Available online at http://www.iorajournal.org/index.php/ijgor/index

International Journal of Global Operations Research

e-ISSN: 2722-1016
p-ISSN: 2723-1739

Vol. 6, No. 3, pp. 163-170, 2025

Analysis of Health Insurance Claims Factors using The Stochastic Restricted Maximum Likelihood Estimation (SRMLE) Binary Logistic Regression Model (Case Study: Health Insurance Claims at XYZ Company in 2023)

Elizabeth Irene Bagariang[1*], Riaman[2], Nurul Gusriani[3.]

[1]*Mathematics Undergraduate Study Program, Faculty of Mathematics and Natural Science, Universitas Padjadjaran, Sumedang, Indonesia*
[2,3]*Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Padjadjaran, Sumedang, Indonesia*

**Corresponding author email: elizabeth20002@mail.unpad.ac.id*

Abstract

The health insurance claim approval process is a crucial aspect for insurance companies. Inaccuracy in predicting claim status can pose financial risks to the company and reduce policyholder trust. This study aims to identify the factors that influence the approval or rejection of health insurance claims. In this type of data analysis, the problem of multicollinearity among predictor variables is often encountered, which can lead to unstable parameter estimates. To address this issue, this study utilizes a binary logistic regression model with the Stochastic Restricted Maximum Likelihood Estimation (SRMLE) method, which is better suited to handle such conditions. The data used in this research includes the variables of total claim amount, premium price, number of insured individuals, employee age, and the number of previous claims recorded at XYZ Company. The results of the factor analysis, through the developed logistic regression model, show that the variables of total claim amount, premium price, and the number of insured individuals are significant factors influencing the probability of claim approval.

Keywords: Health Insurance; Binary Logistic Regression; Newton-Raphson; Stochastic Restricted Maximum Likelihood Estimator (SRMLE); Multicollinearity

1. Introduction

The workforce plays a vital role in Indonesian life and the development of the country (Sayifullah and Emmalian, 2018). Not only is it a human resource that enables the economic sector to function, but it is also a key driver of national economic growth. Indonesia's large population makes the available workforce a substantial economic force if managed properly. For this reason, the role of the workforce in this country is crucial, both through the production and development of goods and services for the welfare of domestic and international consumption.

According to Juniarti et al. (2017), occupational safety and protection are still very minimal in Indonesia, with a high number of workplace accidents reaching 96,400. Therefore, the government has established obligations for each company regarding labor rights. Companies are required to comply with Manpower Law No. 13 of 2003, one of which is Article 88 paragraph (1), which states: "Every worker has the right to receive protection in terms of occupational safety and health, morals and ethics, and treatment in accordance with human dignity and religious values." Based on this article, insurance is a right that employees can have, in accordance with company policy (Sayifullah and Emmalian, 2018). To reduce the risk of potentially greater costs if coverage is not provided, companies provide insurance policies to their employees, including health coverage, with adjustments to company policy and adjustments to the company's premium costs for each employee (Saraswat et al., 2023).

There are several types of insurance, one of which is general liability insurance, which includes health insurance policies. Health insurance policies cover or minimize the costs of losses caused by various hazards (Hassan et al., 2021). Employees who receive health insurance from their company can wisely file claims in accordance with applicable regulations by fulfilling the requirements of the relevant parties for payment, either cashless or through reimbursement.

According to Julia and Fitrianah (2020), a claim is a financial obligation regarding the cost of healthcare services submitted individually or collectively, with a process that can be approved or rejected. Health insurance claims are approved if there are no issues identified during the review by the relevant company. The smoothness of this review process is greatly influenced by the policyholder's understanding of their rights and obligations.

Insurance awareness is a condition where a person understands, comprehends, and is aware of insurance, characterized by openness to accepting and utilizing insurance, including health insurance. If employees in a company understand and are willing to utilize the insurance provided by the company, they are considered insurance-aware. One benefit that employees can utilize is filing claims. Insurance awareness is influenced by several factors, including age, gender, occupation, and education (Eko Siswoyo et al., 2015).

