BAREKENG: Journal of Mathematics and Its Application
December 2022 Volume 16 Issue 4 Page 1399–1410
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss4pp1399-1410

POISSON REGRESSION MODELING OF AUTOMOBILE INSURANCE USING R

Sandy Vantika[1], Mokhammad Ridwan Yudhanegara[2], Karunia Eka Lestari[3*]

[1]Statistics Research Division, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung
Jl. Ganesa No. 10 Coblong, Bandung, 40132, Indonesia

[2,3]Departement of Mathematics Education, Faculty of Teacher Training and Education,
Universitas Singaperbangsa Karawang
Jl. H. S. Ronggowaluyo Teluk Jambe Timur, Karawang, 41361, Indonesia

Corresponding author's e-mail: [3] karunia@fkip.unsika.ac.id*

Abstract. *Automobile insurance benefits are protecting the vehicle and minimizing customer losses. Insurance companies must provide funds to pay customer claims if a claim occurs. Insurance claims can be modeled by Poisson regression. Poisson regression is used to analyze the count data with Poisson distributed data responses. In this paper, the data model of the sample is automobile insurance claims from the companies in one year (in 2021) of observation, which contains three types of insurance products, i.e., Total Loss Only (TLO), All Risk, and Comprehensive. The results of data analysis show that the highest number of claims comes from Comprehensive insurance products, especially if the premium value gets more extensive. In contrast, the least comes from TLO insurance products.*

Keywords: *count data, insurance claim, poisson distribution, prediction.*

Article info:

Submitted: 29th July 2022 *Accepted: 23rd October 2022*

How to cite this article:
S. Vantika, M. R. Yudhanegara and K. E. Lestari, "POISSON REGRESSION MODELING OF AUTOMOBILE INSURANCE USING R", *BAREKENG: J. Math. & App.*, vol. 16, iss. 4, pp. 1399-1410, Dec., 2022..

1. INTRODUCTION

Insurance is a form of risk management to reduce financial losses by transferring risk from one party to another, in this case, the insurance company. Insurance is an agreement between two parties, the insurance company and the customer, which is the basis for receiving premiums by the insurance company in return for providing reimbursement to the customer. The compensation is due to losses, damages, costs incurred, lost profits, or legal liability to third parties that may be suffered by the insured or the customer due to an uncertain event [1].

Human needs develop along with the times. Vehicles that used to be tertiary can now become secondary or even primary needs. The number of unwanted uncertainties that can occur in everyday life, such as accidents, crime, riots, natural disasters, and so on, raises concerns for some people to choose to buy motor vehicle insurance to minimize losses and protect their vehicles.

Each customer is required to pay a certain amount of money, called a premium, to the insurance company so that the risk of customer loss in the future is now the responsibility of the insurance company based on the applicable policy. Therefore, insurance companies must provide funds ready to pay claims submitted by customers. So that these funds can always handle claims from customers, insurance companies need to know the prediction of the number of shares offered. One way to model insurance claims is by Poisson regression [2].

The phenomenon in Indonesia, a paper that discusses modeling the number of automobile insurance claims based on premiums using Poisson regression, has yet to be found. Modeling using Poisson regression is easy to apply. Many studies have been found on automobile insurance modeling with a reasonably complicated method, so companies are reluctant to use it, see [3], [4].

The advantage of this research is the implementation of data modeling on the number of claims in a simple form using Poisson regression practically with the types of insurance products used in insurance companies in Indonesia. This research aims to model the association between the amount of premium and the number of claims per year based on the types of insurance products.

2. RESEARCH METHODS

Poisson regression is often used to analyze count data; in this case, the response of the data is the Poisson distribution with parameters [5]. This parameter highly depends on some particular unit or period, distance, area, volume, etc. This distribution is then used to model an event with relatively rare or rare to occur in specific units. For example, an event occurs randomly and uniformly in a particular time or area. The event occurs with a known average; for example, suppose a random variable represents the number of events in a certain period (region). For the record, the time interval or area here has specific units. Examples of time intervals in hours/days/months/years or regions in the form of a particular line/area/volume or maybe pieces of material [6].

