Safari-Katesari, Hadi; Zaroudi, Samira

Article

Count copula regression model using generalized beta distribution of the second kind

Statistics in Transition New Series

Provided in Cooperation with:
Polish Statistical Association

Suggested Citation: Safari-Katesari, Hadi; Zaroudi, Samira (2020) : Count copula regression model using generalized beta distribution of the second kind, Statistics in Transition New Series, ISSN 2450-0291, Exeley, New York, Vol. 21, Iss. 2, pp. 1-12, https://doi.org/10.21307/stattrans-2020-011

This Version is available at:
https://hdl.handle.net/10419/236761

STATISTICS IN TRANSITION new series, June 2020
Vol. 21, No. 2, pp. 1–12, DOI 10.21307/stattrans-2020-011
Submitted – 13.04.2019; paper accepted for published – 10.03.2020

Count copula regression model using generalized beta distribution of the second kind

Hadi Safari-Katesari[1], **Samira Zaroudi**[2]

ABSTRACT

Modelling claims severity for obtaining insurance premium is one of the major concerns of the insurance industry. There is a considerable amount of literature on the actuarial application of the copula model to calculate the pure premium. In this paper, we model claims severity for computing the pure premium in the collision market by means of the count copula model. Moreover, we apply a regression model using a generalized beta distribution of the second kind (GB2) to compute the premium for an average claim and the conditional computation for all coverage levels. Like many other researchers, we assume that the number of accidents is independent from the size of claims. For real data application, we use a portfolio of a major automobile insurer in Iran in 2007-2008, with a subsample of 59,547 policies available in their portfolio. We then proceed to compare the estimated premiums with the real premiums. The results demonstrate that there is strong positive dependency between the real premium and the estimated one.

Key words: count copula, GB2 regression, pure premium, collision insurance.

1. Introduction

Premium is the payment that a policyholder pays for buying full or partial insurance coverage versus a specified risk. Premium ratemaking is a vital subject to balancing insurance payments (Zhang et al., 2015). In confronting with financial outcomes of the random phenomenon, insurance plays the role of supporting policyholders. It includes the accumulation of a big bunch of policyholder risks such that, within a given time cycle, a number of insurance claims and an accumulated loss to the insurer can be determined. Nowadays, estimating premium plays a pivotal role for insurance companies in the competitive markets. The biased computation may lead to losing the market share and confronting ruin. There is a range of works in this field such as Weisberg (1982), David (2015), Marton et al. (2015), Zhang et al. (2015), Schirmacher (2016), Yang et al. (2017), Shi and Yang (2018), Lesmana et al. (2018), Wolny-Dominiak et al. (2018) and Avanzi et al. (2019). However, using the copula model in ratemaking and actuarial application is to some extent new. Frees et al. (2013) used a multivariate two-part regression model such that the correlation ratio and copula regression for the claims and severity modelling were considered, respectively.

[1]Corresponding Author: Department of Mathematics, Southern Illinois University, Carbondale, IL 62901-4408, USA. E-mail: hadi.safari@siu.edu. . ORCID: https://orcid.org/0000-0003-2630-3133
[2]Department of Mathematics, Southern Illinois University, Carbondale, IL 62901-4408, USA. ORCID: https://orcid.org/0000-0001-8290-6137

For more information, one can refer to Shi (2016). We cannot distinguish risky policyholders beforehand but the severity and the number of claims in a portfolio of an insurance company are predictable. In this paper, our aim is to calculate pure premium by using the insured coverage selection preferences and the number of accidents during the policy period. For this goal, the authors use a generalized beta distribution of the second kind (GB2) regression model to model the average claims for each level of coverage preferences. In the actuarial literature, the assumption between the number of accidents and the size of the claim is common, which is used here as well. The wide variety application of the probabilistic model for claim severities can be justified by the "long-tail" nature of insurance losses, which appear as a result of the delay in reporting and long settlement periods of claims. So, this matter makes it difficult to evaluate the exact price of some liability insurance products and actuary job is to compute the average loss, or pure premium, for different classes of insurance products for fairly rating insurance policies. They use the observed past claim data from a portfolio of an insurance company for predicting the future pure premium for a determined period. Shi and Valdez (2011) and Katesari and Vajargah (2015) used count Copula model for examining asymmetric information in the insurance industry. The former computed pure premium using the information of selected coverage level and loss number for a specific year, and here with following the latter we try to compute the premium. Frees and Valdez (2008) computed premiums under alternative reinsurance coverage. However, Katesari and Zarodi (2016) predicted accident probability after observing the accidents for a specific year by using the copula model in the latter.

