision-making rule is as follows: if $X^2_{critical} > X^2_{tested}$, then H_0 is not rejected; otherwise, if $X^2_{critical} < X^2_{tested}$, then H_0 is rejected.

2.4 Generalized Additive Model

The generalized additive model (GAM) is an extension of the generalized linear model (GLM) that involves the summation of smoothing functions of the covariates on the linear predictor [8,9]. Several smoothing techniques can be used, such as kernel smoothing, smoothing splines, and the locally estimated scatterplot smoothing (Loess) method. Generally, for each observation i, with $i = 1, \dots, n$, the structure of the GAM is as follows:

$$\eta_i = g(\mu_i) = s_0 + \sum_{j=1}^{p} f_j(x_{j(i)}) \tag{4}$$

where x_j denotes the covariate with $j = 1, \dots, p$ and where x_{ji} represents the value of covariate x_j for the i −th observation. The function $f_j(x_{j(i)})$ is a smoothing function for

covariate x_j. In equation (4), $\mu_i \equiv E(Y_i)$ is the expected value of the response variable Y_i, where the distribution of the response variable is a member of the exponential family (μ_i, ϕ).

2.5 Cubic Spline

Suppose a covariate X with observation values $x_{(i)}, i = 1, \dots, n$ and a response variable Y with observation values $y_{(i)}, i = 1, \dots n$, resulting in pairs of points $(x_{(i)}, y_{(i)})$. The values of the covariate X are sorted in ascending order, with $x_j \ , j = 1, \dots, n$ denoting the j-th ordered statistic of $x_{(i)}$ within the interval $[a, b]$. A cubic spline f that interpolates the data points$\{(x_0, v(x_0)), (x_1, v(x_1)), \dots, (x_n, v(x_n))\}$ with $v(x_j) = y_j$ is defined as follows.

Definition 1.

Given a function g defined on $[a, b]$ and a set of knots $a = x_0 < x_1 < \cdots < x_n = b$, a cubic spline interpolant f for g is a function that satisfies the following conditions [15]:

(a) $f(x)$ is a cubic polynomial, denoted as $f_j(x)$ on the subinterval $[x_j, x_{j+1}]$ for each $j = 0,1, \dots, n-1$;

(b) $f_j(x_j) = v(x_j)$ and $f_j(x_{j+1}) = v(x_{j+1}), \forall j = 0,1, \dots, n-1$;

(c) $f_{j+1}(x_{j+1}) = f_j(x_{j+1}) , \forall j = 0,1, \dots, n-2$;

(d) $f'_{j+1}(x_{j+1}) = f'_j(x_{j+1}) , \forall j = 0,1, \dots, n-2$;

(e) $f''_{j+1}(x_{j+1}) = f''_j(x_{j+1}) , \forall j = 0,1, \dots, n-2$;

(f) One of the following sets of boundary conditions is satisfied:
(i) $f''(x_0) = f''(x_n) = 0$, (natural boundary)
(ii) $f'(x_0) = v'(x_0)$ dan $f'(x_n) = v'(x_n)$ (clamped boundary)

We use a natural boundary because it produces the smoothest interpolation [9]. In statistical work, y_i is usually measured with noise. Therefore, it is better to treat $f(x_i)$ as a parameter to be estimated rather than assume that $f(x_i) = y_i$ and interpolating it. The estimation of $f(x_i)$ is performed by minimizing

$$\eta_i = g(\mu_i) = s_0 + \sum_{j=1}^{p} f_j(x_{j(i)}) \tag{5}$$

where the first term is the residual sum of squares (RSS) and the second term is a penalty, with λ denotes a smoothing parameter that can be adjusted and is nonnegative. If $\lambda = 0$, then no penalty is applied. On the other hand, if $\lambda = \infty$, the resulting function approaches linearity [16].

2.6 Iteratively Reweighted Least Squares (IRLS)

Iteratively reweighted least squares (IRLS) [8,9] is a method that can be used to estimate generalized linear model (GLM) parameters. Iterate the steps below until the sequence $\widehat{\boldsymbol{\beta}}^{[t]}$ converges.

1. Estimate the values of $\hat{\mu}_i^{[0]} = y_i^{[0]} + \delta_i$ and $\hat{\eta}_i^{[0]} = g(\mu_i^{[0]})$, where δ_i is typically zero. Then, the following two steps are repeated until convergence.
2. Calculate the pseudodata $z_i = g'(\hat{\mu}_i)\left(y_i - \hat{\mu}_i^{[t]}\right) + \hat{\eta}_i$ and the iterative weights $w_i = \frac{1}{V\left(\hat{\mu}_i^{[t]}\right)g'\left(\mu_i^{[t]}\right)^2}$
3. Minimize the objective function:

$$\sum_{i=1}^{I} w_i (z_i - \boldsymbol{X}_i\boldsymbol{\beta})^2$$

Then, update $\widehat{\boldsymbol{\eta}}^{[t+1]} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{[t]}$ and $\hat{\mu}_i^{[t+1]} = g^{-1}\left(\hat{\eta}_i^{[t+1]}\right)$.

