Report I
MT7052
Inference and prediction for life and health processeses

Malin Andersen - malinandersen1988@gmail.com
Oscar Potts - oscar.n.potts@gmail.com
Majken Gunnarsson - majken.gunnarsson@gmail.com

December 11, 2024

Task 1

In task 1 we have simulated survival data according to a Weibull distribution from whom we calculated the Kaplan-Meier & Nelson-Aalen[1] survival and hazard functions. The true survival and hazard functions can be seen in figure 1 and is the Weibull distribution with shape parameter $a = 5.5$ and scale parameter $b = 22.5$, these are the response functions we want to estimate. Using Kaplan-Meier and the Nelson-Aalen we estimate survival or hazard according to:

$$\hat{S}_{KM}(t) = \Pi_{j:T_j \leq t}(1 - \tfrac{c_i}{Y(T_j)}) \quad \hat{A}_{NA}(t) = \textstyle\sum_{j:T_j \leq t} \tfrac{c_i}{Y(T_j)}. \tag{1}$$

Where $\hat{S}_{KP}$ denotes the Kaplan-Meier survival estimator, $\hat{A}_{NA}$ the Nelson-Aalen hazard estimator, T_j is observed time of death, c_i denote if the observation is censored or not and $Y(T_j)$ is the remaining population at time T_j.

As Nelson-Aalen estimates the hazard function and Keplen-Meier estimates the survival function we use the definitions to compare hazard and survival respectively.

$$S(t) = e^{-A(t)} \quad A(t) = -ln(S(t)) \tag{2}$$

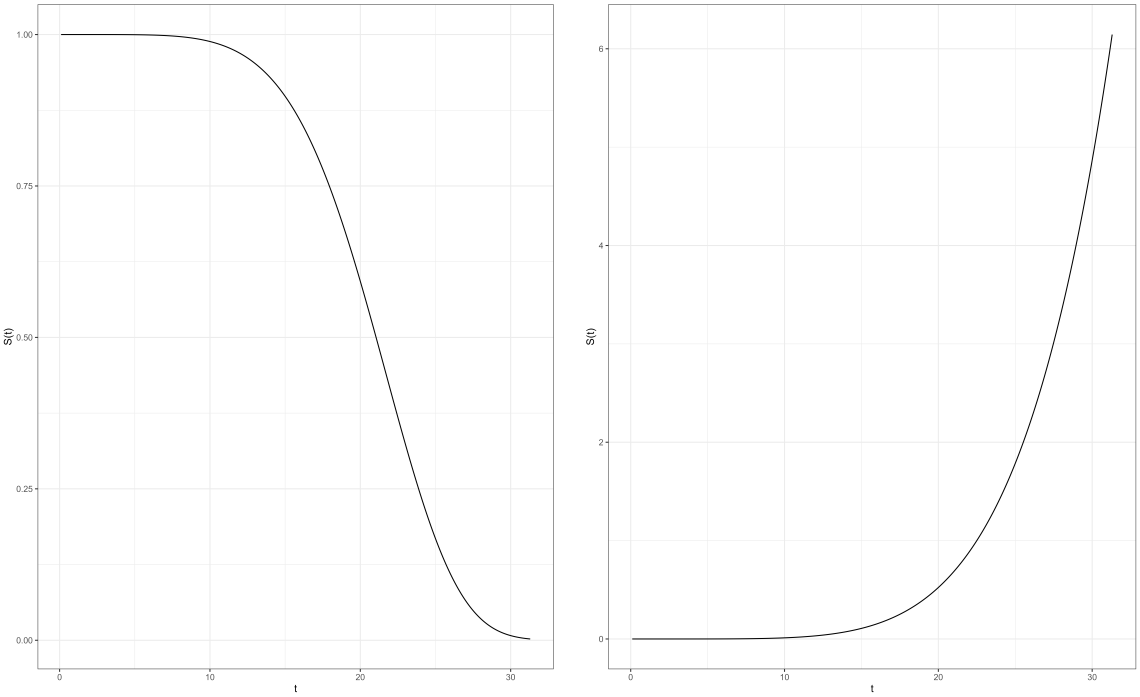
To estimate the variance we use the Greenwood for Kaplan-Meier and standard plug-in variance estimators for the Nelson-Aalen as:

