Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 12, Number 4 (2016), pp. 3569–3575

http://www.ripublication.com/gjpam.htm

Modeling Gamma-Pareto Distributed Data Using GLM Gamma

Herlina Hanum

Department of Mathematics,
Universitas Sriwijaya, Indonesia.

Aji Hamim Wigena and Anik Djuraidah

Department of Statistics,
Bogor Agricultural University, Indonesia.

I Wayan Mangku

Department of Mathematics,
Bogor Agricultural University, Indonesia.

Abstract

Regression modeling is usually based on the probability distribution of the response variable. For exponential family distribution, the model is usually in the form of generalized linear model (GLM). Unfortunately, for data which follows Gamma-Pareto (G-P) distribution computational technique for GLM Gamma-Pareto has not been established yet. Since there is a mathematical relationship between G-P and gamma distribution, there is a possibility to develop model G-P distributed data using gamma distribution. In this paper we study the modeling of data which follows G-P distribution using GLM gamma. The simulated data were used to analyze G-P distributed data using GLM gamma. The response variable was transformed such that data follow gamma distribution. Then the transformed response variable, with the explanatory variable, was analyzed using GLM gamma. Finally the estimator of response variable was obtained by inversely transform the fitted value of GLM gamma. The result shows that GLM gamma fits good model for the data as long as the response variable has a good fit to G-P distribution and has high correlation with the explanatory variable.

AMS subject classification:
Keywords: Gamma-Pareto, gamma, GLM gamma.

1. Introduction

The regression model explains a phenomenon (the response variable) based on other phenomena (explanatory variables). Classical regression model is developed with the assumption that the response variables are normally distributed. This assumption is used for the validity of the test, for both the model and its parameters.

In real data, the response variable does not always normally distributed. For data of exponential family distribution, is developed generalized linear model (GLM). GLM uses link function that link the mean of response variable to the linear form of the explanatory variables. The link function is a monotone differentiable function (Dobson [3]). Since GLM is developed based on the probability distribution of response variable, the form of link function depends on this distribution.

G-P distribution developed by Alzaatreh *et al.* [1] is a member of the exponential family distribution Hanum [6]. Therefore, regression modeling of G-P could be developed in the form of GLM. Hanum [6] develop GLM G-P analytically. The model requires computational program so the model can be applied. Unfortunately, the computation program has not published yet.

Among GLM, GLM gamma which is based on the gamma distribution is often used. The right skew data are often fit when analyzed with GLM gamma. Alzaatreh *et al.* [1] mentioned the mathematical relationship between G-P and gamma distribution. This is reasonable because the G-P distribution developed from the gamma distribution. The existence of this relationship provides a possibility to analyze the G-P distributed data through GLM Gamma. The availability of many computing program for GLM Gamma is helpful for this work. Therefore, in this paper we study the modeling G-P distributed data using GLM gamma.

2. Exponential Form of Gamma-Pareto's PDF

According to Dobson [3] probability density function (pdf) of exponential family distribution is

$$g(y:\tau) = \exp\{a(y)b(\tau) + c(\tau) + d(y)\}. \tag{2.1}$$

Reparameterized τ in (2.1) into ω as canonical parameter for Y, McCullagh & Nelder [8] write the pdf exponential family as

$$g(y:\omega,\phi) = \exp\left\{\frac{y\omega + b(\omega)}{a(\phi)} + c(y,\phi)\right\}. \tag{2.2}$$

In Alzaatreh et al. [1], pdf of G-P distribution is written as

$$g(y) = \frac{\theta^{-1}}{\varrho^{\alpha}\Gamma(\alpha)}\left(\log\frac{y}{\theta}\right)^{\alpha-1}\left(\frac{y}{\theta}\right)^{-\frac{1}{\varrho}-1} \tag{2.3}$$

with $\alpha, \varrho, \theta > 0$ and $y > \theta$. Taking $\tau = \varrho$, this pdf can be written in the form of (2.1)

$$g(y) = \exp\left\{-\varrho^{-1}\log\frac{y}{\theta} - \alpha\log\varrho - \left[\log(y\Gamma(\alpha)) + (\alpha-1)\log\left(\log\frac{y}{\theta}\right)\right]\right\}. \tag{2.4}$$