In this paper, we use the count copula model for computing the pure premium of the severity data from a major insurance company in Iran. Specifically, we consider the generalized beta distribution of the second kind (GB2) regression model for the severity claims. For this, we need the joint distribution of coverage selection and the risk of policyholders. An ordered multinomial model is used to measure the coverage levels and a negative binomial regression model is used to measure the risk of policyholders for the specific year. Moreover, a copula regression model is used to measure the linear and nonlinear dependence between these two margins and the estimated results are presented. The estimation results of the fitted model using Frank copula is available in Katesari and Vajargah (2015). Instead, we use another tow famous members of the Archimedean copula family that is Clayton and Gumbel to measure this dependence. The benefit of our bivariate copula regression model is that it provides the joint distribution of coverage levels and the risk of policyholders. We exploit this joint distribution in conditional expectation for computing the pure premium of the severity data. For real data application, we use a portfolio of major automobile insurer in Iran in the calendar year 2007-2008 with a subsample of 59,547 policies in their portfolio. Also, this dataset was used to the work of Katesari and Vajargah (2015) to test asymmetric information in the collision insurance portfolio of this company.

We have organized the remains of the article as follows. In Section 2, the data description is given. In Section 3, the count copula regression model will be considered for computing the pure premium and the estimation results are given. In Section 4, premium estimation is presented and the results are compared with the actual premium. Finally, in Section 5 we provide some concluding remarks.

2. Data attributes

For fitting the model, we use a portfolio of a major automobile insurer in Iran in 2007-2008 with a total of 59,547 policies available in their portfolio. According to the policy of the company, policyholders buy insurance policy from these main and overall claims: overall accident, overall theft and overall fire. Furthermore, policyholders are able to purchase one or more coverage options from the below items:

1. Damaged caused by flood, earthquake and hurricanes,
2. Broken glass,
3. Stolen parts and accessories of vehicle,
4. Damage caused by spills or splashes of paint, acid and chemicals,
5. Compensation by not using the vehicle in repair period,
6. Slippage (only in minor damage).

For our purpose, we ordered the levels as follows:

1. overall coverage of collision insurance,
2. overall coverage of collision insurance as well as one or two more item(s),
3. (comprehensive) overall coverage of collision insurance plus three, four, five or all of more item(s).

Note that with increasing the levels (from 1 to 3), the insured coverage will increase. The dataset comes from a major insurer in Iran and we use a subsample of 59,547 cases from more than 800,000 recorded cases the portfolio in 2007-2008 for this insurer. One can find frequency statistics of policy selection and the number of losses in Table (1).

Table 1: Frequency statistics of policy selection and number of losses

	Levels				
Claims	1	2	3	Total Number	Percent
0	30176	20033	4879	55088	92.51
1	405	1497	2130	4032	6.77
2	39	161	192	392	0.66
3	2	11	21	34	0.06
4	0	1	0	1	0.00
Total Number	30622	21703	7222	59547	
Percent	51.42	36.45	12.13		100

Like every insurance database, more than 90 percent of the policyholders did not have an accident during the considered year. Moreover, Table (2) elaborates the available covariates

