

Araştırma Makalesi – Research Article

Weibull and Log-logistic yaşlanma modellerinin performansının *Saccharomyces cerevisiae* ömür verisi kullanılarak karşılaştırılması

Emine Güven^{1*}

Geliş / Received: 02/08/2019

Reviz / Revised: 09/01/2020

Kabul / Accepted: 09/01/2020

ÖZ

Ampirik yaşam veri setleri genellikle yaşlanma için en uygun matematiksel modellerle incelenir. Bu çalışmada, dikkatimizi tomurcuklanan maya bakterisi *S. cerevisiae* ömrüne ve bu bakterilerin en uygun yaşlanma modelinin belirlenmesine verdik. Model seçiminin maya bakterisi ömür veri kümelerindeki etkisini ve iki parametrelili Weibull (WE) ve Log-logistic (LL) yaşlanma modellerinin uyum sonuçlarını araştırdık. Bu modellerin her ikisi de yaşlanma araştırmalarında yaygın olarak incelenmekte ve uygulanmaktadır. Bir sağkalım fonksiyonu olarak, zamanla artan ve sonra azalan mortalite oranlarına karşılık gelen benzer bir eğilim gösterirler. Şu ana kadar yapılan çalışmalar genellikle Akdeniz meyve sinekleri, meyve sinekleri, ev sinekleri, un böcekleri ve insanların ömür verisi üzerinde bu modellerle çalışmalar yapılmıştır. Önceki araştırmalardan farklı olarak, dikkatimizi sonuçların ve kalibrasyonların ampirik ömür veri örnekleri üzerindeki etkisine odakladık. Beklediği gibi her iki model de birbirlerinin yerine kullanılabilir. Bununla birlikte, WE modelinin maya ömür verilerine $R^2=0.86$ ile LL modelinden önemli ölçüde daha iyi fit olduğunu gördük. Bu bulgu, tipik olarak hayatta kalma modelleri uygulandığından maya yaşlanma çalışmasında özellikle önemlidir ve bu nedenle hangi modelin maya verilerine daha uygun olduğunu öngörebilir. Bu makalede, karşılaştırmalarla geliştirilen bu yaklaşımın potansiyeli, laboratuvar BY4741 ve BY4742 değişime uğramamış referans suşlarının maya replikatif ömür veri setlerinin model karşılaştırması ile gösterilmiştir. Çalışmamız, deneysel ömürlerin model uyum sonuçlarının yorumlanmasının model seçimini dikkate alması ve sonuçlanan varyasyonu göz önünde bulundurulması gerektiğini vurgulamaktadır.

Anahtar Kelimeler- Weibull Model, Log-logistik Model, sağ-kalım analizi, yaşlanma ve maya bakterisi

*Sorumlu yazar iletişim: emine.guven@duzce.edu.tr (<https://orcid.org/0000-0001-9324-0879>)

Biyoenformatik Anabilim Dalı, Biyomedikal Mühendisliği Bölümü, Mühendislik Fakültesi, Düzce Üniversitesi, Düzce, Türkiye

A comparison between the performance of Weibull and Log-logistic Aging Models on *Saccharomyces cerevisiae* lifespan data

ABSTRACT

Empirical lifespan datasets are often studied with the best-fitted mathematical model for aging. In this study, we focus our attention to the budding yeast *S. cerevisiae* lifespan and the determination of the best-fitted model of aging. We investigate the influence of model selection in yeast lifespan datasets and the fitting outcomes of the two-parameter Weibull (WE) and Log-logistic (LL) models of aging. Both of these models are commonly studied and implemented in aging research. They show similar tendency as a survival function that they correspond to mortality rates that increase, and then decrease, with time. Studies so far has been usually done with medflies, *Drosophila*, house flies, flour beetles, and humans with these models. Different than previous research, we focus our attention on the influence of fitting results and calibrations on empirical lifespan data samples. As expected both of the models could be used as a substitute of each other. However, we also find WE model fits the yeast lifespan data significantly better than LL model with an $R^2 = 0.86$. This finding is especially important in yeast aging study because of typically survival models are applied and therefore one can see which model fits the yeast data better. In this article, comparisons are done and developed and the potential of the approach is demonstrated with a model comparison of yeast replicative lifespan datasets of the laboratory BY4741 and BY4742 wildtype reference strains. Our study highlights that interpreting model fitting results of experimental lifespans should take model selection and resulted variation into account.

