Computer Assignment 1

Jan Alexandersson, Anton Stråhle & Max Sjödin

September 19, 2020

- Argue for the performance of the estimators relative to the true theoretical model used, is any of the two preferable?
- What is the effect of censoring?
- What is the effect of n?

In this computer assignment we aim to estimate the cumulative hazard rate with the Nelson-Aalen estimator and the Kaplan-Meier estimator. However, the Kaplan-Meier estimator estimate the survival funcktion and not the cumulatize hazard rate, but this can be obtained by taking the negative logarithm of the Kaplan-Meier estimator. We chose to present both estimators as cumulative hazard rates so that they can easily be compared.

We begin by generating n Weibull distributed random numbers from Weibull distribution given by the density function

$$f(t;a,b) = \frac{a}{b}(\frac{t}{b})^{a-1}\exp\{-(\frac{t}{b})^a\}, t, a, b \geq 0,$$

where $a = 4.5$ and $b = 22.5$. This is done for $n = 10, 100, 200, 500$ and 1000.

The Nelson-Aalen estimator is given by

$$\hat{A}(t) = \sum_{T_j \leq t} \frac{d_i}{Y(T_j)},$$

where $Y(T_j)$ is the number of individuals at risk at time T_j and d_i is the number of events if there is a tie and 1 if not.

The Kaplan-Meier estimator is given by

$$\hat{S}(t) = \prod_{T_j \leq t} \Big(1 - \frac{d_i}{Y(T_j)}\Big).$$

As written earlier, we get the Kaplan-Meier estimate for the cumulative hazard rate by $-\log(\hat{S}(t))$. In figures 1-5a) we can see the Nelson-Aalen estimate and the Kaplan-Meier estimate compared with the true cumulative hazard rate, in the case of no ties and no censoring. We can also see 95% confidence intervals for the estimates, which are log-transformed Wald intervals. We can see that when increasing n our estimates are closer to the theoretical cumulatie hazard rate and our confidence intervals less wide. We also see that tere is not much difference between using the Nelson-Aalen estimate compared to Kaplan-Meier estimate.

We continued by also computing the Nelson-Aalen and Kaplan-Meier estimates when we only observe total number of events at the end of each interval of length 0.1, therefore creating "ties". We regard these ties as "true ties". We see in figures 1-5b) that the Kaplan-Meier estimate is slightly more accurate compared to the theoretical and is therefore preferred, however the difference to the Nelson-Aalen estimate is not large.

Overall the effect of n is that the Nelson-Aalen and Kaplan-Meier estimates are underestimating the cumulative hazard rate for smaller n and also deviates more from the theoretical cumulative hazard rate.

We then add censoring by generating n exponentially distributed random numbers with mean 80. We denote these by C_i. If $T_i > C_i$ the observation is censored. We can see in figures 1-5c) that when the estimates deviate from the theoretical distribution for when t is large ($t > 25$).

Since the expected value of the C_i's are large compared to the T_i's we can expect more censoring for larger values of t. That is, we can expect the main effect of the censoring to show for larger values of t. For smaller n we can not say much about the effect because of the uncertainty in the estimates. However, for large values of n (see Figure 4d) and 5d) we can see a clear difference. We see that fitting a distribution to the censored data while only taking uncensored datapoints into account will lead to bad estimates with regards to the dataset when we also take censored observations into account. We can see in Figure 4d) and 5d) that fitting without regards to the censoring will clearly underestimate the cumulative hazard rate in the end of the time period and that both Nelson-Aalen and Kaplan-Meier estimate is close to the true cumulative hazard rate, thus performing well also with censored data.

