Available online at https://journal.rescollacomm.com/index.php/ijbesd/index

International Journal of Business, Economics and Social Development

e-ISSN 2722-1156
p-ISSN 27722-1164

Vol. 6, No. 1, pp. 44-51, 2025

Estimation Model of Pure Health Insurance Premiums in Southeast America Using Generalized Linear Model (GLM) with Gamma Distribution

Aulya Putri[1*]

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

**Corresponding author email: aulya21001@mail.unpad.ac.id*

Abstract

Health insurance premiums are on the rise due to increasing medical costs, inflation, and the lingering effects of the COVID-19 pandemic. Accurate premium pricing is crucial for insurance companies to maintain financial stability and offer fair rates to policyholders. Generalized Linear Models (GLM) have been widely used in actuarial science for modeling insurance premiums. This study proposes the use of GLM with a Gamma distribution to model health insurance premiums. The Gamma distribution is suitable for non-negative and positively skewed data, which is characteristic of insurance claim amounts. By analyzing historical data from a Southeast United State insurance company, we aim to identify key factors influencing premium pricing and develop a robust premium model. The model will consider factors such as age, gender, BMI, number of children, and smoking status to predict individual risk profiles and determine appropriate premiums. Our findings indicate that age and smoking status are the most significant factors affecting premium rates. Older individuals and smokers tend to have higher premiums due to their increased risk of health issues. Gender and BMI, however, were found to have no significant impact on premium pricing in this specific dataset. Insurance companies can use the identified factors (age, smoking status, etc.) to create more precise risk profiles for their policyholders.

Keywords: GLM, premiums, health insurance.

1. Introduction

In 2024, health insurance premiums in the US are predicted to increase by over 6%, marking a considerable increase in premiums. Higher medical expenses, inflation, and the COVID-19 pandemic's continuing effects are some of the factors contributing to this increase (Ortaliza et al, 2023). In health insurance, insurance companies need to appropriately determine the premium price for policyholders. This premium determination is very important to maintain financial stability for the insurance company and ensure a fair price for policyholders.

Generalized Linear Models (GLM) is one of the models often used in determining premium prices (Rahmawati et al., 2023). GLM provides the flexibility to model various types of non-normal data, thus allowing the use of various probability distributions to model the relationship between its variables making GLM powerful in statistical analysis and premium modeling (yi et al., 2018).

One of the distributions that can be used is the gamma distribution. The gamma distribution which is a continuous probability distribution is often used to model non-negative and positively skewed data which corresponds to insurance claim amounts or premium amounts where the data cannot be negative (Xin et al., 2018).

Several studies have utilized Generalized Linear Models (GLM) to estimate insurance premiums in various sectors, such as motor vehicle insurance and life insurance. As in the research by Oktaviani et al., (2024) applying GLM to model the number of claims using the Poisson distribution in health insurance. Likewise, Garrido et al. (2015) also used Poisson for the dependent frequency and severity of car insurance claims. As a result, GLM is able to model the number of claims with several risk factors.

In the research of Naufal et al. (2024), using GLM with a different distribution approach, namely binomial. Conducted to find risk factors that influence mortality rates in Indonesia. The results show that gender is a significant predictor of mortality, while age and smoking status issues have no significant effect on the likelihood of death.

The application of GLM development was carried out by Saputra et al., (2024) who used non-Gaussian distributions using GLM and Generalized Linear Mixed Models (GLMM). The results show that GLMM is designed for non-linear and non-Gaussian distributions, the results are not always better than GLM in its practical application. Hence, the importance of model selection based on specific data characteristics and analysis context.

Rahmawati et al., (2023) used GLM to determine pure premium rates for motor vehicle insurance policies. Modeling claim frequency with Poisson distribution and claim severity with Gamma distribution, while identifying influential characteristics such as distance traveled and geographic zone.