Keywords- the Weibull Model, the Log-Logistic Model, Survival Analysis, Aging, Yeast

I. INTRODUCTION

The Weibull (WE) and Log-logistic (LL) models are survival models mostly used in reliability theory of systems such as biological and machineries based on laws of mortality. The WE model of aging is frequently in use to predict a power law mortality rate where the systems of machinery more homogenous. Whereas, the LL model is continuous probability distribution and has a non-monotonic hazard function and is frequently suitable to model cancer survival data. The LL model is understudy for events whose rate exhibits an increase initially and decrease later. Log-logistic distribution is a probability density function of a random variable where its logarithm is a logistic function [1]. For instance, both of the models are frequently used in order to estimate values of cancer aging or related disease hazards from observations [2].

The key purpose of ageing-related research is to determine and analyze the nature of the inevitable and irreversible damage that contributes age-related health disparities and diseases. Selecting between the best-fitting distributions for a given lifespan (aging) data is a significant leverage tool that contributes to understand aging and related health disparities. This issue of selecting the correct model of the given lifespan dataset is still understudy [4, 5].

Previously, no research has been done to compare the WE and LL model fits on yeast lifespan data. Therefore, our goal here is to infer a deeper understanding of how these very similar models can be distinguished and how the model selection helps us to improve aging study.

There are two primary ways to measure lifespan of budding yeast. The replicative lifespan (RLS) assay of a cell is the number of generation can divide. The assay is performed with the necessary lab tools by separation of daughter cells from mother cells manually [6]. The other assay is chronological lifespan (CLS), which refers to the length of time a mother yeast cell culture can survive post-diauxic and stationary phase [7]. It has been shown that chronologically old yeast also have a shorter RLS which indicates RLS and CLS are related in sharing major mechanisms [8, 9]. RLS measurements based on individual cells are often subject to maximal likelihood analysis, which is a significant method of analyzing our study here [10].

Accordingly, the objective of this current study is to fit the yeast lifespan data with the WE and LL models to estimate parameters using maximum likelihood parameter estimation technique. In principle, the maximum likelihood estimates (MLE) of parameters of survival models could be found analytically by solving a set of equations involving first partial derivatives of the logarithm of the likelihood function which is called log-likelihoods or probable outcomes. This approach basically maximizes the likelihood function of the targeted parameter which also corresponds to the joint probability of the observed data over a parameters space of the distribution. Solving the resulting set of the logarithm of the likelihood equations would lead to the parameter estimates. Only in some simple cases, the maximum likelihood problem could reveal an analytical solution. One can write the maximum likelihood estimator explicitly as a function of a given data. However, in many cases there is no explicit solution [11–13]. In order to estimate numerical model parameters which are in one-dimensional space, the numerical technique of maximum likelihood is used.

II. MATERIALS AND METHODS

A. Weibull Distribution

Mortality rate, or failure rate, in machine aging typically follows a power law i.e. the Weibull model of aging. The Weibull model with a given PDF and CDF is defined as

$$f_{\theta,\gamma}(t) = \gamma\theta^\gamma t^{\gamma-1} \exp(-\theta t)^\gamma, \quad \text{and} \quad F_{\theta,\gamma}(t) = 1 - e^{-(\theta t)^\gamma},$$

respectively where $t > 0$, θ is the scale and γ is the shape parameter. Moreover, the survival and hazard functions are given by

$$s(t) = \exp(-\theta t)^\gamma, \text{ and } h(t) = \theta \gamma t^{\gamma-1} \text{ respectively.}$$

B. Log-logistic Distribution

The log-logistic distribution is also a power law can be used as a suitable substitute for Weibull distribution. It can be used to model the lifetime of an object, the lifetime of an organism, or a service time [2]. The log-logistic model with a given PDF and CDF is defined as

$$f_{\lambda,\kappa}(t) = \frac{\lambda\kappa(\lambda t)^{\kappa-1}}{(1+(\lambda t)^\kappa)^2}, \text{ and } F_{\lambda,\kappa}(t) = \left(1 + \left(\frac{\lambda}{t}\right)^\kappa\right)^{-1},$$

respectively where λ is a scale and $\kappa > 0$ is the shape parameter. Further, the survival and hazard functions of Log-logistic distribution is given by

$$S(t) = \frac{1}{1+\lambda t^\kappa}, \text{ and } h(t) = \frac{\lambda\kappa t^{\kappa-1}}{1+\lambda t^\kappa}, \text{ respectively.}$$