Table 2: Descriptive statistics of the covariates

Variable	Explanation			Level 1		Level 2		Level 3	
		Mean	StdDev	Mean	StdDev	Mean	StdDev	Mean	StdDev
Driver attributes									
Sexinsured	=1 F, 0 M	0.2014		0.1694		0.2306		0.2779	
NCD	=1(0-15%)	0.4383		0.4538		0.4289		0.3827	
	=2(15-30%)	0.2156		0.2187		0.2131		0.2066	
	=3 (30-45%)	0.2529		0.2393		0.2684		0.2725	
	=4 ($\geq$ 45%)	0.9320		0.0882		0.0896		0.1382	
Vehicle attributes									
Vage		4.2951	3.4921	4.5838	4.0601	3.8714	2.5323	4.2597	2.8845
Vtype	=Sedan	0.8849		0.8002		0.9882		0.9818	
	=Others	0.1151		0.1998		0.0117		0.0182	
Vapplication	=Personal	0.8632		0.7672		0.9798		0.9741	
	=Non-Personal	0.1368		0.2328		0.0202		0.0259	

Table 3: Severity size by months for the calendar year 2007-2008

	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar
Severity	1.3	4	6.3	9.5	13.1	13.8	19.2	21.1	26.2	28.5	26.8	17.2

in the dataset. One can classify each of these covariates as a driver or vehicle attributes. Vehicle age (Vage), vehicle type (Vtype: Sedan or Non-Sedan) and vehicle application (Vapplication: Personal or Non-Personal) are vehicle attributes while sex (Female and Male) and No Claim Discount (NCD) are driver attributes. As can be seen from Table (2), many of these covariates are categorical, which demonstrates the proportion of an observed variable in each class. Moreover, both mean and standard deviation are presented for vehicle age, which is the only continuous covariate in this dataset. Like Shi and Valdez (2011), we used average claims in the observed calendar year. Table (3) provides summary of the severity claims for different months of the year 2007-2008. As demonstrated in Table (3), the majority of the policyholder's loss, nearly 28.5 in this case, occurred in January and the minority of the policyholder's loss, roughly 1.3, occurred in April. One of our restrictions is that the amounts of these losses are adjusted and we cannot distribute the exact amounts to all.

3. Count copula model fitted to the data

A bivariate copula $C(.,.)$ is a joint cumulative distribution function $C : [0,1] \longrightarrow [0,1]^2$. The application of copula comes from Sklar's theorem. Sklar (1959) says that for random variables y_1 and y_2 with corresponding marginal distributions $F_1(y_1)$ and $F_2(y_2)$, the bivariate distribution $F(y_1,y_2)$ can be stated as

$$F(y_1,y_2) = C(F_1(y_1),F_2(y_2);\theta) \tag{1}$$

where C is a copula function with dependence parameter θ. If the marginal distributions are continuous, then the copula in Equation (1) is unique, otherwise C is uniquely determined on $RanF_1 \times RanF_2$. In Shi and Valdez (2011) and Katesari and Vajargah (2015), count copula models were used for testing asymmetric information and adverse selection in automobile insurance market. Here, according to our database, we take y_{i1} and y_{i2} as selected coverage level and loss number, correspondingly, for each policyholder. y_{i1} shows the selected coverage level such that first level (overall), second level, and third level (comprehensive) coverages are connected with possible values 1, 2 or 3, correspondingly. We use latent variables y_{i1}^* and y_{i2}^*, for modelling y_{i1} and y_{i2} with a parametric copula $C(.,.)$.The joint probability mass function of y_{i1} and y_{i2} can be express as:

$$\begin{aligned} f_i(y_{i1},y_{i2}) =& C(F_{i1}(y_{i1}),F_{i2}(y_{i2})) - C(F_{i1}(y_{i1}-1),F_{i2}(y_{i2})) \\ & - C(F_{i1}(y_{i1}),F_{i2}(y_{i2}-1)) + C(F_{i1}(y_{i1}-1),F_{i2}(y_{i2}-1)) \end{aligned} \tag{2}$$

where F_{i1} and F_{i2} are the CDF of y_{i1} and y_{i2}, correspondingly. Now, we need to calibrate the marginal distribution functions of F_{i1} and F_{i2} for model identification (Shi and Valdez, 2011). For coverage level and catching the connection between y_{i1} and y_{i1}^*, we use an ordered multinomial model as follows:

