5-1-2019

A FLEXIBLE DISTRIBUTION IN MODELING SURVIVAL DATA

Ian Harris

A FLEXIBLE DISTRIBUTION IN MODELING SURVIVAL DATA

by

IAN HARRIS

Presented to the Faculty of the Honors College of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

HONORS BACHELOR OF SCIENCE IN MATHEMATICS

THE UNIVERSITY OF TEXAS AT ARLINGTON

May 2019

ACKNOWLEDGMENTS

I would like to acknowledge the following people.

I would like to thank my family, for their constant encouragement to help me see this project through.

I would like to thank my mentor Dr. Suvra Pal, whose devotion to the subject matter helped in my motivation and whose explanations of the subject helped clear any gaps of understanding. I sincerely appreciate his time and patience towards this project's completion.

I would also like to especially thank Zachary Loucks, who had a huge part in writing the codes that allowed the experiment of this project to happen. I cannot thank him enough for his unfailing help toward this project's completion.

April 15, 2018

A FLEXIBLE DISTRIBUTION IN

MODELING SURVIVAL DATA

ABSTRACT

Ian Harris, B.S. Mathematics

The University of Texas at Arlington, 2018

Faculty Mentor: Suvra Pal

When analyzing survival data, which involves such parameters as lifetime, censoring rate, and any number of covariates, we have several distributions to try to fit the study into a model. Among these are the exponential, the gamma, the lognormal, and the Weibull distributions. The problem with these distributions is that their parameter requirements are quite stiff and not flexible. So, if some parameters are even slightly off (or otherwise unknown), how would we be able to model the data and, better yet, see if the data falls outside the given distributions? That is where the generalized gamma distribution comes in. The beauty of this distribution is how malleable it is and how it can be used as a blanket distribution of sorts to catch datasets that fall outside the commonly used distributions. Using R software, we performed a simulation study in which we generated datasets under the generalized gamma distribution and compared different iterations of the simulated data to models of the different distributions in a likelihood ratio test to show the

rejection rates of models whose parameters differ. As the number of generated generalized gamma datasets increased (50 to 300 to 500), the rejection rates among different parameters (Q=0 vs. Q=0.5 to name one) grew larger and larger whilst the fixed vs. fitted model comparisons of the same parameter grew closer and closer to a 5% rejection rate. With this as a background, we applied the generalized gamma distribution to a real dataset, whose parameters were unknown, to estimate its parameters. Although it didn't fall into any of the special cases, it still could fit in the generalized gamma distribution.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 Survival Analysis

Lifetime data modeling, according to Klein and Moeschberger (2003), has a sizable list of statistical distributions to represent the data of any medical or lifetime study. Of course, what do we mean by "lifetime study"? This can be, interchangeably, "survival analysis." Survival analysis is generally defined as a set of statistical methods for analyzing data in a study where the variable is the time until the occurrence of an event of interest, which is often called a lifetime. The event can be death, occurrence of a disease, and so forth. Survival analysis is essentially a lifetime study where the variable outcome is usually death. Say, for example, a hospital wants to administer a treatment for a certain occurrence of a cancerous tumor. The study would be carried out over a designated block of time (e.g. 15 years), and the observed "lifetimes" for the treatment group are recorded. However, these collected data sets contain what is known as "censored data". Censored data, according to Klein and Moeschberger (2003), shows up when a treatment group individual's lifetime is only known to occur within a certain period of time. Common censoring techniques include right censoring, when all we know is the individual lived through the study period and didn't experience the event (e.g. death), as well as left censoring where the individual experienced the event prior to the start of the survival study. In most survival studies, parametric studies in particular, authors have used statistical distributions to model the survival data. Due to censoring, which comes naturally in any

survival data, there are no formal goodness-of-fit tests to check if the chosen distribution provides an adequate fit to the data. In this regard, one can fit several distributions and use information-based criteria, namely the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), to check which distribution results in the minimum AIC and BIC values. Traditionally, the distribution with the minimum AIC and BIC values is supposed to provide the best fit to the data. There are four distributions that stand out as the most useful methods of best fit modelling the clear majority of obtained lifetime data.