Convergence can be based on the change in deviation from iteration to iteration. Iteration stops when the deviation approaches zero or by testing if the derivative of the log-likelihood is sufficiently close to zero.

2.7 Penalized Iteratively Reweighted Least Squares (PIRLS)

The PIRLS method is a modification of the iterative reweighted least squares (IRLS) method [9]. In PIRLS, a penalty term is considered in the objective function. Suppose that there are p covariates with I observations, the model matrix of p covariates is denoted as $\boldsymbol{X} = \begin{bmatrix} X_1 & X_2 & \dots & X_p \end{bmatrix}$, where each X_j represent the j-th covariate where $j = 1,2,\dots,p$, a column vector of length I, and the vector of parameters or model coefficients is $\boldsymbol{\beta} = [\beta_1, \dots, \beta_p]$. The objective function of PIRLS that needs to be maximized is shown by the following equation.

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2\phi}\sum_{e=1}^{p} \lambda_e \boldsymbol{\beta}_e^{\boldsymbol{T}} \boldsymbol{S}_e \boldsymbol{\beta}_e \tag{6}$$

where the first term is the log-likelihood and the second term is a penalty, with **S** being the penalty matrix and λ denoting a nonnegative adjustable smoothing parameter. By adapting the objective function of IRLS as described in section 2.6, the following are the iteration steps for PIRLS. The following iterative steps are performed until the sequence of $\hat{\beta}^{[k]}$ converges.

1. Estimate the values of $\mu_i^{[0]} = y_i^{[0]} + \delta_i \; and \; \eta_i^{[0]} = g\left(\mu_i^{[0]}\right)$, where δ_i is generally 0. The following two steps are then repeated until convergence is reached.
2. Calculate the pseudodata $z_i = g'\left(\hat{\mu}_i^{[k]}\right)(y\_i - \hat{\mu}_i^{[k]}) + \hat{\eta}_i^{[k]}$ and iteration weights $w_i = 1/\left\{g'\left(\hat{\mu}_i^{[k]}\right)^2 V\left(\hat{\mu}_i^{[k]}\right)\right\}$

We find $\widehat{\boldsymbol{\beta}}^{[k]}$ to minimize the weighted least squares objective function and penalty.

$$S_p = \left\|\boldsymbol{z}^{[\boldsymbol{k}]} - \boldsymbol{X}\boldsymbol{\beta}\right\|_W^2 + \sum_{e=1}^{p} \lambda_e \boldsymbol{\beta}_e^T \boldsymbol{S}_e \boldsymbol{\beta}_e \tag{7}$$

where $\|a\|_W^2 = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{a}$, the value of λ can be chosen arbitrarily with $0 \leq \lambda < \infty$. Then, update $\widehat{\boldsymbol{\eta}}^{[\boldsymbol{k+1}]} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{[\boldsymbol{k}]}$ and $\hat{\mu}_i^{[k+1]} = g^{-1}\left(\hat{\eta}_i^{[k+1]}\right)$. Convergence can be based on observing the change in deviance from iteration to iteration. Iterations stop when the deviance approaches zero or by testing whether the derivative of the log-likelihood is close enough to zero.

2.8 Generalized Cross-Validation

Generalized cross-validation (GCV) is a modification of cross-validation that replaces A_{ii} with its mean value, which is $tr(\boldsymbol{A})/n$, resulting in the following GCV formula [9]:

$$\mathcal{V}_g(\boldsymbol{\lambda}) = \frac{I\sum_{i=1}^{I}(y_i - \hat{\mu}_i)^2}{[I - tr(\boldsymbol{A})]^2} \tag{8}$$

where:

- I is the number of observations,
- X is the $(I \times p)$ covariate matrix,
- p is the number of covariates,
- y_i is the actual observed value at the $i-th$ observation,
- $\hat{\mu}_i$ is the predicted value from the model at the $i-th$ observation,
- $A = [X(X^T X)^{-1} X^T]$

For cases where the response variable follows a non-normal distribution, the residual sum of squares is replaced with the residual deviance, so the GCV formula can be written as:

$$\mathcal{V}_g(\boldsymbol{\lambda}) = \frac{I \times D(\widehat{\boldsymbol{\beta}})}{(I - \gamma\tau)^2} \tag{9}$$

where $D(\widehat{\boldsymbol{\beta}})$ is the residual deviance, τ is the effective degree of freedom of the model, and γ is usually 1.

3. GENERALIZED ADDITIVE MODEL

The construction process of the generalized additive model (GAM) was based on the following three assumptions:

1. The response variable, which is the number of claims, is assumed to be a discrete random variable distributed according to a Poisson distribution. These variables are independent and identically distributed (i.i.d.).
2. The covariates are assumed to be independent of each other.
3. The link function used is the logarithmic link function because the response variable is assumed to follow a Poisson distribution.

