Homework 10

Denis Ostroushko

2022-11-21

20.2

20.2 - 1

In this section we derive KM survival estimates for each treatment arm and present them in the table. Each table will have all detail that goes into the KM estimates: subjects at risk, number of events at time point i, instantaneous hazard, and survival probability estimate.

This list describes each column of the final table that, and provides formulas to be used in calculations:

1. Number of subjects at risk is the number of remaining subjects at time i who were not censored or did not have an event, denote this quantity as n_i
2. Number of events will occurring at time i will be used for calculation of hazard, denote as d_i
3. Hazard at time i is $\lambda_i = \frac{d_i}{n_i}$. We need $1 - \lambda_i$ as the instantaneous probability to survive
4. Survival Probability at time i depends on the time period:

- if it is the first available time period, then $S(1) = 1 - \lambda_1$
- Otherwise, $S(i) = (1 - \lambda_i) \times S(i - 1)$

Clearly, in the case of the first time interval recorded in our data at time = 30 we set survival probability equal to the instantaneous probability of surviving. Technically, we can express it as a conditional probability, where survival probability before the start of observation is set to 1. Noone simply experienced the event yet, so everybody "survived" to that point.

At time = 67, survival probability is now truly a conditional probability, expressed as $S_1 \times (1 - \lambda_2)$, which is 0.9545×0.9524

Full Table is given below:

Table 1: Kaplan Meier Survival Estiamtes for 'RoRx+5-Fu' Case Group

Time	At Risk: n_i	Events: d_i	Instantaneous survival Probability: 1 - d_i/n_i	KM Survival Estimate
30	22	1	0.9545455	0.9545455
67	21	1	0.9523810	0.9090909
79	20	0	1.0000000	0.9090909
82	19	0	1.0000000	0.9090909
95	18	1	0.9444444	0.8585859
148	17	1	0.9411765	0.8080808
170	16	1	0.9375000	0.7575758
171	15	1	0.9333333	0.7070707
176	14	1	0.9285714	0.6565657
193	13	1	0.9230769	0.6060606
200	12	1	0.9166667	0.5555556
221	11	1	0.9090909	0.5050505
243	10	1	0.9000000	0.4545455
261	9	1	0.8888889	0.4040404
262	8	1	0.8750000	0.3535354
263	7	1	0.8571429	0.3030303
399	6	1	0.8333333	0.2525253
414	5	1	0.8000000	0.2020202
446	4	1	0.7500000	0.1515152
464	2	1	0.5000000	0.0757576
777	1	1	0.0000000	0.0000000

We also replicate this method to provide a table for controls:

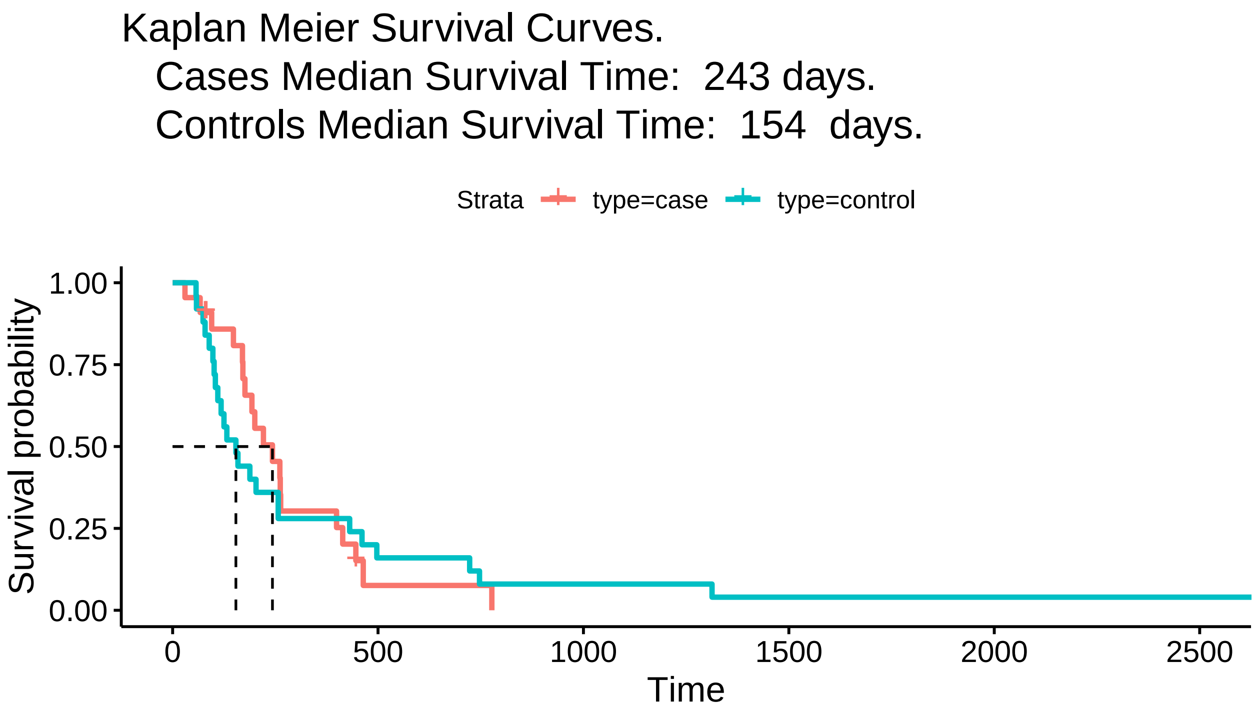
Table 2: Kaplan Meier Survival Estiamtes for Control Group