Winkelmann [7] said that Poisson regression is a simple and robust regression model for count data. In general, the model is expressed by Equation (1) below.

$$\ln Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \quad (1)$$

where Y is a discrete-valued response variable with y_i a discrete value ($y_i = 0,1,2,\cdots$). Meanwhile, the explanatory variable is a random vector $\mathbf{X} = (X_1, X_2, \cdots, X_k)^t$, which is continuous, dichotomous, or ordinal. The first stage of the Poisson regression analysis is to determine the assumptions. The assumptions of the Poisson regression analysis include $Y \sim Poi(\lambda)$: the response variable Y has a Poisson distribution with mean $\mu = \lambda$, the model estimate is $E[\vec{y}] = \lambda = \exp\left(\mathbf{X}_i\vec{\boldsymbol{\beta}}\right)$. It means that the relationship between Y and X_1, X_2 to X_k can be determined functionally through an ln-linear equation. Interestingly, this Poisson distribution's expected value is the same as the variance value, so $Var[y_i] = \lambda = \exp\left(\mathbf{X}_i\vec{\boldsymbol{\beta}}\right)$.

Since the distribution of Y is known or assumed to have a Poisson distribution, we use the maximum likelihood estimation method to estimate the regression parameters [8]. The assessment is done by

maximizing the likelihood function. Note that Y has a Poisson distribution with the parameter $\lambda = \exp\left(\mathbf{X}_i\vec{\boldsymbol{\beta}}\right)$, so the conditional probability function of Y is as follows:

$$P\left(Y = y_i \middle| \mathbf{X}_i, \vec{\boldsymbol{\beta}}\right) = \frac{e^{-\exp\left(\mathrm{X}_i\vec{\beta}\right)}\left(\exp\left(\mathbf{X}_i\vec{\boldsymbol{\beta}}\right)\right)^{y_i}}{y_i!} \tag{2}$$

We have the likelihood function as the product of all the following conditional probability functions,

$$L\left(\vec{\boldsymbol{\beta}}; \vec{y}, \mathbf{X}\right) = \prod_{i=1}^{n} \frac{e^{-\exp\left(\mathbf{X}_i\vec{\boldsymbol{\beta}}\right)}\left(\exp\left(\mathbf{X}_i\vec{\boldsymbol{\beta}}\right)\right)^{y_i}}{y_i!} \tag{3}$$

So we get the following ln-likelihood function,

$$\ell\left(\vec{\boldsymbol{\beta}}\right) = \sum_{i=1}^{n} y_i\mathbf{X}_i\vec{\boldsymbol{\beta}} - \sum_{i=1}^{n} \exp\left(\mathbf{X}_i\vec{\boldsymbol{\beta}}\right) - \sum_{i=1}^{n} \ln y_i! \tag{4}$$

The value $\vec{\boldsymbol{\beta}}$ is obtained from the derivative of the ln-likelihood function,

$$\frac{\partial\ell\left(\vec{\boldsymbol{\beta}}\right)}{\partial\vec{\boldsymbol{\beta}}} = \sum_{i=1}^{n} \mathbf{X}_i{}^t\left[\vec{y} - \exp\left(\mathbf{X}_i\vec{\boldsymbol{\beta}}\right)\right] = 0 \tag{5}$$

and

$$\frac{\partial^2\ell\left(\vec{\boldsymbol{\beta}}\right)}{\partial\vec{\boldsymbol{\beta}}\partial\vec{\boldsymbol{\beta}}'} = -\sum_{i=1}^{n} \exp\left(\mathbf{X}_i\vec{\boldsymbol{\beta}}\right)\mathbf{X}_i{}^t\mathbf{X}_i = \mathbf{H}\left(\vec{\boldsymbol{\beta}}\right) \tag{6}$$

The result of the derivative in Equation (6) is known as the Hessian matrix [9]. Since the entries of this Hessian matrix are exponential or non-linear, then to determine the estimated parameter $\widehat{\boldsymbol{\beta}}_{t+1}$, we use the Newton-Rapson iteration method as follows,

$$\widehat{\boldsymbol{\beta}}_{t+1} = \widehat{\boldsymbol{\beta}}_t - \mathbf{H}\left(\widehat{\boldsymbol{\beta}}_t\right)^{-1}\left(\sum_{i=1}^{n} \mathbf{X}_i{}^t\left[\vec{y} - \exp\left(\mathbf{X}_i\vec{\boldsymbol{\beta}}\right)\right]\right) \tag{7}$$

where iteration $t = 0,1,2,\cdots$.