Saraswat et al. (2023) explain that by providing detailed employee insurance data such as age, gender, number of children, Body Mass Index (BMI), and domicile, modeling can be performed that can help companies reduce unnecessary expenses. However, this is very resource-intensive if done manually. Mathematical modeling can help companies analyze factors that influence the occurrence of health insurance claims provided to their employees. In this case, the claim factor acts as a binary variable.

One statistical analysis related to binary variables is binary logistic regression. Hosmer and Lemeshow (2000) state that binary logistic regression is a data analysis method useful for exploring the relationship between a binary dependent variable (y) and an independent variable (x). The categories contained in the dependent variable (y) are success, denoted by y=1, and failure, denoted by y=0. This closely aligns with the insurance claims process itself, which includes approval (success) and rejection.

Insurance analysis research has been extensively conducted, both on health and non-health insurance. A study using the C4.5 algorithm by Julia and Fitrianah (2020) found that factors influencing claims were cost, gender, status, dependents, and claim type. Another study by Saraswat et al. (2023) used Extreme Gradient Boosting and found that factors influencing claims were age, domicile, Body Mass Index (BMI), and smoking status. Yousof et al. (2024) used the Burr-Hatke (DGBH) distribution to analyze car insurance and found the method to be powerful and flexible for segmenting and capturing insurance claims data.

A study of the relationship between factors and health insurance costs using linear regression analysis and skewness and kurtosis analysis by Tang and Wang (2024) found that age and smoking status significantly influence health insurance prices. However, in insurance claims data analysis, multicollinearity often presents a problem. Senaviratna and Cooray (2019) stated that multicollinearity in logistic regression can lead to unstable estimates and inaccurate variance, which in turn affects confidence intervals and hypothesis testing. When multicollinearity occurs, the model can face problems such as unnecessary standard error inflation, unreasonably low or high t-statistic values, and illogical parameter estimates. Consequently, these problems can lead to invalid statistical conclusions. Furthermore, the presence of multicollinearity can lead to biased coefficient estimates and very large standard errors for logistic regression coefficients. One way to detect the presence of multicollinearity is to calculate the Variance Inflation Factor (VIF). Therefore, it is important to analyze cases of multicollinearity to produce a more stable and reliable model and enable valid conclusions to be drawn from the statistical analysis.

The Stochastic Restricted Maximum Likelihood Estimator (SRMLE) was introduced as a new method that utilizes stochastic linear restrictions to address the problem of multicollinearity in logistic regression models. Nagarajah and Wijekoon (2015) conducted a study on parameter estimation using simulated data and found that SRMLE parameter estimates performed significantly better than MLE parameter estimates. This restriction is a stochastic restriction that aims to reduce the correlation between independent variables that causes multicollinearity (Alheety et al., 2021). SRMLE offers the advantage of introducing a more flexible restriction structure, which relies not only on existing data but also uses prior information to guide parameter estimation.

SRMLE is capable of producing more stable and efficient parameter estimates, even in conditions where multicollinearity is very strong among the independent variables. This approach allows for reducing large standard errors and correcting errors in the interpretation of results, which often occur in models with high multicollinearity (Alheety et al., 2021).

Based on these studies, this study discusses the analysis of health insurance claims using a binary logistic regression model with Stochastic Restricted Maximum Likelihood Estimator (SRMLE) parameter estimation. The use of SRMLE was chosen based on the assumption that the total claims and premium prices experience multicollinearity, which must be demonstrated by the VIF results. While previous research by Nagarajah and Wijekoon (2015) used simulated data, this study uses operational data, namely insurance claims data for 2023 at Company XYZ, with influencing factors being total claims, premium prices, number of insured persons, employee age, and number of previous claims.

2. Literature Review

2.1. Insurance and Claims

Insurance is an agreement, generally written in a policy, containing the rights and obligations agreed upon between the policyholder and the insurance company (Law of the Republic of Indonesia No. 40 of 2014). As an insurance product, health insurance provides health coverage as initially agreed to by the policyholder if an insured event occurs, as outlined in the policy. This coverage can be realized through an insurance claim. Pitacco (2014) explains the types of health insurance products, namely, Disease Insurance, Accident Insurance, Critical Illness Protection, and Long-Term Care Insurance.

