Survival Estimation Using Bootstrap, Jackknife and K-repeated Jackknife Methods

Innocent, Ebun, Kamorudeen, Samuel & Wenjuan

BOWLING GREEN STATE UNIVERSITY

December 8, 2023

Outline of Study

Overview

1. Introduction
   - Review
   - Learning Outcomes
   - Background of the data
2. Methodology and Analysis of Results
   - Bootstrap
   - Jackknife
   - k-Repeated Jackknife
3. Summary and Conclusion
   - Consistencies and Inconsistencies
   - Conclusion and Recommendation

Questions

Questions

1. What comes to mind if you hear about finding Survival Probabilities?
   - Kaplan-Meier Estimator?
   - Parametric Survival Models using (exponential, Weibull, and log-normal distributions)?
2. Do you think we can use the concept learnt in this class to also estimate parameter survival probabilities?

Introduction

Review

1. The field of modern statistics relies on statistical methods and hypothesis tests that typically assume the normal distribution of populations.

1. While these methods are generally robust to deviations from normality, there are situations where it is essential to empirically investigate the distribution of the underlying population or a specific statistic.

1. This need becomes particularly prominent when working with survival data and other time-to-event data, which find applications in diverse fields including biomedical sciences, engineering, economics, and social sciences.

Introduction

Motivation

1. Inspired by these ideals, our team made a collective decision to search for a research paper that specifically tackles these concerns.

Introduction

Overview of the paper

1. The research paper titled "Survival Estimation Using Bootstrap, Jackknife and K-Repeated Jackknife Methods" was authored by Johnson A. Adewara and Ugochukwu A. Mbata from the University of Lagos, Nigeria. It was published in the Journal of Modern Applied Statistical Methods in November 2014.

Overview of the Paper

The paper explores three resampling techniques:

- Bootstrap estimation method (BE)
- Jackknife estimation method (JE)
- k-repeated Jackknife estimation method (KJE)

Overview of the Paper

1. The authors compared the performance of these resampling methods by calculating the mean square error (MSE) and mean percentage error (MPE) based on simulated data. The results indicate that the K-repeated Jackknife method reduces the MSE value compared to the other methods

Learning Outcomes

Main Objectives of the study

1. Replicate the results: Our main aim is to replicate the results of the original research paper by Johnson A. Adewara and Ugochukwu A. Mbata.
2. We also aim to verify the consistency of our results with the claims made by the authors of the original study.

Background to the Data

Background to the Data

1. Our study aim at estimating parameter of exponential distribution based on simulated data. We will consider 12 different samples using 4 different λ and 3 sample sizes

- we would consider $\lambda = 0.5, 1.0, 1.5, 2.0$
- For each sample n= 10, 20,30

Survival Function

When $t \geq t_0$, the probability density function of the exponential distribution is given by:

$$f(t;\theta) = \frac{1}{\theta}\exp\left(-\frac{t-t_0}{\theta}\right), \quad t_0 \geq 0, \quad \theta > 0, \quad t > t_0. \tag{1}$$

$$\begin{aligned}
F(t) = \int_{t_0}^{t} f(t;\theta)dt &= \int_{t_0}^{t} \frac{1}{\theta}\exp\left(-\frac{1}{\theta}(t-t_0)\right) dt \\
&= \frac{1}{\theta}\int_{t_0}^{t} \exp\left(-\frac{1}{\theta}(t-t_0)\, dt\right) \\
&= \frac{1}{\theta}\cdot -\theta\left[\exp\left(-\frac{1}{\theta}(t-t_0)\right)\right]_{t_0}^{t} \\
&= -\left[\exp\left(-\frac{1}{\theta}(t-t_0)\right) - \exp\left(-\frac{1}{\theta}(t_0-t_0)\right)\right]
\end{aligned}$$

Survival Function

Survival Function

$$= 1 - \exp\left(-\frac{1}{\theta}(t - t_0)\right)$$

S(t)=1-F(t)=exp $\left(-\frac{1}{\theta}(t - t_0)\right)$
where $\lambda = \frac{1}{\theta}$

$$S(t;\lambda) = \exp\left(-\lambda(t - t_0)\right) \tag{2}$$

If we set t_0=0.Then we have,

$$S(t;\lambda) = \exp\left(-\lambda t\right) \tag{3}$$

Bootstrap Resampling and Estimation

1. It is a Monte Carlo method that estimates the distribution of a population by resampling.
2. Consider a random sample $t_1, t_2, \ldots, t_n$ drawn from the distribution of a random variable $T \sim exp(\theta)$.
3. An estimator $\hat{\theta} = \hat{\theta}(t_1, t_2, \ldots, t_n)$ provides an estimate for a parameter θ.
4. To generate m random variables from the sampling distribution of $\hat{\theta}$, we repeatedly draw independent random samples $t^{(j)}$ and compute the corresponding estimator values $\hat{\theta}^{(j)} = \hat{\theta}\left(t_1^{(j)}, t_2^{(j)}, \ldots, t_n^{(j)}\right)$ for each sample $t^{(j)}$.
5. The mean of these estimator values, denoted as $\bar{\hat{\theta}}$ is estimated as