1.1.1 The Exponential Distribution

The simplest of the four is the exponential distribution. According to Balakrishnan and Basu (1995), the exponential distribution is mathematically tractable. What makes the exponential distribution unique is its constant hazard rate, λ. Assume for any distribution, f(t) is the probability density function, F(t) is the cumulative distribution function, S(t) is the survival function (probability that an event will happen after time t, given as $S(t) = 1-F(t) = P(T > t)$), h(t) is the hazard function (probability that an event will happen in the next instance of time t, given that the event in question has not yet occurred until time t), and H(t) is the cumulative hazard function. In this particular case, the exponential distribution can be denoted as $T \sim Exp(\lambda)$. For $t > 0$,

$$f(t) = \lambda e^{-\lambda t} \text{ for } \lambda > 0 \text{ (scale parameter)}$$

$$F(t) = 1-e^{-\lambda t}$$

$$S(t) = e^{-\lambda t}$$

$$h(t) = \lambda \leftarrow \text{constant hazard function}$$

$$H(t) = \lambda t.$$

Balakrishnan and Basu (1995) state that "while the exponential distribution's constant hazard rate makes the simple form of the distribution inadequate to describe real-life complexity, it often serves as a bench mark with reference to which effects of departures allow for a specific type of disturbance to be assessed." This distribution is also essentially a watermark model from which other, more complex distributions can stem.

1.1.2 The Weibull Distribution

One such distribution is the Weibull distribution. This distribution, described by Murthy et al. (2004) as well as Rinne (2008), is an extremely flexible distribution with a hazard function that can increase, decrease, or remain constant, making it more applicable to a wide range of survival studies. This distribution's functions of interest are given by:

$$F(t) = 1-e^{(-(\lambda t)^{\wedge}\alpha)}$$

$$f(t) = \alpha(\lambda^{\alpha})(t^{\alpha-1})\, e^{(-(\lambda t)^{\wedge}\alpha)}$$

$$h(t) = \alpha(\lambda^{\alpha})(t^{\alpha-1})$$

$$H(t) = (\lambda t)^{\alpha}$$

In the equation above, $t > 0$, $\lambda > 0$ is the scale parameter, while $\alpha > 0$ is the shape parameter of the distribution.

1.1.3 The Lognormal Distribution

Another special distribution is the lognormal distribution. Crow and Shimizu (1988) describe the lognormal distribution as incredibly flexible in multiple fields of lifetime study, most notably species abundance data, count data, and so forth. It can be modeled with elements of both the exponential and gamma distributions. The lognormal distribution is denoted $LN(\mu,\sigma2) \sim \exp\{N(\mu,\sigma2)\}$. The quantities of interest are defined as follows:

$$F(t) = \Phi((\log(t)-\mu)/\sigma)$$

$$f(t) = \varphi((\log(t)-\mu)/\sigma)/t\sigma$$

$$h(t) = f(t)/F(t)$$

where φ and Φ are the probability density function and cumulative distribution function of the standard normal distribution, respectively.

1.1.4 The Gamma Distribution

Finally, another widely used distribution in model building is the gamma distribution. This distribution, according to Thom (1958), has plenty of flexible applications in data modeling, including maximum likelihood estimators for the data, useful for reliability and weather data to name a few. The gamma distribution can be seen as a generalization of the exponential distribution, since there are real-life phenomena for which an associated survival distribution is approximately gamma and simple functions of random variables can have a gamma distribution. The gamma distribution is given as:

$$f(t) = (\lambda^{\alpha}\, t^{\alpha-1}\, e^{-\lambda t})/\Gamma(\alpha)$$

$$\text{where } \Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1}\, e^{-t}\, dt$$

is known as the gamma function.

Parameters $\lambda > 0$ and $\alpha > 0$ are the scale and shape of this distribution. Note that if the shape parameter was equal to 1, the distribution becomes exponential.

However, there is an underlying problem with all four of these given distributions. Not all of them are nested in each other, so there is a high possibility of a best-fit test result for a survival dataset being misleading as the information-based methods do not give the user any warning of how good or bad the best-fit actually is. For instance, even if all the chosen distributions provide the worst possible fit to the data, the AIC and BIC criteria

would still select one as the best-fit distribution. So, how can we work with the stiff requirements of the above four distributions, or be more confident in our results without the nagging doubt that the best-fit is not our own techniques just picking the lesser of a group of bad models?

1.2 Generalized Gamma Distribution

This is where the generalized gamma distribution becomes important. This particular distribution circumvents the best-fit problem and introduces researchers to a wider class of lifetime distributions which is flexible in the sense that it includes the aforementioned lifetime distributions as its special cases. Since the four above cases, or sub-families, are nested within the bigger family distribution, it allows formal tests of hypotheses to be carried out to select the best distribution within the family. The resulting distribution would model the lifetime data most accurately and would thus allow inference to be drawn with a minimum amount of bias. This would ultimately minimize errors in estimating the survival probabilities, comparing survival trends between two or more groups, and identifying influential covariates.