C. Maximum Likelihood Estimation (MLE), AIC, and LogL values

One question rises up for MLE, unique identification of the parameters depending on the formulation of the model. Determining the unique parameters of a given model must be resolved before estimation can even be considered. Therefore, solving the maximum likelihood function analytically is tedious task for large samples [14]. Thus, an optimization procedure in R environment to obtain the MLEs of θ and γ (or λ and κ) is used. WE and LL models are compared by using AIC (Akaike Information Criterion) approach [15] to evaluate model fittings. AIC is defined as

$$AIC = -2 \text{LogL} + 2K,$$

where K is the number of model parameters, and LogL is a measure of model fit. The higher number LogL means the better fit a model reveals. As one can follow from the definition of AIC, the smaller AIC leads a better model choice [16]. For small sample sizes when $n/K < \approx 40$, the second-order AIC can be used. Since the RLS of the BY4742 strain was measured in a large size experiments, we do not prefer to use the second-order AIC. The definition of the second order AIC is as follows

$$AIC_c = -2\text{LogL} + 2K + (2K(K+1))/(n-K-1),$$

where n is the sample size.

The MLEs is performed on both of the aging models and compared AIC numbers and the maximum log-likelihood as an outcome of the yeast data. In Table 1, sample wild type genotype backgrounds; BY4742 and BY4741 and their fitting results are demonstrated. Population represents the number of yeast cells for the time duration given in minutes. Once we fit the yeast lifespan data with WE and LL aging models, we calculated the log-likelihood of these two models. We obtained the log-likelihood function of the Weibull distribution as

$$\log(L(\theta, \gamma|t_i)) = \text{LogL}(\theta, \gamma|t_i) = \sum_{i=1}^N \log[f_{\theta,\gamma}(t_i)]$$

$$\text{LogL}(\theta, \gamma|t_i) = \sum_{i=1}^N [\log(\gamma\theta^\gamma) - (\gamma - 1) \log(t_i) - (\theta t_i)^\gamma]$$

The log-likelihood function of the Log-logistic distribution can be derived as

$$\log(L(\lambda, \kappa|t_i)) = \text{Log}L(\lambda, \kappa|t_i) = \sum_{i=1}^N \log[f_{\lambda, \kappa}(t_i)]$$

$$\text{Log}L(\lambda, \kappa|t_i) = n \log \lambda - n \log \kappa + (\lambda - 1) \sum_{i=1}^N [\log t_i - n(\lambda - 1) \log \kappa - 2 \sum_{i=1}^N \log[1 + (t_i/\kappa)^\lambda]]$$

Table 1. Sample fitting results of WE and LL Models on the yeast RLS of BY4742 and BY4741 wild type genetic backgrounds.

Genotype	n	std _{LS}	median _{LS}	mean _{LS}	WE _{AIC}	WE _{LogL}	LL _{AIC}	LL _{LogL}
BY4742	80	10.90	18	21.18	598.58	-297.29	594.16	-295.08
BY4741	80	9.75	24	28.7	592.48	-294.24	584.08	-290.04
BY4742	60	8.53	19	21.06	428.49	-212.24	433.47	-214.73
BY4741	60	9.47	25	25.2	443.29	-219.64	450.46	-223.23

D. Data Analysis and Code Availability

We conducted analysis and codes in the R statistical environment. Sample codes of to fit and analysing empirical data can be found at <https://github.com/emineguven/WEvsLLcomparison2019>. Maximum likelihood estimations were performed using the `flexsurvreg()` functions in the `flexsurv` package [17]. RLS of *S. Cereviasive* was shared by the Kaerberlein group (personal communication via e-mail). BY4742 and BY4741 WT genetic backgrounds are pooled from the empirical data since the most populated experiments were these two genetic backgrounds.

Given a genotype lifespan sequence, the data $\{x_1, x_2, \dots, x_n\}$ has a random sample of size n from a known lifetime distribution function. The histogram in Fig.1 (A) presents the lifespan of WT BY4742 genotype. The RLS lifetime sequence for this single WT BY4742 is

18 43 11 22 29 42 18 44 30 22 31 8 13 28 20 24 24 40 44 24 33 22
19 11 18 39 26 33 21 29 48 17 36 12 41 43 40 21 45 26 12 11 7 11
14 9 16 13 19 12 17 8 17 16 17 16 20 11 10 15 12 11 25 21 16 12
14 8 11 15 13 7 18 27 12 9 23 14 15 26

with a population size n=80 in minutes.

