ACTUARIAL RESEARCH CLEARING HOUSE
1995 VOL. 1

Non-parametric Estimation for Joint Survival Distribution Using Interval-Censoring Technique

Yibing Wang, Lijia Guo † and Robert B. Brown

Department of Mathematics
The Ohio State University
Columbus, Ohio 43210
U.S.A.

Abstract

We present a method which first converges a two dimensional data to a univariate one and then uses the interval censoring technique to estimate the probability of failure of a Joint-life status from a heavily censored data. Our study shows that the life annuities of joint-life status calculated assuming independence overestimates the ones evaluated from an actual sample of population.

Keywords: Maximum likelihood estimation, Interval-censoring, Self-consistent estimation, Product-Limit Estimation

† This research has been partially supported by the actuarial faculty fund at the Ohio State University.

1. Introduction

The traditional textbooks assume that the individual lifetime variables of Joint-life status such as a married couple or twins are independent, for simplicity of calculation. However, one may ask if this assumption is valid or not. Recent study on twins has shown the strong correlation between twins, see Philip, Bent and Niels [6]. In our study, we also found some statistical evidence for the dependences in the lifetime of married couples (see [7]).

In the case of dependencies, research has shown that the expected values of annuity can differ by large amounts from values calculated assuming independence in Joint-life status [4]. Therefore modeling the joint survival distribution for the calculation of annuities without assuming independence becomes an important issue. In general, it is very difficult to provide a suitable substitute model without independence assumption. To the best our knowledge, there is no such a joint survival model that fit to the real population experience well and as widely accepted as the one under the independence assumption.

In this paper we study the impact of the independence assumption on the Joint-life or Joint-and last-survivor annuity values, present a non-parametric estimation for joint survival model and provide a method to test how well a joint survival model would fit to the the available data sample.

2. Methodology

Mathematically, for Joint-life status, we are interested in estimating $Prob(T_x^M > t, T_y^F > t)$ from an actual sample of population where T_x^M and T_y^F stand for the future lifetime variables, of a couple aged at x and y for man and women respectively and t is an arbitrary positive number.

The classical approach to study $Prob(T_x^M > t, T_y^F > t)$ is examining the bivariate distribution of the pair of variables T_x^M and T_y^F. It is the proper way if all the observations of the lifetime of people in the sample were complete and the sample size, i.e. number of observations, is not too small. However, since the average lifetime of a human being is more than 70, some of the observations might be terminated before the death occured. In this case, the life time of some people in the sample is not exactly known at the time the data is taken. Therefore, to reflect the interesting survival pattern, we have to deal with incomplete data, or censored data.

The data we use in this study is an incomplete sample of data set. However, our method could be applied to both complete and incomplete (or censored) data to estimate the joint survival distribution and to check if a given model fits a data sample from actual population.

Our data set consists of 4,211 people whose entry ages were between 30 to 65 and who were observed for a maximum of 30 years. So not every person in the study was followed up to death. In fact, among the 1,146 married couples there were only 134 pairs where both partners died at the time the data was taken. We see the censoring rate is heavy. Another important issue about the data is that the people recruited into the study were required to pass certain health examinations, a typical situation for collecting hospital data. This suggests the elder the people recruited into the study, the healthier they might be. So it is a question whether to assume that the sample is from the same distribution. Especially, one has to avoid grouping couples with different entry ages under the assumption that all the couples in the group come from the same bivariate distribution.

We now group couples by fixing x and y, the entry age of man and women respectively and assume they were all from the same bivariate distribution of the bivariate variable (T_x^M, T_y^F). We consider the following two typical age groups.

The first group consist of all the couples with entry ages 42 and 40 for man and women respectively. It has 12 couples in the group. (see appendix 1). Notice that the value “1” in column 3 or in column 5 indicates the observation was exact while a “0” means incomplete. So, only one (Couple No. 4) out of 12 couples whose lifetimes are both exactly known since both people of couple No.4 have death ID “1”, while the rest of the lines has at least one “0”.)

The second group includes couples with entry ages 46 and 44 for man and women respectively. (see appendix 2) Notice that no couple's information on the two individuals lifetimes is complete. (At least one ”0” in each of the seven lines.)