3.1 Generalized Additive Model with One Covariate

Let Y_i denote a random variable representing the number of claims from an insured individual i during one year of observation, where $i = 1, \dots, I$. We assume that Y_i follows a Poisson distribution with the following expectation:

$$E[Y_i] = \mu_i \tag{10}$$

Suppose that we are given a covariate X for the i-th insured individual with observed value x_i, where $i = 1, \dots, I$. The representation of the GAM with one covariate can be expressed by the following equation:

$$g(\mu_i) = intercept + f(x_i) \tag{11}$$

where $g(\cdot)$ is the link function and $f(x)$ is the smoothing function of covariate X constructed via a cubic spline. This formulation establishes the framework for analyzing the relationship between covariate X and the expected number of claims μ_i via the generalized additive model (GAM) approach. The link function $g(\cdot)$ transforms the expected value of the response variable (number of claims) to ensure that it aligns with the linear predictor, which includes the intercept and the smoothed effect of the covariate X. The smoothing function $f(x)$ captures potential nonlinear relationships between the covariate X and the response variable within the GAM framework, employing cubic splines to achieve flexibility and accuracy in modeling.

3.2 Determination of the Number and Location of Knots

Before splines are constructed, it is necessary to first determine the number and location of knots in the point set. According to the findings of Stone [17], several knots greater than 5 is rarely required in natural cubic spline models. The main choice of the number of knots that can be used is between 3, 4, or 5. After the number of knots is determined, it is also necessary to determine the locations of the knots. According to the authors in [18,19], knots can be placed at quantile points (percentiles) of the covariate data distribution. This approach is good enough for most datasets, as it ensures that there are enough data points for each interval. Suppose that we are given a covariate X from the i-th insured individual with observed values x_i, where $i= 1, \dots, I$. Then, values of x that appear more than once in the set of observations are considered only once. Let these values be sorted in ascending order, denoted as $x^*_{(r)}$, for $r = 1, \dots, I^*$, where $I * \leq I$, representing the r-th ordered statistic of unique values $x_{(i)}$ within the interval $[a, b]$. Next, the interval $[a, b]$ is divided into several subintervals based on quantiles with $a = x_0 < x_1 < \cdots < x_n = b$ knots, where $(n + 1)$ is the number of knots predetermined initially. Thus, the sequence of knots is denoted by $\{x_0, x_1, x_2, \dots, x_{n-1}, x_n\}$. These knots are then used to construct a cubic spline.

3.3 Construction of the Covariate Smoothing Function via Cubic Splines

This section discusses the process of constructing the cubic spline function f by applying the conditions outlined in Definition 1. After all these conditions are applied, the cubic spline function can be expressed in equation form as follows:

$$f_j(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)\delta_j + c_j^+(x)\delta_{j+1} \tag{12}$$

for $x_j \le x \le x_{j+1}$, where:

$$\beta_j = f_j(x_j)$$

$$\delta_j = f_j''(x_j)$$

$$h_j = x_{j+1} - x_j$$

$$a_j^-(x) = (x_{j+1} - x)/h_j$$

$$a_j^+(x) = (x - x_j)/h_j$$

$$c_j^-(x) = \left[\frac{(x_{j+1}-x)^3}{h_j} - h_j(x_{j+1} - x)\right]/6$$

$$c_j^+(x) = \left[\frac{(x-x_j)^3}{h_j} - h_j(x - x_j)\right]/6.$$

Cubic splines $f(x)$ for the interval $[a, b]$ can be represented by the following piecewise function:

$$f(x) = \begin{cases} f_0(x), & x_0 \le x < x_1 \\ f_1(x), & x_1 \le x < x_2 \\ f_2(x), & x_2 \le x < x_3 \\ \quad\vdots & \\ f_{n-1}(x), & x_{n-1} \le x \le x_n \end{cases} \tag{13}$$

Furthermore, adjusting the condition that the second derivative of the smoothing function must be continuous at each knot implies that

$$\frac{h_{j-1}}{6}\delta_{j-1} + \frac{(h_{j-1} + h_j)}{3}\delta_j + \frac{h_j}{6}\delta_{j+1} = \left(\frac{1}{h_{j-1}}\right)\beta_{j-1} + \left(-\frac{1}{h_{j-1}} - \frac{1}{h_j}\right)\beta_j + \frac{1}{h_j}\beta_{j+1}, \tag{14}$$

$\forall j, j = 1, \dots, n-1$. Suppose that the vector $\boldsymbol{\delta}^-$ is defined as $\boldsymbol{\delta}^- = (\delta_1, \delta_2, \dots, \delta_{n-1})^{\mathrm{T}}$ with $\delta_0 = \delta_n = 0$ and that $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_n)^{\mathrm{T}}$ is the vector of unknown coefficients. Then, the equation above can be expressed in matrix form as

$$\mathbf{A}\boldsymbol{\delta}^- = \mathbf{B}\boldsymbol{\beta} \tag{15}$$

where $\mathbf{A}$ is an $(n-1) \times (n-1)$ matrix and $\mathbf{B}$ is an $(n-1) \times (n+1)$ matrix with the following elements:

