Comparing Gamma and Log-Normal GLMs in R Using Simulation and Real Data Set

Dr. Nagham Mohammad

Lucinda McGivern

University of Guelph, 50 Stone Rd E, Guelph, ON, Canada

August, 24^{th}, 2020

In regression analysis courses, there are many settings in which the response variable under study is continuous, strictly positive, and right skew. This type of response variable does not adhere to the normality assumptions underlying the traditional linear regression model, and accordingly may be analyzed using a generalized linear model assuming either a lognormal or gamma distribution. As such, students oftentimes become confused about which of these two distributions should be chosen to model a given daset. In this article, we argue that the comparability of these two models should be taught through both simulation and real data analysis. Students will learn to identify the cases in which these two models can be used somewhat interchangeably through this teaching methodology.

KEYWORDS: AIC, Confidence Intervals, Residual Deviance

1 INTRODUCTION

In the analysis of data with non-normal error distributions, generalized linear models (GLMs) assuming either a gamma or lognormal distribution can oftentimes both suitably model the phenomena under study. In particular, the gamma and lognormal distributions are appropriate choices when modelling continuous, positively skewed data with a constant coefficient of variation (CV).

In this article, we demonstrate that students can be taught about the comparable results between these analyses using both real and simulated data. Students will learn about the data features and distributional assumptions that must be met in order for these gamma and lognormal models to be used rather interchangeably. In particular, the sensitivity of these models to the sample size and gamma shape parameter of the given dataset are explored.

The issue of whether or not analyses assuming these distributions will produce similar results has been discussed in the literature (Firth 1988; Atkinson 1982). Nath and Das (2012) demonstrated that, although the assumption of a constant coefficient of variation may be met, regression estimates may be different between gamma and lognormal models. Wiens (1999) explored how censoring, model misspecification, and the presence of outliers can contribute to the discrepancies in lognormal and gamma analysis in real data sets. Atkinson (1982) indicated that, given a non-constant variance, lognormal and gamma analyses may yield dissimilar results. However, these analyses are largely carried out on real datasets that tend to deviate from the assumptions of the gamma and lognormal models; in contrast, our analyses are first carried out on simulated data. As such, we are able to clearly demonstrate how adherence to these model assumptions produces similar results between gamma and lognormal analyses.

The purpose of this article is not to discuss the dissimilarities between gamma and lognormal analyses in certain cases (as is the case in the aforementioned articles). Instead, we want to present a method of instruction by which students will come to understand the situations in which both analyses assuming gamma and lognormal distributions are appropriate.

The structure of this article will be as follows. In section 2, data simulated from a gamma distribution over various shape parameters are produced. Analyses of this dataset assuming gamma and lognormal distributions are shown to yield similar results, and inferences are drawn about the sensitivity of these models to the shape parameter and sample size of the simulated data. In section 3, this instruction is then extended to a real insurance claim severity dataset. Analyses assuming lognormal and gamma distributions are carried out in both interaction and non-interaction models, and are found to give comparable results in both cases. Conclusions about the effects of the variance and gamma shape parameter of the dataset on the similarity of these analyses are presented.

1.1 Teaching Gamma and LogNormal Analysis Using Simulations

Students in regression courses are often taught that the gamma and lognormal distributions are appropriate choices when modelling continuous, positively skewed data with a constant coefficient of variation. These students, however, are often confused as to which distribution they should assume the response data follows. We have found that fitting both gamma and lognormal models to simulated data is an effective way of demonstrating that either model choice is adequate in many cases. The use of simulated data is, in this context, an excellent way to ensure that the data is continuous, unimodal, positively skewed, and adhering to the model assumptions of constant coefficient of variation and constant variance. Real datasets are not guaranteed to meet all of these model assumptions, and may further confuse the issue for these students. Moreover, simulations can readily demonstrate that datasets with large sample sizes and gamma shape parameters will yield similar results between analyses assuming gamma and lognormal distributions, as these data features can be specified in the simulation.