$$\begin{aligned} \sigma\sigma_{KP}(t) = \hat{S}_{KP}(t)^2 \textstyle\sum_{j:T_j \leq t} \tfrac{c_i}{(Y(T_j)-c_I)Y(T_j)} \quad \hat{\sigma}_{NA}(t) &= \textstyle\sum_{j:T_j \leq t} \tfrac{(Y(T_j)-c_i)c_i}{(Y(T_j)-1)Y(T_j)^2} \\ &= A(t) \textstyle\sum_{j:T_j \leq t} \tfrac{Y(T_j)-c_i}{(Y(T_j)-1)Y(T_j)} \end{aligned} \tag{3}$$

In calculating the confidence intervals we use these estimates plain confidence intervals and use equation 2 to translate between response function. We do not use the log-transformation as suggested by Aalen, Borgan and Gjessing[1] but use the standard confidence interval estimate.

In R we use the `survival`[5] package for estimating the Kaplan-Meier and `mvna`[2] package to estimate the Nelson-Aaleen.

The results are first without censoring observations and then with censoring where we discuss the differences between the methods and what happens when we introduce censoring. In presenting the results we will show the deviance from true distribution as seen in figure 1.



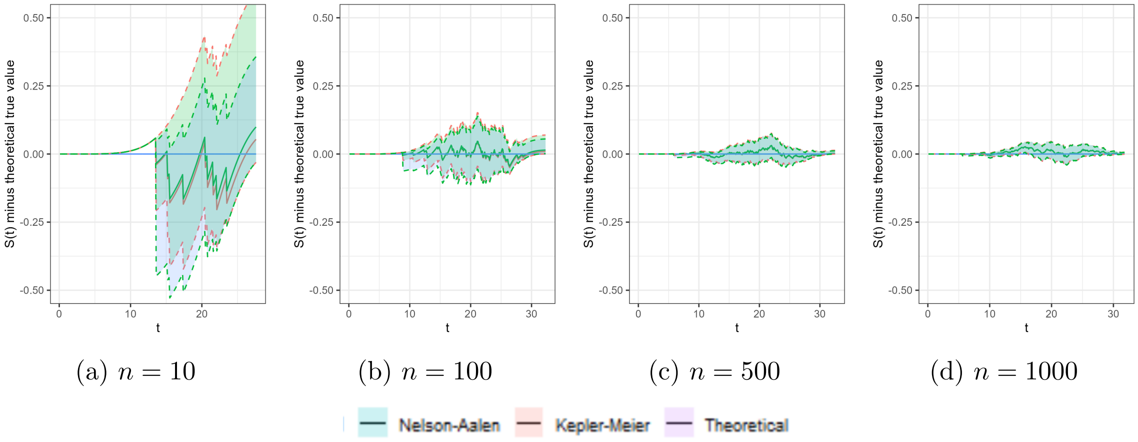
(a) Survival Function

(b) Hazard function

Figure 1: True Survival and Hazard function of Weilbull distribution used in assignment, results will be presented in relation to these true functions.

Esstimation on Non-Censored Data

In figure 2 & 3 we see the difference between our estimated survival and hazard function towards the true underlying distribution for different amount of non-censored data. What is observed is how both methods convert towards the true underlying Weibull distribution. At all levels of simulated data points we see how the confidence interval covers the true distribution and how the interval shrinks as n increases. Generally we see more variance and deviance for the middle part of our time interval, and less to no variance for low amd high values, this is because we have few to none observations in these intervals. One may argue that both methods underestimates the true underlying survival function for smaller sample sizes, and also that Kaplan-Meier shows more an exponential nature compared to Nelson-Aleen, as seen in the hazard plots, but no definite or significant conclusions can be drawn. We also note that the step-like nature of the estimator disapear visualizing the results as the deviance from true underlying survival and hazard function, but it is believed that this gives a better picture of the behavior of our estimators.