$$\mathbf{A} = \begin{bmatrix} \frac{(h_0+h_1)}{3} & \frac{h_1}{6} & 0 & 0 & \cdots & 0 & 0 \\ \frac{h_1}{6} & \frac{(h_1+h_2)}{3} & \frac{h_2}{6} & 0 & \cdots & 0 & 0 \\ 0 & \frac{h_2}{6} & \frac{(h_2+h_3)}{3} & \frac{h_3}{6} & \cdots & 0 & 0 \\ 0 & 0 & \frac{h_3}{6} & \frac{(h_3+h_4)}{3} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \frac{(h_{n-3}+h_{n-2})}{3} & \frac{h_{n-2}}{6} \\ 0 & 0 & 0 & 0 & \cdots & \frac{h_{n-2}}{6} & \frac{(h_{n-2}+h_{n-1})}{3} \end{bmatrix} \quad (16)$$

$$\mathbf{B} = \begin{bmatrix} \frac{1}{h_0} & -\frac{1}{h_0}-\frac{1}{h_1} & \frac{1}{h_1} & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{h_1} & -\frac{1}{h_1}-\frac{1}{h_2} & \frac{1}{h_2} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \frac{1}{h_{n-3}} & -\frac{1}{h_{n-3}}-\frac{1}{h_{n-2}} & \frac{1}{h_{n-2}} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{h_{n-2}} & -\frac{1}{h_{n-2}}-\frac{1}{h_{n-1}} & \frac{1}{h_{n-1}} \end{bmatrix} \quad (17)$$

Let us denote $\mathbf{F}^- = \mathbf{A^{-1}B}$ as a $(n-1)\times(n+1)$ matrix and $\mathbf{F} = \begin{bmatrix} \mathbf{0} \\ \mathbf{F}^- \\ \mathbf{0} \end{bmatrix}$ as a $(n+1)\times(n+1)$ matrix, where $\mathbf{0} = (0,0,\dots,0)$ is an $(n+1)$ row vector; then, $\boldsymbol{\delta} = \mathbf{F}\boldsymbol{\beta}$. Therefore, the cubic spline for all $j, j = 1,\dots,n-1$ can be restated as follows:

$$f_j(x_{(i)}) = a_j^-(x_{(i)})\beta_j + a_j^+(x_{(i)})\beta_{j+1} + c_j^-(x_{(i)})\mathbf{F}_j\boldsymbol{\beta} + c_j^+(x_{(i)})\mathbf{F}_{j+1}\boldsymbol{\beta},$$

for all $i, i = 1,\dots,I$ where $x_j \le x_{(i)} \le x_{j+1}$, $\boldsymbol{\beta} = (\beta_0,\beta_1,\dots,\beta_n)^{\mathrm{T}}$, and $\mathbf{F}_j$ is the $(j+1)$-th row of the matrix $\mathbf{F}$.

3.3. Penalty for the Covariate Smoothing Function

Before the coefficients of the GAM are estimated via PIRLS, a penalty on the smoothing function is required to control the smoothness level of the curve. Fitting the model in the GAM with cubic splines aims to minimize the following expression [8,9]:

$$\sum_{i=1}^{n}\{y_i - f(x_{(i)})\}^2 + \lambda\int[f''(x)]^2dx,$$

The first term is the residual sum of squares (RSS), whereas the second term is the penalty that measures the level of curve fluctuation. Let us denote the above penalty as a function $P(f_x)$ defined as follows:

$$P(f_x) = \lambda \int_a^b [f''(x)]^2 \, dx \tag{18}$$

where λ with $\lambda \geq 0$ is a tunable smoothing parameter that serves to control the curve adaptation to the data and the level of smoothing. Given $\mathbf{A}\boldsymbol{\delta}^- = \mathbf{B}\boldsymbol{\beta} \rightarrow \boldsymbol{\delta}^- = \mathbf{A}^{-1}\mathbf{B}\boldsymbol{\beta}$, the equation above can be restated as follows.

$$\begin{aligned} \int_a^b f''(x)^2 \, dx &= \boldsymbol{\beta}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{A}^{-\mathbf{1}^{\mathbf{T}}}\mathbf{A}\mathbf{A}^{-1}\mathbf{B}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{A}^{-\mathbf{1}^{\mathbf{T}}}\boldsymbol{I}\mathbf{B}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{A}^{-\mathbf{1}}\mathbf{B}\boldsymbol{\beta} \end{aligned} \tag{19}$$

with $\mathbf{A}^{-\mathbf{1}^{\mathbf{T}}} = \mathbf{A}^{-1}$ since $\mathbf{A}$ is a symmetric matrix. Therefore, the final form of the penalty for the smoothing function of covariate X is as follows:

$$P(f_x) = \lambda\boldsymbol{\beta}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{A}^{-\mathbf{1}}\mathbf{B}\boldsymbol{\beta} \tag{20}$$

3.4 Estimating GAM Coefficients with PIRLS

The response variables, Y, follow a Poisson distribution.