$$y_{i1} = \begin{cases} 1, & \text{if } y_{i1}^* \le \alpha_1 \\ 2, & \text{if } \alpha_1 \le y_{i1}^* \le \alpha_2 \\ 3, & \text{if } y_{i1}^* > \alpha_2 \end{cases},$$

where α_1 and α_2 are unknown and should be estimated. Also, for estimating y_{i1}, we fit an ordered logistic regression model as follows:

$$F_{i1}(y_{i1}) = \begin{cases} \dfrac{1}{1+exp(-(\alpha_1 - \mathbf{x_i}'\beta))}, & \text{if } y_{i1}=1 \\ \dfrac{1}{1+exp(-(\alpha_2 - \mathbf{x_i}'\beta))}, & \text{if } y_{i1}=2 \\ 1, & \text{if } y_{i1}=3 \end{cases} \tag{3}$$

where $\mathbf{x_i}$ is the vector of covariates used for the coverage level of the ith policyholder. Another marginal variable y_{i2} can be calibrated by using a negative binomial regression model. Like Shi and Valdez (2011), we define its probability mass function as follows:

$$f_{i2}(y_{i2}) = Pr(Y_{i2}=y_{i2}) = \frac{\Gamma(y_{i2}+\psi)}{\Gamma(\psi)\Gamma(y_{i2}+1)}\left(\frac{\psi}{\psi+\lambda_i}\right)^{\psi}\left(\frac{\lambda_i}{\psi+\lambda_i}\right)^{y_{i2}} \tag{4}$$

where ψ is the dispersion parameter for policyholder i, and we use a log link function for the conditional mean that is $Y_{i2}|\mathbf{z}_i$. Note that $\mathbf{z}_i$ is the vector of covariates used for the risk of the ith policyholder. For estimating this model, we can use maximum likelihood method. The copula functions of the Gumbel and Clayton can be expressed, respectively, as follow:

$$C(u_1,u_2;\theta) = exp\{-[(-logu_1)^{\theta} + (-logu_2)^{\theta}]^{1/\theta}\}, \theta \ge 1 \tag{5}$$

Table 4: Estimation results of Clayton copula model for all reported accidents

Choice-Cumulative Logit			Risk-Negative Binomial		
	Estimate	StdErr		Estimate	StdErr
Choice -α_1	-0.9033	0.0131			
Choice-α_2	0.3767	0.0125	Risk-intercept	-2.1585	0.0194
Choice-sex (F)	0.4938	0.0194	Risk-sex (F)	0.0189	0.9339
Choice-Vage	0.0614	0.0012	Risk-Vage	0.0235	0.0015
Choice-(NCD=2)	0.0675	0.0199	Risk-(NCD=2)	-0.4443	0.0389
Choice-(NCD=3)	0.2301	0.0190	Risk-(NCD=3)	-0.7799	0.0441
Choice-(NCD=4)	0.1362	0.0269	Risk-(NCD=4)	-1.3939	0.0826
Choice-Vapplication (2)	-0.0858	0.1036	Risk-Vapplication (2)	-0.3052	0.2911
Choice-Vtype (2)	-0.1704	0.0698	Risk-Vtype (2)	-0.3052	0.2911
			Dispersion	0.8326	0.0734
Dependence parameter θ	**0.2615**	0.0184			
-2Loglikelihood	160165.1				

$$C(u_1,u_2;\theta) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}, \theta > 0 \tag{6}$$

where θ is the dependence parameter that shows the amount of association between two marginals. For more details about application of copula in finance and actuarial science, see Frees and Valdez (1998), Cherubini et al. (2004), Joe (2014), Zaroudi et al. (2018a), Zaroudi et al. (2018b), and Shi and Yang (2018). In the similar work of Katesari and Vajargah (2015), they explained the problems arising from adverse selection based on copula model. They estimated parameter of Frank copula with $\theta = 1.3$ referred to the existence of adverse selection in their dataset. In this paper, we are interested in modelling the severity of claims and we will use the modelled copula for computing the pure premium with the same database.

