University of Richmond
UR Scholarship Repository

Honors Theses | Student Research

4-20-2023

Length Bias Estimation of Small Businesses Lifetime

Simeng Li
University of Richmond, simeng.li@richmond.edu

Follow this and additional works at: https://scholarship.richmond.edu/honors-theses

 Part of the Mathematics Commons, and the Statistical Methodology Commons

Recommended Citation

Li, Simeng, "Length Bias Estimation of Small Businesses Lifetime" (2023). *Honors Theses*. 1693.
https://scholarship.richmond.edu/honors-theses/1693

This Thesis is brought to you for free and open access by the Student Research at UR Scholarship Repository. It has been accepted for inclusion in Honors Theses by an authorized administrator of UR Scholarship Repository. For more information, please contact scholarshiprepository@richmond.edu.

Length Bias Estimation of Small Businesses Lifetime

by

Simeng Li

Honor Thesis

Submitted to

Department of Mathematics & Statistics
University of Richmond
Richmond, VA

Apr 20, 2023

Primary Advisor: Dr. Paul Kvam
Secondary Advisor: Dr. M. Saif Mehkari

1 Abstract

Small businesses, particularly restaurants, play a crucial role in the economy by generating employment opportunities, boosting tourism, and contributing to the local economy. However, accurately estimating their lifetimes can be challenging due to the presence of length bias, which occurs when the likelihood of sampling any particular restaurant's closure is influenced by its duration in operation. To address this issue, this study conducts goodness-of-fit tests on exponential/gamma family distributions and employs the Kaplan-Meier method to more accurately estimate the average lifetime of restaurants in Carytown St. By providing insights into the challenges of estimating the lifetimes of small businesses, this study contributes to our understanding of the broader economic impact of these businesses and the development of policies that support small businesses.

2 Introduction

a. Background

Small businesses can contribute greatly to our economy. For example, they employ half of the American private sector workforce[MM16]. More specifically, small businesses like restaurants are vital. According to National Restaurant Association, restaurants added nearly 2.2 million jobs in 2021 and 2022[Ass]. Besides job creation, restaurants also help to maintain local supply chains, and attract visitors from other cities.

Restaurants are generally considered risky businesses. Luo and Stark[LS14] mentioned in their study that for restaurants that have been open for 10 years, the cumulative survival rate is only about 38%, which means about 62% restaurants close in the first 10 years of their life span. While there are many internal reasons that lead to failure, such as poor management, poor product and financial volatility[PAR+05], other external reasons such as a financial crisis and unemployment also play major roles in restaurants' failure. Therefore, studying restaurants' lifetimes is crucial, as economists and analysts can then understand an area's economic development, and make predictions on the area's future development.

There is a common problem with business lifetime data. When we observe the lifetimes at a certain point of time, it is more likely for people to see restaurants that have lasted for a long time compared to those that are not, since if a restaurant is already closed, we will not get its data in most cases. We will look into the detail reasons later in this article.

In this article, we look at some restaurants located in Carytown, Richmond, VA, as well as the number of years they have been in operation. We will show the reasons why they can not be directly used to analyze any overall operation patterns of the area in the presence of survival bias. We will also use this data to construct estimators to examine the effects of bias.

b. Data Description

In this research, we picked 24 restaurants on the same commercial street: Carytown, Richmond, Virginia. Most of them are not chain restaurants, or only have one or two branches in Richmond. 19 of them where in normal operation up until December 2022, and the other 5 have already closed. The data I collected included their name, starting year, end year, and length of operation. Here are the restaurants that were open:

Pho Luca's RVA	Baker's Crust Carytown	Carytown Burger and Fries
Mom's Siam Restaurant	Galaxy Diner	CanCan
New York Deli	The Mantu	Carytown Sushi
Ginger Thai Taste	Tulsi	East Coast Provision
Home Sweet Home	Stella's	Don't Look Back
Mellow Mushroom Carytown	The Daily Kitchen and Bar	Les Crepe
Citizen Burger Bar		

Here are the restaurants that have already closed:

Weezie's Kitchen	Portrait House	Mezzanine
Broken Tulip	Xtra's Café	

Among all these restaurants, 18 out of 20 have been open for 10 or more years, and the other two have also been open for more than 5 years. The statistic here also shows the problem: I collected data over a brief interval of time, but most restaurants' lifespan data show that they have been in business for a relatively long time. As a consequence of this type of sampling, the data are less likely to include newer restaurants that have been in operation for a short time.