The next step is to perform a regression parameter test procedure. The following is the regression parameter test procedure for simultaneous testing with the null hypothesis $\mathrm{H}_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ and the alternative hypothesis $\mathrm{H}_1: \exists\beta_j \neq 0,\ j = 1,2,\cdots,k$. The H_0 here states that $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ or not significant. So that is significant only β_0. In other words, a model without explanatory variables fits our data and our model. Meanwhile, H_1 states that there is at least one significant regression coefficient. It shows that the model we build fits the data compared to the model that only involves the intercept or β_0. The test statistic uses the likelihood ratio of a simple model without involving explanatory variables with a complete model involving explanatory variables. The test criteria, reject H_0 if $\lambda < \lambda_{\alpha,\nu_1,\nu_2}$, where $\nu_1 = k$ and $\nu_2 = n - (k+1)$. Alternatively, we can use the following G-test likelihood ratio statistic; this test is carried out with the chi-square approximation,

$$G = -2\ln\left[\frac{L(\widehat{\omega})}{L(\widehat{\Omega})}\right] \tag{8}$$

with reject criteria H_0 if $G > \chi^2_{\nu,\alpha}$, where $\nu = k - 1$.

The next step is to perform a partial regression parameter test. When the simultaneous test H_0 is rejected, this indicates that at least one explanatory variable is significant in predicting the response variable. Further analysis is needed, and we may be interested in determining each explanatory variable's effect on the response variable. Therefore, a partial test was conducted with the null hypothesis $\mathrm{H}_0: \beta_j = 0$ and $\mathrm{H}_1: \beta_j \neq$

0 for $j = 1,2 \cdots, k$. The H_0 states that $\beta_j = 0$ for $j = 1,2 \cdots k$, which means that β_j is not significant. Meanwhile, H_1 says $\beta_j \neq 0$, which means the effect is significant. For the test statistics using t-test as follows,

$$t = \frac{\hat{\beta}_j}{SE\left(\hat{\beta}_j\right)} \tag{9}$$

Where

$$SE\left(\hat{\beta}_j\right) = \sqrt{\left(\sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i^t\right)^{-1}} \tag{10}$$

with reject criteria H_0 if $|t| > t_{(\alpha/2,\nu)}$ where $\nu = n - (k+1)$.

The last stage is to perform a regression model evaluation procedure. After getting the model we are looking for, we evaluate whether the model represents the relationship between the response variable and the explanatory variable. The goodness of fit test uses the chi-square test when assessing the model. In this test, we evaluate whether the sample comes from a population with a Poisson distribution.

Previously we knew that in this Poisson distribution, the mean value of Y is equal to the variance or equidispersion. Therefore, we need to check whether this assumption is met. It can be seen from the phi dispersion parameter. This dispersion parameter is the ratio between the dispersion value and the degree of freedom. The dispersion value is obtained through the following G^2 statistic,

$$G^2 = 2\sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{\lambda_i}\right) \tag{11}$$

It's taking into account the value of ϕ as

$$\phi = \frac{G^2}{n-k} \tag{12}$$

with criteria if $\phi > 0$ indicates overdispersion, while if $\phi < 0$ indicates underdispersion occurs.

Alternatively, the test uses the chi-square test as follows.

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(y_i - \hat{\lambda}_i\right)^2}{\hat{\lambda}_i}, \tag{13}$$

where

$$\phi_{alt} = \frac{\chi^2}{n-k}. \tag{14}$$

The criteria, if $\phi_{alt} > 1$ indicates overdispersion, while if $\phi_{alt} < 1$ indicates underdispersion occurs. Another criterion is to determine the coefficient of determination, R-square with the following equation,

$$R^2_{dev} = \frac{\sum_{i=1}^{n} y_i \ln\left(\frac{\hat{\lambda}_i}{\bar{y}}\right) - \left(y_i - \hat{\lambda}_i\right)}{\sum_{i=1}^{n} y_i \ln\left(\frac{\hat{y}_i}{\bar{y}}\right)}. \tag{15}$$

The R-square represents the proportion of variability in Y that explanatory variables can explain by the given model. So, the more significant the R-square value, the better the model. We also look at the AIC (Akaike Information Criterion) value,

$$\text{AIC} = -2\ln L(\hat{\Omega}) + 2k. \quad (16)$$

The AIC deals with the trade-off between fit and model simplicity. If we are faced with several choices of models, choose the best model with the smallest AIC value. Next, we also have to pay attention to the value of VIF (Variance Inflation Factor),

$$\text{VIF} = \frac{1}{1 - R_k^2}. \quad (17)$$

The VIF value is used to detect the presence or absence of multicollinearity between the explanatory variables. Multicollinearity itself states a condition where two or more explanatory variables have a high or linear solid relationship. The consequence if a model contains multicollinearity is that the variance will continue to increase so that the standard error of the parameters also increases. Multicollinearity can be detected from the VIF. The value is more than 10 indicates the presence of multicollinearity. In the other, limit the VIF value to more than 5.

