Journal of Statistical Modelling: Theory and Applications
Vol. 1, No. 1, 2020, pp. 169-177
Yazd University Press 2020

JSMTA

Research Paper

Modeling insurance data using generalized gamma regression

Hossein Zamani*, Marzieh Shekari, Zohreh Pakdaman
Department of Statistics, University of Hormozgan

Received: November 15, 2019/ Revised: January 3, 2020 / Accepted: January 29, 2020

Abstract: The generalized gamma (GG) is a flexible distribution in statistical literature with the special cases of exponential, gamma, Weibull and lognormal distributions. This paper investigates the GG additive model for modeling hospital claim costs. In comparison to other models, the GG is more flexible and has a better performance in modeling positively skewed data. The proposed model was fitted to the hospital costs data from the nationwide inpatient sample of the health care cost and utilization project, a nationwide survey of hospital costs conducted by the U.S. Agency for healthcare research and quality. The results indicate that the claim cost is affected by the given explanatory variables and based on the AIC and BIC criteria, the GG has a better performance for the given data compared to the alternatives.

Keywords: Generalized additive models; Generalized gamma distribution; Insurance.
Mathematics Subject Classification (2010): 62J02.

1 Introduction

The regression approach is used to investigate the relationship between a response variable and one or more explanatory variables that can be expressed through a mathematical formula. The regression model is applied to predict the response and discover the effect of the covariates on the response variable. The generalized linear model (GLM) is an extension of the linear regression model where the response variable departs from the normal to an exponential distribution. The response variable may have a discrete distribution or follows a continuous distribution with the right-skewed shape. The researchers used the GLM in many areas such as economic, insurance, engineering, and other fields . For instance, the Tweedie model was applied by Czado (2005), Jorgenson and Souza (1994) and Smyth and Jorgenson (2002) for modeling claim costs. Brockman and Wright (1992), MacCullagh and Nelder (1989), Hogg and Klugman

*Corresponding author: zamani.huni@hormozgan.ac.ir

(2009) used the GLM for insurance data. Tong et al. (2013) used the GG additive model for modeling the mortgage loan loss.

Due to the common properties of loss and severity data that commonly follows positive support and right skewness, the GG distribution is a good candidate for modeling data in the economic and insurance area. In this paper, we propose the GG additive model for modeling hospital claim costs. Compares to the gamma, Weibull and inverse Gaussian models, the GG is more flexible and has a better performance in modeling positively skewed data. The proposed model is fitted to the hospital costs data from the nationwide inpatient sample of the health care cost and utilization project (NIS-HCUP), a nationwide survey of hospital costs conducted by the U.S. agency for healthcare research and quality (AHRQ). The results indicated that the claim coast is affected by the given explanatory variables and the GG has a better fit for the given data compared to the alternatives.

The paper is organized as follows. In Section 2, the GG distribution and the GG regression model is presented. In Section 3, we describe the data set used in this paper. Also, the generalized gamma distribution regression and its nested models is used to model the hospital costs. The results showed that the generalized gamma model has better fit than its nested models such as Weibull and gamma distributions.

2 Statistical models

In this section, we present the generalized gamma regression model and the method of inserting this model into the generalized linear model approach.

2.1 Generalized Gamma distribution

Let $f(y|\alpha,\tau,\lambda)$ denotes the probability density function (pdf) of GG distribution with the parameters α,τ and λ . The pdf of the GG distribution (denoted by $GG(\alpha,\tau,\lambda)$) is given by

$$f(y|\alpha,\tau,\lambda)=\frac{\tau}{\lambda\Gamma(\alpha)}\left(\frac{y}{\lambda}\right)^{\alpha\tau-1}e^{\left(\frac{y}{\lambda}\right)^{\tau}},\quad y>0,\quad \alpha,\tau,\lambda>0, \tag{1}$$

where the mean and variance are given, respectively, by

$$E(Y)=\frac{\lambda\Gamma(\frac{1}{\tau}+\alpha)}{\Gamma(\alpha)},\qquad Var(Y)=\frac{\lambda^2\Gamma(\frac{2}{\tau}+\alpha)}{\Gamma(\alpha)}-[\frac{\lambda\Gamma(\frac{1}{\tau}+\alpha)}{\Gamma(\alpha)}]^2. \tag{2}$$

The GG distribution reduces to the exponential distribution if $\alpha=\tau=1$, to gamma distribution if $\tau=1$, and to Weibull distribution if $\alpha=1$. Also, the log-normal distribution is a limiting case of the GG distribution when $\alpha\to\infty$.