The histogram in Fig.1 (B) presents the lifespan of WT BY4741 genotype. The RLS lifetime sequence for this single WT BY4741 is

34 47 39 25 25 6 52 19 21 27 23 26 22 28 38 30 21 32 46 53 22 25
19 34 22 18 38 48 53 34 17 25 35 29 36 28 36 37 29 35 19 15 33 28
18 22 24 13 21 21 27 19 24 40 22 19 24 18 29 24 20 24 14 15 7 26
18 26 15 22 29 23 18 20 18 27 21 27 20 18

with a population size n=80 in minutes.

The RLS of the BY4742 strain was measured in 2108 experiments, and the RLS of BY4741 wild type strain was measured in 381 experiments. We had fit the empirical data with both of the models under study where estimated the model parameters using MLE method. We then compared fitted lifespan parameter outcomes with the empirical lifespan by using Linear Regression and Coefficients of Variations values.

III.RESULTS

A.Comparison of RLS data of Initial Mortality by WE and LL Models

It is witnessed that the WE model fits the data using MLE method better than LL model. Previously, it has been found that this method of empirical data modeling under discrete sets of observations could describe the data well [1]. We further compared both models based on AIC and LogL fitting estimation values. There are two approaches to investigate the better model fit we used. In the first approach, we have compared the WE and LL models using the linear regression analysis of the LogL values model fits.

We find WE model fits the yeast lifespan data significantly better than LL model with an $R^2 = 0.86$. The other approach we performed is linear regression analysis between LogL and AIC of each of the model fits which reveals an $R^2 > 0.90$. Our results suggests that in contrast to the WE model, the LL model is more sensitive to variation in the initial mortality rate independently of aging-related mortality. This could be the reason for comparisons between wild type strains appear to support the intrinsic-causes such as the experimental procedure for especially yeast lifespan.

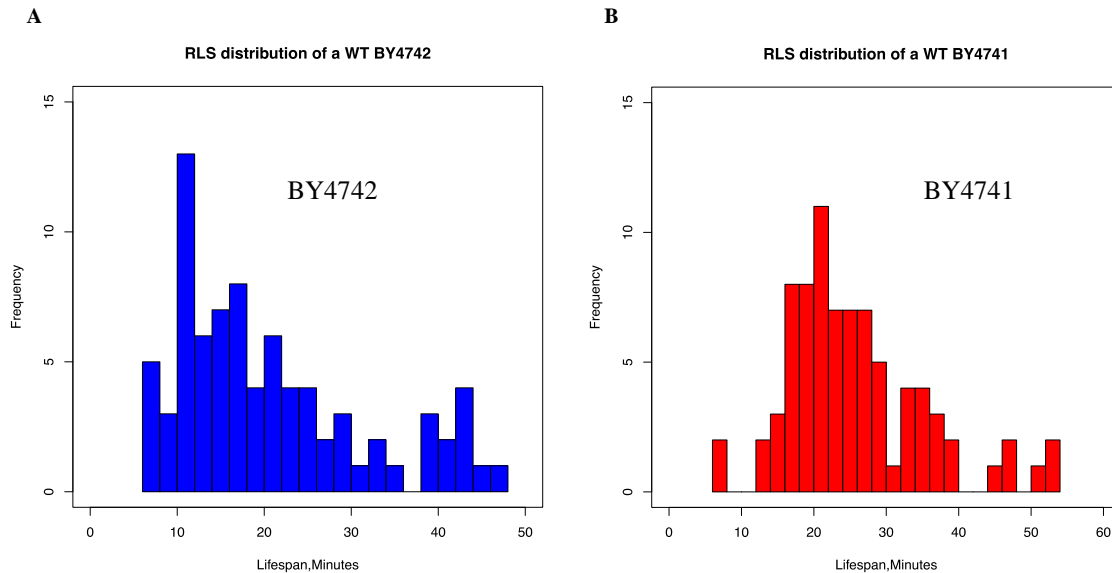


Figure 1. Distribution of (A) Replicative lifespan (RLS) of yeast data WT BY4742 genetic background. (B) Replicative lifespan (RLS) of yeast data WT BY4741 genetic background. Both of the histogram distribution follows a positive (left) skew. Thus, the WE and LL models are suitable models to fit the yeast lifespan data.