To provide health protection to their employees, companies in Indonesia generally include insurance as an employee benefit. Several health insurance products are available to companies when determining appropriate benefits for their employees. The products used by companies are typically group-based, so the costs incurred are not as high as for individuals.

2.2. Linear Regression Analysis

Regression analysis is an analysis that examines the relationship between a quantitative dependent variable and one or more explanatory variables and explores the conditional distribution of Y (Fox, 2015). There are two types of variables in regression analysis: the dependent variable, which is influenced by another variable, denoted Y, and the independent variable, which is independent and unaffected by another variable, denoted X. A single independent variable indicates simple regression analysis, while more than one indicates multiple regression analysis.

Classical assumptions underlying the use of linear regression models include that the data measurement scale for the dependent and independent variables must be interval or ratio, and that the error must be normally distributed. For multiple linear regression, it is necessary to ensure the absence of multicollinearity, which is a high correlation between the independent variables.

The linear regression model in matrix form can be described as follows,

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} \tag{1}$$

with matrix **Y** of size $(N \times 1)$, matrix **X** of size $(N \times (k+1))$, matrix **β** of size $((k+1) \times 1)$, and matrix **ε** of size $(N \times 1)$ where $\varepsilon_i \sim N(0, \sigma\text{^}2\ )$ for $i = 1,2,\dots,N$.

2.3. Logistic Regression Analysis

As a statistical method used to analyze the relationship between one or more numeric or categorical independent variables and two-category dependent variables, logistic regression analysis is an option for modeling several conditions (Hosmer and Lemeshow, 2000). The outcome variable in logistic regression is binary or dichotomous, which differentiates the logistic regression model from the linear regression model. Dichotomy itself means containing two conflicting values, namely 0 for a condition that does not occur/failure and 1 for a condition that occurs/success.

$E(Y|X)$ which is the expectation of Y conditional on X implies that $E(Y|X)$ can take any value as long as X is in the interval $-\infty < X < \infty$. $E(Y|X)$ is written as follows:

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \tag{2}$$

The conditional mean of the dependent variable y given the value of x is π(x)=E(Y|X) (Hosmer and Lemeshow, 2000). The logistic regression model with k variables (the number of independent variables) according to Hosmer and Lemeshow (2000) can be written as follows:

$$\pi(x_i) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}. \tag{3}$$

2.4. Multicolinearity

Multicollinearity is a condition where two or more independent variables (X) have a strong relationship/correlation with each other in a regression model. Williams (2015) stated that Variance Inflation Factor (VIF) is better than Correlation Matrix Analysis in testing multicollinearity. VIF = 10 is considered to have no multicollinearity and VIF> 10 is considered to have an indication of multicollinearity between the independent variables. VIF can be calculated using the following formula:

$$VIF_j = \frac{1}{1 - R_j^2}\ ,j = 1, \dots, k. \tag{4}$$

where R_j^2 is the coefficient of determination between X_j and the other independent variables in the model equation. The value of R_j^2 is obtained from the following formula:

$$R_j^2 = \frac{\text{JK Regresi}\,j}{\text{JK Total}\,j} = \frac{\Sigma_{i=1}^{n}\left(\hat{X}_{ij} - \bar{X}_j\right)^2}{\Sigma_{i=1}^{n}\left(X_{ij} - \bar{X}_j\right)^2} VIF_j = \frac{1}{1 - R_j^2}\ ,j = 1, \dots, k. \tag{5}$$

2.5. Maximum Likelihood Estimator

The Maximum Likelihood Estimator method is a method that optimizes the likelihood function to obtain estimates of unknown parameters. In the application of MLE, it begins by building the likelihood as a function of the parameters in this case is a logistic regression model. Suppose there are n pairs of data (x_i, y_i) $i = 1, 2, \ldots, n$ are mutually independent, with y_i being the value of the dependent variable, namely 0 or 1 and x_i the value for the i-th independent variable. Because the independent variable is binary, if $y_i = 0$ then its involvement in the likelihood function is $1 - \pi(x_i)$ and if $y_i = 1$ then its involvement in the likelihood function is $\pi(x_i)$. The likelihood function of the pair (x_i, y_i) is as follows:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(Y = y_i) = \prod_{i=1}^{n} [\pi(x_i)]^{y_i} [1 - \pi(x_i)]^{1-y_i} \tag{6}$$

