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 using both the Nelson-Aalen 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 choose to present both estimators as cumulative hazard rates to allow for easier comparisons.

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

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

As mentioned earlier, we obtain the Kaplan-Meier estimate of the cumulative hazard rate from $-\log(\hat{S}(t))$. 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 see that by increasing n our estimates move closer to eachother as well as to the theoretical cumulative hazard rate whilst our confidence intervals become more narrow. Lastly we also note that the differences between the two estimates are minor.

We continued by computing the Nelson-Aalen and Kaplan-Meier estimates whilst only observing the 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 counterpart and is therefore preferred, however the difference compared to the Nelson-Aalen estimator is once again not very large.

We can see that when increasing n the confidence intervals gets more narrow but still capture the true distribution until t gets high. We notice that the fit is worse for the last observations, for all values of n, which we can expect since the jumpsize is larger when we have less observations left. That is, the estimators have higher variance the fever observations we have left.

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 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

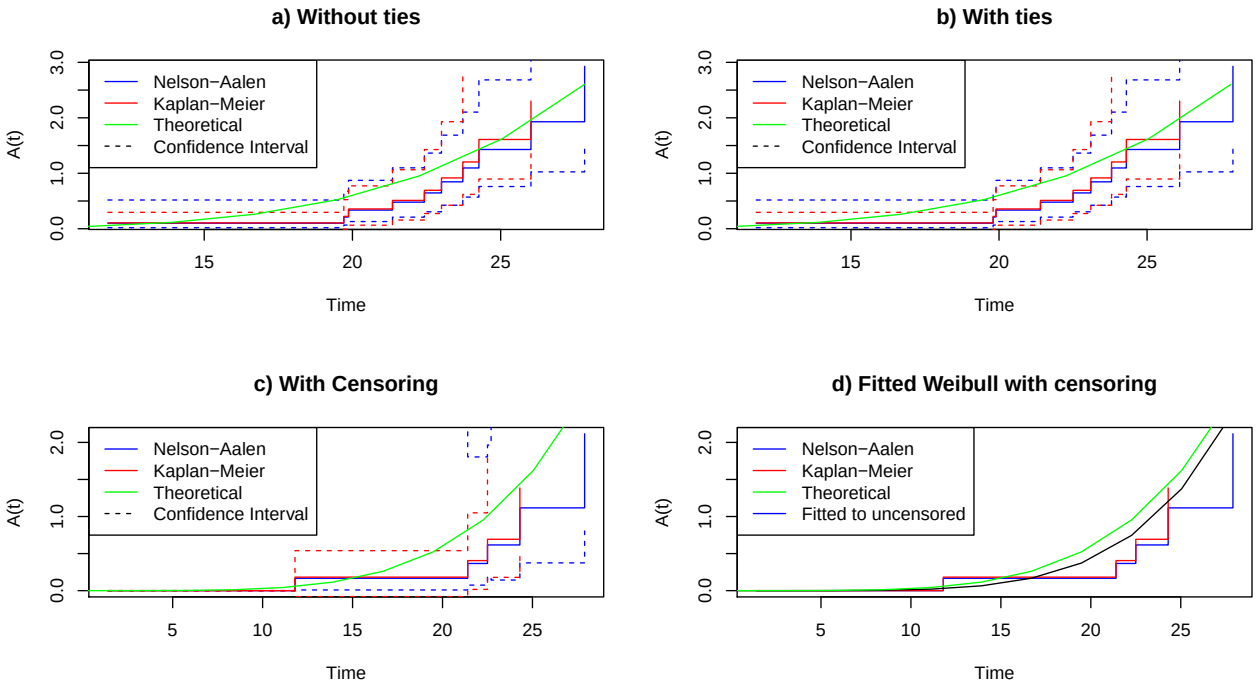

Figure 2 : For 100 observations

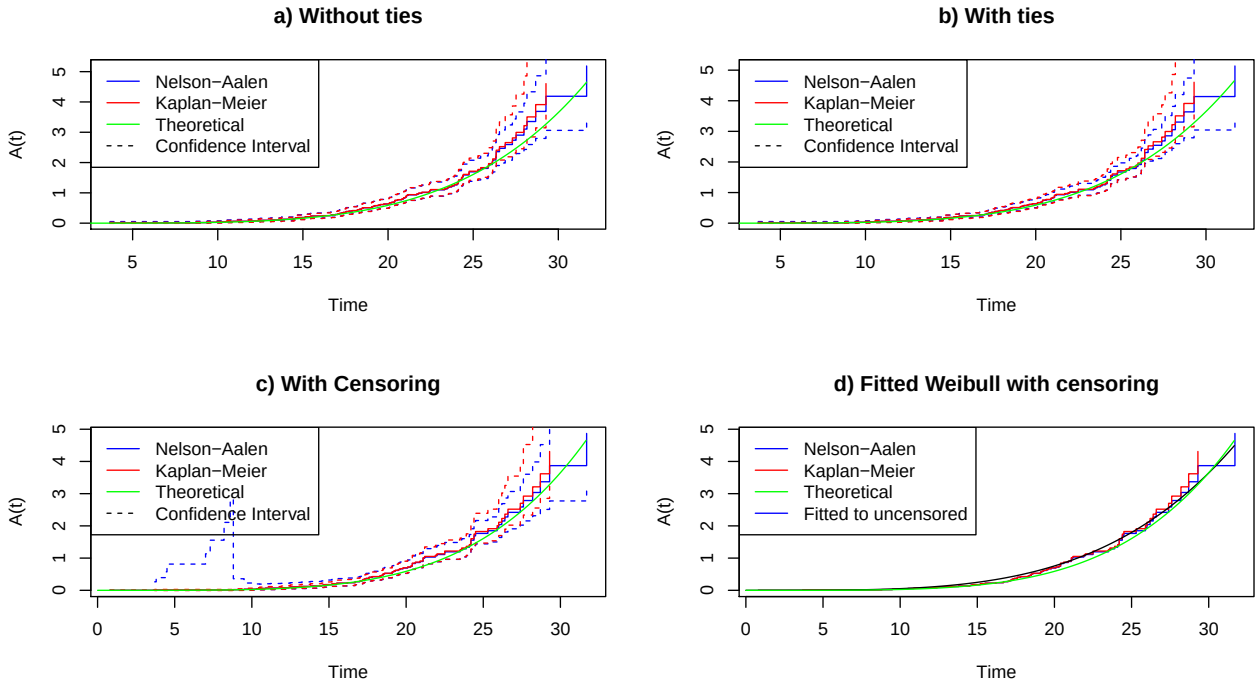

Figure 3 : For 200 observations

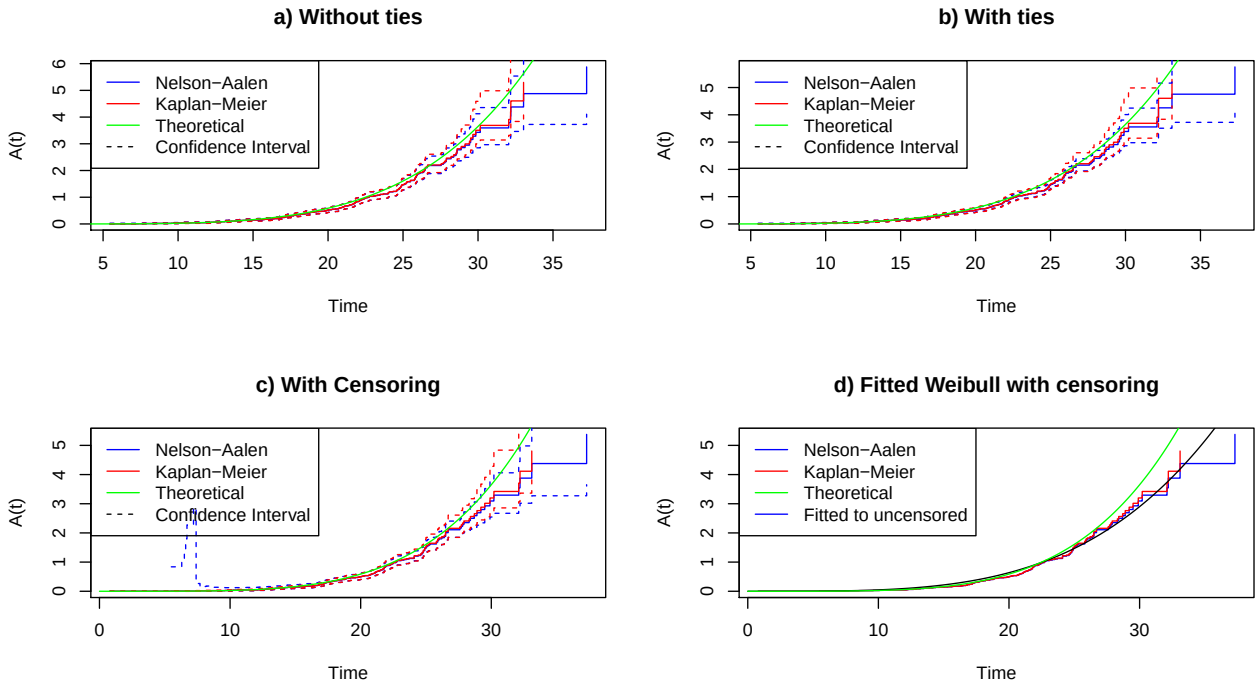

Figure 4 : For 500 observations

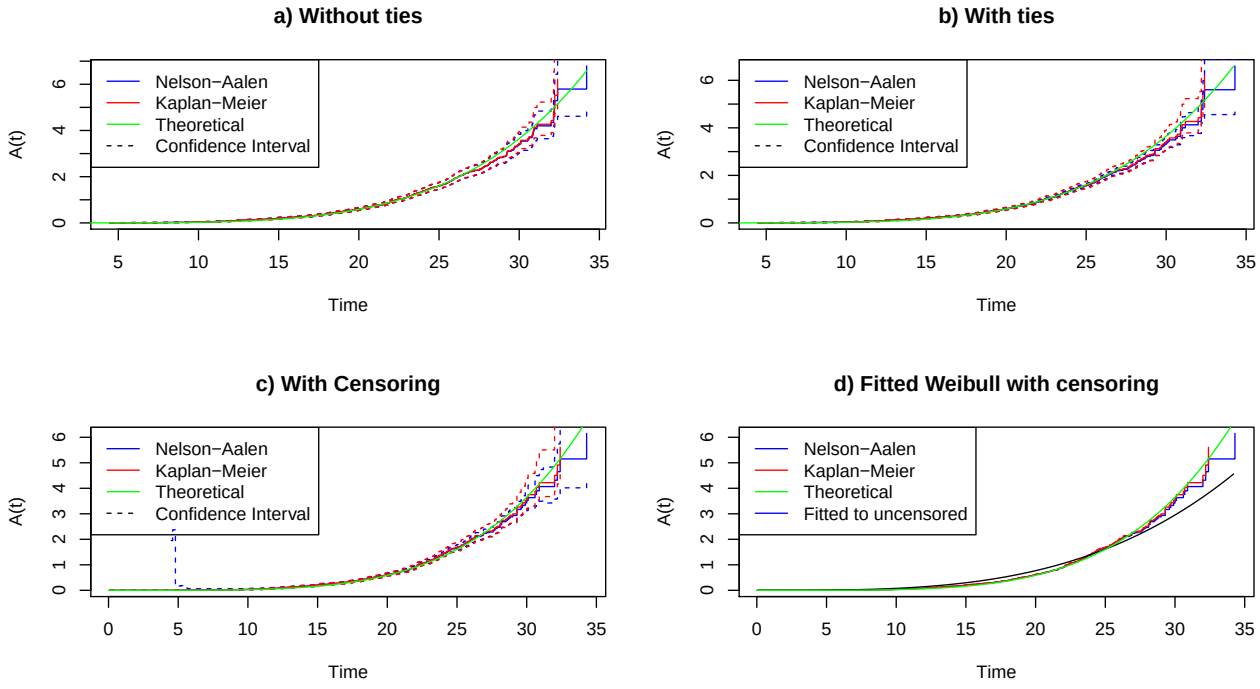

Figure 5 : For 1000 observations