Fig. 2 demonstrates the lifetime of the yeast data for the genetic backgrounds BY4742 and BY4741 of the Kaplan Meier Survival curve in black. The results of the targeted lifespan datasets clearly show that both models recover values for initial mortality reasonably well regardless of whether the data were fitted by WE or LL models (Figs. 2A and 2B). However, as one can follow from the blue curve (WE) demonstrates a much better fit in contrast to the red curve (LL). In the initial mortality LL model under fits the data whereas in the late mortality survival fraction shows an over fit for the LL model (Fig. 2).

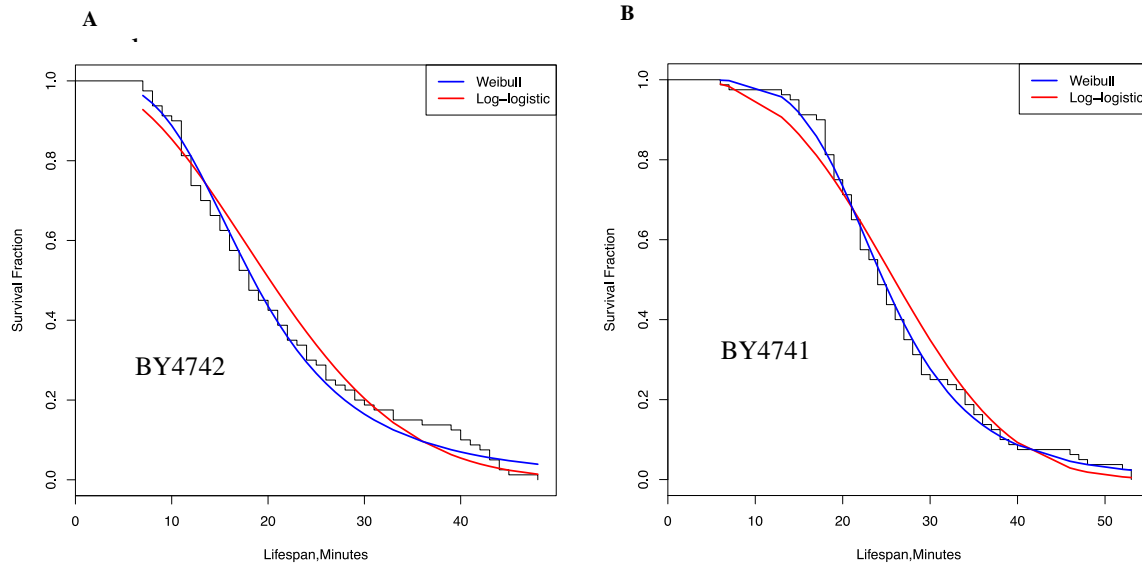


Figure 2. (A) Plot of Kaplan Meier Survival analysis (black) curve of lifespan of only one BY4742 wild type genetic background. Blue curve represents WE model fit of the same lifespan data whereas red curve represents the LL model fit. (B) Plot of Kaplan Meier Survival analysis (black) curve of only one lifespan of BY4741 wild type genetic background. Blue curve represents WE model fit of the same lifespan data whereas red curve represents the LL model fit.

B. Comparison of Estimates of R^2 relation between both of the models and RLS

Fig.3 shows the R^2 relation between both of the models and empirical data. The tendency to a Gaussian curve in Fig.3 A leads to a better fit of WE model on the yeast lifespan samples. Because the WE and LL rates of aging are comparable, we can decide whether either is biased when one equation is used to fit data produced by the other model. However, rates of aging estimated by the WE model tended to be less variable than those estimated by the LL model when the rate of aging was high (Figs. 2 and 3). As confirmed with the LogL value comparison of both models, the WE model is one of the most popular distributions in analyzing skewed data well (Fig.3).