To insert the GG distribution in the generalized linear model, we adjust the parameters of distribution such that $E(Y)=\mu$. Thus a re-parametrization is applied as follows

$$\tau=\nu,\quad \alpha=\frac{1}{\sigma^2\nu^2},\qquad \lambda=\mu\,(\sigma^2\nu^2)^{\frac{1}{\nu}}.$$

With the above re-parameterization, the pdf of GG denotes by $g(y|\mu,\sigma,\nu)$ and becomes to

$$g(y|\mu,\sigma,\nu)=\frac{\nu\mu^{-1}}{\Gamma(\frac{1}{\sigma^2\nu^2})\,(\sigma^2\nu^2)^{\frac{1}{\sigma^2\nu^2}}}\left(\frac{y}{\mu}\right)^{\frac{1}{\sigma^2\nu}-1}e^{-\frac{(y\mu^{-1})^{\nu}}{\sigma^2\nu^2}},\qquad y>0, \tag{3}$$

where $\mu>0$ and $\sigma>0$. The mean and the variance of the new form of (3) are

$$E(Y)=\mu,\qquad Var(Y)=\mu^2\sigma^2.$$

2.2 Generalized gamma regression model

Let assume the response variable y_i affected by covariates $X_i=(x_1,...,x_r)$. As an example, in the insurance literature, the claim cost is affected by age, gender, occupation and so on.

The GG distribution contains three parameters including the location parameter (μ), the scale or dispersion parameter (σ) and the shape parameter (ν). The whole or some of these parameters can be incorporated in the regression model via a suitable link function. To apply the GG regression model on given data, the vector of co-variates x_i, can be incorporated through some appropriate link functions in the model. Here a log link function is used for the μ and σ and an identity link function is applied for incorporating the ν in the model. Thus, the generalized additive regression model can be represented as

$$\begin{aligned}\log(\mu)&=X_1^T\beta+\sum_{j=1}^{m_1}h_{j1}(x_{j1}),\\ \log(\sigma)&=X_2^T\gamma+\sum_{j=1}^{m_2}h_{j2}(x_{j2}),\\ \nu&=X_3^T\delta+\sum_{j=1}^{m_3}h_{j3}(x_{j3}),\end{aligned} \tag{4}$$

where β, γ, and δ are the vector of regression coefficients corresponding to μ,σ and ν, respectively and $h_{jk}(x_{jk})$ are the penalized spline functions as the non-parametric smoothing terms. Also, for $i=1,2,3$ and $m_i=1,\ldots,r$, $X_i=(x_{1i},...,x_{m_ii})$ are the sets of the predictor variables that are similar at least in one variable. The regression parameters β, γ, and δ can be estimated using the maximum likelihood procedure.

The penalized spline function $h_{jk}(x_{jk})$ is modeled using penalized B-spline which allows the estimation of the smoothing parameters using a local maximum likelihood minimizing Akaike Information Criterion (AIC) which is defined as $AIC=-2\log(L)+2p$ with L being the log of penalized likelihood and p the number of parameters in the fitted model.

3 Data and results

In this section, we introduce the dataset used in this paper and fit the generalized gamma regression model beside its nested models and compare the results based on several statistical criterions.

3.1 Data

The data (Hospital Costs) which is considered in this paper, is a standard dataset in healthcare insurance. The dataset is from the nationwide inpatient sample of the health care cost and utilization project (NIS-HCUP), a nationwide survey of hospital costs conducted by the U.S. Agency for healthcare research and quality (AHRQ). This dataset considered by Frees (2010). The data used for analysis was restricted to Wisconsin hospitals and contain a random sample of $n = 500$ claims from 2003 data. Although the data comes from hospital records, it is organized by individual discharge and includes the information about the age, gender, and length of stay (Los) in hospitals of the patient discharged. The data will be used to model the severity of hospital charges (response) by age, gender and Los in hospitals. The histogram of the total charges is given in Figure 1. It can be found that the distribution of hospital costs is positively skewed. The mean and the standard deviation of the variables are given in Table 1.