1.2 Determination of Gamma Shape Parameter

Simulations produced from the gamma distribution with a small shape parameter α may be subject to some numerical inaccuracies. In particular, as α tends to zero, the gamma distribution similarly converges to a concentrated point mass at zero. The algorithm employed by the rgamma function in R encounters this particular problem; for small values of α, the gamma distribution may return values so small that they will be represented as zero in computer arithmetic (Liu et al. 1993). The glm function in R is fitted using the iteratively reweighted least squares method, which in this case maximizes the log-likelihood of the gamma distribution function.

$$\ell(\alpha, \beta) \;=\; -n\alpha \ln \beta \;-\; n \ln \Gamma(\alpha) \;-\frac{1}{\beta}\sum_{i=1}^{n} x_i + \;(\alpha - 1)\sum_{i=1}^{n} \ln x_i \tag{1}$$

Exact values of zero in the data sample will then necessarily cause the simulation to fail. Omission of these zero values will render the simulation viable, but will also effectively reduce the sample size of the simulation (a result that becomes untenable for small sample sizes as the number of parameters increases). Several issues were encountered when fitting gamma and lognormal GLMs to data generated from gamma distributions with a shape parameter $\alpha \leq 0.4$. The glm function in R failed to converge for many of these cases. In the few instances that the glm function did converge, the confidence intervals associated with the predicted values for both the gamma and lognormal models were unreasonably large. As a result of these practical limitations, one-parameter analyses were carried out using a shape parameter of $\alpha = 0.5$, 0.6, 0.7, 0.8, 0.9, 1, 5, and 10. Two-parameter simulations were analyzed only in the cases with a shape a parameter of $\alpha = 0.7$, 0.8, 0.9, 1, 5, and 10. Alternative algorithms, which promise improved accuracy and efficiency when simulating from a gamma distribution with a small shape parameter, have been established by Liu et al. (1993) and Kundu and Gupta (2007).

1.3 Simulation Method

We began the simulation process by drawing explanatory variables from a uniform distribution $X_i \sim Uniform(n, 0, 1)$ using the runiform function in R. We carried this process out across sample sizes n=15, 30, 50, 100, and 1000, as well as for one and two parameter trials. Response data was simulated from the gamma distribution $X_i \sim Gamma(\alpha, \beta)$, where the values of the gamma scale parameter β were taken to be

$$\beta = \alpha / exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p) \tag{2}$$

and the values of the shape parameter α were established as above. We used the rgamma function to draw samples of the appropriate size according to these parameters α and β. Gaussian and gamma family GLMs, both using log-links, were fit to this data set assuming the model $y \sim X_1 + \ldots + X_i$ according to the number of parameters, i, under simulation.

1.4 Analyses of the Simulated Data

When presenting these analyses in class, we would start by discussing the single parameter simulations using a small shape parameter ($\alpha < 1$). In many of the smaller sample cases ($n = 30$), the gamma and lognormal regression estimates fell within ±0.25 units of one another. However, as sample size was increased to n=1000, these estimates began to approximate both one another as well as the true covariate values (see Table 1). We would here note that the proximity of regression estimates to the true covariate values appeared to be very sensitive to changes in sample size when using response data with a small shape parameter.

Table 1. Regression estimates of Gamma and Lognormal models given response data with a shape parameter $\alpha = 0.5$ and true values of $\beta_0 = 0.5$ and $\beta_1 = 1.2$.

	Gamma	Lognormal	Sample Size
$\hat{\beta}_0$	0.74579	0.84002	30
$\hat{\beta}_1$	0.82767	0.64991	
$\hat{\beta}_0$	0.47852	0.49283	1000
$\hat{\beta}_1$	1.39043	1.36859	

Once the students have grasped the behaviours of these comparable models over datasets with small gamma shape parameters, we extend our discussion into the single parameter simulations using a larger shape parameter ($\alpha > 1$). The lognormal and gamma GLM regression estimates in these cases converged to both one another as well as to the true covariate values, even at smaller sample sizes (see Table 2 and 3).

Table 2. Regression estimates of gamma and lognormal models given response data with a shape parameter $\alpha = 10$ and true values of $\beta_0 = 0.5$ and $\beta_1 = 1.2$.