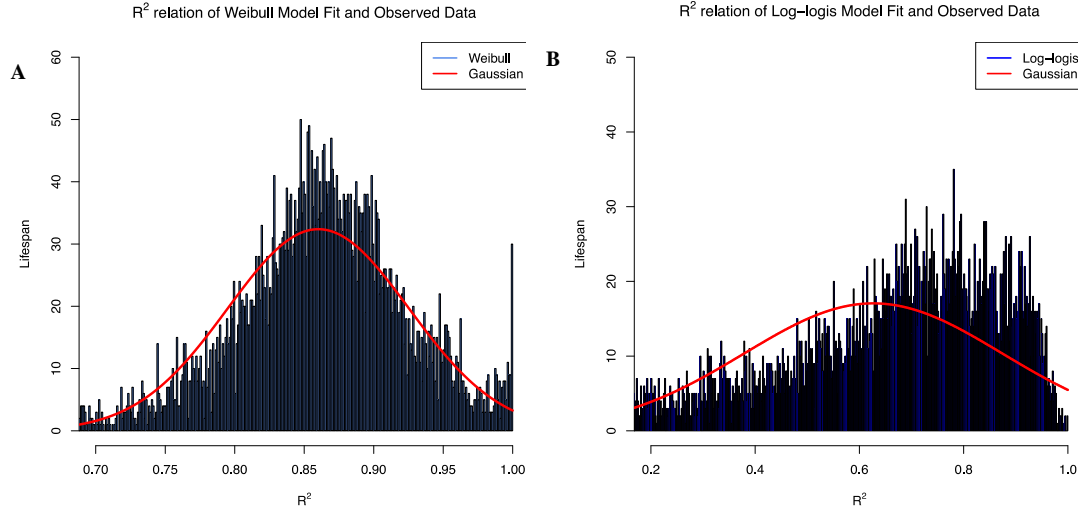


Figure 3. (A) Histogram of $R^2 \sim [70\%,100\%]$ between WE model and empirical lifespan of each RLS of all the genetic backgrounds in the data set. (B) Histogram of $R^2 \sim [20\%,100\%]$ between Log-logis and lifespan of each RLS of all the genetic backgrounds in the same data set. This relation between the WE model and empirical data shows a tendency to a Gaussian curve which shows a better fit and model for the yeast lifespan samples.

C. Retrieving Parameters From RLS Data Sets Generated by the WE and LL Models

The WE model retrieves RLS lifespan data using the following parameter fit space as follows; shape parameter $\gamma = [0.78, 8.72]$, scale parameter $\theta = [11.3, 44.88]$. Similarly, the LL model retrieves RLS lifespan data using the following parameter fit space; shape parameter $\kappa = [1.02, 12.64]$, and scale parameter $\lambda = [6.24, 37]$. Because both of the models generally in the literature used as a suitable alternative of each other, our estimation of fit parameter space also confirms this approach to WE and LL models after parameter estimations. However, on the other hand, these parameter space results suggests that the WE model is generally more tolerant to the experimental noises than the LL model.

IV.DISCUSSION

One objective of our study was to determine how well the WE and LL models fit the same data sets. We found that both equations could be fitted reasonably well to RLS data. Thus, the two models of mortality-rates are roughly equivalent in their ability to characterize aging-related health issues. We also found that variation in the WE estimates for a given RLS lifespan i.e. BY4742 and BY4741 wild type strains was generally lower than that for LL estimates, regardless of the genetic backgrounds on target. However, alternative mortality indices for the WE and LL models could be defined in order to characterize aging dependent mortality resulting from intrinsic causes such as the noise during the experimental procedures.

Based on our fitting studies, one can argue that attention should be taken when determining which mortality model better describes the biological nature of the aging process. For the populations studied, random sampling approach could be used to determine which model fits the data best [17,18]. Of course, there is no known reason why a given dataset in survival analysis should be fit by any certain curve or why a model fitting of a biological population necessarily fit another part of lifespan well. Previous and earlier studies shows that mortality models explores the nature of underlying and extrinsic causes of aging [19,20].

This study had a limitation. Using only two genetic backgrounds is the main limitation of this study. Future studies involving more species from multiple genetic background are required to validate the best model of lifespan samples. In a future study, the intrinsic causes of aging can be investigated on different lifespan data sets with more biological species. Further, the loss of information due to censoring would be compared for these two distributions. The analysis of more datasets could be performed for illustrative purposes.

In this paper we have used mainly the method of maximum likelihood approach to choose among these two distributions. We have used the maximized likelihood method to discriminate the correct model and computed the asymptotic probability of correct selection.