Table 1: Research Gap

Researcher	Research Title	Data Objects	Distribution Used	Generalized Linear Models	Research Gap
Oktaviani et al., (2024)	Determination of Pure Health Insurance Premiums Using Generalized Linear Models (GLM) with Poisson Distribution	Health insurance claims data obtained from Sumit Kumar Shukla	Poisson	Yes	Gamma Distribution is not used
Saputra et al., (2024)	Generalized Linear Mixed Models for Predicting Non-Life Insurance Claims	Ausprivauto0405 (motor vehicles from Australia)	Non-Gaussian	Yes and Generalized Linear Mixed Models	The study was conducted in the vehicles sector, it did not focus on health insurance and Gamma Distribution is not used
Naufal et al., (2024)	Generalized linear model (GLM) to determine life insurance premiums	Life claims data covering various risk factors that influence mortality rates in Indonesia	Binomial	Yes	Gamma Distribution is not used
Rahmawati et al., (2023)	Determining Pure Premium of Motor Vehicle Insurance with Generalized Linear Models (GLM)	Swedish motor vehicle insurance	Poisson and Gamma	Yes	The study was conducted in the vehicles sector, it did not focus on health insurance and not focus on Gamma Distribution
Garrido et al., (2015)	Generalized linear models for dependent frequency and severity of insurance claims	Canadian car insurance data	Poisson	Yes	Gamma Distribution is not used
Aulya (2024)	Estimation of Pure Health Insurance Premiums in Southeast America Using Generalized Linear Model (GLM) with Gamma Distribution	Insurance Premium Charges in Southeast United State	Gamma	Yes	Use Gamma Distribution on Insurance Premium Charges in United State

Table 1 shows previous research on this research. This indicates that there is no research that directly use GLM with Gamma. This study uses GLM with gamma as the distribution to determine the pure premium determination model based on historical data in Southeast United State. The goal is to find out the factors that affect premium pricing and provide recommendations for premium models according to individual risk, which can be implemented by insurance companies to determine premium prices.

2. Literature Review

2.1. Gamma Distribution

The gamma distribution is a probability distribution used to model real-valued measurements that are always positive, such as time and length. For example, if you are the third customer in line for a taxi, this distribution shows that as the line length increases, the average waiting time increases, the uncertainty relative to the average decreases, and the distribution becomes more symmetric. At very large values of the shape parameter, the gamma distribution approaches the normal distribution (Dagpunar, 2019).

Bain and Engelhardt (1992) explain that the gamma distribution has an important role in queuing theory and various measurement problems, such as the waiting time distance to reach a service facility (e.g. bank, toilet, or train station) and the time until the breakdown of components such as spare parts or lamps. The gamma distribution is also widely used in survival data analysis, making it very suitable for estimating the life time of an object or object. Variables such as claim size and annual income are often described by a gamma distribution. Gamma random variables are continuous, non-negative, and skewed to the right (Jong & Heller, 2008).

2.2. Generalized Linear Models

Generalized Linear Models (GLM), introduced by Nelder and Wedderburn in 1972, were designed to handle heterogeneity of characteristics in insurance by modeling systematic differences using explanatory variables. GLMs have become standard in many fields, such as actuarial science for predictive modeling and economics or social science for exploring causal relationships (Wüthrich & Merz, 2023). GLM is an expansion of the linear regression model, where the distribution of the response variable is put into an exponential family to measure the effect of the explanatory variable (X) on the response variable (McCulloch et al., 2011).

In applications to create property insurance rates, response variables in GLMs are represented by distributions that belong to the family of exponential distributions, such as Poisson, Gamma, Binomial, Inverse Gaussian, Normal, and Negative Binomial distributions. Response variables usually consist of claim frequency, claim intensity, pure premium, or loss ratio (Goldburd et al., 2019).

3. Materials and Methods

3.1. Materials

The object of this study is historical health insurance data that contains information on individual risk factors and premiums charged. The data was taken from the kaggle.com website and downloaded on December 13 2024.This data includes individuals of various age groups, Body Mass Index (BMI), sex, number of dependent children and smoking status, with a focus on the Southeast region of the United States.