Then we will look for the natural logarithm of the likelihood function (6) as follows:

$$L(\boldsymbol{\beta}) = \ln l(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) - \ln\{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})\}] \tag{7}$$

In this method, the principle of estimating parameters in the logistic regression model is to maximize equation (7). To find the maximum value of $L(\beta)$, equation (7) will be derived partially with respect to the parameters $\boldsymbol{\beta}_j = 0,1, \ldots, k$ and equated to zero.

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}(y_i - \pi(x_i)) = 0 \tag{8}$$

Because equation (8) is not a linear equation, a method is needed to obtain the estimated values of the parameters of the logistic regression model so that it is easier to obtain the formula derivation and also determine the size of each parameter. The method that can be used is Newton-Raphson Iteration (Hosmer and Lemeshow, 2000). According to Winkelmann (2003), the new parameters based on Newton-Raphson are:

$$\widehat{\boldsymbol{\beta}}^{(a+1)} = \widehat{\boldsymbol{\beta}}^{(a)} - [\mathbf{X}^{\mathbf{T}}\mathbf{U}\mathbf{X}]^{-1} \cdot \mathbf{X}^{\mathbf{T}}[\boldsymbol{y} - \pi(\boldsymbol{x})]. \tag{9}$$

where $\boldsymbol{H} = \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{U}\boldsymbol{X}$.

2.6. Stochastic Restricted Maximum Likelihood Estimator

The Maximum Likelihood Estimator is commonly used to estimate parameters in logistic regression. However, Siray (2015) stated that the drawback of the MLE is that it cannot address multicollinearity in the data, which can lead to very high variance values. The Stochastic Restricted Maximum Likelihood Estimator (SRMLE), which is a development of the Restricted Maximum Likelihood Estimator (RMLE) method, can address the shortcomings of previous methods in addressing multicollinearity (Nagarajah and Wijekoon, 2015). The RMLE method is used when parameters must comply with certain constraints and can be used when the data has multicollinearity, but it needs to be done carefully by considering appropriate constraints but does not consider the stochastic elements in the data. Meanwhile, the SRMLE can address problems in the RMLE model (Nagarajah and Wijekoon, 2015).

The $\boldsymbol{\mathcal{H}}$ is a design matrix used to identify the relationships between variables in a logistic regression model. In the context of SRMLE, the $\boldsymbol{\mathcal{H}}$ matrix serves to assess the structure of restrictions given in the model. The $\boldsymbol{\mathcal{H}}$ matrix has size $(q \times (p + 1))$, where q is the number of restrictions and p is the number of independent variables. This matrix must satisfy the full rank requirement for both columns and rows, ensuring that the restriction system can be solved properly.

The **h** is a vector containing the linear constraints imposed on the model parameters. The vector **h** has size $(q \times 1)$, where q is the number of constraints imposed on the parameters in the model. These constraints govern the relationships between different parameters, such as constraints involving variables or relationships between coefficients.

We suppose that $\mathbf{C} = \mathbf{X}^{\mathbf{T}}\mathbf{U}\mathbf{X}$,

$$\widehat{\boldsymbol{\beta}}_{\text{RMLE}} = \widehat{\boldsymbol{\beta}}_{\text{MLE}} + \mathbf{C}^{-1}\boldsymbol{\mathcal{H}}^{\mathrm{T}}\left(\boldsymbol{\mathcal{H}}\mathbf{C}^{-1}\boldsymbol{\mathcal{H}}^{\mathrm{T}}\right)^{-1}\left(\mathbf{h} - \boldsymbol{\mathcal{H}}\widehat{\boldsymbol{\beta}}_{\text{MLE}}\right). \tag{10}$$