$$\bar{\hat{\theta}}_B = \frac{1}{m}\sum_{j=1}^{m}\hat{\theta}^{(j)} \qquad (4)$$

Methodology-Bootstrap Resampling and Estimation

Efficiency of the Bootstrap Estimator

1. The estimate of the mean squared error (MSE) and Mean Percentage error (MPE) of $\hat{\theta}_B$ are given as:

$$\text{MSE}(\hat{\theta}_B) = \frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{\theta}^{(j)} - \bar{\hat{\theta}}_B\right)^2 \quad (5)$$

$$\text{MPE}(\hat{\theta}_B) = \frac{\sum_{j=1}^{m}\left|\frac{\hat{\theta}^{(j)} - \bar{\hat{\theta}}_B}{\bar{\hat{\theta}}_B}\right|}{m} \quad (6)$$

2. The estimate of the survival function is given as:

$$\hat{S}_B(t) = \exp\left(-\frac{t}{\hat{\theta}_B}\right) \quad (7)$$

Methodology

Jackknife Resampling and Estimation

- The jackknife is like a "leave-one-out" type of cross-validation.
- Consider a random sample $t = (t_1, t_2, \ldots, t_n)$ drawn from the distribution of a random variable $T \sim exp(\theta)$, and define the i^{th} jackknife sample $t_{(i)}$ to be the subset of t that leaves out the i^{th} observation t_i. That is, $t_{(i)} = (t_1, \cdots, t_{i-1}, t_{i+1}, \cdots, t_n)$
- If $\hat{\theta} = T_n(t)$, define the i^{th} jackknife replicate

$$\hat{\theta}_{(i)} = T_{n-1}(x(i)),\ i = 1, \cdots, n$$

$$\bar{\hat{\theta}}_{jack} = \frac{\sum_{i=1}^{n} \hat{\theta}_{(i)}}{n} \tag{8}$$

Methodology-Jackknife Resampling and Estimation

Efficiency of the Jackknife Estimator

1. The estimate of the mean squared error (MSE) and Mean Percentage error (MPE) of $\hat{\theta}_{jack}$ are given as:

$$\text{MSE}(\hat{\theta}_{jack}) = \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{\theta}_{(i)} - \bar{\hat{\theta}}_{(\cdot)} \right)^2 \quad (9)$$

$$\text{MPE}(\hat{\theta}_{jack}) = \frac{\sum_{i=1}^{n} \left| \frac{\hat{\theta}_{(i)} - \bar{\hat{\theta}}}{\bar{\hat{\theta}}} \right|}{m} \quad (10)$$

2. The estimate of the survival function is given as

$$\hat{S}_{jack}(t) = \exp\left(-\frac{t}{\bar{\hat{\theta}}_{jack}} \right) \quad (11)$$

Methodology

K-Repeated Jackknife Resampling and Estimation

1. The K-repeated jackknife method is a resampling technique aimed at minimizing the Mean Square Error (MSE).
2. This involves jackknifing the observed data k times, where k equals the same size of the observed data.
3. The stopping rule for the repeated replications depend on the size of the original data.
4. The procedure converges before or at the k^{th} time, where the estimate from the jackknife replication is the same as the estimator of the parameter θ based on the complete sample of size n

Methodology-K-Repeated Jackknife Resampling and Estimation

Algorithm for the K-repeated Jackknife procedure

1. Step 1: Observe a random sample $T = (t_1, t_2, \ldots, t_n)$
2. Step 2: Compute $\hat{\theta}(t)$ for θ

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} t_i \qquad i = 1, 2, \ldots, n \tag{12}$$