3 Right Censoring

In this study, most restaurant data are right-censored. By definition, right-censoring means that the data we have are collected before a certain time t and at this point, the event we are looking for (the closure of restaurant) has not yet occurred. The event occurs after time t, but how long is unknown. For example, in this study, 19 of the restaurants are open at the time t data were collected (December 2022), but we don't know how long after t they will close. Thus they are right-censored data.

This is important because if we ignore the censoring problem[DC+96], it will lead to an underestimation of the lifespan data, and may result in potential useful data being misused or thrown away.

The most well-known estimation method for right-censored data is Kaplan Meier method[Nai84], which is a non-parametric estimation method. We will use it for further analysis and estimation.

4 Length Bias

In this section, I will introduce length bias and its affects regarding estimation.

a. Definition

Survival bias, or survivorship bias, is commonly discussed in many fields of study such as epidemiology, business, and military. It is an error occuring when people only focus on subjects that are selected by certain criteria, but ignore those that are not. Therefore, survival bias may cause problems due to incomplete information.

We can understand length bias to be a special case survival bias. While they all affect our estimation of distributions by giving invisible bias, length bias leads to overestimation and survival bias leads to underestimation.

b. Cancer Screening Example

A good example to understand length bias is cancer screening, which is also the most frequently discussed context of this topic. When people look at cancer screening data in a limited period of time, one may conclude that patients who discover their cancer during screening tend to survive for a longer period of time than those who didn't discover cancer during screening. This phenomenon seems reasonable, because those patients who are diagnosed to have cancer may actively seek treatment.

However, the presence of length bias is misleading us. As Figure 1 shows, darker blue arrows represent slow-growing cancers, and lighter blue ones represent fast-growing cancers. When we do screening at a specific time point, it is more likely to discover a cancer if it has lasted for a longer period of time. These are usually more slow-growing cancers.

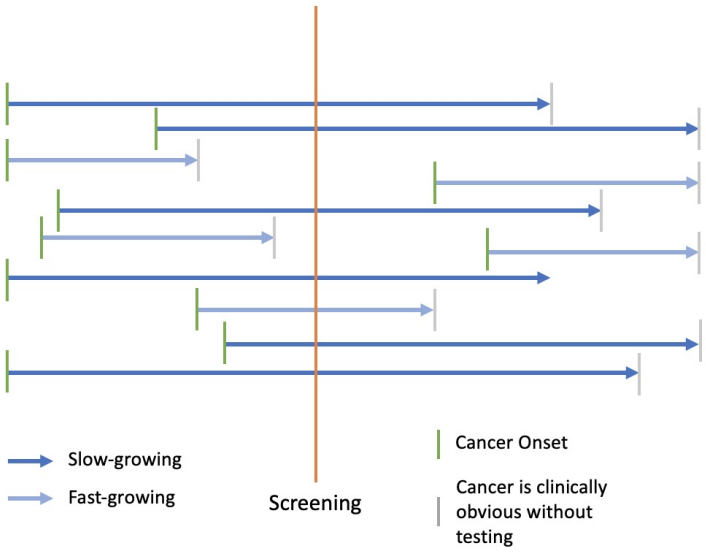

Figure 1: Length Bias in Cancer Screening

Also, It is more likely that patients will go to cancer screening if they start to have symptoms, and this is why cancer screening usually detects more slow-growing cancers, as they are in patients' bodies for long enough time to lead to symptoms. In addition, slow-growing cancers are usually less fatal that fast-growing ones, thus patients who get fast-growing cancers are less likely to get cancer screening during the period when the cancer is still treatable.

c. Length Bias' Effects on Restaurants Lifetime Data

In this study, the problem presenting in restaurants lifetime data is similar to the problem in cancer screening example. After collected the data, if people see many restaurants have been in business for a long time, they may just assume all restaurants here have last for that long. However, we need to consider the probability theory behind our sampling method: as Figure 2 shows, it is always more likely to see restaurants that last for a long time.