(a) $n = 10$

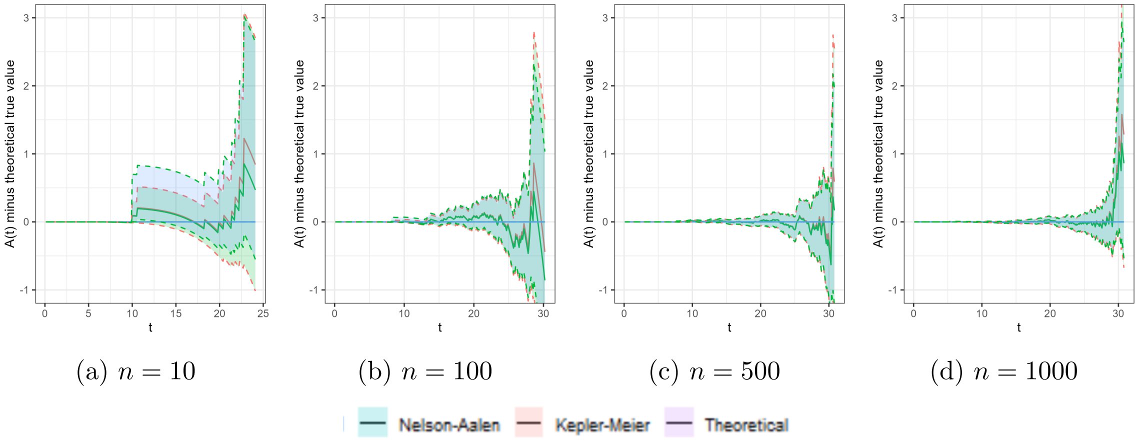
(b) $n = 100$

(c) $n = 500$

(d) $n = 1000$

Figure 2: Difference between theoretical and estimated survival functions according to underlying Weibull distribution towards the Kaplan-Meier estimator and Nelson-Aalen estimator with 95% percentile confidence for different amount of data-points generated.

Estimating the median of the distribution as seen in figure 4 we see that they cover the true median value of $m = bln(2)^{1/a} = 21.049$, even for low values of n (even as the variance is quite large). One may also observe that the Nelson-Aalen is always larger than the Kaplan-Meier, but this is by no ways significant. We also note that the upper confidence interval for $n = 10$ disapear when estimating the median due to a numerical error.



(a) $n = 10$ (b) $n = 100$ (c) $n = 500$ (d) $n = 1000$

Figure 3: Difference between theoretical and estimated hazard functions according to underlying Weibull distribution towards the Kaplan-Meier estimator and Nelson-Aalen estimator with 95% percentile confidence for different amount of data-points generated.

```
          Median Lower CI Upper CI
n=10    20.60597 15.40320
n=100   21.36780 19.98221 22.02791
n=500   21.25135 20.76050 21.84473
n=1000  21.11619 20.72275 21.45620
```

(a) $Kaplan - Meier$

```
          Median Lower CI Upper CI
n=10    20.73730 15.40320
n=100   21.41457 19.98221 22.32173
n=500   21.25549 20.76050 21.88808
n=1000  21.11991 20.72992 21.45635
```

(b) $Nelson - Aalen$

Figure 4: The median survival time and the 95% upper and lower confidence interval for the median survival time for both the Kaplan-Meier and Nelson-Aalen estimate. n is the number of samples used in the estimates.