Time	At Risk: n_i	Events: d_i	Instantaneous survival Probability: 1 - d_i/n_i	KM Survival Estimate
57	25	1	0.9600000	0.96
58	24	1	0.9583333	0.92
74	23	1	0.9565217	0.88
79	22	1	0.9545455	0.84
89	21	1	0.9523810	0.80
98	20	1	0.9500000	0.76
101	19	1	0.9473684	0.72
104	18	1	0.9444444	0.68
110	17	1	0.9411765	0.64
118	16	1	0.9375000	0.60
125	15	1	0.9333333	0.56
132	14	1	0.9285714	0.52
154	13	1	0.9230769	0.48
159	12	1	0.9166667	0.44
188	11	1	0.9090909	0.40
203	10	1	0.9000000	0.36
257	9	2	0.7777778	0.28
431	7	1	0.8571429	0.24
461	6	1	0.8333333	0.20
497	5	1	0.8000000	0.16
723	4	1	0.7500000	0.12
747	3	1	0.6666667	0.08
1313	2	1	0.5000000	0.04
2636	1	1	0.0000000	0.00

20.2 - 2

In this section we put the data from two tables on the same graphs. We also add median survival time for both groups. From slide #5 in Lecture file #21 it appears that the median time is defined as the time point when the value of survival function $S(t)$ drop below 0.5 for the first time. We can't just take the median of observed survival times due to cencoring and incomplete data that it imposes.

Thus, as we can see in the tables, median survival time for cases who are subject to treatment is 243 days, while it is 154 days for controls.



In appears that we have the phenomenon of crossing survival curves. That means that he hazards are not proportional, and if we are to analyze these curves and compare survival times we would need parametric methods that can handle different scale and shape parameters for the two groups.

Judging from the location of the median survival times, cases' survival time should be higher, "on average" by around 89 days.

That means that the drug we are researching may be effective at extending survival time of patients. However, as we noted, at some point survival curves cross. Which means that the drug must have some adverse effects on the subjects after some prolonged period of time.

Alternatively, the drug potency may simply decrease over time and may no longer deter the unfavorable outcomes.