	Gamma	Lognormal	Sample Size
$\hat{\beta}_0$	0.43801	0.44984	30
$\hat{\beta}_1$	1.21189	1.19330	
$\hat{\beta}_0$	0.48510	0.47447	1000
$\hat{\beta}_1$	1.19923	1.21693	

Table 3. Some of the predicted values of the gamma and lognormal models given response data with a shape parameter $\alpha = 10$, sample size of $n = 30$, and true values of $\beta_0 = 0.5$ and $\beta_1 = 1.2$.

Prediction Output	
Gamma	Lognormal
2.508154	2.588879
2.404936	2.500109
3.079315	3.069642
2.031268	2.173056
2.870337	2.895657
2.907224	2.92652

We would note that the proximity of these regression estimates to the true covariate values appeared far less responsive to changes in sample size than did those fitted to small-α response data.

As a part of this instruction, we also present the measures of goodness of fit by which a student can assess how each model fits the given dataset. In particular, we compare model fit using deviance residuals (see Table 4). We also present a comparison of the ninety-five percent confidence intervals for the predicted values between the models (Figures 1 and 2).

Table 4. Gamma and Lognormal Deviance Residuals over various sample sizes given $\alpha = 10$

Sample Size	Gamma Deviance Residual	Lognormal Deviance Residual
30	0.016923	0.09876944
50	0.35656	0.97976655
100	0.508188	2.49771439
1000	0.048169	0.13175174

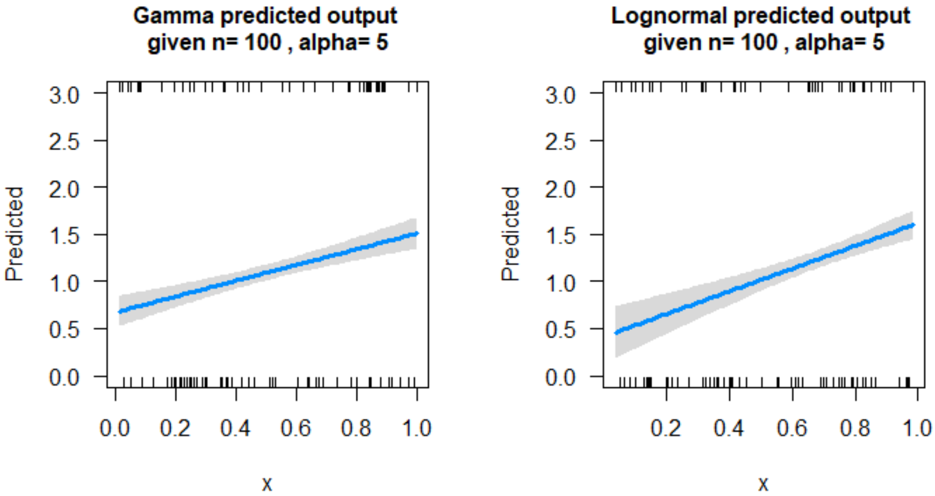

Figure 1. Gamma (left) and lognormal (right) prediction plots and associated confidence intervals for a dataset sample of size n=100 simulated from a gamma distribution with a shape parameter of $\alpha = 5$.