3. RESULTS AND DISCUSSION

3.1. Data Description

The software used to process the data using R. The program script using R is available in the attachment, see also Indratno [10]. The data used in this study are sample data on automobile insurance claims in the companies for one year (in 2021) of observation in Indonesia, see [11]. The information has four variables: identity (id), many claims, types of insurance products, and the value of insurance premiums. The id column contains the unique number of each policyholder. The premium value contained in this data is denominated in USD and observed every month. This data product consists of three types: Total Loss Only (TLO), Comprehensive, and All Risk. TLO is an insurance product that guarantees losses due to loss or damage that causes the vehicle to not function or the value of vehicle repairs reaches 75% due to accidents. Comprehensive is an insurance product caused by vehicle accidents such as abrasions dents, up to significant damage. All Risk is an insurance product that is a complete package of comprehensive automobile insurance, which is added with protections such as driver accidents, floods, earthquakes, terrorism, and liability to third parties.

The variables of Y, X_1, and X_2 is variable of customer that has insurance product. Then, X_3 is a variable that states the customer's premium amount. The response variable Y was chosen from these data, namely the number of automobile insurance claims in the company for one year in 2021. There are two explanatory variables in the model, namely X_1 and X_2, which states the type of insurance product (nominal data), and X_3, which displays the premium amount. The data summary o is shown in Figure 1.

```
> summary(p)
       id          claim             product       premium     
 1      :  1   Min.   :0.00   TLO          : 45   Min.   :33.00  
 2      :  1   1st Qu.:0.00   Comprehensive:105   1st Qu.:45.00  
 3      :  1   Median :0.00   AllRisk      : 50   Median :52.00  
 4      :  1   Mean   :0.63                       Mean   :52.65  
 5      :  1   3rd Qu.:1.00                       3rd Qu.:59.00  
 6      :  1   Max.   :6.00                       Max.   :75.00  
 (Other):194
```

Figure 1. Data Summary

Figure 1 provides information that there are 200 observations or customers. The data in the claims column has a minimum value 0, the first quartile is 1, and the median is 0. Furthermore, the mean is 0.63, the third quartile is 1, and the maximum value is 6. So, in a year, customers can make a maximum of 6 claims. For the insurance product column, 45 customers have TLO products, 105 have Comprehensive products, and the rest have All Risk products. So, most customers choose Comprehensive products. For the premium large column, the minimum value is \$33. Then, the value of the first quartile is \$45, the median is \$52, the mean is \$52.65, the third quartile is \$59, and the maximum value is \$75.

In Figure 2, it can be seen that the mean number of claims by product varies for each product. Thus, the product is a good candidate explanatory variable for predicting multiple claims. Meanwhile, the mean

and variance in each product have almost the same value, so it can be assumed that the data has a Poisson distribution to be modeled using Poisson regression.

```
> with(p, tapply(claim, product, function(x) {
+   sprintf("M (var) = %1.2f (%1.2f)", mean(x), var(x))
+ }))
                     TLO            Comprehensive                  AllRisk
"M (var) = 0.20 (0.16)" "M (var) = 1.00 (1.63)" "M (var) = 0.24 (0.27)"
```

Figure 2. Output Mean and Variance of Each Product

The data histogram of the number of claims for each insurance product is shown in Figure 3.