By adding the $\boldsymbol{\Psi}$ matrix, namely the positive definite identity matrix $(q \times q)$ as the uncertainty in the parameter estimation with the parameter constraints set in RMLE in equation (3.2), the SRMLE parameter estimation model can be written as follows (Nagarajah and Wijekoon, 2015):

$$\widehat{\boldsymbol{\beta}}_{\text{SRMLE}} = \widehat{\boldsymbol{\beta}}_{\text{MLE}} + \mathbf{C}^{-1}\boldsymbol{\mathcal{H}}^{\mathrm{T}}(\boldsymbol{\Psi} + \boldsymbol{\mathcal{H}}\mathbf{C}^{-1}\boldsymbol{\mathcal{H}}^{\mathrm{T}})^{-1}\left(\mathbf{h} - \boldsymbol{\mathcal{H}}\widehat{\boldsymbol{\beta}}_{\text{MLE}}\right). \tag{11}$$

The SRMLE covariance matrix is written as follows (Nagarajah and Wijekoon, 2015):

$$Var(\widehat{\boldsymbol{\beta}}_{\text{SRMLE}}) = \mathbf{C}^{-1} - \mathbf{C}^{-1}\boldsymbol{\mathcal{H}}^{T}(\boldsymbol{\Psi} + \boldsymbol{\mathcal{H}}\mathbf{C}^{-1}\boldsymbol{\mathcal{H}}^{\mathrm{T}})^{-1}\boldsymbol{\mathcal{H}}\mathbf{C}^{-1} \tag{12}$$

2.7. Significance Test of Logistic Regression Model Coefficients

Overall parameter testing can use the Likelihood Ratio Test which is carried out by comparing the observed value of the dependent variable value with the predicted value derived from the model (Hosmer and Lemeshow, 2000).

$$G = 2\left\{\sum_{i=1}^{n}\left(y_i \ln \hat{\pi}(x_i) + (1 + y_i)\ln\left(1 - \hat{\pi}(x_i)\right)\right) - (n_1 \ln n_1 + n_0 \ln n_0 - n \ln n)\right\} \tag{13}$$

with the hypothesis used are

H_0: $\hat{\beta}_0 = \hat{\beta}_1 = \cdots = \hat{\beta}_k = 0$ (there are no independent variables that contribute)

H_1: at least one $\hat{\beta}_k \neq 0$ (at least one independent variable that contributes).

chi-square comparison will be used with $\chi_{(\alpha,db)}$;α: significance level and db: $(n-1)$. With the available hypothesis above, reject H_0 if $G \geq \chi^2_{(\alpha,db)}$.

To test the significance of individual coefficients, the Wald Test can be used (Hosmer and Lemeshow, 2000). By squaring the ratio of parameter predictions/estimates ($\hat{\beta}_j$) with the standard error of each parameter prediction/estimate (SE($\hat{\beta}_j$j)), the value of W_j is obtained. The Wald test for logistic regression can be determined as follows:

$$W_j = \left(\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}\right)^2 \tag{14}$$

$$SE(\hat{\beta}_j) = \sqrt{\left((X^T U X^{-1})\right)_{jj}}$$

with the hypothesis used are

H_0: $\hat{\beta}_{\mathrm{j}} = 0$ where $j = 1,2,\ldots,k.$ (variable X_j does not make a significant contribution to $\pi(x)$)

H_1: $\hat{\beta}_{\mathrm{j}} \neq 0 \neq 0$ (variable X_j make a significant contribution to $\pi(x)$).

3. Materials and Methods

3.1. Materials

This study uses health insurance claims data obtained from XYZ Company. The dependent variable is whether or not a health insurance claim occurred, and independent variables that influence the dependent variable include total claim costs, premium price, employee age, number of previous claims, and number of dependents. This study analyzes the insurance claims data using a binary logistic regression model using the Stochastic Restricted Maximum Likelihood Estimator.

The dependent variable, insurance claims, is categorized, with a value of 0 for rejected claims and 1 for successful or accepted claims. The independent variables, total claim costs (X_1) in Rupiah, premium price (X_2) in Rupiah, number of insured persons (X_3), employee age (X_4), and number of previous claims (X_5), are numerical data and therefore require no data adjustment.