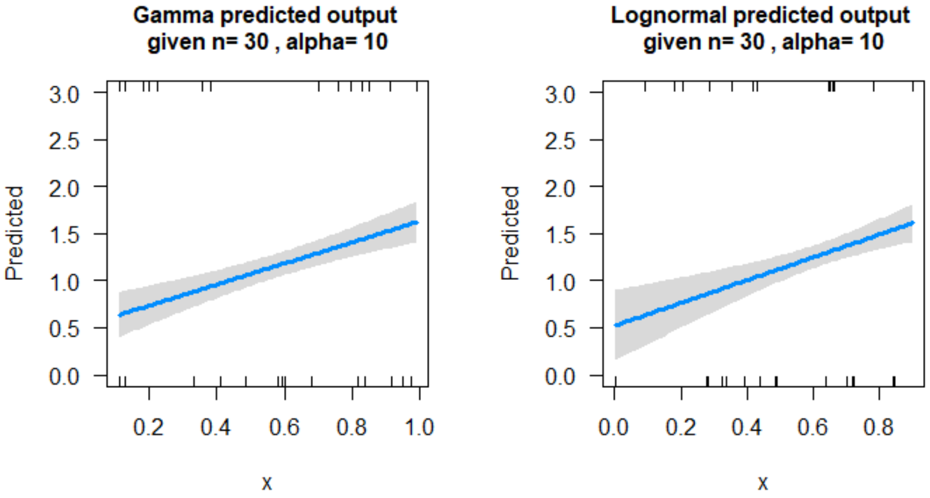

Figure 2. Gamma (left) and lognormal (right) prediction plots and associated confidence intervals for a dataset sample of size n=30 simulated from a gamma distribution with a shape parameter of $\alpha = 10$.

Two parameter simulations with a large shape parameter ($\alpha > 1$) produced covariate estimates that were similar within ±0.15 units across simulations with large sample sizes (n=1000). Instruction of this section should be concluded by explaining that, given a large shape parameter α, the gamma distribution approximates the normal distribution with a mean of μ=$\alpha\beta$. In particular, as the shape parameter α increases, the skew of the distribution decreases. As such, the student should grasp that these two analyses will yield similar results given a dataset with a sufficiently large estimate of the gamma shape parameter, or given a sufficiently large sample size.

2 EXAMPLE: INSURANCE CLAIM SEVERITY

In this section, we carry out both a gamma and a lognormal analysis on a real data set from the French Motor Personal Line dataset five (freMPL5) within the R package CASdatasets (see Figure 3). This data set includes claim amount, claim history, and risk predictors for a set of autoinsurance policies from the year 2004.

The purpose of this analysis was to determine how the number of claims made in the past four years would impact the current claim severity, after discriminating for previous claim type. The explanatory variables under study were as follows:

- ClaimNbResp – Number of responsible claims in the 4 preceding years.

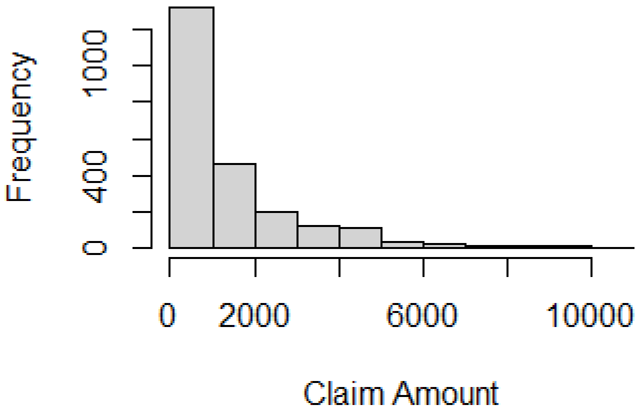

Figure 3. Autoinsurance claim amounts for a sample of n= 2297 from the freMPL5 dataset in CASdatasets

- ClaimNbNonResp – Number of non-responsible claims in the 4 preceding years.

- ClaimNbParking – Number of parking claims in the 4 preceding years.

- ClaimNbFireTheft – Number of fire-theft claims in the 4 preceding years.

- ClaimNbWindscreen – Number of windscreen claims in the 4 preceding years.

Interactions between these predictors were also tested for. The response variable, 'ClaimAmount', was the total claim amount of the guarantee. An initial model assuming a gamma distribution with a log-link was fit using selection by AIC. This same model was then fit assuming a gaussian distribution with a log-link.

Presentation of this dataset should indicate that, while the assumption of a constant coefficient of variation was not met for this dataset, regression estimates between the models were still comparable (Table 5). Across both models, the number of responsible, non-responsible, and windscreen claims in the past 4 years appeared to significantly affect the current claim size. After removing the interaction term from this model, analysis was again carried out assuming both gamma and lognormal distributions with a log-link (see Figure 4). Covariate estimates were again comparable between the models (Table 6).

Table 5. Regression estimates from interaction models assuming a gamma (top) and lognormal (bottom) distribution.