3. Step 3: For i up to n + generate a jackknife sample T_{-i} by leaving out the i_{th} observation. + calculate $\hat{\theta}_{(-i)}$ from each of the Jackknife sample T_{-i} by

$$\hat{\theta}_{(-i)} = \frac{1}{n-1}\sum_{i=1}^{n-1} T_{-i} \tag{13}$$

Methodology-K-Repeated Jackknife Resampling and Estimation

Algorithm cont'd

1. Step 4: Repeat step 3 using the estimates from $\hat{\theta}_{(-i)}$ to form pseudo samples. The new pseudo samples are used to generate another set of jackknife estimates; this is continued until the k^{th} time. This implies that the process is repeated k times, and at any given stage the preceding jackknife estimates are used as new samples in the next stage until the k^{th} time.
2. Step 5: At the k^{th} time, the K-repeated Jackknife estimate is calculated as

$$\bar{\hat{\theta}}^{K} = \frac{1}{k}\sum_{i=1}^{n}\hat{\theta}_{i-1}^{k} \qquad (14)$$

Methodology-K-Repeated Jackknife Resampling and Estimation

R codes implementation for K-repeated Jackknife

```
kjack.theta.stat = function(dat, ind){
  n = length(dat)
    survJ = function(dats){ # function to compute jackknife estimate
    n = length(dats)
    d = numeric(n)
    for (i in 1:n) {
      q = dats[-i]
      d[i] = mean(q)
    }
    return(d)
  }

 e = survJ(dat[ind]) # first estimate of theta based on jackknife estimate
using original sample
  f = matrix(data = NA, nrow = n, ncol = n)
  f[1, ] = survJ(e) # first estimate of theta based on original sample
  for (k in 1:(n-1)) { # now estimate thetas based on the previous
jackknife replicates
    f[k+1, ] = survJ(f[k,])
  }
  t = mean(f[n,]) # value of theta on the Kth replication
  return(t)
}
```

Methodology-K-Repeated Jackknife Resampling and Estimation

Efficiency of the K-Repeated Jackknife Estimator

1. The estimate of the mean squared error (MSE) and Mean Percentage error (MPE) of $\hat{\theta}$ are given as:

$$\text{MSE}\left(\bar{\hat{\theta}}^K\right) = \frac{1}{k(k-1)} \sum_{i=1}^{n} \left(\hat{\theta}_{i-1}^{k} - \bar{\hat{\theta}}^K\right)^2 \quad (15)$$

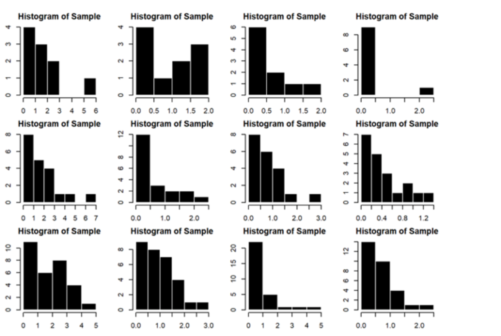
$$\text{MPE}\left(\bar{\hat{\theta}}^K\right) = \frac{\sum_{i=1}^{n} \left| \frac{\hat{\theta}_{i-1}^{k} - \bar{\hat{\theta}}^K}{\bar{\hat{\theta}}^K} \right|}{k} \quad (16)$$

2. The estimate of the survival function is given as

$$\hat{S}^K(t) = \exp\left(-\frac{t}{\bar{\hat{\theta}}^K}\right) \quad (17)$$

Results

Histograms



1. The histogram shows the distribution is right skewed.
2. This confirmed one of the properties of exponential distribution.

Results

1. Our findings demonstrate that the mean values closely align with the reciprocal of λ.

Table: Descriptive Statistics of Samples

Sample index	N	λ	Mean	Median	Std.Dev.
Sample 1	10	0.5	1.6852444	1.2948647	1.6623883
Sample 2	10	1.0	0.8960689	0.8706322	0.6891788
Sample 3	10	1.5	0.4981274	0.2746070	0.5275997
Sample 4	10	2.0	0.4677158	0.3405017	0.6077853
Sample 5	20	0.5	1.8014286	1.4236665	1.7248715
Sample 6	20	1.0	0.5985835	0.3544807	0.6210249
Sample 7	20	1.5	0.7564254	0.5756086	0.6670282
Sample 8	20	2.0	0.4102627	0.3474443	0.3763015
Sample 9	30	0.5	1.8323527	1.6049523	1.3055313
Sample 10	30	1.0	0.9591738	0.8576519	0.6634190
Sample 11	30	1.5	0.9960805	0.6416917	1.0357711
Sample 12	30	2.0	0.6233302	0.5521193	0.5431154

Results-Estimate of Lambda

1. We estimate the lambda using the three methods and notice that as the sample sizes increase the estimates are getting to the original lambda.