Appendix

Figure 1 : For 10 observations

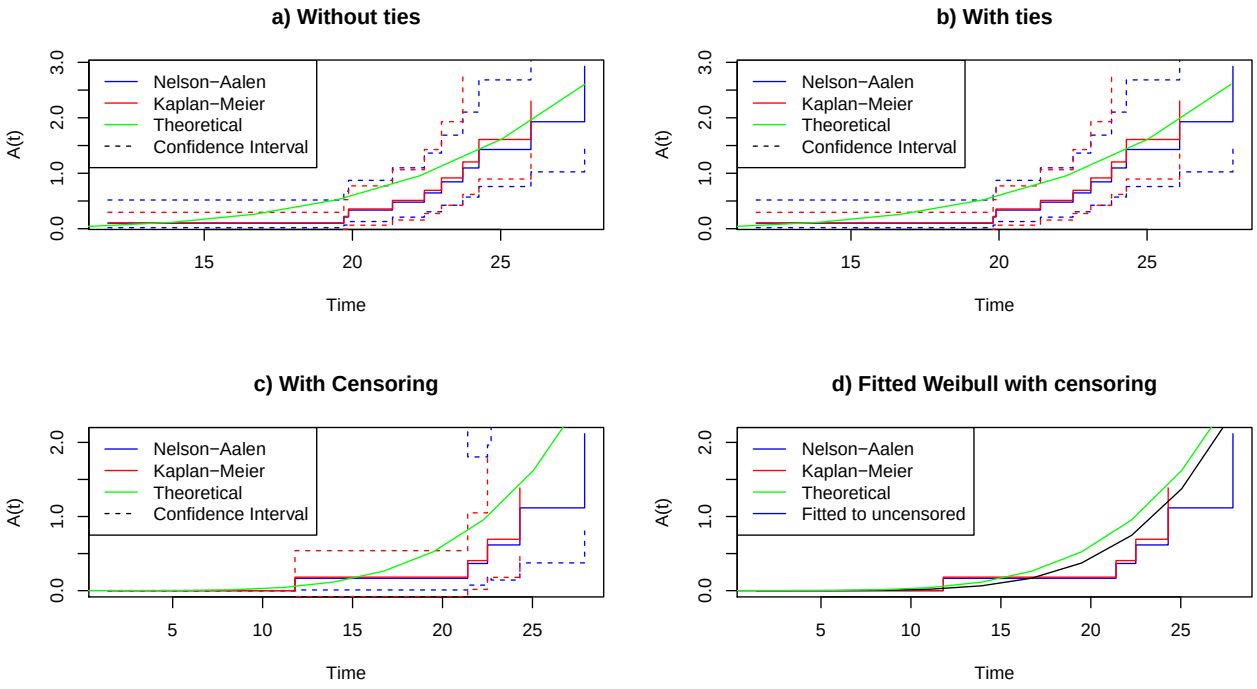

Figure 2 : For 100 observations

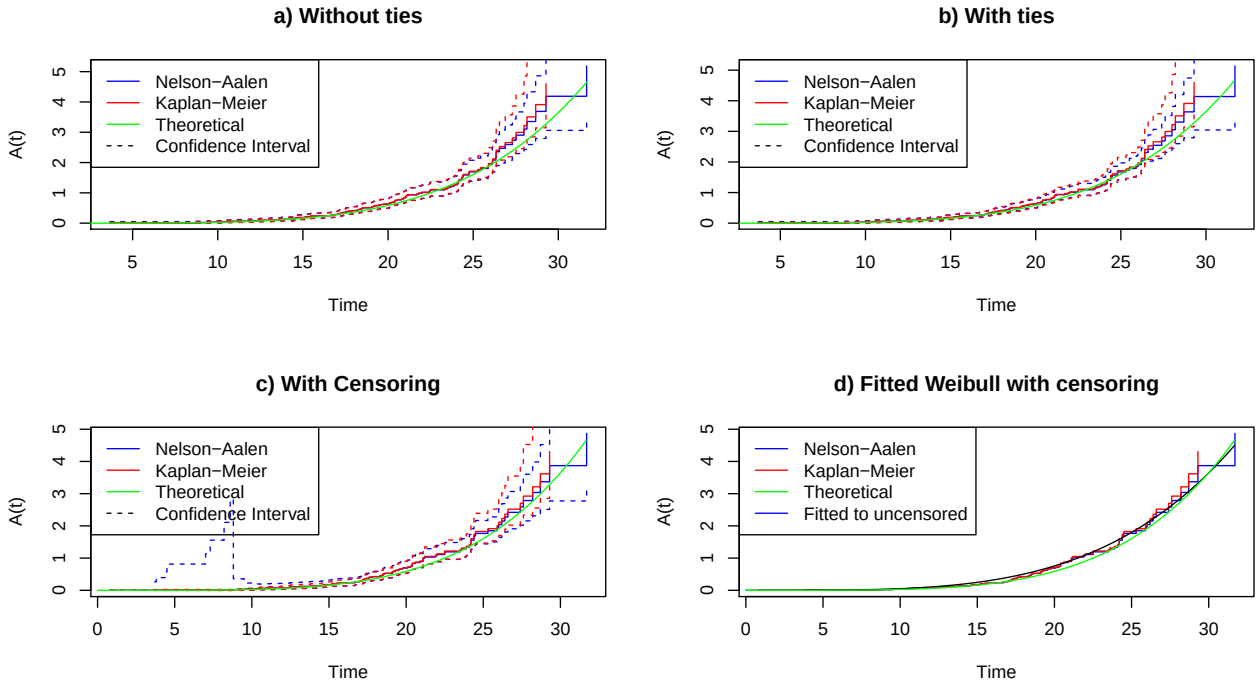