	Gamma Model			
	Estimate	Std. Error	t value	Pr(> |*t*|)
(Intercept)	7.18943	0.0367	195.899	< 2e-16 ***
ClaimNbResp	0.15235	0.04375	3.482	0.000506 ***
ClaimNbNonResp	0.06684	0.04374	1.528	0.126629
ClaimNbParking	0.10293	0.07159	1.438	0.150612
ClaimNbWindscreen	-0.16272	0.03758	-4.33	1.56e-05 ***
ClaimNbNonResp:ClaimNbWindscreen	0.07184	0.05214	1.378	0.168389
Shape Parameter	0.886829			
	Lognormal Model			
	Estimate	Std. Error	t value	Pr(> |*t*|)
(Intercept)	7.20493	0.03629	198.541	< 2e-16 ***
ClaimNbResp	0.14954	0.03604	4.15	3.45e-05 ***
ClaimNbNonResp	0.04522	0.0374	1.209	0.226709
ClaimNbParking	0.07317	0.06484	1.129	0.259203
ClaimNbWindscreen	-0.18419	0.04857	-3.793	0.000153 ***
ClaimNbNonResp:ClaimNbWindscreen	0.10698	0.05046	2.12	.034122 *

Table 6. Regression estimates from non-interaction models assuming a gamma (top) and lognormal (bottom) distribution.

	Gamma Model			
	Estimate	Std. Error	t value	Pr(> |*t*|)
(Intercept)	7.17831	0.03582	200.391	< 2e-16 ***
ClaimNbResp	0.15157	0.04371	3.468	0.000535 ***
ClaimNbNonResp	0.10151	0.03735	2.718	0.006626 **
ClaimNbParking	0.10072	0.07152	1.408	0.159177
ClaimNbWindscreen	-0.13692	0.03219	-4.253	2.19e-05 ***
Shape Parameter	0.886115			
	Lognormal Model			
	Estimate	Std. Error	t value	Pr(> |*t*|)
(Intercept)	7.19085	0.0357	201.446	< 2e-16 ***
ClaimNbResp	0.14911	0.03613	4.127	3.81e-05 ***
ClaimNbNonResp	0.07548	0.03182	2.372	0.017778 *
ClaimNbParking	0.06833	0.06473	1.056	0.291212
ClaimNbWindscreen	-0.12919	0.03835	-3.369	0.000767 ***