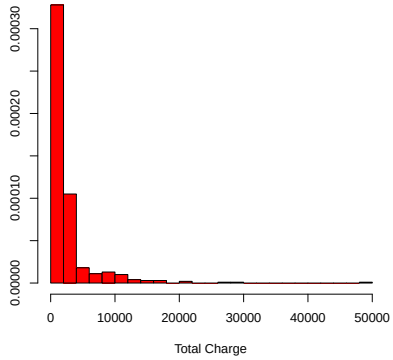

Figure 1: Histogram of total charge in the hospital costs data

Table 1: Mean and standard deviation of the variables

Variable	Type	Mean	Std
Total Charge	Continuous	2774.388	3888.407
Age	Continuous	5.086	6.95
Los	Discrete	2.82	3.363
Gender(Female=1)	Categorical	0.512	0.500

3.2 Results and discussion

The plots of GG distribution for some values of parameters are displayed in Figure 2. In Figure 3 the GG model and its special cases are fitted to the hospital costs values. In addition to generalized gamma, the gamma, Weibull and log-normal distributions are candidates for fitting on hospital costs as the positively skewed distributions. From Figure 3, it can be seen that the generalized gamma and the Weibull have a better fit for the histogram of costs compared to the gamma and log-normal distributions. Here it should be considered that these performances are given in the absence of explanatory

variables and the result maybe change once the explanatory variables were included in the model.

The GG model is implemented using the generalized additive models for location, scale and shape (GAMLSS) framework (Rigby and Stasinopoulos, 2007). The additive models allow all parameters of location, scale, and the shape to be included in the regression model and represented as functions of predicted variables. The GG and its nested models are fitted to the hospital costs data using the R package *R.3.3.1.* The fitted GG regression model can be compared using several measures such as AIC and BIC. The smaller AIC or BIC results in the better fitting.

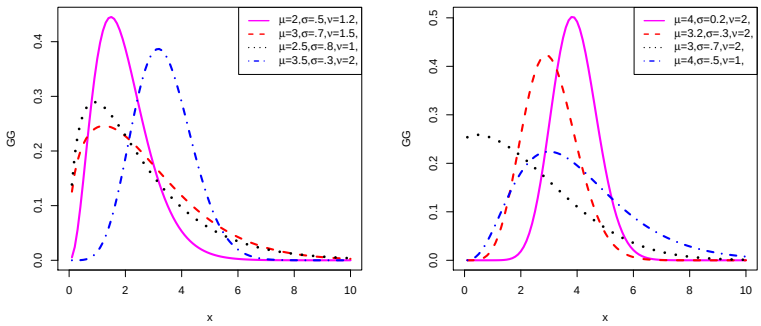

Figure 2: Plots of GG pdf for some values of parameters

Table 2: Likelihood ratio statistic,-2log$L$, AIC and BIC of the fitted models

Test/Criteria	GG	Gamma	Weibull	Log-normal
-2log$L$	7568.55	7985.81	8207.9	7867.74
Likelihood ratio statistic	-	400.52	622.61	282.45
AIC	7650.204	8049.39	8268.39	7934.03
BIC	7822.26	8183.38	8395.87	8073.73

Table 2 provides the $-2\log L$, the AIC and BIC of the fitted models. It can be seen that the GG model has a better performance compared to its nested models based on all criterions. The likelihood ratio statistic can be employed to assess the adequacy of the nested models. Here, the LRT is implemented for testing the adequacy of the GG model over the gamma, Weibull and log-normal models since all of these models are nested in GG distribution. As an example, for testing the GG against the gamma, the hypothesis may be stated as $H_0 : \tau = 1$ against $H_1 : \tau \neq 1$, and the likelihood ratio has an asymptotic chi-square distribution. In this case, the likelihood ratio statistic is $\chi^2 = 400.52$ which should be compared with a chi-square distribution with one degree of freedom (difference in the number of parameters). Based on the significance level $\alpha = 0.05$ the $\chi^2_{0.05}(1) = 3.841$ which indicates that the GG model is adequate versus the gamma and also the Weibull and log-normal models.