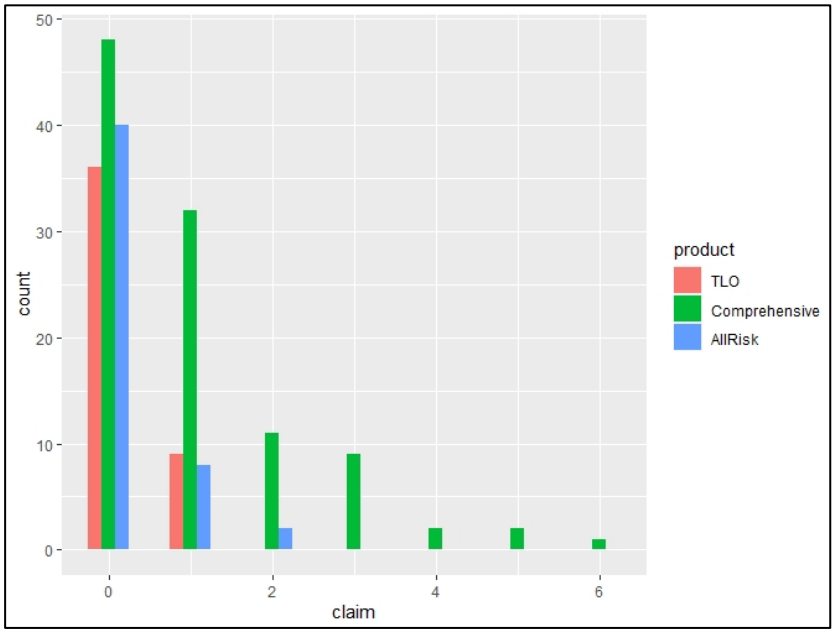

Figure 3. Hostogram Data

3.2. The Best Model

From the histogram in Figure 3, it can be seen that it has a shape like a Poisson distribution. In addition, customers with Comprehensive insurance products make the most claims for every many claims. Customers with TLO insurance products only submit at most one claim in 1 year of observation, while customers with All Risk insurance products submit two claims. Customers with Comprehensive insurance products submit six claims in one year of observation.

Furthermore, based on Figure 4, we have estimated coefficients of the Poisson regression model. The Poisson regression model is represented at Equation (18),

$$\ln(y) = -5.24712 + 1.08386x_1 + 0.36981x_2 + 0.07015x_3. \qquad (18)$$

Based on Equation (18), x_1 is variable of customer that has comprehensive insurance product (value 1 if the customer has comprehensive insurance product, value 0 if the customer has other insurance product). Next, x_2 is variable of customer with All Risk insurance product (value 1 if the customer has All Risk insurance product, value 0 if the customer has other insurance product). Then, x_3 is variable that states the customer's premium amount. The TLO insurance product is used as a reference category contained in the intercept.

```
> summary(y1 <- glm(claim ~ product + premium, family="poisson", data=p))

Call:
glm(formula = claim ~ product + premium, family = "poisson",
    data = p)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2043  -0.8436  -0.5106   0.2558   2.6796

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -5.24712    0.65845  -7.969 1.60e-15 ***
productComprehensive  1.08386    0.35825   3.025  0.00248 **
productAllRisk        0.36981    0.44107   0.838  0.40179
premium               0.07015    0.01060   6.619 3.63e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 287.67  on 199  degrees of freedom
Residual deviance: 189.45  on 196  degrees of freedom
AIC: 373.5

Number of Fisher Scoring iterations: 6
```

Figure 4. The Output of Poisson Regression Model Using GLM (Generalized Linier Model)

Figure 4 can also be seen in the column Pr ($> |z|$), which displays the value of the test statistic and the sign *, which indicates that the estimated value of the coefficient is significant with a particular significance level. The more signs *, the better the estimated coefficient. Furthermore, a description of the residual deviation is also presented. It can be seen that the minimum value is -2.2043, and the maximum is 2.6796, while the median is -0.5106. Therefore, it can be assumed that the distribution is not symmetrical or the residual deviation does not approach the normal distribution. In a case like this, Cameron and Trivedi [2] suggest using a robust model; in this case, it can be seen in Figure 5.

```
> r.est
                      Estimate  Robust SE     Pr(>|z|)          LL          UL
(Intercept)         -5.2471244 0.64599839 4.566630e-16 -6.51328124 -3.98096756
productComprehensive  1.0838591 0.32104816 7.354745e-04  0.45460476  1.71311353
productAllRisk        0.3698092 0.40041731 3.557157e-01 -0.41500870  1.15462716
premium               0.0701524 0.01043516 1.783975e-11  0.04969947  0.09060532
```

Figure 5. The Output of Poisson Regression Model Using Robust Model

The regression coefficient for TLO insurance products is used as a reference category whose effects are combined with the intercept. Reference categories in R are selected alphabetically (earliest or last). Based on Figure 5, the coefficient of the regression model is five decimal places behind the comma. In part marked in orange, which is the standard error value of the coefficient estimate using the robust model, it can be seen that the standard error value is smaller than the previous modeling with GLM. It can happen because the robust model estimates parameters by minimizing errors, so it is relatively robust than the classical model.