3.2. Methods

1) Input health insurance claim data
2) Multicollinearity testing with VIF based on equations (4) and (5)
3) Estimate the $\boldsymbol{\beta}_{\text{MLE}}$ parameters using Newton-Raphson iteration
4) Determine the $\boldsymbol{\Psi}$ matrix and the full rank matrix $\boldsymbol{\mathcal{H}}$ with order $q \times (p+1)$) and the vector **h**.
5) Estimate the $\widehat{\boldsymbol{\beta}}_{\text{SRMLE}}$ parameters by substituting the $\boldsymbol{\mathcal{H}}$, $\mathbf{C}$, and **h** matrices based on equation (11).
6) Determine the value of L($\widehat{\boldsymbol{\beta}}_{\text{SRMLE}}$) based on equation (7).
7) Perform an overall test using the Likelihood Ratio Test based on equation (13).
8) Perform individual or partial tests using the Wald test based on equation (14).
9) Substitute $\widehat{\boldsymbol{\beta}}_{\text{SRMLE}}$ into the binary logistic regression model equation.

4. Results and Discussion

4.1. Multicollinearity Test

The coefficient of determination and VIF were calculated using the Python programming language. The VIF results for each variable are presented in Table 1

Table 1. Multicolinearity test

Variables	Determination Coefficient Value	VIF Value
claim costs	0.981549	54.196431
premium price	0.981074	52.837980
number of insured persons	0.273074	1.375657
employee age	0.452419	1.826214
number of previous claims	0.066431	1.071158

Based on the table, variables claim costs and premium price contain multicollinearity because $VIF \geq 10$, so the insurance claim data above can be used to determine logistic regression estimates using the SRMLE method.

4.2. Maximum Likelihood Estimator

Newton-Raphson iteration is the process of finding a convergent beta (**β**) estimate as in equation (9). By carrying out four iterations using Newton-Raphson iteration, the estimated beta (**β**) value is obtained to find the MLE parameter estimate as shown in Table 2.

Table 2. Estimaed value of $\boldsymbol{\beta}_{MLE}$

Variables	Estimated Value of Regression Coefficient
intercept	-6.01939×10^{-1}
claim costs	-7.06039×10^{-8}
premium price	9.71942×10^{-8}
number of insured persons	1.29164×10^{-1}
employee age	1.09453×10^{-5}
number of previous claims	-5.67055×10^{-2}

4.3. Hessian Matrix Calculation

Before searching for parameter estimates using the SRMLE method, it is necessary to determine the values of the **U** (1000×1000) matrix and the **H** (6×6) matrix based on equation (9) using Python, obtained:

$$\mathbf{U} = \begin{bmatrix} 0{,}2485042 & 0 & \dots & 0 \\ 0 & 0{,}2377412 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0{,}24937303 \end{bmatrix},$$

$$\mathbf{H} = \begin{bmatrix} 2.44667 \times 10^{2} & 3.21719 \times 10^{10} & 2.31381 \times 10^{10} & 7.58941 \times 10^{2} & 9.61309 \times 10^{3} & 4.71039 \times 10^{2} \\ 3.21719 \times 10^{10} & 4.33633 \times 10^{18} & 3.11838 \times 10^{18} & 1.02698 \times 10^{10} & 1.30073 \times 10^{11} & 6.35566 \times 10^{10} \\ 2.31381 \times 10^{10} & 3.11838 \times 10^{18} & 2.24356 \times 10^{18} & 7.38445 \times 10^{10} & 9.35265 \times 10^{11} & 4.57459 \times 10^{10} \\ 7.58941 \times 10^{2} & 1.02698 \times 10^{10} & 7.38445 \times 10^{10} & 2.83664 \times 10^{3} & 2.98327 \times 10^{4} & 1.46680 \times 10^{3} \\ 9.61309 \times 10^{3} & 1.30073 \times 10^{11} & 9.35265 \times 10^{11} & 2.98327 \times 10^{4} & 4.11383 \times 10^{5} & 1.89269 \times 10^{4} \\ 4.71039 \times 10^{2} & 6.35566 \times 10^{10} & 4.57459 \times 10^{10} & 1.46680 \times 10^{3} & 1.89269 \times 10^{4} & 1.39348 \times 10^{3} \end{bmatrix}$$