In Table 3 the results of fitting the GG model on the hospital costs data are given.The pb(.) function represents the penalized beta spline (P-spline). The results show that all of the regression coefficients are significance except the coefficient of the Age in the parameter of location (μ). As another interpretation, the coefficients of the Los covariate for the location parameter are positive which means the amount of hospital costs is increasing by increasing the level of Los. While the negative coefficient of

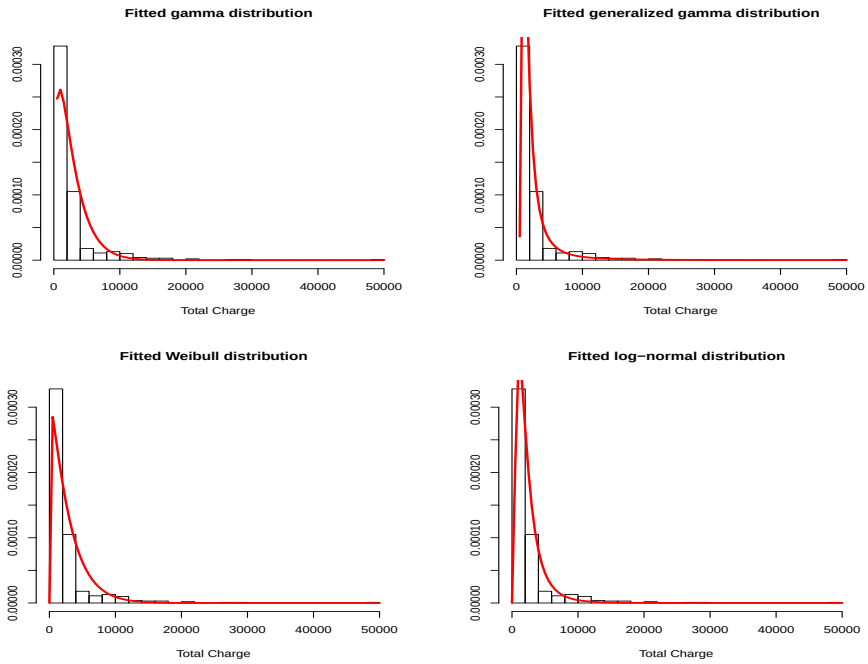

Figure 3: Fitted distributions on total charge values

Table 3: The estimation of coefficients of GG regression model

Model Parameters	Estimates	SE	P-value
$\log(\mu)$			
Intercept	7.012	0.0078	0.00
pb(Age)	-0.001	0.0001	0.3
Female=1	-0.100	0.0044	0.00
pb(Los)	0.130	0.0001	0.00
$\log(\sigma)$			
Intercept	-1.485	0.0453	0.00
pb(Age)	0.026	0.0036	0.00
Female=1	-0.487	0.051	0.00
pb(Los)	0.1011	0.0083	0.00
ν			
Intercept	-3.507	1.241	0.00
pb(Age)	-0.269	0.092	0.003
Female=1	-7.622	1.335	0.00
pb(Los)	-2.136	0.487	0.00

FEMALE (Gender covariate) shows that the hospital costs for the FEMALE are lower than the MALE at the same level of Age and Los covariates. Based on the estimated parameters, the regression models can be written as

$$\log(\mu) = 7.012 - 0.001\,(Age) + 0.130\,(Los) - 0.100\,(Female),$$
$$\log(\sigma) = -1.485 + 0.026\,(Age) - 0.101\,(Los) - 0.487\,(Female),$$
$$\nu = -3.507 - 0.269\,(Age) - 2.136\,(Los) - 7.622(Female).$$

For further analysis the partial effect plots are considered. Figures 4-5 display the partial effects plots for the log scales of the mean, $\log(\mu)$, and dispersion $\log(\sigma)$ for the GG distribution versus the covariates Age, Female and Los. The solid line represents the smoothing estimates based on the penalized beta spline, while the dashed lines represent the standard errors. From Figure 4, it can be found that the relationship between Age and μ is increasing at first and decreases later. In the other words the hospital costs is incraesing for the patients greater than 7 years when the Age increases. It also can be observed that the variable Los has a positive effect on the log of the mean.