1. Estimate the initial values $\mu_i^{[0]} = y_i^{[0]} + \delta_i$ and $\eta_i^{[0]} = g(\mu_i^{[0]})$, where δ_i is usually 0. Then, the following two-step iteration is performed until convergence.
2. Calculate the pseudodata $z_i = \left(\frac{1}{\mu_i^{[t]}}\right)\left(y_i - \mu_i^{[t]}\right) + \mathbf{X}_i\boldsymbol{\beta}^{[t]}$ and the iterative weights $w_i = \left(\mu_i^{[t]}\right)$.
3. Find $\widehat{\boldsymbol{\beta}}^{[t]}$ that minimizes the following weighted least squares objective function:

   $$S_p = \left\|\boldsymbol{z}^{[t]} - \boldsymbol{X\beta}\right\|_W^2 + \lambda\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{S\beta}$$

   where $\|a\|_W^2 = \mathbf{a}^{\mathrm{T}}\mathbf{W}\mathbf{a}$, and λ can be chosen arbitrarily with $0 \leq \lambda < \infty$. Then, update $\widehat{\boldsymbol{\eta}}^{[t+1]} = \mathbf{X}\widehat{\boldsymbol{\beta}}^{[t]}$ and $\hat{\mu}_i^{[t+1]} = g^{-1}(\hat{\eta}_i^{[t+1]})$.

After the GAM coefficients are obtained, the generalized cross-validation (GCV) value is calculated as described in subsection 2.6:

$$\mathcal{V}_g(\boldsymbol{\lambda}) = \frac{I \times D\left(\widehat{\boldsymbol{\beta}}\right)}{(I - \gamma\tau)^2} \tag{21}$$

where $D\left(\widehat{\boldsymbol{\beta}}\right)$ is the residual deviance, τ is the effective degree of freedom of the model, and γ is usually 1. To obtain the optimal smoothing parameter, the GAM coefficients are iteratively estimated with different λ values until the smallest GCV value is achieved.

3.5 GAM Construction with Two Covariates

This subsection discusses the construction and estimation of the GAM coefficients with two covariates. Consider the X_1 representing the distance traveled by the i-th policyholder, with observation values $x_{1(i)}, i = 1, \dots, I$. Additionally, consider the covariate X_2, which represents the duration of the insurance contract for the i-th policyholder, with observation values $x_{2(i)}, i = 1, \dots, I$. Slightly different from the representation of a GAM with one covariate, the representation of a GAM with two covariates is as follows:

$$g(\mu_i) = intercept + f_1(x_{1(i)}) + f_2(x_{2(i)}), \tag{22}$$

where $f_1(x_1)$ is the smoothing function for covariate X_1 (distance traveled) and where $f_2(x_2)$ is the smoothing function for covariate X_2 (duration of the insurance contract).

The steps or process for constructing a GAM with two covariates are the same as those for constructing a GAM with one covariate. The determination of the number and placement of knots for two covariates is analogous to what has been explained in the previous subsections.

3.6 Determining Pay-As-You-Drive Premium Rates

The form of the GAM in this case is as follows:

$$\log(\hat{\mu}) = intercept + \hat{f}_1(x_1) + \hat{f}_2(x_2). \tag{23}$$

$$\hat{\mu} = \exp(intercept) \times \exp\left(\hat{f}_1(x_1)\right) \times \exp\left(\hat{f}_2(x_1)\right) \tag{24}$$

where $\hat{\mu}$ is the estimated average number of claims and where the intercept is the model constant. Since the model has been estimated, the price relativity for the premium rate can be obtained as follows:

1. The reference premium or base value is equal to $\exp(intercept)$;
2. The price relativity for the distance traveled is equal to $\exp\left(\hat{f}_1(x_1)\right)$;
3. The price relativity for the duration of the insurance contract is equal to $\exp\left(\hat{f}_2(x_2)\right)$.

Therefore, the formula for calculating the premium rate is

$$premium\ rate = reference\ premium \times \exp\left(\hat{f}_1(x_1)\right) \times \exp\left(\hat{f}_2(x_2)\right) \tag{25}$$

4. APPLICATION

4.1 Dataset

The dataset used for the application of the Generalized Additive Model (GAM) is a motor vehicle insurance dataset obtained from the study by So, Boucher, and Valdez (2021). It contains information on 10,000 car insurance policyholders over a one-year observation period, comprising 52 variables relevant to telematics, such as the duration of the insurance contract within the observation period, the total distance traveled by the car, the car's age, and the number of claims. However, for this discussion, only two covariates are included in the model: mileage and the duration of the insurance contract.

4.2 Determination of Knot Placement and Number

Denote covariates X_1 as mileage and X_2 as the insurance contract duration. For the i-th policyholder, the observed value of distance traveled is denoted by $x_{1(i)}, i = 1, \dots, 10{,}000$, and the observed value of insurance contract duration is denoted by $x_{2(i)}, i = 1, \dots, 10{,}000$. We use the smallest number of knots, 3, to construct the cubic spline for both covariates. The knots are placed at quantiles 0, 0, 5, and 1 of the unique observed values of the distance traveled covariate. Let $x_{1,k}, k = 0,1,2$ denote the first knot to the third knot for covariate X_1, and let $x_{2,l}, l = 0,1,2$ denote the first knot to the third knot for covariate X_2; then, the knots can be written as follows:

$$\begin{aligned} x_{1,0} &= 0.193 & x_{2,0} &= 0.2049 \\ x_{1,1} &= 5{,}681.247 & x_{2,1} &= 0.689 \\ x_{1,2} &= 72{,}725.53 & x_{2,2} &= 1 \end{aligned}$$