After obtaining the candidate model, it is sometimes necessary to consider other models to get the best model to predict the response variable, such as considering a model without an intercept which can be seen in Figure 6. In this case, we will model/predict the number of claims only based on the product and the number of insurance premiums without considering other factors. Therefore, the same analysis is carried out by assuming the model is y_0. Furthermore, compared with the previous results, the parameter estimation result automatically shows that parameter estimation appears for the type of TLO insurance product but without intercept. The following is a model with an intercept (y_0) shown in Figure 6.

```
> summary(y0 <- glm(claim ~ 0 +product + premium, family="poisson", data=p))

Call:
glm(formula = claim ~ 0 + product + premium, family = "poisson",
    data = p)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2043  -0.8436  -0.5106   0.2558   2.6796

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
productTLO          -5.24712    0.65845  -7.969 1.60e-15 ***
productComprehensive -4.16327    0.66288  -6.281 3.37e-10 ***
productAllRisk      -4.87732    0.62818  -7.764 8.21e-15 ***
premium              0.07015    0.01060   6.619 3.63e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 319.24  on 200  degrees of freedom
Residual deviance: 189.45  on 196  degrees of freedom
AIC: 373.5

Number of Fisher Scoring iterations: 6
```

Figure 6. The Output of Poisson Regression Model without Intercept (y_0)

Based on Figure 4 and Figure 6, the parameter estimates the coefficients of the types of Comprehensive and All Risk insurance products, the value changes between the model without intercept and the model with intercept. From these two models, it can be seen that the intercept is the estimated effect for the reference category and the coefficient of the other category is the deviation for that effect. Then when compared to the standard error value, for the model without an intercept, the value is greater than the model with an intercept. Therefore, it is necessary to consider other models.

Another alternative model is without a variable type of insurance product, so the number of claims is predicted only based on the size of the premium. For that, define y_2 as the response variable of the model without the insurance product type variable (in Figure 7). Model 2 is the model involving products and premiums previously defined as y_1 (in Figure 4).

```
> anova(y2, y1, test="Chisq")
Analysis of Deviance Table

Model 1: claim ~ premium
Model 2: claim ~ product + premium
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       198     204.02
2       196     189.45  2   14.572 0.0006852 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> AIC(y2)
[1] 384.0762
```

Figure 7. The Alternative Model without Insurance Product Type (y_2).

Based on Figure 6 and Figure 7, if we look at the AIC value, the AIC value of Model y_2 is greater than the AIC value of Model y_1. Thus, the model with a better intercept. Therefore, there are three candidate models to predict the number of claims. It is y_0, which is a model without an intercept. Then, y_1 is a model that involves an intercept and explanatory variables consist of the type of insurance product and the amount of premium. The last is y_2, which is a model that involves intercept and premium. From the three models, when viewed from the standard error and AIC value, it is concluded that the y_1 is the best model.

3.3. Model Evaluation

Previously, the best model choice obtained was y_1. The model is then evaluated. The first step is to test the model's fit using the chi-square test, as shown in Figure 8. The chi-square test is carried out with the null hypothesis being the model fits, and the counter hypothesis is that the selected model does not match the data held based on a particular significance level.

```
> with(y1, cbind(res.deviance = deviance, df = df.residual,
+                p = pchisq(deviance, df.residual, lower.tail=FALSE)))
     res.deviance  df         p
[1,]     189.4496 196 0.6182274
```

Figure 8. Output of Goodness of Fit Test

Figure 8 shows that the $p - value$ = 0.6182274, which if the 5% significance level is used, then the $p - value > \alpha$. Therefore, H_0 is not rejected. So, based on the chi-square test with a significance level of 5%, the model chosen is suitable to describe the data held. The overdispersion test was carried out, with the null hypothesis being that there was no overdispersion in the model. The counter hypothesis was overdispersion in the model, as shown in Figure 9.

```
> library(AER)
> update.packages("AER")
> dispersiontest(y1,trafo=1)

        Overdispersion test

data:  y1
z = 0.53224, p-value = 0.2973
alternative hypothesis: true alpha is greater than 0
sample estimates:
     alpha
0.04725442
```

Figure 9. Output of Overdispersion Test

From Figure 9, it can be seen that the $p - value$ = 0.2973, which if the 5% significance level is used, the $p - value > \alpha$, so H_0 is not rejected. It means that the mean is the same as the variance, or there is no overdispersion. Next, it will be checked for multicollinearity. Based on Figure 10, the GVIF value of the model is less than 10, so it can be concluded that the multicollinearity is low.

```
> vif(y1)
             GVIF Df GVIF^(1/(2*Df))
product 1.153435  2        1.036331
premium 1.153435  1        1.073981
```

Figure 10. GVIF Value for Multicollinearity

Then the coefficient of determination is also determined from this model—furthermore, the results are obtained in Figure 11.

```
> library(rsq)
> update.packages("rsq")
> rsq(y1)
[1] 0.3154051
```

Figure 11. Coefficient of Determination

There is no standard rule regarding what percentage of variability can be explained in the model, but the more significant the percentage, the better the model. Based on Figure 11, information is obtained that 31.54% of the variability in the number of claims (Y) can be explained by using the product (X_1 and X_2) and claim size (X_3). In comparison, the remaining 68.46% of the variability in Y can be explained by other factors that are not observed or included in the model.