Table: Estimated Lambda values

S. Index	N	orig. lambda	Bootstrap	Jackknife	RJackknife
S1	10	0.5	0.5964	0.5934	0.5934
S2	10	1.0	1.1058	1.1160	1.1160
S3	10	1.5	2.0489	2.0075	2.0075
S4	10	2.0	2.0956	2.1381	2.1381
S5	20	0.5	0.5561	0.5551	0.5551
S6	20	1.0	1.6776	1.6706	1.6706
S7	20	1.5	1.3159	1.3220	1.3220
S8	20	2.0	2.4079	2.4375	2.4375
S9	30	0.5	0.5453	0.5457	0.5457
S10	30	1.0	1.0444	1.0426	1.0426
S11	30	1.5	1.0075	1.0039	1.0039
S12	30	2.0	1.5855	1.6043	1.6043

Results-Survival Probabilities

1. Comparing the results, we can see that all the three methods are very closed to the original survival this means that all the three methods are effective.

Table: Survival Probabilities

S index	N	λ	Orig. Surv.	SB	SJ	SRepJ
Sample 1	10	0.5	0.5335338	0.4877985	0.4891339	0.4891339
Sample 2	10	1.0	0.4976241	0.4715786	0.4692232	0.4692232
Sample 3	10	1.5	0.5920405	0.5270536	0.5312230	0.5312230
Sample 4	10	2.0	0.5360183	0.5237277	0.5184016	0.5184016
Sample 5	20	0.5	0.5216188	0.4953224	0.4957582	0.4957582
Sample 6	20	1.0	0.6348561	0.5181701	0.5191189	0.5191189
Sample 7	20	1.5	0.4373413	0.4731838	0.4719109	0.4719109
Sample 8	20	2.0	0.5438038	0.4985848	0.4956277	0.4956277
Sample 9	30	0.5	0.4792352	0.4551163	0.4548756	0.4548756
Sample 10	30	1.0	0.4607897	0.4484329	0.4489331	0.4489331
Sample 11	30	1.5	0.4008689	0.5011062	0.5020032	0.5020032
Sample 12	30	2.0	0.4220864	0.4826063	0.4795086	0.4795086

Results- Estimated Mean Square Error

1. Measures average squared distance between estimators and true value.
2. Bootstrap and Jackknife yieldsimilar results.
3. K-Repeated Jackknife is significantly smaller. Supports authors claim that K-Repeated Jackknife method is more accurate.

Table: Comparison of Bootstrap, Jackknife, and K-Rep. Jackknife

S. Index	N	λ	Bootstrap	Jackknife	K-Rep. Jackknife
S1	10	0.5	0.2596	0.2764	4.26×10^{-5}
S2	10	1.0	0.0413	0.0475	7.30×10^{-6}
S3	10	1.5	0.0258	0.0278	4.30×10^{-6}
S4	10	2.0	0.0364	0.0369	5.70×10^{-6}
S5	20	0.5	0.1439	0.1488	1.10×10^{-6}
S6	20	1.0	0.0181	0.0193	1.00×10^{-7}
S7	20	1.5	0.0204	0.0222	2.00×10^{-7}
S8	20	2.0	0.0072	0.0071	1.00×10^{-7}
S9	30	0.5	0.0568	0.0568	1.00×10^{-7}
S10	30	1.0	0.0148	0.0147	$0.00 \times 10^{+0}$
S11	30	1.5	0.0355	0.0358	1.00×10^{-7}
S12	30	2.0	0.0093	0.0098	$0.00 \times 10^{+0}$

Results-Mean Percentage Error

1. Measures average distance between estimator and the true value as a percentage of the true value.
2. Similar to before, Bootstrap and Jackknife methods produce similar results.
3. K-Repeated Jackknife has significantly lower MPE

Table: MPE Comparison Table

Sample	N	λ	Bootstrap	Jackknife	K-Rep. Jackknife
1	10	0.5	0.2410	0.0763	0.0009535
2	10	1.0	0.1845	0.0748	0.0009349
3	10	1.5	0.2588	0.0970	0.0012129
4	10	2.0	0.3155	0.0800	0.0010003
5	20	0.5	0.1654	0.0377	0.0001047
6	20	1.0	0.1828	0.0434	0.0001205
7	20	1.5	0.1517	0.0336	0.0000932
8	20	2.0	0.1644	0.0373	0.0001035
9	30	0.5	0.1051	0.0208	0.0000247
10	30	1.0	0.1014	0.0192	0.0000229
11	30	1.5	0.1517	0.0257	0.0000306
12	30	2.0	0.1244	0.0224	0.0000267

Summary-Consistencies with the original paper

1. We successfully reproduced the results in Tables 1, 2, and 3 of the original paper.

Table 3. Estimation Using the Three Methods Bootstrap, Jackknifing and K repeated jackknifing