In both groups, since little exact information are available, it is impossible to derive a reasonable estimation for the bivariate distribution of $(T_x^M, \ T_y^F)$, i.e. $Prob(T_x^M > t, T_y^F > t)$.

Notice that

$$Prob(T_x^M > t, T_y^F > t) = Prob(min(T_x^M,\ T_y^F) > t).$$

We now consider the distribution of a one dimensional variable $Z_{xy} = min(T_x^M,\ T_y^F)$ instead of the bivariate one.

For the first group 1, the following three situations arise:

i) both of the individual lifetimes of the couple are exactly known. Then Z_{xy}, the minimum, is exactly known. For example, Couple No. 4.

ii)both of the life time of the couple are both censored. Then all we know is that the joint of the couple was not failed yet at the minimum of the two censored observations of the couple since a joint-life status is alive if both are alive.) So the value for Z_{xy} is a simply censored data. For example, couples No.1, 3, 6,10 and 12.

iii) if the observed lifetime of one person is exact and less than the one of the spouse which is censored, then we also know the exact failure time of the joint status. For example, couples No. 2, 5, 7, 8, 9 and 11.

For the concern of data for Z_{xy}, the situations i) and iii) are considered as death class since the exact failure time for Z_{xy} is known in both cases while the values given by situation ii) is considered as simple censoring class.

Our approach reduce the degree of incompleteness dramatically. One can see that 7 out of 12 values for a one-dimensional variable Z_{xy} are exactly known comparing to that 1 out of 12 for a two-dimensional case. the rate of exactly known data increases from the two dimensional $(T_x^M,\ T_y^F)$ to the one dimensional Z_{xy} in all of the 13 age groups. In fact, in one of the groups, the rate is increased from 0 out of 8 to 7 out of 8.)

Now that the data set for Z_{xy} from Group 1 consists of two mutually exclusive subsets, the death class and the simple loss class, we can apply the well known Kaplan Meier product limit estimation to calculate the Non-parametric Maximum likelihood estimation (MLE) ([2]) of the joint survival distribution, i.e..

$$Prob(min(T_x^M,\ T_y^F) > t)$$

or

$$Prob(Z_{xy} > t)$$

Now consider the second age group (see Appendix 2). We see that the first three observations give values to Z_{xy} that belong to death class while the

5th and 6th produce values for Z_{xy} which are in the simple censoring class. However, the 4th and 7th observation belong neither of the two classes. In fact, the situation is :

iv) — the observed lifetime of the man is exact but great than the one of his spouse which had not been observed until the death occured. In this case, the joint-life status was not only known to be alive at the moment the wife was censored but also known to be failed before the moment the man was dead.

In this case, which the value of Z_{xy} is known to be falling into an interval with finite ends, we call the incomplete data to be a **interval-censoring data** which has been encountered in medical or correctional follow-up or industrial life-testing when there is periodic inspection. Hence data set for Z_{xy} from Group 2 can be divided into three mutually exclusive classes, i.e., in addition to the regular death class and simple censored class, the interval censored data. Notice that we receive more information from the interval censored data than the simple censoring data.

However, the Kaplan Meier method can not be applied for the MLE of the Joint survival distribution from the interval-censoring data. Peto, R ([1]) proves the existence and the uniqueness of the MLE for the interval censored data sample and provides a computer algorithm to calculate the MLE directly. The criticism of the algorithm is that it is too cumbersome and hence seldom be used. Bruce W. Turnbull [5] provides a less cumbersome algorithm to calculate the so called self-consistent estimator which was claimed to be equivalent to the maximum likelihood estimator. Despite the fact that the method is often be referred, the proof of the equivalence is only given for the case when the algorithm is convergent. But the convergence of the algorithm is proved only if the initial point is picked close enough to the MLE. In fact, we have an counterexample where the equivalence between the consistent estimator(s) and the Maximum likelihood estimator is not true. In our counterexample, there are two self consistent estimators, one is the maximum likelihood estimator, and the other is a saddle point. We can show that the Turnbull's algorithm will always converge to the maximum likelihood if we pick any initial staring point other than the saddle point. So the self-consistent estimator is not always equivalent to the maximum likelihood estimator. In this particular example, the algorithm still enable us to find the right MLE as long as one could avoid to pick up the saddle point as the initial point. However, the algorithm would not guarantee a MLE for any interval censored data in general.