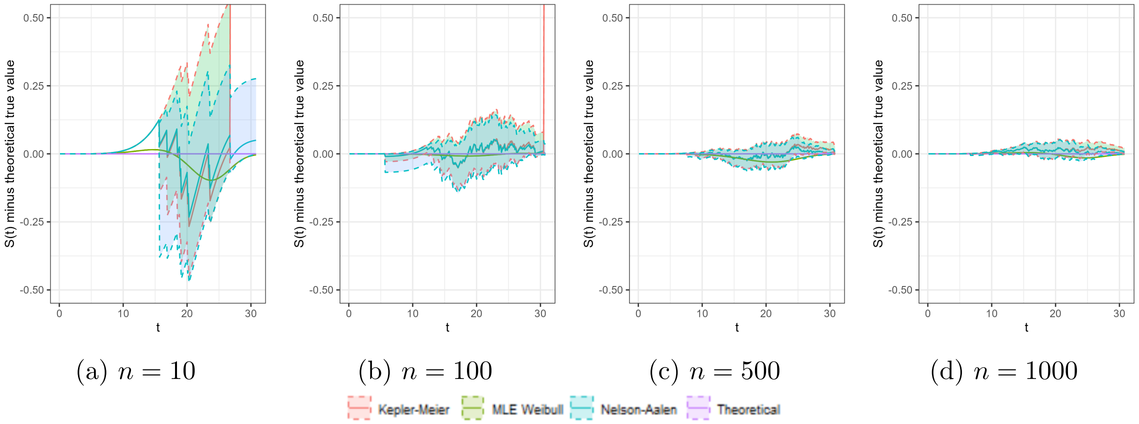
Estimation on Censored Data

Introducing censoring of observations, being uniformly distributed $C \sim U(20, 60)$ we allow observations to be non-observed, meaning in an applied setting that we have not seen a occurrence of said event yet for observation c_i.

We have also used the *EnvStats*[3] package and *eweibull* to make an MLE estimate on the underlying weibull distribution ignoring censored observations.

As seen in figure 5 & 6 we see the same analysis as made in the initially and that our estimators converge towards the true underlying distribution. We notice for the smaller sample sizes, specifically for the hazard function how the estimates are ~~quite~~ highly varied.

We also observe that by ignoring censored observations and conducting a Maximum Likelihood Estimation (MLE), we tend to underestimate the true hazard rate, underestimate early survival rates and overestimate longer survival times.~~We also see how ignoring censored observation and conducting MLE estimate, that we underestimate the true hazard rate and underestimate early survival rates and overestimate long survival times.~~ This over and underestimation is observed across all sample sizes, but we do see how the margin decreases as the amount of data-points increase.



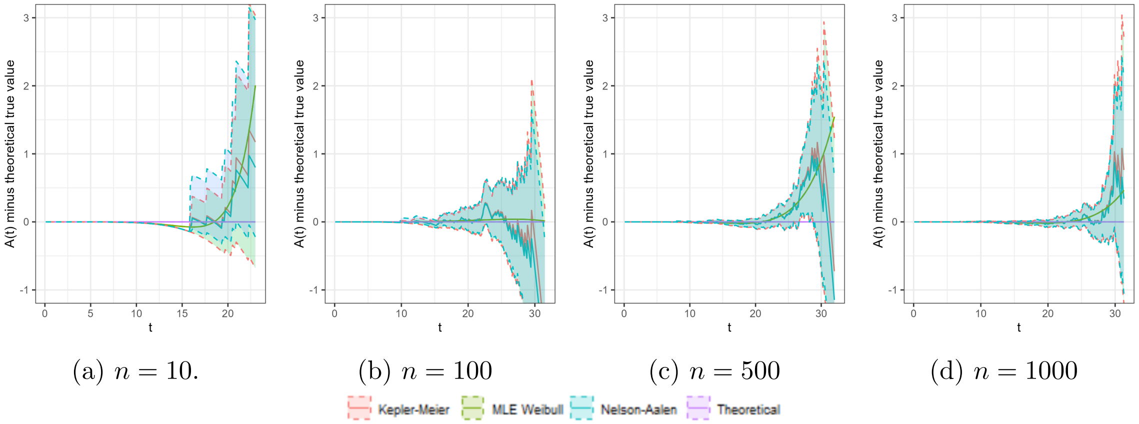
(a) $n = 10$

(b) $n = 100$

(c) $n = 500$

(d) $n = 1000$

Figure 5: Comparison with theoretical and estimated survival functions according to underlying Weibull distribution and the Kaplan-Meier and Nelson-Aalen estimator with 95% percentile confidence across different amount of data-points generated.



(a) $n = 10.$

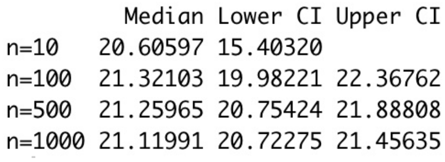
(b) $n = 100$

(c) $n = 500$

(d) $n = 1000$

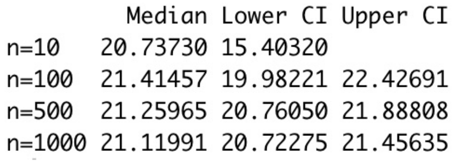
Figure 6: Comparison with theoretical and estimated hazard functions according to underlying Weibull distribution and the Kaplan-Meier and Nelson-Aalen estimator with 95% percentile confidence across different amount of data-points generated