3.2. Methods

There are two data method, namely:

3.2.1. Gamma Distribution

Taken from (Jong & Heller, 2008), suppose Y is a random sample from a population that follows a Gamma distribution with parameters α and β denoted $G(\alpha, \beta)$. The probability density function for the Gamma distribution is as follows:

$$f(Y|\alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} y^{\alpha-1} e^{-\frac{y}{\beta}}, \tag{1}$$

with,

$f(Y)$: Probability Density Function of the Gamma,
y : amount of insured premiums,
α, β : Gamma distribution parameters,

for $y > 0$, $\alpha > 0$ dan $\beta > 0$ with mean and variance $E(Y) = \alpha\beta$ and $Var(Y) = \alpha\beta^2$.

The Exponential Family distributions and their parameters used in this study are shown in the following table.

Table 2: Gamma distribution and its parameters

Distribution	θ	$b(\theta)$	ϕ	$E(Y)$
Gamma $G = (\mu, v)$	$-\frac{1}{\mu}$	$-\ln(-\theta)$	$\frac{1}{v}$	μ

3.2.2. Generalized Linear Models

The purpose of this Generalized Linear Models (GLM) model is to estimate the response variable (Y) which depends on the explanation of the explanatory variables (X). The variable variable Y that has an exponential family distribution has the following probability function as follows (Jong & Heller, 2008):

$$f(y|\theta,\phi) = exp(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)), y \epsilon S, \quad (2)$$

with,

y : response variable,
θ : canonical parameter,
ϕ : scale parameter,
S : is a subset of the set of natural numbers or real numbers,
$b(\theta)$ and $c(y,\phi)$: known fungtion.

Generalized Linear Models aims to determine the conditional expected value of the response variable using existing observational data (Putra et al., 2021). Parameters will be determined $\beta_1, \beta_2, \ldots, \beta_n$ through the log link function of the explanatory average value (μ_i), which can be written as follows:

$$g(\mu_i) = ln(\mu_i) = \beta_0 + \sum_{j=1}^{k} \beta_K x_{ij}, \quad (3)$$

4. Results and Discussion

4.1. Data Preparation

There are 364 pieces of historical premium price data in Southeast America along with information on age, sex, BMI, number of dependent children, and smoking status. Next, we will detect and remove outliers in the data using Isolation Forest, an ensemble-based machine learning method for anomaly or outlier detection. So that 345 outlier-free data are obtained which will be processed later.