Reparameterized (2.4) with $\omega = -1/(\alpha\varrho)$ and $\phi = 1/\alpha$, then we get the form (2.2) of (2.4)

$$g(y) = \exp\left\{\frac{\omega \log\frac{y}{\theta} - \log(-\frac{1}{\omega})}{\phi} - \left[\frac{1}{\phi}\log\phi + \right]\right\}. \tag{2.5}$$

Using (2.5) we can form GLM of G-P with $\log\dfrac{y}{\theta}$ as response variable. Despite of developing GLM G-P, it may be easier to transform $\log\dfrac{y}{\theta}$ into another existing distribution that already has GLM.

3. Relationship between G-P and Gamma Distributions

Let $U = \log\dfrac{Y}{\theta}$, so $Y = \theta e^{U}$, and $\dfrac{dY}{dU} = \theta e^{U}$. Rewrite pdf (2.3) in U

$$\begin{aligned}
g(u) &= \frac{\theta^{-1}}{\varrho^{\alpha}\Gamma(\alpha)} u^{\alpha-1}(e^{u})^{\frac{-1}{\varrho^{-1}}}\theta e^{u} \quad \text{or}\\
g(u) &= \frac{\theta^{-1}}{\varrho^{\alpha}\Gamma(\alpha)} u^{\alpha-1}(e^{u})^{\frac{-1}{\varrho}} \quad \text{or}\\
g(u) &= \frac{\theta^{-1}}{\varrho^{\alpha}\Gamma(\alpha)} u^{\alpha-1}e^{\frac{-1}{\varrho}}, \quad u > 0.
\end{aligned} \tag{3.6}$$

Equation (2.1) is the pdf of $\Gamma(\alpha, \varrho)$. This result confirms Alzaatreh *et al.* [1] that if $Y \sim$ G-P$(\alpha, \varrho, \theta)$ then $U = \log\dfrac{Y}{\theta} \sim \Gamma(\alpha, \varrho)$. As the inverse, if we have $U \sim \Gamma(\alpha, \varrho)$ then $Y = \theta e^{U} \sim$ G-P$(\alpha, \varrho, \theta)$.

It is clear that we can transform data Y from G-P distribution into U which follows gamma distribution. This transformation may follows with analyzing the transformed data through GLM gamma. Finally, estimation of Y can be determined using inverse transformation $\hat{U}$ to $\hat{Y}$.

4. Simulation, Analysis, and Verification Methods

The data used in this paper is simulation data. The simulation aimed to get the response variable Y that follows G-P distribution. We also need explanatory variable X which highly correlated to Y. In order to fulfill these two conditions of X and Y, we use equation $Y = a + bX + \varepsilon$ to generate Y, were a and b are constants, X is fix variable, and $\varepsilon \sim$ G-P$(\alpha, \varrho, \theta)$. Variables X and ε are each of length n. With this equation, X and Y will be highly correlated. Using $\varepsilon \sim$ G-P$(\alpha, \varrho, \theta)$, we expect Y will follow G-P distribution. The steps for this simulation are

1. Determine n, a, b, and X, also α, ϱ, and θ.

2. Generate $\varepsilon \sim$ G-P$(\alpha, \varrho, \theta)$ and determined $Y = a + bX + \varepsilon$.

3. Check the correlation between Y and X.

4. Fit the response variable Y to G-P distribution using minimum value of Y, $y_{(\min)}$, as the estimator of θ.

5. Check the goodness of fit using Kolmogorov-Smirnov test (KST).

To generate $\varepsilon \sim$ G-P$(\alpha, \varrho, \theta)$ we use quantile function of G-P$(\alpha, \varrho, \theta)$ as described in Hanum *et al.* [5]. Fitting Y to G-P distribution, we use maximum likelihood as describe by Alzaatreh *et al.* [1] and Hanum *et al.* [5]. If the correlation is good enough and Y follows G-P, then we can continue to analyze the data.