Finally, estimating the median time from our estimates we have in figure 7 same results as in the non-censored data. Both estimates correctly converge to and estimates the true underlying median, even as data is being censored. Again, the upper confidence interval for $n = 10$ disapears due to numeric calculation errors, but one may context that it high.



```
         Median Lower CI Upper CI
n=10   20.60597 15.40320
n=100  21.32103 19.98221 22.36762
n=500  21.25965 20.75424 21.88808
n=1000 21.11991 20.72275 21.45635
```

(a) $Kaplan - Meier$



```
         Median Lower CI Upper CI
n=10   20.73730 15.40320
n=100  21.41457 19.98221 22.42691
n=500  21.25965 20.76050 21.88808
n=1000 21.11991 20.72275 21.45635
```

(b) $Nelson - Aalen$

Figure 7: The median survival time and the 95% upper and lower confidence interval for the median survival time for both the Kaplan-Meier and Nelson-Aalen estimate used on censored data. n is the number of samples used in the estimates.

Part II

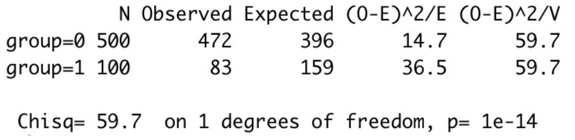
In the second part we used the Weibull random numbers with $n = 500$ that was generated in part 1. This group is denoted as group 1 and have the parameters $a = 5.5$ and $b = 22.5$ in the density:

$$f(t;a,b) = \frac{a}{b}(\frac{t}{b})^{a-1}exp(\frac{t}{b})^{a}, t, a, b >= 0 \tag{4}$$

In addition, we generated another set of $n = 100$ Weibull random numbers which is denoted as group 2. This group have the parameters $a = 4.5$ and $b = 28$. For the first part in part 2 we applied censoring on both groups that were uniform on [20,60]. In the second part of part 2 we changed the censoring for group to being uniform on [30,60]. With the resulting censored groups we fit a Cox regression model and compared it to the Kaplan-Meier estimation of the survival curves.

Same censoring interval

In this part we carried out a Logrank test to check that the cumulative hazards are not equal between the groups. The result from the test can be seen in figure 8. In figure 9 we can see the result from the fitted Cox model and in figure 10 we can see the difference on how the Cox model[4] estimate the survival curves for the groups compared to Kaplan-Meier. A broader discussion about the plots can be seen under the Questions section.



```
            N Observed Expected (O-E)^2/E (O-E)^2/V
group=0 500      472      396      14.7      59.7
group=1 100       83      159      36.5      59.7

 Chisq= 59.7  on 1 degrees of freedom, p= 1e-14
```

Figure 8: Logrank test when group 1 and 2 are uniform censored on [20,60].

Smaller censoring interval on group 2

In this part the same kind of output were carried out as when the censoring were in the same interval for the groups. The only difference is that the second group have a censoring that are uniform distributed between [30,60] in this part. Figure 11 show the result from the Logrank test, figure 12 show the result from the fitted Cox model and figure 13 shows the estimated survival curves for the groups regarding the Cox model and Kaplan-Meier.

```
  n= 600, number of events= 555

         coef exp(coef) se(coef)      z Pr(>|z|)
group -1.0000    0.3679   0.1335 -7.491 6.84e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

      exp(coef) exp(-coef) lower .95 upper .95
group    0.3679      2.718    0.2832    0.4779

Concordance= 0.554  (se = 0.01 )
Likelihood ratio test= 69.32  on 1 df,   p=<2e-16
Wald test            = 56.11  on 1 df,   p=7e-14
Score (logrank) test = 59.72  on 1 df,   p=1e-14
```

Figure 9: The fitted Cox regression model when group 1 and 2 are uniform censored on [20,60].