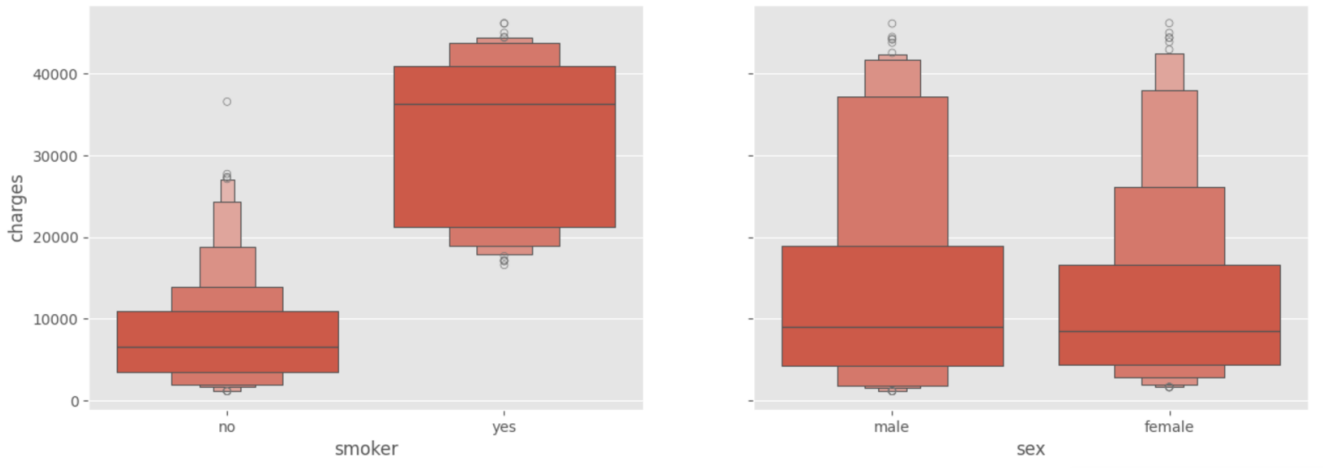

Figure 1: Boxplot Graph

In the case of smokers, the range of premium values is very wide. There are some smokers who pay very high premiums, well above the average. The higher median premium for smoker status indicates that in general, smokers tend to pay higher premiums than non-smokers. This suggests that insurance companies assess higher health risks in smokers.

Whereas in non-smoker status, the range of premium values is narrower. Most non-smokers pay premiums within a more limited range. The median premium is also lower. On average, non-smokers pay lower premiums than smokers. This is consistent with the assumption that the health risk of non-smokers is generally lower.

On the right graph, premiums by sex, the difference is not very significant. In general, both men and women have a similar range of premium values. Thus, smoking is a strong factor affecting premium costs. Figure 2 shows the effect of smoking status coupled with other factors.

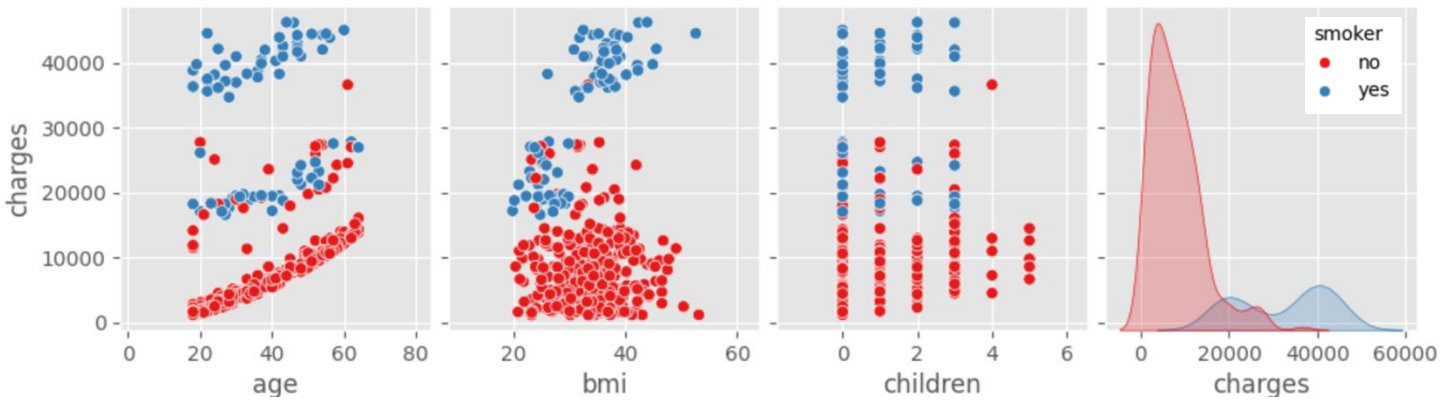

Figure 2: Pairplot of Numerical Features Colored by Smoking Status

If we look at Figure 2 in the age and charges section, the premium cost increases with age. This reflects the greater health risks in older age groups. The blue group (smokers) has significantly higher premium costs than the red group (non-smokers), across all age groups. This difference indicates that smoking is considered a significant risk factor that increases health insurance premiums.

Furthermore, the relationship between BMI and premiums. Individuals with higher BMI coupled with smoking tend to have higher premium costs. This is likely due to the association of high BMI with the risk of health conditions such as obesity or cardiovascular disease.