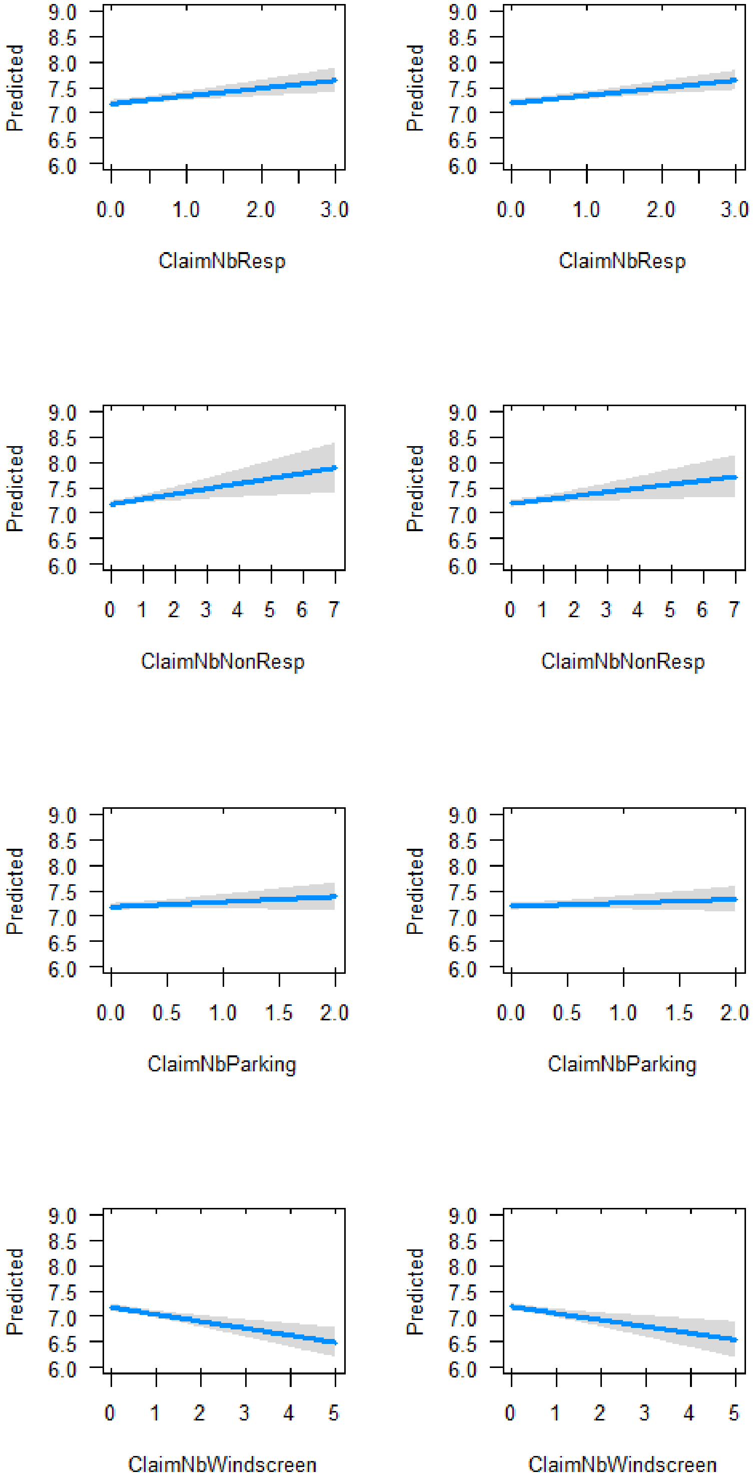

Figure 4. Prediction plots and associated confidence intervals for each predictor in the gamma (left) and lognormal (right) non-interaction models.

Atkinson's criterion states that two analyses should provide similar results given the condition that the variance $\sigma^2 \leq 0.6$ (Atkinson 1982). This dataset does not adhere to the assumption of constant variance, and as such this modest discrepancy between analysis results is consistent with Atkinson's criterion not being met.

Students presented with this example will learn that, even in the cases where the assumption of a constant coefficient of variation is not met, the gamma and lognormal model will oftentimes produce comparable results when analyzing continuous, positively skewed data. An emphasis will be placed on the fact that the proximity of these regression estimates can especially be predicted by the size of the dataset under study (as in agreement with the simulation analyses).

3 CONCLUDING REMARKS

Simulation analysis can unambiguously convey the manner in which certain dataset parameters produce similar regression estimates under comparable models. As such, teaching methods that make use of simulations help to demonstrate the conditions under which analyses assuming gamma and lognormal distributions will produce the same results. Following these methods, students will learn about the data features and distributional assumptions that must be met in order for these gamma and lognormal models to be used fairly interchangeably.

References

[1] Atkinson, A. C. Regression Diagnostics, Transformations and Constructed Variables. *Journal of the Royal Statistical Society: Series B (Methodological) 44*, 1 (1982), 1–22.

[2] Firth, D. Multiplicative Errors: Log-Normal or Gamma? *Journal of the Royal Statistical Society. Series B, Methodological 50*, 2 (1988), 266–268.

[3] Kundu, D., and Gupta, R. D. A convenient way of generating gamma random variables using generalized exponential distribution. *Computational Statistics and Data Analysis 51*, 6 (2007), 2796–2802.

[4] Liu, C., Martin, R., and Syring, N. Efficient simulation from a gamma distribution with small shape parameter. *Computational Statistics 32*, 4 (2017), 1767–1775.

[5] Nath Das, R., and Park, J. Discrepancy in regression estimates between log-normal and gamma: some case studies. *Journal of Applied Statistics 39*, 1 (2012), 97–111.

[6] Wiens, B. When Log-Normal and Gamma Models Give Different Results: A Case Study. *The American statistician 53*, 2 (1999), 89–93.