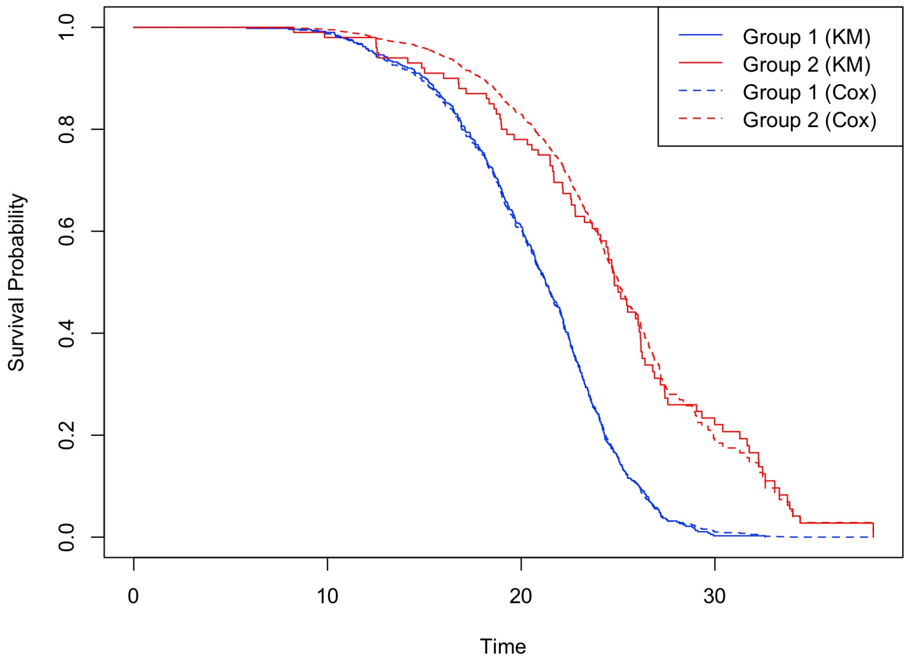
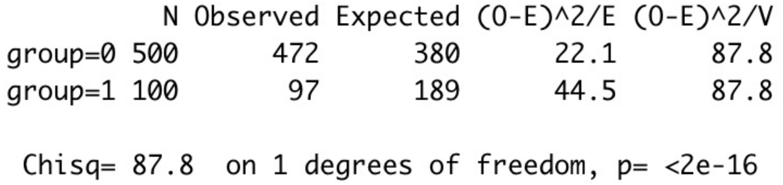


Figure 10: Estimated survival curves for the two groups regarding the fitted Cox model and Kaplan-Meier. Both groups use uniform censoring on [20,60].

Questions

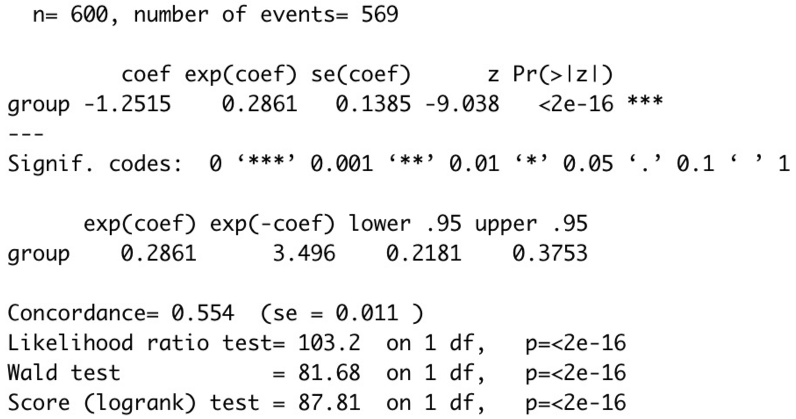
1. *Since you know that data is generated using certain Weibull distributions, you know whether or not the proportional hazards assumption is fulfilled—can you, based on the*



```
           N Observed Expected (O-E)^2/E (O-E)^2/V
group=0 500      472      380      22.1      87.8
group=1 100       97      189      44.5      87.8

 Chisq= 87.8  on 1 degrees of freedom, p= <2e-16
```

Figure 11: Logrank test when group 1 are uniform censored on [20,60] and group 2 are uniform censored on [30,60].



```
  n= 600, number of events= 569 

         coef exp(coef) se(coef)      z Pr(>|z|)    
group -1.2515    0.2861   0.1385 -9.038   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

      exp(coef) exp(-coef) lower .95 upper .95
group    0.2861      3.496    0.2181    0.3753

Concordance= 0.554  (se = 0.011 )
Likelihood ratio test= 103.2  on 1 df,   p=<2e-16
Wald test            = 81.68  on 1 df,   p=<2e-16
Score (logrank) test = 87.81  on 1 df,   p=<2e-16
```

Figure 12: The fitted Cox regression model when group 1 are uniform censored on [20,60] and group 2 are uniform censored on [30,60].