The relationship between the number of children and premium costs is not as clear or significant. There is some variation in costs, but the overall distribution is flatter, suggesting that the number of children does not have a large impact on health costs.

The non-smoker group (red) has a distribution of premium costs that is more concentrated at low values, reflecting relatively less health risk. The smoker group (blue) shows a wider distribution with higher average premiums. This indicates that smokers are more likely to have serious health conditions that increase their premiums.

Thus, when looking at historical premium price data in Southeast America, the smoker variable has the most significant influence on premiums (charges), followed by age and BMI.

4.2. Gamma Distribution

After conducting data preparation and understanding the data, the next step is to see whether the gamma distribution is suitable for the historical premium data that we have. After trying to match the distribution on the changes variable, there is no distribution that fits the data perfectly, but these five distributions are the most suitable among the others. The top five distributions that fit were obtained and Gamma is included in the top five distributions that fit the data.

Based on the figure 3, the histogram shows that most of the data is concentrated in the lower values, with little data having very high values. This is a typical characteristic of a right-skewed distribution and Gamma is a distribution that is often used to model non-negative, right-skewed data, such as waiting times or insurance claim size.

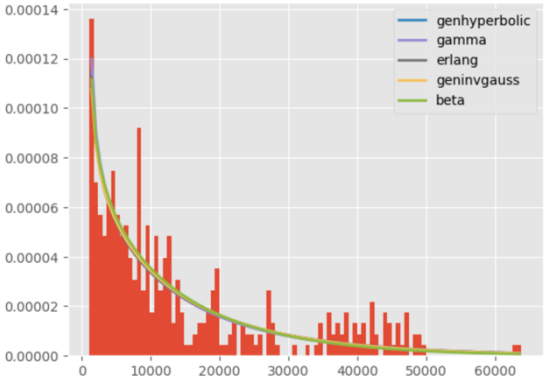

Figure 3: Fitted Distribution

4.3. Generalized Linear Models

Before performing Generalized Linear Models, Table 3 displays the variable information.

Table 3: Variable Information

	Age	**BMI**	**Children**	**Charges**
Count	345	345	345	345
Mean	38.417391	33.148580	1.063768	13090.076999
Std. Deviation	13.888666	6.470869	1.186892	12204.646721
Min	18	19.8	0	1121.8739
25%	26	28.16	0	4340.4409
50%	38	33.33	1	8596.8278
75%	50	37.62	2	18218.16139
Max	64	53.13	5	46200.9851

Table 4: Generalized Linear Model Regression Results

	Coefficient Regression	**Standard Error**	***z***	$P > \|z\|$
Intercept	7.2746	0.254	28.674	0
Age	0.0310	0.003	10.566	0
BMI	0.0076	0.006	1.204	0.228
Children	0.0836	0.034	2.432	0.015
Sex	0.0521	0.081	0.640	0.522
Smoker	1.6270	0.1	16.344	0

The regression coefficient shows the effect of the predictor variable on the response variable. The coefficient for the age variable of 0.0310 means that if a person's age increases by one year, the average cost of insurance will increase by 0.0310 units (in the same units as the charges variable).

The standard error of the regression coefficient measures the uncertainty or variability in the coefficient estimate. The smaller the standard error, the more accurate the coefficient estimate. The standard error for the age variable of 0.003 indicates that the coefficient estimate of 0.0310 has a relatively low level of uncertainty.

Z-score shows how far a coefficient is from zero in standard deviation units. Z-score is used to test the statistical significance of the coefficient. The z-score for the age variable of 10.566 indicates that the age coefficient is highly

statistically significant, meaning that the effect of age on insurance costs is highly significant.