4.3 Smoothing Function Construction

The cubic spline function for the covariate distance traveled is expressed by the following equation:

$$\begin{aligned} f_{1,k}(x_1) = a^-_{1,k}(x_1)\beta_{1,k} + a^+_{1,k}(x_1)\beta_{1,k+1} + c^-_{1,k}(x_1)\mathrm{F}_{1_{(k+1),1}}\beta_{1,0} \\ + c^-_{1,k}(x_1)\mathbf{F}_{1_{(k+1),2}}\beta_{1,1} \\ + c^+_{1,k}(x_1)\mathrm{F}_{1_{(k+2),1}}\beta_{1,0} + c^+_{1,k}(x_1)\mathrm{F}_{1_{(k+2),2}}\beta_{1,1} \\ + \cdots + c^+_k(x_1)\mathrm{F}_{1_{(k+2),5}}\beta_{1,4} \end{aligned}$$

On the basis of this equation, the functions for each subinterval are as follows:

$$\begin{aligned} f_{1,0}(x_1) = \frac{(5{,}681.247 - x_1)}{5{,}681.055}\beta_{1,0} + \frac{(x_1 - 0.193)}{5{,}681.055}\beta_{1,1} \\ + \frac{\left[\frac{(x_1 - 0.193)^3}{5{,}681.055} - 5{,}681.055\,(x_1 - 0.193)\right]}{6} \\ \left((7.261 \times 10^{-9})\beta_{1,0} + (-7.877 \times 10^{-9})\beta_{1,1} + (6.153 \times 10^{-10})\beta_{1,2}\right) \end{aligned}$$

if $0.193 \le x_1 < 5{,}681.247$.

$$\begin{aligned} f_{1,1}(x_1) = \frac{(72{,}725.53 - x_1)}{67{,}044.287}\beta_{1,1} + \frac{(x_1 - 5{,}681.247)}{67{,}044.287}\beta_{1,2} \\ + \frac{\left[\frac{(72{,}725.53 - x)^3}{67{,}044.287} - 67.044{,}287(72{,}725.53 - x)\right]}{6} \\ \left((7.261 \times 10^{-9})\beta_{1,0} + (-7.877 \times 10^{-9})\beta_{1,1} + (6.153 \times 10^{-10})\beta_{1,2}\right) \end{aligned}$$

if $5{,}681.247 \le x_1 \le 72{,}725.53$. The cubic spline function for the covariate insurance contract duration is expressed by the following equation:

$$f_{2,l}(x_2) = a_{2,l}^{-}(x_2)\beta_{2,l} + a_{2,l}^{+}(x_2)\beta_{2,l+1} + c_{2,l}^{-}(x_2)\mathrm{F}_{2_{(2+1),1}}\beta_{2,0} + c_{2,l}^{-}(x_2)\mathrm{F}_{2_{(l+1),2}}\beta_{2,1} + c_{2,l}^{+}(x_2)\mathrm{F}_{2_{(l+2),1}}\beta_{2,0} + c_{2,l}^{+}(x_2)\mathrm{F}_{2_{(l+2),2}}\beta_{2,1} + \cdots + c_{l}^{+}(x_2)\mathrm{F}_{2_{(l+2),5}}\beta_{2,4}$$

Based on this equation, the functions for each subinterval are as follows.

$$f_{2,0}(x_2) = \frac{(0.689 - x_2)}{0.485}\beta_{2,0} + \frac{(x_2 - 0.205)}{0.485}\beta_{2,1} + \frac{\left[\frac{(x_2 - 0.205)^3}{0.485} - 0.485\,(x_2 - 0.205)\right]}{6}\left(7.78\beta_{2,0} - 19.947\beta_{2,1} + 12.167\beta_{22}\right)$$

if $0.205 \leq x_1 < 0.689$ and

$$f_{2,1}(x_2) = \frac{(1 - x_2)}{0.31}\beta_{2,1} + \frac{(x_2 - 0.689)}{0.31}\beta_{2,2} + \frac{\left[\frac{(x_2 - 0.689)^3}{0.31} - 0.31\,(x_2 - 0.689)\right]}{6}\left(7.78\beta_{2,0} - 19.947\beta_{2,1} + 12.167\beta_{2,2}\right)$$

if $0.689 \leq x_2 < 1$.

4.4 Penalty

The penalty smoothing function of the travel distance covariate can be expressed as follows.

$$P(f_1) = \lambda_1 \boldsymbol{\beta_1^{\mathrm{T}}} \mathbf{B_1^{\mathrm{T}}} \mathbf{A^{-1}} \mathbf{B} \boldsymbol{\beta}$$

$$P(f_1) = \lambda_1 [\beta_{1,0}\ \ \beta_{1,1}\ \ \beta_{1,2}] \begin{bmatrix} 1.278 \times 10^{-12} & -1.386 \times 10^{-12} & 1.083 \times 10^{-13} \\ -1.386 \times 10^{-12} & 1.504 \times 10^{-12} & -1.175 \times 10^{-13} \\ 1.083 \times 10^{-13} & -1.175 \times 10^{-13} & 9.177 \times 10^{-15} \end{bmatrix} \begin{bmatrix} \beta_{1,0} \\ \beta_{1,1} \\ \beta_{1,2} \end{bmatrix}$$

