Two-sample problem under random truncation

Adrián Lago Balseiro

This document contains a summary of the Master's Thesis entitled *Two-sample problem under random truncation*, written by Adrián Lago Balseiro and supervised by Jacobo de Uña Álvarez and Juan Carlos Pardo Fernández. This work is the beginning of a Doctoral Thesis, thus its publication is not authorised.

Summary

Survival Analysis arises to study the time needed for some event to happen; for example, lifetimes of individuals or electronic components lifespan are studied. Then, Survival analysis is relevant for many different scientific fields. The event of interest is usually known as failure and the time needed for this event to happen is the time to event. When collecting data, two main problems can arise: censorship and truncation. In this work, we will deal with the second one. Truncation occurs when some individuals are not observed as they do not satisfy certain condition. In the particular case of left-truncation, individuals whose event occurs before certain threshold are not observed. For example, this is the case of individuals that are not large or old enough. That is, lack of information comes from the impossibility to observe individuals, but for every observed individual, the time to event will be known. Moreover, left-truncation causes the empirical quantiles to be shifted to the right respect to the theoretical quantiles of the distribution from which they were generated. This has to be taken into consideration when estimating the survival function. That is, a new estimator has to be defined, as ignoring the truncation when estimating the survival function makes the estimate to be biased. The maximum likelihood estimate of the survival function for left-truncated data is defined in Lynden-Bell (1971) and it is analogous to the Kaplan-Meier estimate of the survival function for right-censored data.

The two-sample problem consists of testing the null hypothesis of equality between two distributions. This problem has been studied since the beginning of the last century and many statistics have been proposed to address it. One of the first ones to be studied is the well-known Kolmogorov-Smirnov statistic. One property of that test is that it is distribution-free, that is, the distribution of the statistic does not depend on the distribution of any of the samples.

Let us now consider two left-truncated samples; we want to compare the survival functions of the times to event. In this work, the Kolmogorov-Smirnov statistic is adapted for left-truncated data, considering the Lynden-Bell estimator. The behaviour of this statistic has not been studied in the literature yet and that is the main goal of this work. For right-censored data, the two-sample Kolmogorov-Smirnov test was studied in Schumacher (1984). Under left-truncation, we

are only aware of the work by Guilbaud (1988), who considered the Kolmogorov-Smirnov test but in the one-sample problem.

First of all, a simulation study was carried out to analyse the distribution of the test statistic under the null hypothesis. More in detail, by defining different simulation scenarios, values of the statistic under the null hypothesis were simulated. Some scenarios had the same time-to-event distribution and other scenarios shared the same truncation distributions. The results of this simulation study suggest the test is no longer distribution free, in contrast to the test for complete data. Now the distribution of the test statistic depends on both the time-to-event and the truncation distributions. According to this conclusion, it looks appropriate to employ bootstrap to approximate the $p-$value of the test. For that sake, the bootstrap resampling plan proposed in Gross and Lai (1996) for left-truncated and/or right-censored data was adapted to the two-sample problem under random left-truncation. By means of a Monte Carlo study, the bootstrap resampling plan has been proved to be adequate, since it leads to a correctly calibrated test. In conclusion, the proposed bootstrap resampling plan can be employed to approximate the $p-$value of the test.

As the bootstrap works properly, the proposed test can be compared to other existing tests to investigate their relative performance. In this work, the Kolmogorov-Smirnov test for left-truncated data was compared to the log-rank test, one of the most common tests in Survival Analysis to address the two-sample problem. One should recall the log-rank test is optimal under proportional hazards. This property of the log-rank test is also seen in simulations carried out in this work, as in proportional-hazards scenarios under the alternative hypothesis, the log-rank test has power and can detect differences between the theoretical survival functions already with moderate sample sizes. On the other hand, the Kolmogorov-Smirnov-type test has only power for large sample sizes. However, if we consider nonproportional-hazards scenarios, the results have to be carefully interpreted, as now the truncation variable has more effect on them. If the upper bound of the support of the truncation random variable in both samples is close to or over the intersection point of the density functions being compared, the log-rank test will outperform the Kolmogorov-Smirnov test. On the other hand, with a less aggressive truncation model, the Kolmogorov-Smirnov test for left-truncated data will outperform the log-rank test, even for small sample sizes. Moreover, in this work we also investigate scenarios under the alternative hypothesis for which the log-rank test achieves power as low as the significance level, whereas the Kolmogorov-Smirnov test is able to detect the difference between the distributions.

A real dataset analysis was carried out. That dataset consists of times of pregnancy, from the time women are aware of their pregnancy and the follow-up begins, to the time they experiment an spontaneous abortion (Meister and Schaefer, 2008). This data is left-truncated due to the delayed entry into study. In other words, some women can experiment an spontaneous abortion before being aware of their pregnancy. These women were split into two groups: women in one group are exposed to coumarin derivatives and the rest belong to the control group. The main goal of the study was to determine whether the coumarin derivatives have effect on the distribution of the times to the spontaneous abortion. From all of the above, the Kolmogorov-Smirnov test for left-truncated data and the log-rank test are suitable for this situation. The

$p-$values of both tests are greater than the usual significance levels, which lead to conclude that there is no statistical evidence to reject the null hypothesis of equality of the survival functions in both groups, that is, the coumarin derivatives have no effect on the pregnancy duration up to an spontaneous abortion.

This work concludes with an appendix that includes some functions programmed in `R`, such as the Lynden-Bell estimate of the survival function, the Kolmogorov-Smirnov statistic for left-truncated data, a function to estimate the $p-$value of the test with bootstrap and the log-rank test. It is important to recognize that that the Lynden-Bell estimate and the log-rank test can be obtained with functions already included in the `survival` package of `R`, but they were independently programmed to obtain a deeper understanding of them.

References

[1] Gross TS, Lai TL (1996) Bootstrap methods for truncated and censored data. *Statistica Sinica* 6:509-530.

[2] Guilbaud O (1988) Exact Kolmogorov-Smirnov-type tests for left-truncated and/or right-censored data. *Journal of the American Statistical Association* 83:213-221.

[3] Lynden-Bell D (1971) A method of allowing for known observational selection in small samples applied to 3RC quasars. *Monthly Notices of the Royal Astronomical Society* 155:95-118.

[4] Meister R, Schaefer C (2008) Statistical methods for estimating the probability of spontaneous abortion in observational studies-analyzing pregnancies exposed to coumarin derivatives. *Reproductive Toxicology* 26:31-35.

[5] Schumacher M (1984) Two-sample test of Cramér-von Mises and Kolmogorov-Smirnov type for randomly censored data. *International Statistical Review* 52:263-281.