Data analysis begins with confirming the relationship between G-P and gamma distribution. That is, we check that $U = \log \dfrac{Y}{\theta}$ follows gamma distribution. The next step is running GLM gamma with U as response variable and X as explanatory variable to obtain the estimator $\hat{U}$. Finally transform $\hat{U}$ to get the estimator of $\hat{Y}$. The steps for analyzing the data are

1. Transform Y to $U = \log \dfrac{Y}{\theta}$ or $u_i = \log \dfrac{y_i}{y_{(1)}}$.

2. Fit U with gamma distribution, check the fitness using KST.

3. Model U and covariate X using GLM gamma, determine the estimator $\hat{U}$.

4. Determine the estimator $\hat{Y} = y_{(1)}e^{\hat{U}}$ or $\hat{Y}_i = y_{(1)}e^{\hat{U}_i}$.

Analysis of GLM gamma using R according to the theory in Balajari [2].

In order to diagnosis the goodness of the estimation of Y we need to compare Y and $\hat{Y}$. The goodness of estimation can be analysis through Mean Absolute Percentage Error (MAPE), correlation, and KST between Y and $\hat{Y}$. Estimation with smaller MAPE is better. MAPE less than 10% means very good estimation (Lewis [7]). On the other hand, KST with p-value larger than significance level means both variable come from similar distribution (Crutcher [3]). The closer p-value to 1, the better $\hat{Y}$ fit to Y distribution.

Finally we do correlation analysis to analyze the relationship between data characteristics and the goodness of the estimation. For this goal we do correlation analysis between data characteristics those are goodness of fit Y to G-P, and correlation between X and Y, and estimation properties those are MAPE, correlation between Y and $\hat{Y}$,and goodness of fit $\hat{Y}$ fit to Y distribution.

5. Results and Discussion

In order to obtain pairs of X and Y which meet our conditions, we have run many times of simulations. Some results are failed to fulfill the conditions due to low correlation or the distribution of Y. Finally we have 50 pairs of X and Y those meet the conditions. Here we show some of them in Table 1.

Table 1: Some of simulation and analysis results

No	n, a, b, X		ε		Y		U		Correlation		Estimation	
1	n	100	α	7	α	6.6653	α	6.6648	(y, x)	0.923	MAPE	0.0938
	a	5	ϱ	0.1	ϱ	0.135	ϱ	0.1347	(u, x)	0.919	Cor(y, yh)	0.9124
	b	0.5	θ	10	θ	19.906					p-val	0.8106
	X	1:100			p-val	0.3621	p-val	0.3621			b hat	0.5308
2	n	100	α	7	α	2.4352	α	2.4345	(y, x)	0.978	MAPE	0.1332
	a	5	ϱ	0.1	ϱ	0.3805	ϱ	0.3807	(u, x)	0.956	Cor(y, yh)	0.9693
	b	1	θ	10	θ	27.88					p-val	0.1509
	X	1:100			p-val	0.1509	p-val	0.1509			b hat	0.7424
3	n	100	α	7	α	5.4289	α	5.4311	(y, x)	0.936	MAPE	0.0849
	a	5	ϱ	0.1	ϱ	0.1627	ϱ	0.1627	(u, x)	0.944	Cor(y, yh)	0.9171
	b	1	θ	20	θ	37.72					p-val	0.9053
	X	1:100			p-val	0.5765	p-val	0.5675			b hat	1.069
4	n	100	α	2	α	4.3484	α	4.3466	(y, x)	0.667	MAPE	0.1929
	a	5	ϱ	0.5	ϱ	0.2628	ϱ	0.2629	(u, x)	0.76	Cor(y, yh)	0.6625
	b	1	θ	10	θ	29.221					p-val	0.8106
	X	10:109			p-val	0.5765	p-val	0.5765			b hat	1.231
5	n	100	α	2	α	3.2864	α	3.2871	(y, x)	0.803	MAPE	0.9357
	a	5	ϱ	0.5	ϱ	0.3045	ϱ	0.3046	(u, x)	0.849	Cor(y, yh)	0.7684
	b	1	θ	10	θ	32.412					p-val	0.00
	X	10:109			p-val	0.052	p-val	0.3621			b hat	1.231
6	n	100	α	0.5	α	3.0437	α	3.0428	(y, x)	0.781	MAPE	0.1323
	a	5	ϱ	0.5	ϱ	0.2538	ϱ	0.2653	(u, x)	0.856	Cor(y, yh)	0.7278
	b	0.5	θ	10	θ	21.719					p-val	0.2765
	X	10:109			p-val	0.2064	p-val	0.2064			b hat	0.5963