Nelson-Aalen (or Kaplan-Meier) estimates, say something about the assumption regarding proportional hazards?

As the hazard function of a Weibull distribution is proportional, the underlying data follows the Cox-Model assumption of hazard proportionality. We can observe this from the results in assignment 1, where our estimates hazard are monotonic and follows exponential path, meaning that the hazard is continuously proportional towards current timestamp. In applied situations on real data, this pattern is not always true, but data can be non-monotonic and have certain areas of decreasing/increasing hazard, for instance in life insurance where intensity is higher for young and old ages compared to ages in the mid 20s[**?**].

2. *What does the logrank test say?*

   The null hypothesis in Logrank test is that there is no difference in survival for the groups. In figure 8 we can see that the p-value is smaller than 0.05. Therefore we can reject the null hypothesis and conclude that the groups have different survival curves to

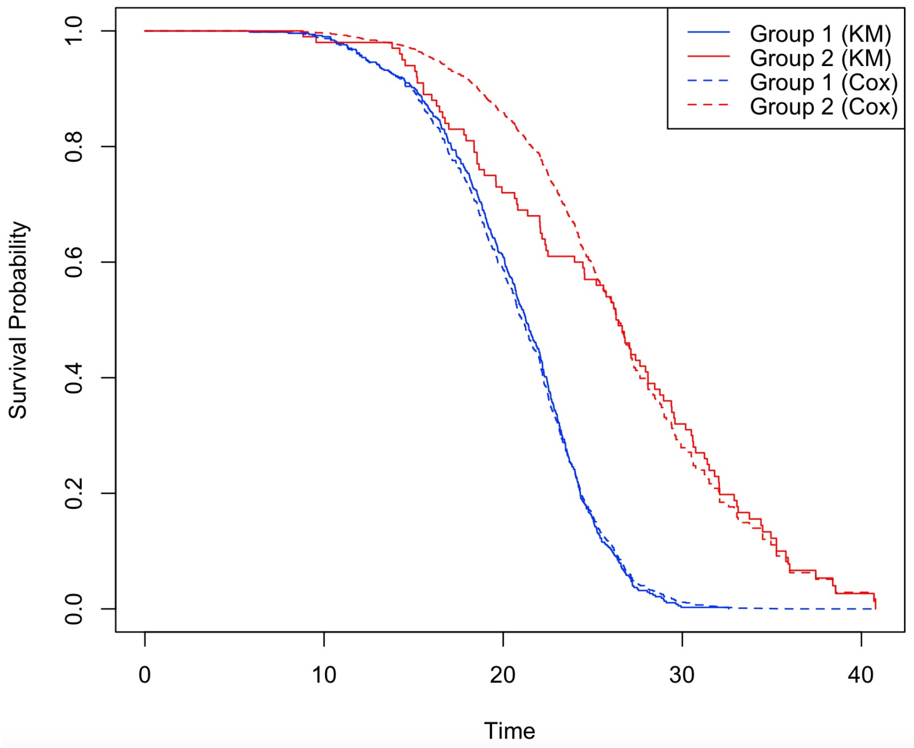


Figure 13: Estimated survival curves for the two groups regarding the fitted Cox model and Kaplan-Meier. Group 1 are uniform censored on [20,60] and group 2 are uniform censored on [30,60].

a 95% confidence level.

3. *Is there a difference between the two groups with respect to estimated regression coefficients?*

   From figure 9 we can see that the regression coefficient is $\beta_2 = -1.000$ for group 2. Since the coefficient is negative, the hazard risk is lower for group 2 compared to group 1. Meaning that individuals in group 2 have a higher probability of living longer. With $e^{\beta_1} = 0.3679$ we can interpret the coefficient as that the hazard rate is reduced by a factor of 0.3679 (or 63%) for group 2. This is also shown significant to a z-score of $z = -7.491$ meaning that we observe a more than 7 times standard deviations if assuming the groups to be equal, meaning it to be a near 0% probability of the coefficient being 0 if we would repeat the experiment.