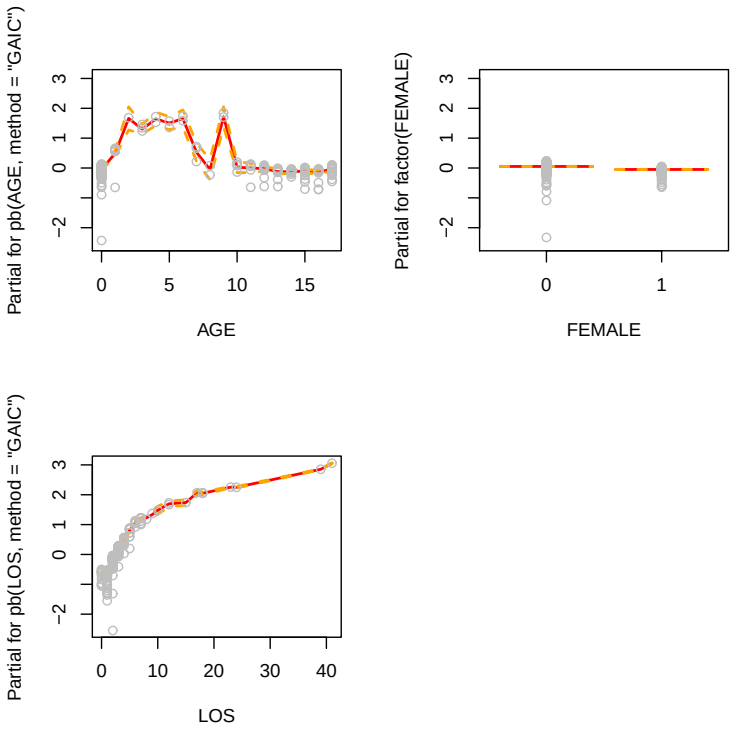

Figure 4: Partial effect plots $\log(\mu)$ versus Age, Female and Los

Further, Figure 5 displays the partial effect plot for the $\log(\sigma)$ versus the covariates. The plots represent that both of the Age and Los have a linear relationship with the log of dispersion in an increasing and decreasing fashion respectively. Figure 6 represents the diagnostic plots of fitted GG regression model. The plots are given based on quantile residuals which for the GG model is defined as $r_i = \Phi^{-1}\{F(y_i, \hat{\mu}, \hat{\sigma}, \hat{\nu})\}$ where $\Phi^{-1}(.)$ is the inverse of standard normal distribution function and $F(.)$ is distribution function of the GG model. If μ, σ and ν are consistently estimated then the distribution of r_i converges to the standard normal distribution. From Figure 6, it can be concluded that there is no trend in the plot of quantile residual versus index (top right). This shows that the residuals are randomly dispersed around the horizontal axis that results in the adequacy of the GG model. In addition, both of the density and normal q-q plots confirm the normality of quantile residuals.