20.2 - 3

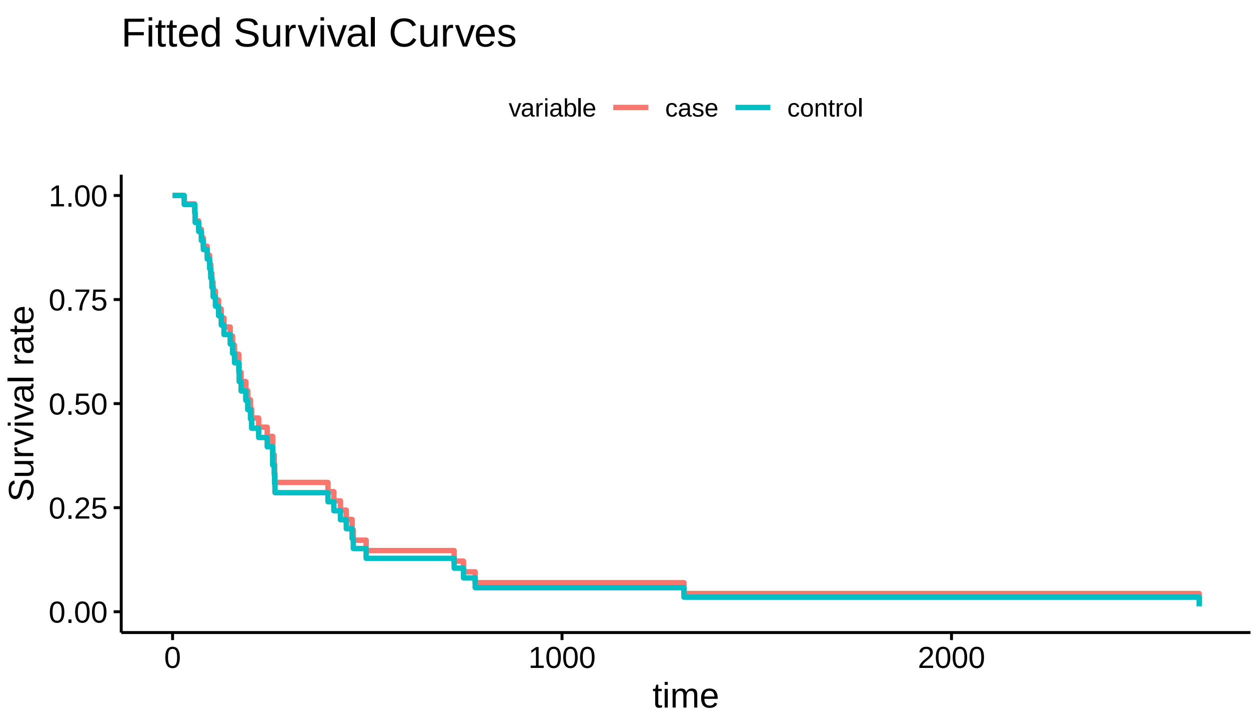
Output of the Cox Proportional Hazard Model is given below:

```
##                   coef exp(coef)  se(coef)         z  Pr(>|z|)
## typecontrol 0.06830808  1.070695 0.3168571 0.2155801 0.8293151
```

Exponentiated coefficient is 1.0707 meaning that the controls' instantaneous hazard at any time is 7.07 % higher at any time, when compared with cases.

However, low Wald Z statistic, high standard error, and low p-value indicate that this difference is not statistically significant.

We can also obtain fitted survival curves from the Cox Proportional Hazards Regression Model. As we can see, the two curves are hard to distinguish.



21.1

21.1 - 1

Using the same methods we used in problem #20, we provide the two tables, one for each group, with KM survival estimates at each point in time.