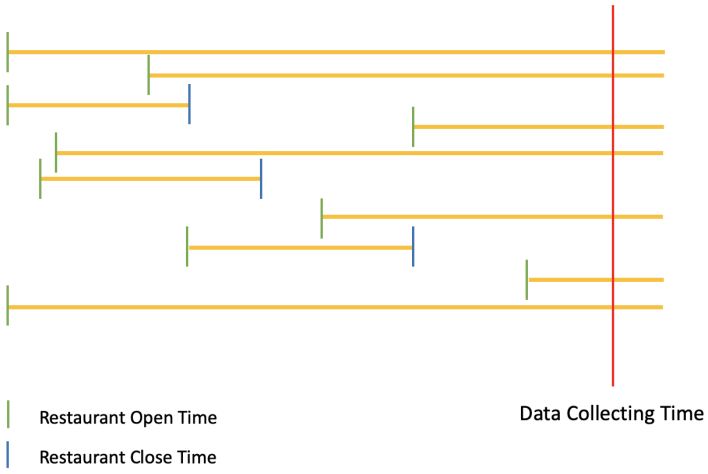

Figure 2: Length Bias in Restaurants Lifetime

In Figure 2, the yellow arrows represent the length of time since restaurants are in operation, and the red line is the time when we collect data. We can see that the probability of hitting yellow arrows that represent restaurants with longer lifetime is bigger.

d. Statistical Model for Length Bias

Asgharian and Wolfson[AW05] showed in their work that if density $f(x)$ describes the population length of the objects being sampled, which are restaurants lifetimes in this study, then set off the equation of the density we actually observe by sampling:

$$g(x) = \frac{xf(x)}{\mu}$$

We can use a simple example to see how length bias affects our sampling process. Suppose the underlying lifetime data are exponentially distributed and we have the lifetime density:

$$f(x) = \lambda e^{-\lambda x}$$

According to the definition of exponential distribution, we know $E(x) = \frac{1}{\lambda}$. Under length-biased sampling, the density of the observed data is

$$g(x) = \frac{x\lambda e^{-\lambda x}}{1/\mu} = x\lambda^2 e^{-2\lambda}$$

which is a gamma distribution $\Gamma(2, \frac{1}{\lambda})$. Now, for g(x), we know $E(x) = \frac{2}{\lambda} = 2(\frac{1}{\lambda})$. This means that we overestimate f(x) by 100% if not considering the length bias. Since restaurants with longer lifespan are easier to be observed, it is likely for us to assume that most restaurants in an area have longer lifespan.

5 Kaplan-Meier Estimator

As mentioned above, Kaplan-Meier non-parametric estimator is used for censored lifetime data analysis. In this section, we will use the Kaplan-Meier method to estimate our restaurants lifetime in R.

The Kaplan-Meier estimator is also called a "product limit estimate"[GKK10], since it requires the computation of probabilities that event will occur at a certain point of time. The final estimator is obtained by multiplying each probability we calculated, and we can refer to the cumulative probability function formula below:

$$F_{KM}(t) = 1 - \prod_{x_j \leq t} (1 - \frac{d_j}{m_j})$$

where d_j = number of failures at x_j, m_j = number of observations that had survived up to x_j.

To build a Kaplan-Meier model in this study, we need to specify whether a restaurant has failed or not given its lifetime data. If a restaurant is still in operation, this specific lifetime data is right-censored.

Figure 3 is the survival plot made using Kaplan-Meier estimation as described above. The x-axis shows restaurants' length of operation, and y-axis shows the probability of survival. The red line is the actual estimation we estimated for different lifetimes, and we also include two dark green dotted lines representing the its confidence interval. For example, for a restaurant with 10 years of operation, we estimate that the probability of its survival is about 80%. Also, we are confident to say that the probability should be between about 65% to 100%.