The p-value is the probability of obtaining the least extreme result as observed, assuming the null hypothesis is true (i.e., the true coefficient is zero). A very small p-value (usually less than 0.05) indicates that we can reject the null hypothesis and conclude that the coefficient is statistically significant. The p-value for the age variable of 0.000 indicates that we are very confident that age has a significant influence on insurance costs. So that it produces a model for calculating pure premiums:

$$\boldsymbol{\mu = exp(7.2746 + 0.0310 \cdot age + 0.0076 \cdot bmi + 0.0836 \cdot children + 0.0521 \cdot sex + 1.6270 \cdot smoker)}, \quad (4)$$

with μ is pure premium.

5. Conclusion

In the age factor, the older a person is, the higher the premium. This relationship is highly statistically significant. The effect of BMI on premiums is not statistically significant. The more children, the higher the premium tends to be, although the effect is not very strong. There is no statistically significant difference between male and female premiums. Smokers pay significantly higher insurance premiums than non-smokers. This relationship is highly statistically significant.

Thus, the GLM shows that smoker status is the most significant factor in determining insurance costs, followed by age and number of children. Gender and BMI have no statistically significant effect in this model.

References

Bain, L.J. dan Engelhard B. 1992. Introduction to Probability and Mathematical Statistic Second Edition. Belmant: Duxbury Press.

Dagpunar, John. (2019). The gamma distribution. Significance. The Royal Statistical Society, 16(1):10-11.

Garrido, José., Genest, Christian., Schulz, Juliana. (2015). Generalized linear models for dependent frequency and severity of insurance claims. 1-17.

Goldburd, M. et al. (no date) Casualty Actuarial Society CAS MONOGRAPH SERIES NUMBER 5 Second Edition GENERALIZED LINEAR MODELS FOR INSURANCE RATING Second Edition. Available at: www.casact.org.

Jong, P. and Heller, G. (2008) Generalized Linear Models for Insurance Data. New York.

McCulloch, Searle and Neuhaus (2011) Generalized, Linear, and Mixed Models, 2nd Edition. Second. Wiley.

Naufal, N., Devila, S., Lestari, D. (2019). Generalized linear model (GLM) to determine life insurance premiums. 2168(1):020036-. doi: 10.1063/1.5132463.

Oktaviani, Devani., Zahra, Nabila., Halim, Nurfadhlina Abdul. (2024). Determination of Pure Health Insurance Premiums Using Generalized Linear Models (GLM) with Poisson Distribution. International Journal of Global Operations Research, 5(2):88-92.

Ortaliza, Jared., McGough, M., Salaga, M., Amin, K., Cox, C. (2023). How Much And Why 2024 Premiums Are Expected To Grow In Affordable Care Act Marketplaces. https://jrreport.wordandbrown.com.

Putra, T. A. J., Lesmana, D. C., & Purnaba, I. G. P. (2021). Penghitungan Premi Asuransi Kendaraan Bermotor Menggunakan Generalized Linear Models dengan Distribusi Tweedie. Jambura Journal of Mathematics, 3(2), 115-127.

Rahmawati, T., Susanti, D., & Riaman, R. (2023). Determining Pure Premium of Motor Vehicle Insurance with Generalized Linear Models (GLM). International Journal of Quantitative Research and Modeling, 4(4), 207-214.

Saputra, Kie Van Ivanky., Margaretha, Helena., Ferdinand, Ferry Vincenttius., Budhyanto, Johana Daniella. (2024). Generalized Linear Mixed Models for Predicting Non-Life Insurance Claims. Inferensi: Jurnal Statistika, 7(2):141-141.

Wüthrich, M., Merz, M. (2023). Generalized Linear Models. Statistical Foundations of Actuarial Learning and its Applications, Springer Actuarial. 111-205.

Xin, Chen., Aleksandr, Y., Aravkin., R., Douglas, Martin. (2018). Generalized Linear Model for Gamma Distributed Variables via Elastic Net Regularization. arXiv: Methodology,

Yi, Yang., Wei, Qian., Hui, Zou. (2018). Insurance Premium Prediction via Gradient Tree-Boosted Tweedie Compound Poisson Models. Journal of Business & Economic Statistics, 36(3):456-470.