By the same method, the penalty function for smoothing the covariate of insurance contract duration can be expressed by the following equation:

$$P(f_2) = \lambda_2 \boldsymbol{\beta_2^{\mathrm{T}}} \mathbf{B_2^{\mathrm{T}}} \mathbf{A_2^{-1}} \mathbf{B_2} \boldsymbol{\beta_2}$$

It can then be rewritten as the following equation:

$$P(f_2) = \lambda_2 [\beta_{2,0} \quad \beta_{2,1} \quad \beta_{2,2}] \begin{bmatrix} 16.042 & -41.131 & 25.088 \\ -41.131 & 105.455 & -64.324 \\ 25.088 & -64.324 & 39.235 \end{bmatrix} \begin{bmatrix} \beta_{2,0} \\ \beta_{2,1} \\ \beta_{2,2} \end{bmatrix}$$

4.5 Estimation of the GAM Coefficients

Upon obtaining the smoothing basis functions and penalties for both covariates, the estimation of the coefficient basis functions is conducted via the penalized iterative reweighted least squares (PIRLS) method, as explained in subsection 3.1.5. The selection of smoothing parameter values (λ) is performed via generalized cross validation (GCV).

Leveraging the "mgcv" package within the R software environment, the fitted GAM results against the dataset are obtained as follows:

Table 2
GAM Fitting Results with Number of Knots $X_1 = 3$ and $X_2 = 3$

Parametric	Estimate		*t value*	*p value*
$intercept$	-3.508		-53.07	$< 2 \times 10^{-16}$
Nonparametric	EDF	λ	*F Value*	*p Value*
$f_1(x_1)$	1.993	0.0307	128.859	$< 2 \times 10^{-16}$
$f_2(x_2)$	1.965	0.0834	6.519	0.00129
GCV	0.24793			

Table 2 shows that the p value for both smoothing functions is less than 0.05, meaning that both covariates are significant to the model. The minimum GCV value is obtained when the value of the smoothing parameter (λ) for the travel distance covariate smoothing function is 0.0307 and the insurance contract duration is 0.0834. According to the findings of Stone [17], several knots greater than 5 is rarely required in natural cubic spline models. The main choice of the number of knots that can be used is between 3, 4, or 5. Therefore, a GCV comparison of different possible pairs of knots was conducted. Based on the results of the smallest GCV calculation, the selected pair of knots is 5 for each covariate.

4.6 Premium Rates

In this subsection, the calculation of the reference premium and the relativity of premium prices based on the characteristics of the policyholder's risk are discussed. The form of the generalized additive model (GAM) in this case is as follows.

$$\hat{\mu} = \exp(intercept) \times \exp\left(\hat{f}_1(x_1)\right) \times \exp\left(\hat{f}_2(x_2)\right)$$

where $\hat{\mu}$ is the estimated average claim amount and the intercept is the model constant. Since the model has been estimated, the relative price for the premium rate is obtained as follows:

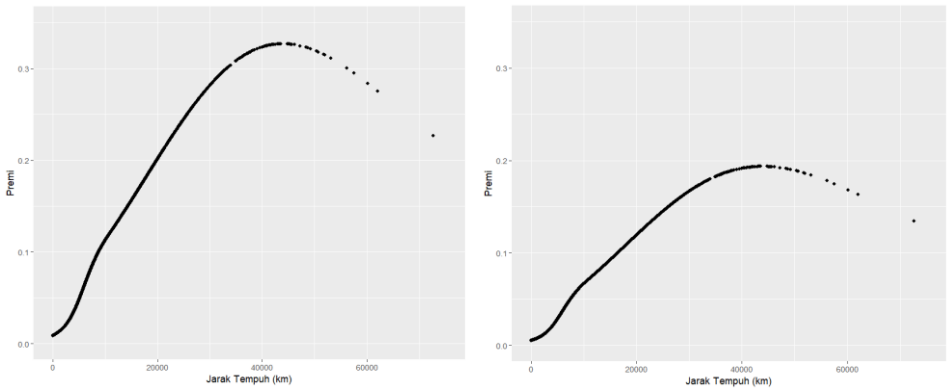
- The reference premium is equal to $\exp(intercept) = \exp(-3.6402)$;
- The price relativity for the distance traveled is equal to $\exp\left(\hat{f}_1(x_1)\right)$;
- The price relativity for the insurance contract duration is equal to $\exp\left(\hat{f}_2(x_2)\right)$.