Our study use a different approach. We develop an algorithm to calculate estimators of conditional probability instead of the empirical distribution, which are calculated in both of the two papers mentioned.

Our algorithm provide an explicit formula for the estimator as well as a method to study the large sample property of the MLE, for example, the consistency. In the case of small sample with simple structure, the exact self-consistent estimators can be computed easily even by hand, and could be easy to verify if it is the MLE. On the other hand, for large sample data, we developed an unique block decomposition technique for calculating the estimator for the efficiency and economical reason.

3. Applications

We now use our method to handle the interval censored data for our study. We group couples by fixing x and y, the entry age of man and women respectively and assume they were all from the same bivariate distribution of the bivariate variable (T_x^M, T_y^F). There are only 13 such age groups whose size is greater than or equal to seven and where x and y are both greater than or equal to 40. We then apply the algorithm and find the estimator for the survival curve for each of the 13 age groups as well as a general data with many incomplete observations.

By applying the interval censored technique to the 13 age groups from this sample data, we reach the conclusion that, for 11 (out of 13) age groups, the traditional method with the independence assumption overestimates the survival probability of Joint-life status. Furthermore, comparing to our estimated distribution with a constant interest rate of 6% in the computation, the net single premium A_{xy} of life insurance of a Joint-life status is underestimated while the annuity part a_{xy} is overestimated by assuming the independence assumption in each of the 11 (out of 13) age groups. On the other hand, in each of the 11 (out of 13) age groups, the net single premium A_{xy} of life insurance of a Joint-and -last-survivor status is overestimated while the annuity part a_{xy} is underestimated by the independence assumption. Therefore, our study shows that insurance companies should charge more premium for the joint-life insurance policy but less for the Joint-and -last-survivor life insurance policy.

References

[1] Peto,R., "Experimental Survival Curves for Interval censored data," Applied Statistics, 22, No. 1(1973),86-89.

[2] Kaplan E.L. and Meier Paul, " Non-parametric estimation from incomplete observations", American Statistical Association, Vol. 53, pp. 457 – 481, 1958.

[3] Klien J.P., "Semi-parametric estimation of random effects using the Cox model based on the EM algorithm", Biometrical Journal, Vol. 34, 1992.

[4] Carriere J. F. and Chen L.K. ,"the bounds of bivariate distributions that limit the last-survivor annuities," Transaction of the Society of Actuaries, Vol.38, 1986.

[5] Turnbull B.W., " the empirical distribution function with arbitrarily grouped, censored and truncated data," J. R. Statist. soc., B, 38, 209-295, 1976.

[6] Philip H., Bent H. and Niels V.H., "Measuring the similarities between the lifetimes of adult Danish twins born between 1881-1930," Journal of Amer. Statis. Assoc., Vol. 87, No. 417, 17-24, 1992.

[7] Wang, Y, "Dependencies in Joint-Life Status", present to the Joint Meeting of AMS and AMA, Cincinnati, January 1994.

Appendix 1

Group 1.

(Entry ages: man=x=42, woman=y=40)

No.	Man's	Dth ID (M)	Woman's	DthID(W)
1	32.021585	0	35.021585	0
2	26.542214	1	34.999668	0
3	33.96954	0	35.002407	0
4	15.25191	1	29.451775	1
5	26.416187	1	30.391492	0
6	31.722986	0	33.6079	0
7	21.021718	1	31.881889	0
8	16.040944	1	31.881889	0
9	23.709364	1	31.766822	0
10	31.838054	0	31.838054	0
11	15.331361	1	31.950381	0
12	1.964365	0	1.964365	0

Appendix 2

Group 2.

(Entry ages: man=x=46 , woman=y=44)

No.	Man's	Dth ID (M)	Woman's	DthID(W)
1	18.660097	1	35.246241	0
2	32.470924	0	27.681929	1
3	20.947746	1	34.983229	0
4	29.799746	1	26.783307	0
5	24.621684	0	24.621684	0
6	17.871063	0	33.808158	0
7	29.194243	1	25.542223	0