4.4. SRMLE Estimation Calculation

After obtaining the MLE estimate, the next step is to find the estimator using the SRMLE method using equation (11). Previously, the ψ, H, and h vector matrices were determined to apply linear restrictions in the model based on Nagarajah and Wijekoon (2015), namely:

$$\boldsymbol{\psi} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} \end{bmatrix},$$

$$\boldsymbol{\mathcal{H}} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{-1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{-1} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

$$\boldsymbol{h} = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$$

From the three matrices, the estimator results using the SRMLE method are presented in Table 3 below.

Table 3. Estimaed value of $\boldsymbol{\beta}_{SRMLE}$

Variables	Estimated Value of Regression Coefficient
intercept	-6.01939×10^{-1}
claim costs	-7.06039×10^{-8}
premium price	9.71942×10^{-8}
number of insured persons	1.29164×10^{-1}
employee age	1.09453×10^{-5}
number of previous claims	-5.67055×10^{-2}

4.5. Significance Test of Binary Logistic Parameters

With the previous SRMLE estimate, it is necessary to test the significance of the binary logistic parameters using the likelihood ratio test. In conducting this test, the G statistic value is required using equation (13) to obtain the following:

$$\boldsymbol{G = 21.30248}$$

The number of independent variables used in this study is 5, therefore the degrees of freedom used are 5. Based on the Chi-Square table in Appendix 1 with $\alpha = 0.05$ and $db = 5$, $\chi^2_{(0,05,5)} = 11{,}070$.

Based on the hypothesis, H_0 is rejected because $G = 21.34715 > \chi^2_{(0,05,5)} = 11{,}070$, concluding that there is at least one independent variable that influences the binary regression model.

Next, we will conduct a significance test for individual parameters using the Wald test to determine which independent variables have an influence on the regression model. All variables will have their W values searched according to equation (14) with a Chi-Square distribution where $\alpha = 0{,}05$ and $db = 1$, thus obtaining $\chi^2_{(0.05,1)} = 3{,}84146$.

The following are the Wald values of the 5 variables in Table 4:

Table 4. Wald value of $\boldsymbol{\beta}_{SRMLE}$

Variables	Wald value
claim costs	$\mathbf{9.74812}$
premium price	$\mathbf{9.69686}$
number of insured persons	$\mathbf{5.98186}$
employee age	$\mathbf{2.29689}$
number of previous claims	$\mathbf{1.44539}$

From the Wald results, only variables A, B, and C are significant, so these variables are eliminated and we repeat the process by eliminating variables A and B and obtain a significant SRMLE value for both the whole and the individual along with the odds ratio as Table 5:

Table 5. Estimated value of $\boldsymbol{\beta}_{SRMLE}$ and odds ration

Variables	$\widehat{\boldsymbol{\beta}}$	exp ($\widehat{\boldsymbol{\beta}}$)
Intercept	-5.80073×10^{-1}	0.50791079
claim costs	-6.53867×10^{-8}	0.99999994
premium price	9.37982×10^{-8}	1.00000009
number of insured persons	1.08732×10^{-1}	1.11406362

5. Conclussion

Based on the analysis, it can be concluded that the outcome of insurance claims at XYZ Company categorized as either accepted (1) or rejected (0) can be effectively modeled using binary logistic regression. To address the issue of multicollinearity within the dataset, the Stochastic Restricted Maximum Likelihood Estimation (SRMLE) method was employed for parameter estimation, and the resulting logistic regression model proved to be statistically significant, as validated by both the likelihood ratio test and the Wald test at a 5% significance level.

The study identified three key factors that significantly influence the probability of a claim being accepted: the total claim cost, the premium price, and the sum insured. An interpretation of the model using Odds Ratios (Exp(β)) reveals the specific impact of each variable. The sum insured has a positive and significant effect, where each one-unit increase raises the odds of a claim being accepted by a factor of 1.114. Conversely, the total claim cost has a statistically significant, albeit very slight, negative effect, decreasing the odds of claim acceptance to 0.99999994 times their original value for each unit increase. The premium price was also found to be a significant factor, exerting a minor positive influence on the likelihood of claim acceptance. These findings provide a clear quantitative understanding of the primary drivers behind claim approval at XYZ Company.