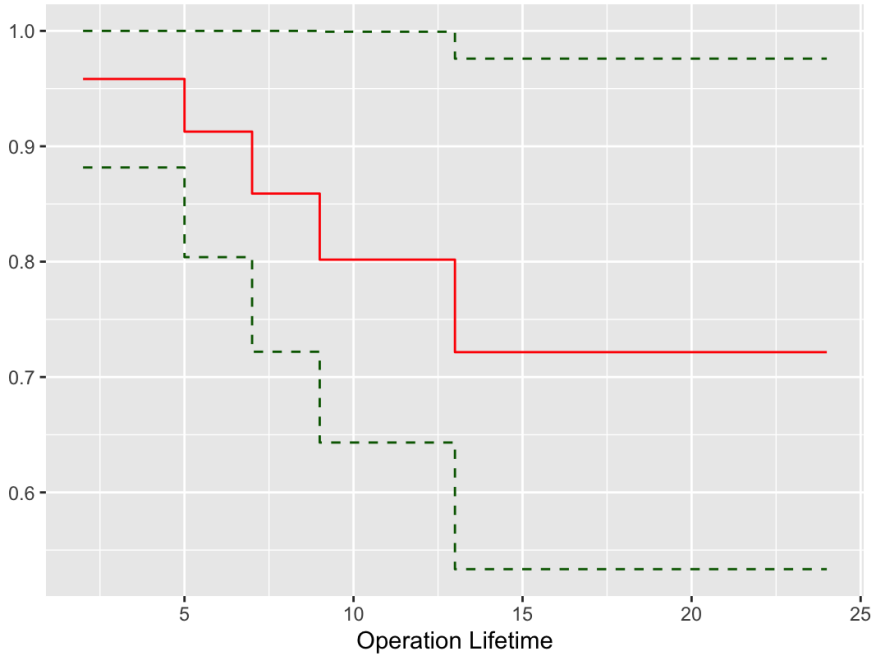

Figure 3: Kaplan-Meier Estimation

In section 4 we calculated that the average lifetime of all restaurants is $\bar{X} = 11.208$. Because of the presence of right censoring, this value should be underestimating our data, since we don't know how long some restaurants will continue to be in operation. Therefore, we will construct a new $\hat{\mu}$ estimator using Kaplan-Meier method by computing the area under our survival plot, which is $\int F(x)dx$. As expected, the area under curve, or the Kaplan-Meier estimator, is 17.56 and it is bigger than the average lifetime sample mean $\bar{X} = 11.208$ we observed.

6 Exponential Model Data Fit

We are interested in finding the real distribution of restaurants lifetime data without length bias, and we will assume the restaurants data are exponentially distributed. This is a reasonable hypothesis because most other distributions have increasing failure rate, and this seems not the case of restaurant lifetimes, as restaurants will not become more likely to fail as time goes on.

Because of the presence of length bias, if restaurant life times were exponentially distributed, then after we sample them, we actually obserbe data that are distributed as $\Gamma(2, \beta)$, as shown in section 4(d). Therefore, we will fit our data into gamma distribution $\Gamma(2, \beta)$. If our data have a good fit, this means that the lifetime data are indeed distributed exponentially.

We learned in Math 330 (Mathematical Statistics) that the exponential distribution is "memoryless", meaning that the probability of an event occurring is independent of how long it has been since the previous event occurred. Thus, our goodness-of-fit test for the

exponential distribution (or for the gamma distribution if the data are length-biased) are not negatively affected by right-censoring.

To calculate β, we know that $E(X) = 2\beta$ for the distribution $\Gamma(2, \beta)$, and the Kaplan-Meier estimator of average lifetime of all restaurants is 17.56. We can estimate $\hat{\beta} = \frac{1}{2}\bar{X} = 8.78$. Thus, we will use the distribution $\Gamma(2, 8.78)$ to check for goodness-of-fit.

Cramer-von Mises Goodness-of-Fit Test (CvM Test)

To check whether assuming restaurants data are exponential distributed is reasonable for our estimation, we can use the Cramer-von Mises Goodness-of-Fit Test[CF96] for the gamma family using R.

According to the test, we get the p-value = 0.97. This means that we cannot reject the null hypothesis that the data follow a gamma distribution. So if the underlying lifetime data is distributed exponentially, the observed length-biased data should have the gamma distribution $\Gamma(2, 8.78)$, and the Cramer-von Mises Goodness-of-Fit Test suggests it fits well.