3.4. Prediction

After obtaining the best model, it will predict the expected value of y for a specific explanatory variable if the other variables have a fixed value. In this case, we want to know the average number of claims for a specific insurance product if the premium value is the average of the premiums. Previously, it was known that the average premium was $52.645. The syntax shown in Figure 12 is used to solve this problem.

Then, the predictive function of the mean y value and standard error for each product is determined if the average premium is 52.645. Next, the predict function is used, whose results are described in Figure 12.

```
> (s1 <- data.frame(premium = mean(p$premium),
+                   product = factor(1:3, levels = 1:3, labels = levels(p$product))))
  premium       product
1  52.645           TLO
2  52.645 Comprehensive
3  52.645       AllRisk
> predict(y1, s1, type="response", se.fit=TRUE)
$fit
        1         2         3
0.2114109 0.6249446 0.3060086

$se.fit
         1          2          3
0.07050108 0.08628117 0.08833706

$residual.scale
[1] 1
```

Figure 12. Output of Predict Function

From these results, it can be seen that based on predictions, the average number of claims for All Risk products is $e^{0.3060086} = 1.358$, for Comprehensive products is $e^{0.6249446} = 1.868$, and for TLO products is $e^{0.2114109} = 1.235$, if the premium amount is \$52.645.

3.5. The Interpretation

Next, the regression plot is described. The data are sorted by product type and premium. The plot is presented in Figure 13. The x-axis represents the premium amount; the y-axis represents the value of $\hat{y}$ obtained from the y_1 model, which is the expected number of claims.

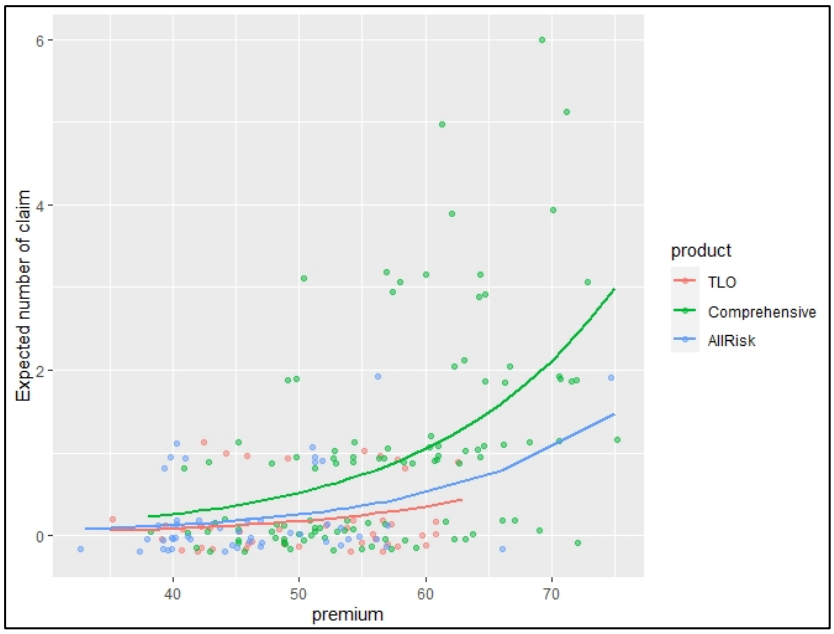

Figure 13. Plot of Expected Number of Claims

From Figure 13, it can be seen that the expectation of the most significant number of claims comes from the type of Comprehensive insurance product, especially if the premium value is getting bigger. Meanwhile, most of the claims came from TLO insurance products, at least.

4. CONCLUSIONS

Based on the data analysis, it can be concluded that the best model to describe and to predict the data is Equation (18). Based on Equation (18), x_1 is variable of customer that has Comprehensive insurance product. Next, x_2 is variable of customer with All Risk insurance product. Then, x_3 is variable that states the customer's premium amount. The predicted mean number of claims for each type of insurance product in the next one year of observation is one claim for TLO and All Risk insurance products and two claims for Comprehensive insurance products.