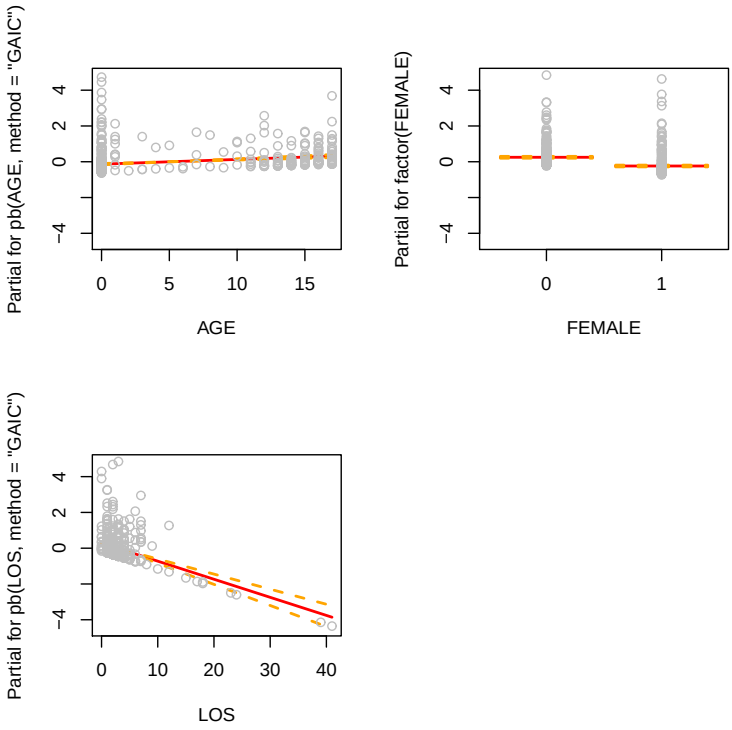

Figure 5: Partial effect plots $\log(\sigma)$ versus Age, Female and Los

Conclusion

This study proposed the generalized gamma regression model referred as GG model that parametrically nests the gamma, Weibull and the log-normal models for modeling the hospital costs. The GG model and its nested models are fitted and compared to the hospital costs as the response versus several covariates. Based on the results, all models produce similar estimates for the regression coefficients. However, based on the likelihood ratio tests given in Table 2 it can be concluded that the GG model is more flexible and has a better performance versus the alternatives. The fitted GG includes the additive components using log-link for both of the mean and dispersion and an identity link for the shape parameters. These components can be fitted with the same or different covariates. To fit the models, the beta spline method was used. The advantage of the spline method over the other approaches is that it considers the non-linearity between the response and covariates. Therefore the bias of the estimated parameters will be reduced which results in decreasing the deviance of the model. Further, the effected plots given in Figures 4 and 5 for the covariates within each parameter show that the relationship between the covariates and given link functions are non-linear and hence the beta spline method is a suitable choice for fitting the model. The diagnostic plots in Figure 6 also confirm the adequacy of the fitted GG regression on the given dataset.

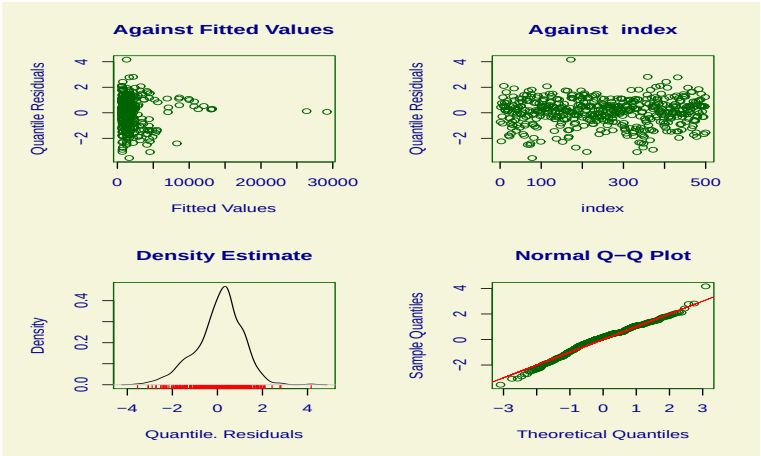

Figure 6: Diagnostic plots for the fitted GG on hospital costs data

References

Czado, G.(2005). Spatial modelling of claim frequency and claim size in insurance. *Insurance for statistics*. Sonderforschungsbereich 386, Paper 461.

Frees E.W. (2010). *Regression Modeling with Actuarial and Financial Applications*, Cambridge University Press.

Jorgensen, B. and Souza, M.C.P.D. (1994). Fitting Tweedie's compound Poisson model to insurance claims data. *Scandinavian Actuarial Journal*, **1**, 69–93.

MacCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Edn., Chapman and Hall, Boca Raton.

Smyth, G.K. and Jorgensen, B. (2002). Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modeling. *ASTIN Bulletin: The Journal of the IAA*, **32**, 143–157.

Brockman, M.J. and Wright, T.S. (1992). Statistical motor rating: Making effective use of your data. *Journal of the Institute of Actuaries*, **119**, 457–543.

Hogg, R.V. and S.A. Klugman.(2009). *Loss Distributions*. First Edn., John Wiley and Sons, New York.

Tong, E.N., Mues, C., and Thomas, L. (2013). A zero-adjusted gamma model for mortgage loan loss given default. *International Journal of Forecasting*, **29**, 548–562.

Rigby, R.A. and Stasinopoulos, D. M. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, **23**, 1-46.