References

Alheety, M.I., Månsson, K. and Golam Kibria, B.M. (2021) '*A new kind of stochastic restricted biased estimator for logistic regression model*', *Journal of Applied Statistics*, 48(9), pp. 1559–1578. Available at: https://doi.org/10.1080/02664763.2020.1769576.

Eko Siswoyo, B. *et al.* (2015) Kesadaran Pekerja Sektor Informal Terhadap Program Jaminan Kesehatan Nasional Di Provinsi Daerah Istimewa Yogyakarta Awareness Of The Informal Sector Workers Towards National Health Insurance Program In Province Of Yogyakarta, *λ Jurnal Kebijakan Kesehatan Indonesia*. Kesadaran Pekerja Sektor Informal.

Fox, J. (2015) *Applied Regression Analysis And Generalized Linear Models*. SAGE Publications, Inc.

Hassan, C.A. ul *et al.* (2021) '*A Computational Intelligence Approach for Predicting Medical Insurance Cost*', *Mathematical Problems in Engineering*, 2021. Available at: https://doi.org/10.1155/2021/1162553.

Hosmer, D.W. and Lemeshow, S. (2000) *Applied Logistic Regression*. 2nd edn.

Julia, A. and Fitrianah, D. (2020) 'Analisa History Data Claim Member Asuransi Kesehatan Menggunakan Algoritma C4.5 Untuk Menentukan Proses Pembayaran Persetujuan Pengobatan', *Jurnal Teknologi Informasi dan Ilmu Komputer* [Preprint].

Juniarti, N., Halin, H. and Roswaty (2017) 'Pengaruh Keselamatan Dan Kesehatan Kerja Terhadap Kinerja Karyawan Pt Putera Sriwijaya Mandiri Palembang', *JURNAL ILMIAH EKONOMI GLOBAL MASA KINI*, 8.

Nagarajah, V. and Wijekoon, P. (2015) '*Stochastic Restricted Maximum Likelihood Estimator in Logistic Regression Model*', *Open Journal of Statistics*, 05(07), pp. 837–851. Available at: https://doi.org/10.4236/ojs.2015.57082.

Pitacco, E. (2014) *Health Insurance Basic Actuarial Models*. Available at: http://www.springer.com/series/7879.

Saraswat, B.K. *et al.* (2023) 'Insurance Claim Analysis Using Traditional Machine Learning Algorithms', in. Institute of Electrical and Electronics Engineers (IEEE), pp. 623–628. Available at: https://doi.org/10.1109/icdt57929.2023.10150491.

Sayifullah and Emmalian (2018) 'Pengaruh Tenaga Kerja Sektor Pertanian Dan Pengeluaran Pemerintah Sektor Pertanian Terhadap Produk Domestik Bruto Sektor Pertanian Di Indonesia', *Jurnal Ekonomi-Qu*, 8.

Senaviratna and Cooray (2019) '*View of Diagnosing Multicollinearity of Logistic Regression Model*'.

Tang, Z. and Wang, X. (2024) *Research On the Factors That Affect the Pricing of Healthcare Insurance*, *Highlights in Science, Engineering and Technology AMMSAC*.

Undang-Undang Republik Indonesia Nomor 13 Tahun 2003 Tentang Ketenagakerjaan.

Undang-Undang Republik Indonesia Nomor 40 Tahun 2014 tentang Perasuransian.

Williams, R. (2015) '*Multicollinearity*'. Available at: https://www3.nd.edu/~rwilliam/.

Winkelmann, R. (2003) '*Econometric Analysis of Count Data Fourth Edition*'. Jerman: Springer.

Yousof, H.M. *et al.* (2024) '*A discrete claims-model for the inflated and over-dispersed automobile claims frequencies data: Applications and actuarial risk analysis*', *Pakistan Journal of Statistics and Operation Research*, 20(2), pp. 261–284. Available at: https://doi.org/10.18187/pjsor.v20i2.4535.