Table 1 column 2 and 3 contain the simulation seeds as mention in step 1. While column 4 and 5 are the result of fitting Y to G-P and fitting U to gamma including the p-value of KST. Correlation between Y and X, and correlation between X and U are in column 6. The last column contains MAPE, correlation, p-value of KST between Y and its estimator, and estimate of b in linear regression between $\hat{Y}$ and X.

In Table 1 we can see that transformation of Y which follows G-P$(\alpha, \varrho, \theta)$ to $U = \log \frac{Y}{\theta}$ which follows $\Gamma(\alpha, \varrho)$. The differences of α and ϱ values are only due to rounding. KST gives similar p-value for fitting both Y to G-P$(\alpha, \varrho, \theta)$ and U to $\Gamma(\alpha, \varrho)$. This result confirms the analytical result in (3.1) and Alzaatreh *et al.* [1]. It also gives us more confidence to use gamma distribution in modeling data which having G-P distribution.

As the coefficient of X in linear model to generated Y, the value of b shoud be estimated to show the goodness of simulation. The estimates are the coefficient of X in linear regression between $\hat{Y}$ and X. Table 2 shows the result of 100 run of simulation to estimate b. It contains the seed parameters of G-P, means of p-value of KST and correlation between Y and $\hat{Y}$, means of b estimate, and lower and upper bound of b. In Table 2 we can see that b is always in the range of its lower and upper bound. This means good estimation of b.

Table 2: The estimation of b

No	ε	p-val	Correlation	b hat	b	Lower	upper
1	2, 0.1, 10	0.8604	0.7032	0.0997	0.1	0.0968	0.1016
2	2, 0.05, 10	0.3640	0.8292	0.0500	0.05	0.0489	0.0512
3	2, 0.3, 10	0.189	06998	0.0564	0.5	0.4911	0.5285
4	2, 0.3, 15	0.7546	0.6660	0.4938	0.5	0.4735	0.5144
5	2, 0.3, 20	0.8051	0.6607	0.4910	0.5	0.4752	0.5180
6	2, 0.5, 7	0.1078	0.6874	0.9988	1	0.9416	1.0480
7	2, 0.5, 15	0.2346	0.6128	1.5589	1.5	1.4732	1.6245
8	2, 0.5, 15	0.3126	0.6008	2.0566	2	1.9935	2.1681
9	2, 0.5, 35	0.3884	0.6910	5.1114	5	4.9341	5.4012
10	2, 0.5, 60	0.3288	0.6263	10.3067	10	9.9516	10.7282
11	2, 0.6, 90	0.2091	0.6124	25.3377	25	24.2139	26.2113

Next we want to know whether GLM gamma yield a proper model and estimation for G-P distributed data. We also need to know in what condition of data that gives good model and estimation. So we do the correlation analysis between data condition and the estimation properties. Both represented by p-value of KST and the correlations those presented in Table 1 respectively. This correlation analysis is based on 50 pairs of X and Y those we got in simulation.