The estimation results of the fitted model by using the maximum likelihood method for Frank copula is available in Table 2 of Katesari and Vajargah (2015). Here, we fit the aforementioned model using the maximum likelihood method for two other members of the Archimedean copula family, which are Gumbel and Clayton in equations (5) and (6), respectively. The estimation results of these two famous copulas are presented in Table (4) and Table (5). As can be seen from the results of Table (4) and Table (5), the dependence parameter θ for Clayton and Gumbel copula is 0.2615 and 1.10207, respectively. These results show a strong dependence between coverage level and the risk of policyholders in the portfolio of the insurance company in Iran.

4. Computing premium

Here, we describe, discuss and compute the pure premium formula from a mathematical viewpoint and then compare it with the gross premium in the original data. We define the premium by $\prod_X$ that an insurance company charges to pay a loss X, which is a random variable. Thus, a premium formula is of the form $\prod_X = \phi(X)$ where ϕ is some function. At first, we consider the mean of X and the simplest premium, which is called pure risk premium ($\prod_X = E(X)$), which means the pure premium is equal to the insurer's expected claims under the considered risk (Dickson, 2016). Additional statistical properties of the

Table 5: Estimation results of Gumbel copula model for all reported accidents

Choice-Cumulative Logit			Risk-Negative Binomial		
	Estimate	StdErr		Estimate	StdErr
Choice-α_1	-0.9032	0.0131			
Choice-α_2	0.3768	0.0126	Risk-intercept	-2.1590	0.0193
Choice-sex (F)	0.4947	0.0194	Risk-sex (F)	0.0188	0.0151
Choice-Vage	-0.0194	0.2013	Risk-Vage	0.0172	1.6611
Choice-(NCD=2)	0.0671	0.0209	Risk-(NCD=2)	-0.4435	0.0391
Choice-(NCD=3)	0.2299	0.0191	Risk-(NCD=3)	-0.7780	0.0413
Choice-(NCD=4)	0.1633	0.0276	Risk-(NCD=4)	-1.3928	0.1828
Choice-Vapplication (2)	-1.7275	0.0765	Risk-Vapplication (2)	0.3179	0.0280
Choice-Vtype (2)	-1.8164	0.1371	Risk-Vtype (2)	-0.5796	0.2811
			Dispersion	0.8321	0.0739
Dependence parameter θ	**1.10207**	0.4579			
-2Loglikelihood	160164.8				

premium computation were explored in Dickson (2016). Our data comes from a big insurance company in Iran. In this section, we use the severity of losses for one year (in the year 2007-2008) for this insurer with a sample of 62,602 policyholders out of total 800,769 recorded cases. Here, we compute the pure premium by using the selected coverage level and the number of losses for the year in the work of Katesari and Vajargah (2015). The common method for price evaluation in the automobile insurance market is modelling the number and severity of losses separately. In reality, the independence assumption between the number and severity of losses is straightforward and we need to model the size of claims to compute the pure premium. So, we compute the claims mean for each of the three levels of coverage using a regression model of Generalized Beta distribution of the second kind (GB2) . The density function of GB2 with four positive parameter goes as follows (Kleiber and Kotz, 2003):

$$f(x) = \frac{ax^{ap-1}}{b^{ap}B(p,q)\{1+(x/b)^a\}^{p+q}}, \qquad x>0, \qquad a,b,p,q>0, \tag{7}$$

where b is a scale parameter and a, b, c are shape parameters and $B(p,q)$ is the usual Euler beta function. For more information about GB2, one can refer to McDonald and Butler (1987), Sun et al. (2008), Frees and Valdez (2008) and Shi and Valdez (2011). Here, we follow the same way of Shi and Valdez (2011) with taking as $b_i = exp(l_i^{'}\beta)$, where $l_i^{'}$ and β show covariates vector for each policyholder and the coefficients, respectively. In our GB2 regression model, sets of parameters for estimation purpose are (β^j, a^j, p^j, q^j) , with possible values $j = 1,2,3$, which show the three selected coverage levels, correspondingly. Table (6) shows the results of estimating the three sets of parameters by using the likelihood-based estimation method. Figure (1) demonstrates the pp-plots of the residuals from the three regression models of GB2 for showing the quality of the fitted model. According to the copula method that was used in Shi and Valdez (2011), we can additionally compute the impact of the policyholder's coverage preference y_{i1} on the number of losses (accidents) y_{i2},