CHAPTER 2

SIMULATION STUDIES AND METHODOLOGY

2.1 R Software and Generalized Gamma

Using R software, the aim of the given experiment is to demonstrate the flexibility of the generalized gamma distribution through the *flexsurv* package. The beauty of this package is its ability to generate multiple randomized datasets through commands. The command used in this experiment's coding was rgengamma(n, μ, σ, Q). In this case, n is the number of observations per dataset. For example, n=500 equals a survival generalized gamma dataset with 500 participants, whether or not they're censored. μ is generally the value of the location parameter of the distribution. σ is the scale parameter and Q is the shape parameter.

Modeled by Stacy (1962) as well as Prentice (1974), the flexsurv package in R views the generalized gamma in the following way. If $\gamma \sim \text{Gamma}(Q^{-2},1)$, and $w = \log(Q^2\gamma)/Q$, then $x = e^{(\mu+\sigma w)}$ follows the generalized gamma distribution with probability density function (exp(.) means $e^{(.)}$):

$$f(x|\mu,\sigma,Q) = (|Q|(Q^{-2})^{Q\text{^}-2} / \sigma x \Gamma(Q^{-2})) / \exp(Q^{-2}(Qw - \exp(Qw))).$$

The kicker to this experiment is that depending on the shape parameter Q's value, the generalized gamma can be simplified to the other major distributions. Within the R data package, the altering Q on the density function results in the following special case behaviors:

$$\text{Dgengamma}(x, \mu, \sigma, Q=0) = \text{dlnorm}(x, \mu, \sigma) \text{ [Lognormal]}$$

$$\text{Dgengamma}(x, \mu, \sigma, Q=1) = \text{dweibull}(x, \text{shape}=1/\sigma, \text{scale}=\exp(\mu)) \text{ [Weibull]}$$

$$\text{Dgengamma}(x, \mu, \sigma, Q=\sigma) = \text{dgamma}(x, \text{shape}=1/\sigma^2, \text{rate}=\exp(-\mu) / \sigma^2) \text{ [Gamma]}$$

2.1.1 Simulated Datasets

So in order to fully demonstrate the flexibility of the generalized gamma distribution, we first constructed a code to generate around 500 datasets of a decently large size per dataset (500 participants per set in our case). In this code, every dataset contains a single covariate (in our case, a simple binomial covariate of size 1 with probability 0.5 was used for the sake of minimizing the amount of coding required). The code only counts the datasets that converge in its calculation of the overall average of the parameter estimates, as well as the parameter confidence intervals (95%), standard error, and bias.

2.1.2 Likelihood Ratio Tests

Next, we wrote a sizeable code to perform a likelihood ratio test of differing sample sizes of randomly generated datasets in order to construct a four-by-four matrix of rejection rates between fixed versus fitted values of Q, those values being Q=0 (lognormal), Q=0.5 (σ in our whole experiment, gamma), Q=1 (Weibull), and Q=1.5 (generalized gamma). If the code were to go without error, the rejection rates should approach 1 the farther away from each other the compared Q values are whilst comparing the same values of Q should yield rejection rates approaching 5% (0.05). The higher the number of generated datasets, the more often the code should recognize when the averages of the shape parameters differ and thus reject the null hypothesis that the generated sets aren't different from a fixed set.

2.1.3 Real Dataset

Finally, with this ratio test as a back, we applied an altered version of the first code to an actual dataset in order to estimate its unknown model parameters rather than a set of

simulated sets whose parameters were entered into the function. This dataset came included in the flexsurv package in R and was called “bc”, which stood for breast cancer.

CHAPTER 3

RESULTS AND DISCUSSION

3.1 Simulated Data Parameter Estimates

Upon entering the parameters 1, 0.5, and 0 for μ, σ, and Q respectively and running the first code which randomly generated 500 generalized gamma lifetime datasets with an entered 0.2 censoring rate, the resultant output shown in R is indicated by the table below.

Table 3.1: Generalized Gamma Parameter Estimates

Simulated	Estimate	Lower 95%	Upper 95%	Std. Error	Bias
μ	0.996	0.892	1.099	0.053	-0.004
σ	0.497	0.456	0.542	0.022	-0.003
Q	-0.010	-0.315	0.295	0.155	-0.010

As seen in the above output, the estimates of the parameters come close enough to the actual parameter values to be considered valid estimates.

3.2 Likelihood Ratio Test Results

The second larger code took increasing numbers of generated datasets, each with increasing size per set, and performed a series of likelihood ratio tests, which involved the use of chi-square with log-likelihood values to compare the goodness of fit of two models. The results of the tests are indicated by the figures below.