Some 5 examples of policyholder risk profiles along with their estimated premium rates are shown in Table 3 as follows:

Table 3
Example of Simple pay-as-you-drive Premium Rates

$(\mathbf{x_1}; x_2)$	Relativity for x_1 (a)	Relativity for x_2 (b)	Total relativity $(c) = (a) \times (b)$	Premium $\exp(-3.6402) \times (c)$
$(7,000; 0.5)$	0.7954	0.5629	0.849	0.02228
$(30,000; 0.5)$	5.4295	0.5629	3.0562	0.0802
$(30,000; 1)$	5.4295	1.1704	6.3544	0.1668
$(20,000; 0.5)$	3.8862	0.5629	2.1875	0.0574
$(40,000; 0.5)$	6.2351	0.5629	3.5097	0.0921

In Table 3, the price premium relativity of each policy characteristic to the reference premium (base value) is shown. The reference premium is the premium when there is no influence from covariates or when it is considered to be 0. When the reference premium is $\exp(-3.689787) = 0.02498$,, then for a policy with an insurance contract duration of 0.5 years and a distance traveled of 20,000 km, the premium is 2.1875 times the reference premium or equal to 0.0546. Based on the dataset used, increasing the insurance contract duration does not always increase the premium rate. Since the dataset used, increasing the insurance contract duration does not always increase the premium rate. Below, a plot of the predicted premium rates when the insurance contract duration is 0.6 years and 1 year is presented.



a) 0.6 year b) 1 year

Figure 1: Predicted Premiums with Constant Insurance Contract Duration

Figure 1 shows that the predicted premium value for an insurance contract duration of 0.6 years is greater than the predicted premium for an insurance contract duration of 1 year. Therefore, the longer the insurance contract duration is, the greater the claim frequency.

5. CONCLUSION

The process of constructing a generalized additive model (GAM) to determine vehicle insurance premium rates, using distance traveled and the duration of the insurance contract as covariates, starts with creating a smoothing function via cubic splines. The determination of the number and location of knots for each covariate is essential for generating the basis functions of cubic splines. GAM coefficients are estimated via the penalized iterative reweighted least squares (PIRLS) method, with the optimal smoothing parameter (λ) determined since the smallest generalized cross-validation (GCV) value. Upon successful construction of the GAM, the model can be utilized to forecast claim frequencies and subsequently utilized as the relative premium rate against the reference premium. Based on vehicle insurance claim data obtained from the study conducted by [12], a simplified premium rate for Pay-As-You-Drive Insurance is derived. Assuming a reference premium value of $exp(-3.689787) = 0.02498$, for a policy with an insurance contract duration of 0.5 years and a distance traveled of 20,000 km, the premium is calculated to be 2.1875 times the reference premium, which is equivalent to 0.0546. According to this dataset, the premium for a 1-year insurance contract duration is lower than the premium for a 0.6-year contract duration. Therefore, it can be inferred that longer insurance contract durations do not necessarily entail greater risks than shorter contract durations do.

CONFLICT OF INTERESTS

The authors declare that there are no conflicts of interest.

REFERENCES

[1] Government of Indonesia, Law of the Republic of Indonesia Number 40 Year 2014 on Insurance, 2014.

[2] Litman, T. (2005). Pay-As-You-Drive Pricing and Insurance Regulatory Objectives. *Journal of Insurance Regulation*, 23(3), 35-53.

[3] Boucher, J.P., Pérez-Marín, A.M. and Santolino, M. (2013). Pay-as-you-drive insurance: the effect of the kilometers on the risk of accident. In *Anales del Instituto de Actuarios Españoles*, 19(3), 135-154.

[4] Langford, J., Koppel, S., McCarthy, D. and Srinivasan, S. (2008). In defence of the 'low-mileage bias'. *Accident Analysis & Prevention*, 40(6), 1996-1999.

[5] Lemaire, J., Park, S.C. and Wang, K.C. (2016). The use of annual mileage as a rating variable. *ASTIN Bulletin: The Journal of the IAA*, 46(1), 39-69.

[6] Boucher, J.P., Côté, S. and Guillen, M. (2017). Exposure as duration and distance in telematics motor insurance using generalized additive models. *Risks*, 5(4), 54.

[7] Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

[8] Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC Press. ISBN: 9781315370279

[9] Wood, S.N. (2017). *Generalized Additive Models: An Introduction with R* (Second Edition). Chapman and Hall/CRC Press.

[10] Ramsay, J.O. (1988). Monotone regression splines in action. *Statistical Science*, 3(4), 425-441.
[11] Arumugam, S. and Bhargavi, R. (2019). A survey on driving behavior analysis in usage based insurance using big data. *Journal of Big Data*, 6, 1-21.
[12] So, B., Boucher, J.P. and Valdez, E.A. (2021). Synthetic dataset generation of driver telematics. *Risks*, 9(4), 58.
[13] Ohlsson, E. and Johansson, B. (2010). *Non-life insurance pricing with generalized linear models*. Berlin: Springer.
[14] Klugman, S.A., Panjer, H.H. and Willmot, G.E. (2012). *Loss Models: From Data to Decisions*. John Wiley & Sons.
[15] Burden, R.L. and Faires, J.D. (2011). *Numerical analysis* (9th Edition) Thomson Brooks/Cole.
[16] Hastie, T., Tibshirani, R., Friedman, J.H. and Friedman, J.H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Second Edition). New York: Springer.
[17] Stone, C.J. (1986). Comment: Generalized additive models. *Statistical Sci.*, 1(312-314), 26-28.
[18] Harrell, Jr, F.E. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis,* (Second Edition). Springer.
[19] Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge University Press, New York.