Table 6: Estimate results of the GB2 regressions for all coverage levels

	1st level		2nd level		3rd level	
	Estimate	StdError	Estimate	StdError	Estimate	StdError
a	3.1465	0.0016	4.6151	0.0100	3.7952	0.7432
Intercept	2.3920	0.0037	3.3120	0.0071	3.3122	0.0650
sexinsured (F)	0.0124	0.0095	0.0088	0.0003	- 0.0445	0.0286
NCD=2	0.0357	0.0024	0.0421	0.0416	- 0.0107	0.0099
NCD=3	0.1068	0.0052	0.0449	0.0003	0.0741	0.0683
NCD=4	0.2026	0.0010	0.1367	0.1164	- 0.0204	0.0835
Vapplication (2)	0.1193	0.0088	- 0.0802	0.0016	- 0.0056	0.0477
Vage	- 0.0001	0.0001	- 0.0063	0.0000	- 0.0063	0.0059
Vtype (2)	- 0.1443	0.0117	0.0736	0.0014	0.0256	0.1407
p	9.9716	0.0001	0.8890	0.0429	1.0804	0.2853
q	0.3398	0.0091	0.2567	0.0079	0.3388	0.0808
-2Loglikelihood	10127.15		14029.72		21469.40	

conditionally by using Bayes' formula:

$$Pr(Y_{i2} = y_{i2}|Y_{i1} = y_{i1}) = f_{i2|1}(y_{i2}|y_{i1},x,z) \times \frac{f_i(y_{i2},y_{i1}|x,z)}{f_{i1}(y_{i1}|x)}. \tag{8}$$

By applying this conditional formula, we can anticipate the likelihood of the number of claims, condition on the policy selection. In the above equation, the joint probability distribution in the numerator can be computed by copula distribution in equation (2) and obviously the marginal distribution of y_{i1} in the denominator by equation (3). According to the coverage selection for y_{i1}, we can conditionally compute the pure premium for the ith policyholder as follows:

$$\begin{aligned} \prod_i &= E(Y_{i2}|Y_{i1} = y_{i1}) \times E(X_i|Y_{i1} = y_{i1}) \\ &= \sum_{y_{i2}=0}^{\infty} y_{i2} f_{i2|1}(y_{i2}|y_{i1},x,z) \times \frac{exp(l_i'\beta^{y_{i1}})B(p^{y_{i1}} + (1/a)^{y_{i1}}, q^{y_{i1}} - (1/a)^{y_{i1}})}{B(p^{y_{i1}}, q^{y_{i1}})} \end{aligned} \tag{9}$$

where $B(p,q) = \dfrac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$, $\prod_X$ is the pure premium for the ith policyholder and $\Gamma(.)$ is the Gamma function. Using the above formula, we are able to compute the pure premium for each policyholder in our dataset.

Dependency coefficients among the real gross premium and the estimated one for all coverage levels have been computed by Spearman's rho and demonstrated in Table (7), which shows strong positive dependency. This strong positive correlation shows that the actual premium paid by the policyholder is according to the conditional computation. In comparison with the results of Shi and Valdez (2011), one can see the same positive dependency in the portfolio of automobile insurance in Singapore. More precisely, the dependency between real and computed premiums for the first, second and third levels of our work is 0.6636, 0.2328, and 0.8372, respectively. This is while that the dependency between real and computed premiums for the first, second and third levels of the work of Shi and Valdez (2011) is 0.58282, 0.62215, and 0.80632, respectively. Also, descriptive statistics of the real