REFERENCES

- [1] Bennet S., “Log-Logistic Regression Models for Survival Data Author (s): Steve Bennett Published by : Wiley for the Royal Statistical Society Stable URL : <http://www.jstor.org/stable/2347295> Log-logistic Regression Models for Survival Data,” *J. R. Stat. Soc. Ser. C*, vol. 32, no. 2, pp. 165–171, 1983.
- [2] Al-Shomrani A. A., Shawky A. I., Arif O. H., and Aslam M., “Log-logistic distribution for survival data analysis using MCMC,” *Springerplus*, vol. 5, no. 1, 2016.
- [3] Weitz J. S., and Fraser H. B., “Explaining mortality rate plateaus,” *Proc. Natl. Acad. Sci.*, vol. 98, no. 26, pp. 15383–15386, 2002.
- [4] D. Wei *et al.*, “Data Descriptor: Structural and functional brain scans from the cross-sectional Southwest University adult lifespan dataset,” *Sci. Data*, 2018.
- [5] Fire M. and Elovici Y., “Data Mining of Online Genealogy Datasets for Revealing Lifespan Patterns in Human Population,” *ACM Trans. Intell. Syst. Technol.*, 2015.
- [6] Longo V. D., Shadel G. S., Kaerberlein M., and Kennedy B., “Replicative and chronological aging in *saccharomyces cerevisiae*,” *Cell Metab.*, vol. 16, no. 1, pp. 18–31, 2012.
- [7] Carmona-Gutierrez S., Didac and Buttner, “The many ways to age for a single yeast cell,” *Yeast*, vol. 31, no. January, pp. 289–298, 2014.
- [8] P. Fabrizio *et al.*, “Superoxide is a mediator of an altruistic aging program in *Saccharomyces cerevisiae*,” *J. Cell Biol.*, vol. 166, no. 7, pp. 1055–1067, 2004.
- [9] Postnikoff S. D. L., Johnson J. E., and Tyler J. K., “The integrated stress response in budding yeast lifespan extension,” *Microb. Cell*, vol. 4, no. 11, pp. 368–375, 2017.
- [10] Korlakai Vinayak R., Kong W., Valiant G., Kakade S. M., and Allen †, “Maximum Likelihood Estimation for Learning Populations of Parameters,” 2019.
- [11] Lenart A., “The Gompertz distribution and Maximum Likelihood Estimation of its parameters - a revision,” *MPDIR Work. Pap.*, vol. 49, no. 0, pp. 0–19, 2012.
- [12] Odell P. M., Anderson K. M., and Agostino R. B. D., “Maximum Likelihood Estimation for Interval-Censored Data Using a Weibull- Based Accelerated Failure Time Model,” *Int. Biometric Soc.*, vol. 48, no. 3, pp. 951–959, 1992.
- [13] Rockette H., Antle C., and Klimko L. A., “Maximum likelihood estimation with the weibull model,” *J. Am. Stat. Assoc.*, vol. 69, no. 345, pp. 246–249, 1974.

- [14] Hill Carter T., and Fomby, *Maximum simulated likelihood methods and applications*. Emerald Group Publishing, 2010.
- [15] deLeeuw J., “Introduction to Akaike (1973) Information Theory and an Extension of the Maximum Likelihood Principle,” no. 1973, pp. 599–609, 2011.
- [16] Cole S. R., Chu H., and Greenland S., “Maximum likelihood, profile likelihood, and penalized likelihood: A primer,” *Am. J. Epidemiol.*, vol. 179, no. 2, pp. 252–260, 2014.
- [17] Jackson C. H., “flexsurv: a platform for flexible parametric survival modelling in R,” *J. Stat. Softw. (in Press.)*, no. Latimer 2013, 2015.
- [18] Wilson D. L., “A comparison of methods for estimating mortality parameters from survival data,” *Mech. Ageing Dev.*, vol. 66, no. 3, pp. 269–281, 1993.
- [19] Güven E., Akçay S., and Qin H., “The Effect of Gaussian Noise on Maximum Likelihood Fitting of Gompertz and Weibull Mortality Models with Yeast Lifespan Data,” *Exp. Aging Res.*, 2019.
- [20] Juckett D. A. and Rosenberg B., “Comparison of the Gompertz and Weibull functions as descriptors for human mortality distributions and their intersections,” *Mech. Ageing Dev.*, vol. 69, no. 1–2, pp. 1–31, 1993.
- [21] Lestienne R., “On the thermodynamical and biological interpretation of the Gompertzian mortality rate distribution,” *Mech. Ageing Dev.*, vol. 42, no. 3, pp. 197–214, Mar. 1988.