4. *Comment on the comparison between the Kaplan-Meier curves and the Cox curves.*

In figure 10 we can see the same thing that were observed in figure 9, which is that group 2 have a lower hazard rate. In the plot we can also see that the Cox model follow the Kaplan-Meier estimation better for group 1. For group 2, the hazard-rate is a bit higher between year 15-25 than the Cox model capture. From the figure we can also see that the Kaplan-Meier estimation is more even for group 1, which is most likely because that group consist of more individuals (500 compared to 100). What is also observed is the inherit nature of a parametric and non-parametric approaches, where the Kaplan-Meier follows a clear step function and the cox-model follows a smooth function, as is to be expected through the baseline hazard rate.

5. *What is the effect if you apply censoring times that are uniform on [30, 60] to group 2?*

   If we apply censoring that are uniform on [30, 60] to group 2 we can see that the survival curves still differ between the groups. Which can be seen in figure 11, since the p-value is smaller than 0.05.

   The hazard rate is reduced by a factor of 0.2861 for group 2, which can be seen in figure 12. The hazard rate is now 71% smaller for group 2 compared to group 1. Moreover, the difference between the hazard rate have become larger between the groups compared to when the censoring were uniform on [20, 60] for group 2. Then the hazard rate were 63% smaller for group 2 compared to group 1.

   In figure 13 we can see that the difference between the Cox model and Kaplan-Meier is even bigger between year 15-25 for group 2 than before the new censoring interval. This means that the new censoring interval contributes to that group 2 have less proportional hazard than it had before. We can also see that the survival curves differ more between the groups than before, which we could also see in exp(coef) in figure 12. Likely the difference in censoring increase the concentration of $T \in [20, 30]$ observations which makes the groups almost intersect on these observations. This increased exposure to t is captured by the parametric model and not the cox-model.

6. *Analyze the changes over time and give a short summary of your conclusions based on suitably chosen plots.*

   Over time we can see that group 1 follow the proportional hazard assumption the whole time. Meanwhile group 2 only follow the proportional hazard assumption after around 25 years. Before that, group 2 follow the proportional hazard assumption a bit more when the censoring is uniform on [20, 60] compared to [30, 60]. This can be seen in figure 10 and 13, and how good Kaplan-Meier follow the Cox proportional hazard model. Moreover, since the Cox regression model assume a proportional hazard the Kaplan-Meier estimations will follow these curves when the survival curves follow a proportional hazard.

7. *Give a short summary of your conclusions based on suitably chosen plots.*

In part 2 we can conclude that the survival curves differ between the two groups, which can be seen in the Logrank tests in figure 8 and 11. We can also conclude that group 2 have a lower hazard rate than group 1 and the difference is even bigger when the censoring is uniform on [30, 60] for group 2, this can be seen both under exp(coef) in figure 9 and 12 and in the plots in figure 10 and 13.

The conclusions on this is that while the Cox-model works well because of our underlying data generation process follows its assumption, the non-parametric Kaplan-Meier is more sensitive to distortions in the data. This is expected out of non-parametric methods, and in some situations it is a good property as the world rarely follows parametric assumptions. But in this scenario it falls out the scope and inaccurately estimates the underlying survival function and the cox-model would be preferable if there was a choice to be made.

References

[1] O. O. Aalen. *Survival and Event History Analysis: A Process Point of View.* Springer-Verlag, 2008.

[2] A. Allignol, J. Beyersmann, and M. Schumacher. Mvna: an r package for the nelson–aalen estimator in multistate models. *The Newsletter of the R Project Volume 8/2, October 2008*, 8:48, 2008.

[3] S. P. Millard. *EnvStats: An R package for environmental statistics.* Springer Science & Business Media, 2013.

[4] T. M. Therneau, P. M. Grambsch, T. M. Therneau, and P. M. Grambsch. *The cox model.* Springer, 2000.

[5] T. M. Therneau and T. Lumley. Package ‘survival’. *R Top Doc*, 128(10):28–33, 2015.