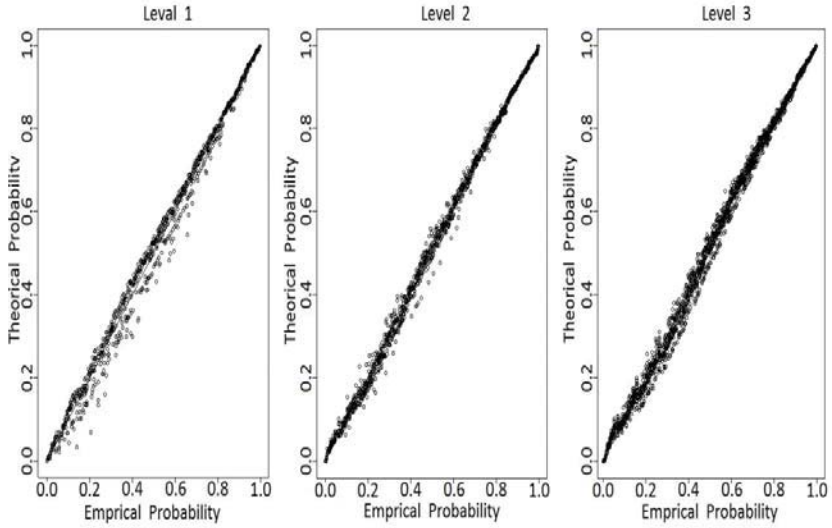

Figure 1: pp-plot for GB2 regression models.

Table 7: Dependency between real and computed premiums

	dependency	p-value
first level	0.6636	0.0098
second level	0.2328	0.0223
third level	0.8372	0.0025

and computed premiums for all coverage levels have been shown in Table (8). These results are not surprising at all and we expected the positive difference between the two premiums. This positive difference can be justified by covering loading expenses such as profits, taxes and other administrative charges, which the policyholder should pay for them as well.

Table 8: Comparison of real and computed premiums

	first level		second level		third level	
	Mean	StdDev	Mean	StdDev	Mean	StdDev
Real	23.3024	17.4888	17.6657	14.7873	19.6919	18.4127
Estimated	15.6222	9.1653	13.4348	7.8841	14.3349	16.8840

5. Conclusions

The main focus of this paper is to compute pure premium by using copula models in the automobile insurance market. We applied a GB2 regression model to compute the claims mean and conditional computation for all coverage levels. This model permits us to compare

real and estimated premiums. For this comparison, the coverage level of policyholders is fitted using an ordered multinomial model and the risk of the policyholder is measured with a negative binomial regression model in the specific year. The difficulty of this method is to modelling two count variables for finding the joint distribution, which is useful in computing the pure premium for the *i*th policyholder. To address this problem, we used a copula regression model, which builds a bivariate distribution function and measures both linear and nonlinear dependency between marginal distributions. For testing the quality of our model we used pp-plots of residuals of the fitted model. The estimation results of our model showed a strong positive dependence between real and estimated premiums.

One of our restrictions in this research is that we used a cross-sectional dataset to fit our model. If we could use a longitudinal dataset that followed each policyholder's records during the years, we would reach out to more knowledgeable results.

REFERENCES

AVANZI, B., TAYLOR, G., WONG, B., YANG, X., (2019). A Multivariate Micro-Level Insurance Counts Model With a Cox Process Approach, *UNSW Business School Research Paper*, (2019ACTL02).

CHERUBINI, U., LUCIANO, E., VECCHIATO, W., (2004). *Copula methods in finance*. John Wiley and Sons.

DAVID, M., (2015). Auto insurance premium calculation using generalized linear models, *Procedia Economics and Finance*, 20, pp. 147–156.

DICKSON, D. C., (2016). *Insurance risk and ruin*. Cambridge University Press.

FREES, E. W., VALDEZ, E. A., (1998). Understanding relationships using copulas, *North American actuarial journal*, 1; 2(1), pp. 1–25.

FREES, E. W., VALDEZ, E. A., (2008). Hierarchical insurance claims modeling, *Journal of the American Statistical Association*, 103(484), pp. 1457–1469.