	λ	$\hat{S}_B(t)$	$\hat{S}_{jack}(t)$	$\hat{S}^K(t)$
10	0.5	0.568858094	0.568879887	0.568879887
	1	0.476453626	0.476456925	0.476456925
	1.5	0.461343936	0.461328523	0.461328523
	2.0	0.529933691	0.529937819	0.529937819
20	0.5	0.491722891	0.491777729	0.491777729
	1	0.490229963	0.490240047	0.490240047
	1.5	0.544947075	0.544930402	0.544930402
	2.0	0.553586925	0.553580134	0.553580134
30	0.5	0.527441921	0.527445588	0.527445588
	1	0.491819455	0.491882638	0.491882638
	1.5	0.491085203	0.491099760	0.491099760
	2.0	0.528125037	0.528118624	0.528118624

SampleSize	Lambda	Bootstrap	Jackknife	JackknifeRe
10	0.5	0.5695832	0.5688799	0.5688799
10	1.0	0.4756130	0.4764569	0.4764569
10	1.5	0.4623369	0.4613285	0.4613285
10	2.0	0.5296122	0.5299378	0.5299378
20	0.5	0.4914082	0.4917777	0.4917777
20	1.0	0.4907075	0.4902400	0.4902400
20	1.5	0.5437886	0.5449304	0.5449304
20	2.0	0.5550153	0.5535801	0.5535801
30	0.5	0.5275474	0.5274456	0.5274456
30	1.0	0.4928460	0.4918826	0.4918826
30	1.5	0.4906251	0.4910998	0.4910998
30	2.0	0.5277365	0.5281186	0.5281186

Summary-Inconsistencies

1. In the original paper, Tables 4 and 5 display identical MSE and MPE for the Jackknife and K-Repeated Jackknife methods.

Table 4. Estimation to the Bootstrap, Jackknifing and K repeated jackknifing using MSE methods

	λ	$\hat{S}_B(t)$	$\hat{S}_{jack}(t)$	$\hat{S}^K(t)$
10	0.5	0.004741437	0.004744439	0.004744439
	1	0.000554432	0.000554276	0.000554276
	1.5	0.001494291	0.001495483	0.001495483
	2.0	0.000896026	0.000896273	0.000896273
20	0.5	0.000068511	0.000067606	0.000067606
	1	0.000095454	0.000095257	0.000095257
	1.5	0.002020240	0.002018741	0.002018741
	2.0	0.002871559	0.002870831	0.002870831
30	0.5	0.000753059	0.000753260	0.000753260
	1	0.000066921	0.000065892	0.000065892
	1.5	0.000079474	0.000079214	0.000079214
	2.0	0.000791018	0.000790657	0.000790657

SampleSize	Lambda	Bootstrap	Jackknife	JackknifeRe
10	0.5	1.322555528	0.1779106373	1.819541e-03
10	1.0	0.054312811	0.0074964448	7.666819e-05
10	1.5	0.030090984	0.0043877462	4.487468e-05
10	2.0	0.071224806	0.0095296511	9.746234e-05
20	0.5	0.149652842	0.0090345787	2.270595e-05
20	1.0	0.057082432	0.0033322834	8.374786e-06
20	1.5	0.019415600	0.0011849137	2.977958e-06
20	2.0	0.047364242	0.0028326288	7.119041e-06
30	0.5	0.084409312	0.0031509108	3.509079e-06
30	1.0	0.022455391	0.0008183606	9.113847e-07
30	1.5	0.010851713	0.0004112341	4.579796e-07
30	2.0	0.008449317	0.0003154647	3.513240e-07

Summary

Inconsistencies

1. This does not support the paper's conclusion that the K-Repeated Jackknife method is superior in performance
2. Our results provide evidence to support their conclusion.
3. We reached out to the authors by email, but have not received a meaningful response

Conclusion

Conclusion

1. Our project provides a thorough examination of the three methods, confirming their utility in survival analysis and offering insights into their relative strengths and limitations.
2. Bootstrap, Jackknife, and K-repeated Jackknife methods are efficient in estimating the population parameters and their mean square errors (MSE).
3. The K-Repeated Jackknife method offers an improvement over the standard Jackknife method for reducing MSE and MPE, resulting in a narrower Confidence Interval.

Recommendation

Addressing Censored Data

1. Censored data is a critical aspect often encountered in medical research. Incorporating censored data could enhance the reliability and applicability of these methods in clinical studies.

Applying these methods to survival regression and other survival estimates

1. Nonparametric Survival Curve Estimation (Kaplan-Meier estimate)
2. Cox Proportional Hazards Model
3. Other parametric models in survival analysis

The End