To visually assess the goodness of fit, we can create some plots using R. The upper left density plot in Figure 4 suggests the fitted distribution(red curve) overlaid on a histogram of the data. The lower left plot is a cumulative distribution function (CDF) plot, showing the cumulative distribution function of the fitted distribution as a curve(red curve) and the distribution of our data (black circles). The lower right plot is a probability-probability (P-P) plot, comparing the probabilities of the fitted distribution (black line) with the corresponding probabilities of our data (black circles). We can see every plot is showing a good match and fit, same as what the Cramer-von Mises Goodness-of-Fit Test indicates.

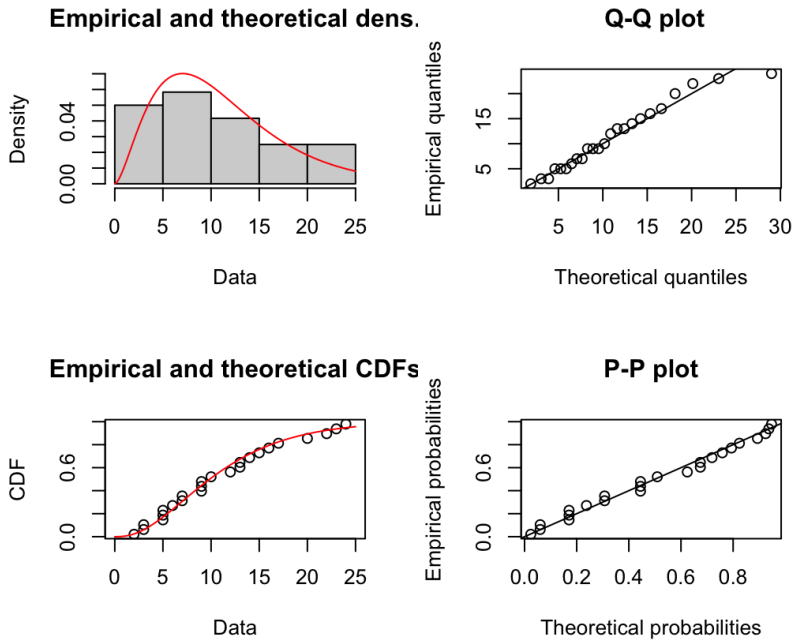

Figure 4: Goodness-of-Fit with Gamma

References

[Ass] National Restaurant Association. *Restaurants added jobs in 24 consecutive months*. URL: https://restaurant.org/research-and-media/research/economists-notebook/analysis-commentary/restaurants-added-jobs-in-24-consecutive-months/.

[AW05] Masoud Asghrian and David B. Woldson. "Asymptotic Behavior of the Unconditional NPMLE of the Length-biased Survivor Function from Right Censored Prevalent Cohort Data". In: *The Annals of Statistics* (2005). DOI: 10.1214/009053605000000372.

[CF96] Sándor Csörgo and Julian J. Faraway. "The Exact and Asymptotic Distributions of Cramer-von Mises Statistics". In: *Journal of the royal statistical society series b-methodological* 58 (1996), pp. 221–234.

[DC+96] Watt DC et al. "Survival analysis: the importance of censored observations." In: (1996). DOI: 10.1097/00008390-199610000-00005.

[GKK10] Manish Kumar Goel, Pardeep Khanna, and Jugal Kishore1. "Understanding survival analysis: Kaplan-Meier estimate". In: (2010). DOI: 10.4103/0974-7788.76794.

[LS14] Tian Luo and Philip B Stark. "Only the Bad Die Young: Restaurant Mortality in the Western US". In: *arXiv.org* (2014).

[MM16] Karen Gordon Mills and Brayden McCarthy. "The State of Small Business Lending: Innovation and Technology and the Implications for Regulation". In: (2016).

[Nai84] Vijayan N. Nair. "Confidence Bands for Survival Functions with Censored Data: A Comparative Study". In: *Technometrics* 26.3 (1984), pp. 265–275. ISSN: 00401706. URL: http://www.jstor.org/stable/1267553.

[PAR+05] H. G. PARSA et al. "Why Restaurants Fail". In: *Cornell Hotel and Restaurant Administration* 46.3 (2005).