FREES, E. W., JIN, X., LIN, X., (2013). Actuarial applications of multivariate two-part regression models, *Annals of Actuarial Science*, 7(2), pp. 258–287.

JOE, H., (2014). *Dependence modeling with copulas*. Chapman and Hall/CRC.

KATESARI, H. S., VAJARGAH, B. F., (2015). Testing Adverse Selection Using Frank Copula Approach in Iran Insurance Markets, *Mathematics and Computer Science*, 15, pp. 154–158.

KATESARI, H. S., ZARODI, S., (2016). Effects of Coverage Choice by Predictive Modeling on Frequency of Accidents, *Caspian Journal of Applied Sciences Research*, 5, pp. 28–33.

KLEIBER, C., KOTZ, S., (2003). *Statistical size distributions in economics and actuarial sciences*, Vol. 470, John Wiley and Sons.

LESMANA, E., WULANDARI, R., NAPITUPULU, H., SUPIAN, S., (2018). Model estimation of claim risk and premium for motor vehicle insurance by using Bayesian method, *In IOP Conference Series: Materials Science and Engineering* (Vol. 300, No. 1, p. 012027), IOP Publishing.

MARTON, J., KETSCHE, P. G., SNYDER, A., ADAMS, E. K., ZHOU, M., (2015). Estimating premium sensitivity for children's public health insurance coverage: selection but no death spiral, *Health services research*, 50(2), pp. 579–598.

MCDONALD, J. B., BUTLER, R. J. (1987). Some generalized mixture distributions with an application to unemployment duration, *The Review of Economics and Statistics*, pp. 232–240.

SCHIRMACHER, E., (2016). Pure Premium Modeling Using Generalized Linear Models, *Predictive Modeling Applications in Actuarial Science*: Volume 2, Case Studies in Insurance, 1.

SHI, P., VALDEZ, E. A., (2011). A copula approach to test asymmetric information with applications to predictive modeling, *Insurance: Mathematics and Economics*, 49(2), pp. 226–239.

SHI, P., (2016). Insurance ratemaking using a copula-based multivariate Tweedie model, *Scandinavian Actuarial Journal*, 2016(3), pp. 198–215.

SHI, P., YANG, L., (2018). Pair copula constructions for insurance experience rating, *Journal of the American Statistical Association*, 113(521), pp. 122–133.

SUN, J., FREES, E. W., ROSENBERG, M. A., (2008). Heavy-tailed longitudinal data modeling using copulas, *Insurance: Mathematics and Economics*, 42(2), pp. 817–830.

SKLAR, M., (1959). Fonctions de repartition an dimensions et leurs marges, *Publ. inst. statist. univ. Paris*, 8, pp. 229–231.

WEISBERG, H. I., TOMBERLIN, T. J., (1982). A statistical perspective on actuarial methods for estimating pure premiums from cross-classified data, *Journal of Risk and Insurance*, pp. 539–563.

WOLNY-DOMINIAK, A., WANAT, S., SOBIECKI, D., (2018). *Modelling Quantile Premium for Dependent LOBs in Property/Casualty Insurance*, In Finance and Sustainability, Springer, Cham, pp. 265–272.

YANG, Y., QIAN, W., ZOU, H., (2017). Insurance premium prediction via gradient tree-boosted Tweedie compound Poisson models, *Journal of Business and Economic Statistics*, pp. 1–15.

ZAROUDI, S., BEHZADI, M. H., FARIDROHANI, M. R., (2018a). Application of Copula in Life Insurance, *International Journal of Applied Mathematics and Statistic*, 57(3), 162–168.

ZAROUDI, S., FARIDROHANI, M., BEHZADI, M., (2018b). A Copula Approach for Finding the Type of Dependency with Mortality Force Function in Insurance Market, *Journal of Advances and Applications in Statistics*, 53(2), pp. 103–121.

ZHANG, X., YIN, W., WANG, J., YE, T., ZHAO, J., (2015). Crop insurance premium ratemaking based on survey data: a case study from Dingxing county, China, *International Journal of Disaster Risk Science*, 6(3), pp. 207–215.