Figure 3 : For 200 observations

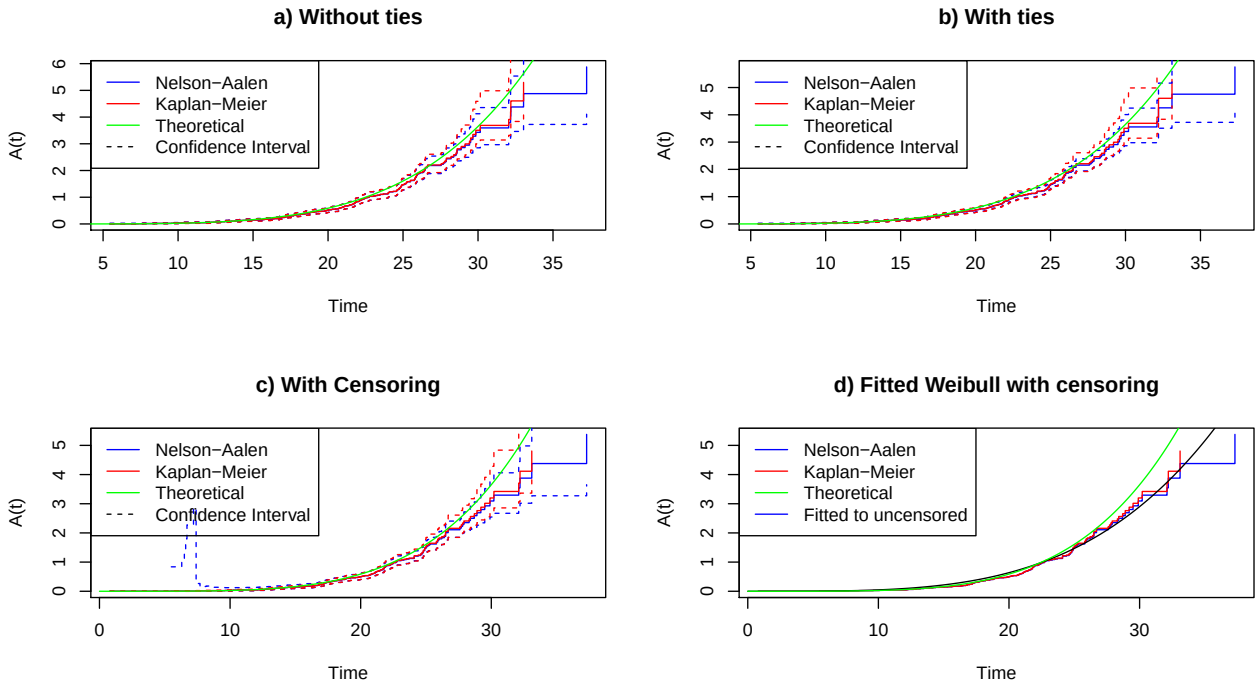

Figure 4 : For 500 observations

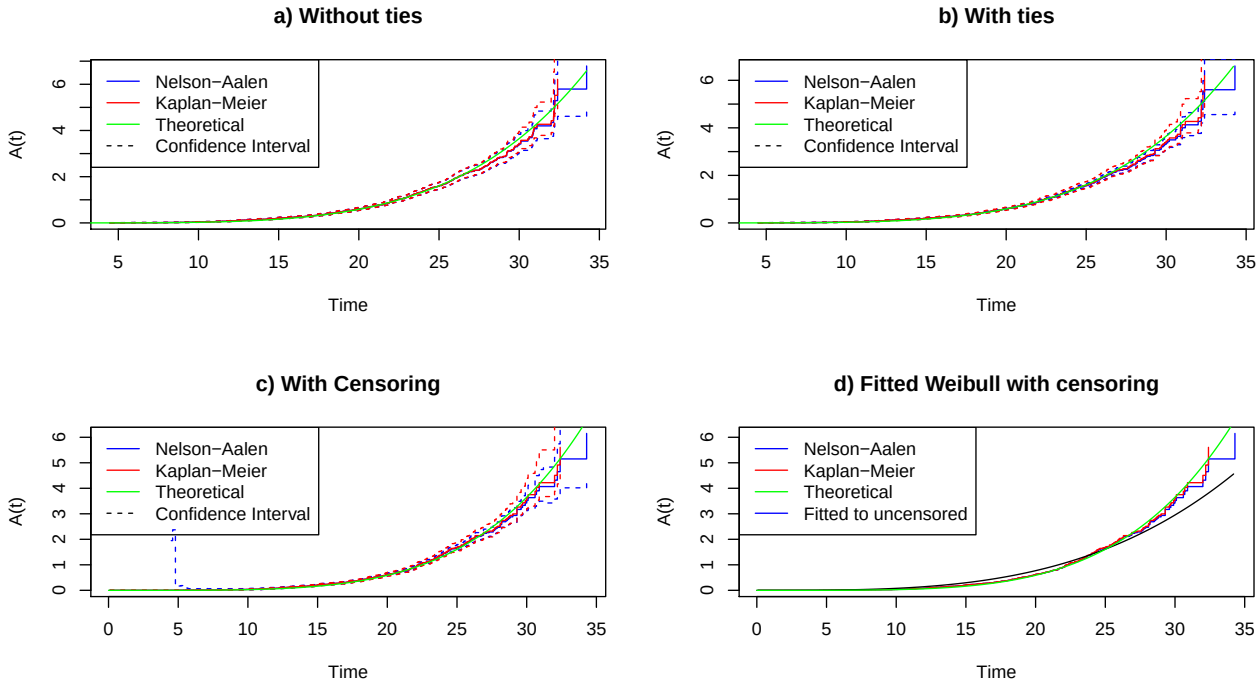

Figure 5 : For 1000 observations