We will use the clustering method to get the best prediction model in future research [12]–[15]. It also compares the prediction results between the Poisson regression and the predictive distribution methods [16] and statistical inference for online decision [17].

AKNOWLEDGEMENT

The authors would like to thank PPMI ITB 2022 for research funding.

REFERENCES

[1] European Insurance and Occupational Pensions Authority, *Open Insurance: Accessing and Sharing Insurance-Related.* Luxembourg: Publications Office of the European Union, 2021, doi: 10.2854/013491.
[2] C. A. Colin and T. Pravin, *Regression Analysis of Count Data,* 2nd ed. Cambridge: Cambridge University Press, 2013, doi: 10.1017/CBO9781139013567.
[3] M. Karim and A. K. Mutaqin, "Modeling Claim Frequency in Indonesia Auto Insurance Using Generalized Poisson-Lindley Linear Model," *J. Mat. Stat. dan Komputasi*, vol. 16, no. 3, pp. 428-439, 2020, doi: 10.20956/jmsk.v16i3.9315.
[4] J. S. K. Chan, S. T. B. Choy, U. Makov, A. Shamir, and V. Shapovalov, "Variable Selection Algorithm for a Mixture of Poisson Regression for Handling Overdispersion in Claims Frequency Modeling Using Telematics Car Driving Data," *Risks*, vol. 10, no. 4, p. 83, 2022, doi: 10.3390/risks10040083.
[5] J. M. Hilbe, *Modeling Count Data*. Cambridge: Cambridge University Press, 2014, doi: 10.1017/CBO9781139236065.
[6] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability & Statistics for Engineers & Scientists*, 6th ed. Hoboken: Pearson, 2012.
[7] R. Winkelmann, *Econometric Analysis of Count Data,* 5th ed. Berlin: Springer, 2008, doi: 10.1007/978-3-540-78389-3.
[8] A. C. Rencher and W. F. Christensen, *Methods of Multivariate Analysis,* 3rd ed. Hoboken: John Wiley and Sons, 2012.
[9] L. Zhang, W. Zhang, Y. Li, J. Sun, and C. X. Wang, "Standard Condition Number of Hessian Matrix for Neural Networks," *IEEE Int. Conf. Commun.*, vol. 2019-May, pp. 1–6, 2019, doi: 10.1109/ICC.2019.8761740.
[10] S. W. Indratno, "Asuransi Kendaraan Bermotor," *Rpubs*. 2018, [Online]. Available: https://rpubs.com/saptoWI/401535.
[11] R. D. Kurnia, "16 Produk Asuransi Mobil Terbaik dan Terpercaya di," *Qoala*. 2022, [Online]. Available: https://www.qoala.app/id/blog/asuransi/mobil/daftar-asuransi-mobil-terbaik/
[12] M. R. Yudhanegara and K. E. Lestari, "Clustering for multi-dimensional data set: A case study on educational data," *J. Phys. Conf. Ser.*, vol. 1280, no. 4, p. 042025, 2019, doi: 10.1088/1742-6596/1280/4/042025.
[13] M. R. Yudhanegara, S. W. Indratno, and R. R. K. N. Sari, "Clustering for Items Distribution Network," *J. Phys. Conf. Ser.*, vol. 1496, no. 1, p. 012019, 2020, doi: 10.1088/1742-6596/1496/1/012019.
[14] M. R. Yudhanegara, S. W. Indratno, and R. K. N. Sari, "Clustering for Item Delivery Using Rule-K-Means," *J. Indones. Math. Soc.*, vol. 26, no. 02, pp. 185–191, 2020, doi: 10.22342/jims.26.2.871.185-191.
[15] M. R. Yudhanegara, S. W. Indratno, and R. K. N. Sari, "Dynamic items delivery network: prediction and clustering," *Heliyon*, vol. 7, no. 5, p. e06934, 2021, doi: 10.1016/j.heliyon.2021.e06934.
[16] M. R. Yudhanegara, S. W. Indratno, and K. N. Sari, "Role of clustering method in items delivery optimization," *J. Phys. Conf. Ser.*, vol. 2084, no. 1, p. 012011, 2021, doi: 10.1088/1742-6596/2084/1/012011.
[17] S. W. Indratno, K. N. Sari, and M. R. Yudhanegara, "Optimization in Item Delivery as Risk Management: Multinomial Case Using the New Method of Statistical Inference for Online Decision," *Risks*, vol. 10, no. 6, p. 122, 2022, doi: 10.3390/risks10060122.