There are 2 significant correlations. Both of them are positive. The first correlation is the correlation between degree of fitness Y to G-P and degree of fitness estimator $\hat{Y}$ to Y. With positive correlation, its mean the better Y fit to G-P the better confirmation of distribution between Y and its estimator. Table 1 shows that with p-value as low as 0.108, with high correlation between X and Y, the goodness measurement indicate good estimation, except for the distribution confirmation between Y and $\hat{Y}$.

The second significant correlation is correlation between X and Y correlation between Y and $\hat{Y}$. With positive correlation, its means that the higher correlation between X and Y, the higher correlation between Y and its estimator. This correlation did not effect by the goodness of fit of Y to G-P. Even with p-value of fitting Y achieves 0.052, the correlation of estimation still close to the correlation of the data.

On the other hand, the distance between Y and $\hat{Y}$ which is described by MAPE is correlated individually neither with the correlation between response and explanatory variable nor with the goodness of fit response variable to G-P. They may explain MAPE simultaneously. We analyzed the simultant effect of goodness of fit Y to G-P and Cor(x, y) using multiple regression. Table 3 shows the result of multiple regression.

Table 3: Regression analysis for MAPE

Coefficients	Estimate	Std. Error	T	value $Pr(>\|t\|)$
Intercept	0.57501	0.10617	5.416	2.74e−06 ***
Goodness of fit	−0.31146	0.09395	−3.315	0.00189 **
Cor(x, y)	−0.33391	0.09841	−3.393	0.00152 **

The regression has p-value: 0.002352 for F-statistic. It means the degrees of fitness of Y to G-P and correlation between X and Y simultaneously have significant effect to MAPE value. Negative values of the estimate of goodnessof fit and Cor(x, y) indicate that the better Y fit to G-P and the higher correlation between X and Y, the smaller MAPE value. It means the better condition of data for modeling the better model and estimation of response variable.

6. Conclusion

Data which is G-P distributed can be analyzed using GLM gamma. The goodness of estimation of response variable depends on two factors. The first factor is the goodness of fit of the response variable to G-P distribution. The other factor is correlation between response and explanatory variable. Good fit response variable to G-P and high correlation between response and explanatory variable will yield good model. Just like common modeling, as long as the data fulfill the good condition of modeling, the model and the estimation will be good.

References

[1] A. Alzaatreh, F. Famoye and C. Lee, Gamma-Pareto distribution and application, Journal of Modern Applied Statistical Methods, 11 (2012), Issue 1, Article 7.

[2] Balajari, GLM with a Gamma-distributed Dependent Variable. 2013, http://civil.colorado.edu/ balajir/CVEN6833/lectures/GammaGLM-01.pdf, downloaded at 11 August 2015.

[3] H.L. Crutcher, A note on the possible misuse of Kolmogorov-Smirnov Test, Journal of Applied Meteorology (1975) 14:1600–1603

[4] A.J. Dobson, An Introduction to Generalized Linear Models, second edition, Chapman & Hall/CRC, Florida, 2002.

[5] H. Hanum, A.H. Wigena, A. Djuraidah, and I.W. Mangku, Fitting extreme rainfall with Gamma-Pareto distribution, Applied Mathematical Sciences, Vol. 9, 2015, no. 121, 6029–6039 HIKARI Ltd, http://dx.doi.org/10.12988/ams.2015.57489.

[6] H. Hanum, A.H. Wigena, A. Djuraidah, and I.W. Mangku, Developing generalized linear model of Gamma-Pareto Distribution, Far east Mathematical Sciences Journal, 2016, to be published.

[7] C.D. Lewis, Industrial and Business Forecasting Methods, Butterworths, London, 1982 in J.J.M. Moreno, A.P. Pol, A.S. Abad, and B.C. Blasco, Using the R-MAPE index as a resistant measure of forecast accuracy, Psicothema 2013. Vol 25, No 4, 500–506 doi:10.7334/psicothema2013.23.

[8] P. McCullagh and J.A. Nelder, Generalized Linear Models, second edition, Chapman & Hall, London, 1989.