Figure 3.1: Rejection Rates Fixed vs. Fitted (Size 50)

Samplesize, Iterations = 50

		Fixed		
Fitted	Q=0	Q=0.5	Q=1	Q=1.5
Q=0	0.080	0.041	0.483	0.741
Q=0.5	0.241	0.120	0.241	0.401
Q=1	0.542	0.260	0.143	0.182
Q=1.5	0.916	0.679	0.405	0.422

Given the small sample sizes and number of iterations, the resultant large variability resulted in almost completely random rejection rates. However, once the sample size per dataset and the number of datasets increased to 300, the resultant rejection rates where shown as output in the below figure.

Figure 3.2: Rejection Rates Fixed vs. Fitted (size 300)

Samplesize, iterations = 300

		Fixed		
Fitted	Q=0	Q=0.5	Q=1	Q=1.5
Q=0	0.080	0.763	1.000	1.000
Q=0.5	0.713	0.053	0.740	0.993
Q=1	1.000	0.777	0.063	0.607
Q=1.5	1.000	0.993	0.667	0.057

Now a pattern is shown, as the farther away the shape parameters are from each other, the more often the code rejects the models as a best fit for each other. Finally, in the figure below, the sample size and iterations are increased to 500, so as to show with more certainty what happens to our rejection rates.

Figure 3.3: Rejection Rates Fixed vs. Fitted (Size 500)

Samplesize, iterations = 500

	Fixed			
Fitted	Q=0	Q=0.5	Q=1	Q=1.5
Q=0	0.078	0.908	1.000	1.000
Q=0.5	0.926	0.056	0.758	1.000
Q=1	1.000	0.960	0.056	0.842
Q=1.5	1.000	1.000	0.864	0.058

Given that the special cases Q=0 (lognormal), Q=0.5 (sigma) and Q=1(Weibull) are not nested in each other, the rate of rejection increased along with sample size and iterations, while the test came close to a 5% rejection of best fit when comparing a model to itself.

3.3 Real Dataset Estimates

With these results, we ran the first code again for a real dataset (bc) in order to estimate its parameters. The resultant output is shown in the table below.

Table 3.2: Real Dataset Parameter Estimates

BC	Estimate	Lower 95%	Upper 95%	Std. Error
μ	7.838	7.582	8.095	0.131
σ	1.056	0.962	1.159	0.050
Q	-0.593	-1.059	-0.127	0.238
GroupMedium	-0.649	-0.868	-0.431	0.112
GroupPoor	-1.283	-1.502	-1.064	0.112

The covariate in this dataset's case was a qualitative group that a patient of the treatment group fell in; Good, Medium, or Poor. GroupMedium and groupPoor are percentage differences when compared to groupGood. The results indicate a 64% decrease in survival

rate for medium along with a whopping 128% decrease in survival rate for the poor group. The output shows that the μ is quite close to 7.8, the scale parameter σ is close to 1, and the shape parameter Q is about -0.6, which falls outside the realm of our special cases of lognormal, Weibull, gamma, etc. However, given that the generalized gamma distribution can allow our shape parameter to be any value, this real dataset still has a best fit in generalized gamma.

3.4 Future Research

So, the results ultimately show that the generalized gamma distribution is incredibly flexible, capable of fitting an extremely wide range of lifetime data sets in and outside the parameter bounds of other distributions. This would allow for more legitimate models and the study of far more complex phenomena in statistics; air pollution, drought analysis, reliability studies, the list is quite large.

REFERENCES

Balakrishnan, N. and Basu, A. P. (Eds.) (1995). The Exponential Distribution: Theory, Methods, and Applications. The Netherlands: Gordon and Breach Publishers.

Crow, E. L. and Shimizu, K. (1988). Lognormal Distributions: Theory and Applications. Marcel Dekker, New York.

Klein, J. P. and Moeschberger, M. L. (2003). Survival Analysis: Techniques for Censored and Truncated Data, Second Edition. Springer, New York.

Murthy, D. N. P., Xie, M. and Jiang, R. (2004). Weibull Models. Hoboken, New Jersey: John Wiley & Sons.

Prentice, R. L. (1974). A log gamma model and its maximum likelihood estimation. Biometrika 61, 539-544.

Rinne, H. (2008). The Weibull Distribution: A Handbook. New York: CRC Press.

Stacy, E. W. (1962). A generalization of the gamma distribution. Annals of Mathematical Statistics 33, 1187-1192.

Thom, H. C. S. (1958). A note on the gamma distribution. Monthly Weather Review 86, 117-122.

BIOGRAPHICAL INFORMATION

Ian Harris will graduate from UT Arlington in Spring 2019 with an Honors Bachelor of Science in Mathematics, with a focus on statistics. He is currently interested in both statistical research and academia, and is planning on applying to graduate school at UT Arlington in Fall 2019 in hopes of earning a Ph.D. in statistics, as well as becoming a mathematics professor. His long term goal is to teach mathematics and statistics at the college level and to pursue further research in statistical model-building.