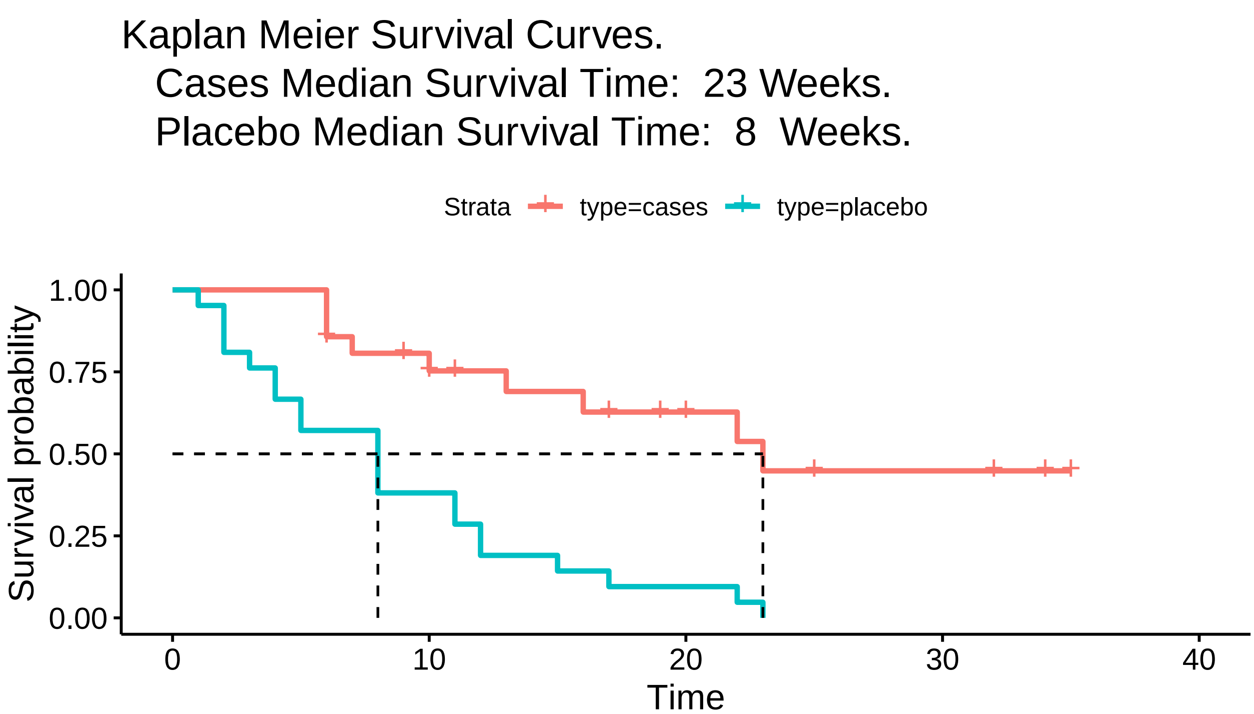
Table 3: Kaplan Meier Survival Estiamtes for Placebo Group

Time	At Risk: n_i	Events: d_i	Instantaneous survival Probability: 1 - d_i/n_i	KM Survival Estimate
1	21	1	0.9523810	0.9523810
2	20	3	0.8500000	0.8095238
3	17	1	0.9411765	0.7619048
4	16	2	0.8750000	0.6666667
5	14	2	0.8571429	0.5714286
8	12	4	0.6666667	0.3809524
11	8	2	0.7500000	0.2857143
12	6	2	0.6666667	0.1904762
15	4	1	0.7500000	0.1428571
17	3	1	0.6666667	0.0952381
22	2	1	0.5000000	0.0476190
23	1	1	0.0000000	0.0000000

Table 4: Kaplan Meier Survival Estiamtes for 6-MP Cases Group

Time	At Risk: n_i	Events: d_i	Instantaneous survival Probability: 1 - d_i/n_i	KM Survival Estimate
6	21	3	0.8571429	0.8571429
7	17	1	0.9411765	0.8067227
9	16	0	1.0000000	0.8067227
10	15	1	0.9333333	0.7529412
11	13	0	1.0000000	0.7529412
13	12	1	0.9166667	0.6901961
16	11	1	0.9090909	0.6274510
17	10	0	1.0000000	0.6274510
19	9	0	1.0000000	0.6274510
20	8	0	1.0000000	0.6274510
22	7	1	0.8571429	0.5378151
23	6	1	0.8333333	0.4481793
25	5	0	1.0000000	0.4481793
32	4	0	1.0000000	0.4481793
34	2	0	1.0000000	0.4481793
35	1	0	1.0000000	0.4481793

We can also visualize the survival curves for the placebo and cases groups



21.1 - 2

Exponential model assumes that the survival probability at time t is given by: $S(t) = exp(-\rho t)$

parameter ρ is estimated to be 0.0251, with a (0.013, 0.0482) 95% confidence interval. I suppose that the confidence interval is not equal width around the estimate because we to obtain the estimate we use the process of exponentiation.

Table below presents estimated survival rates for cases in the 6-MP trial under the assumption that the survival times are distributed exponentially. For comparison, we also keep the Kaplan-Meier estimates in the table. We also present the data only for those times t where we observe at least one event, so those time record where we have a record, but only censored ones, are not in the table. This is specifically requested in the problem statements in the slides.

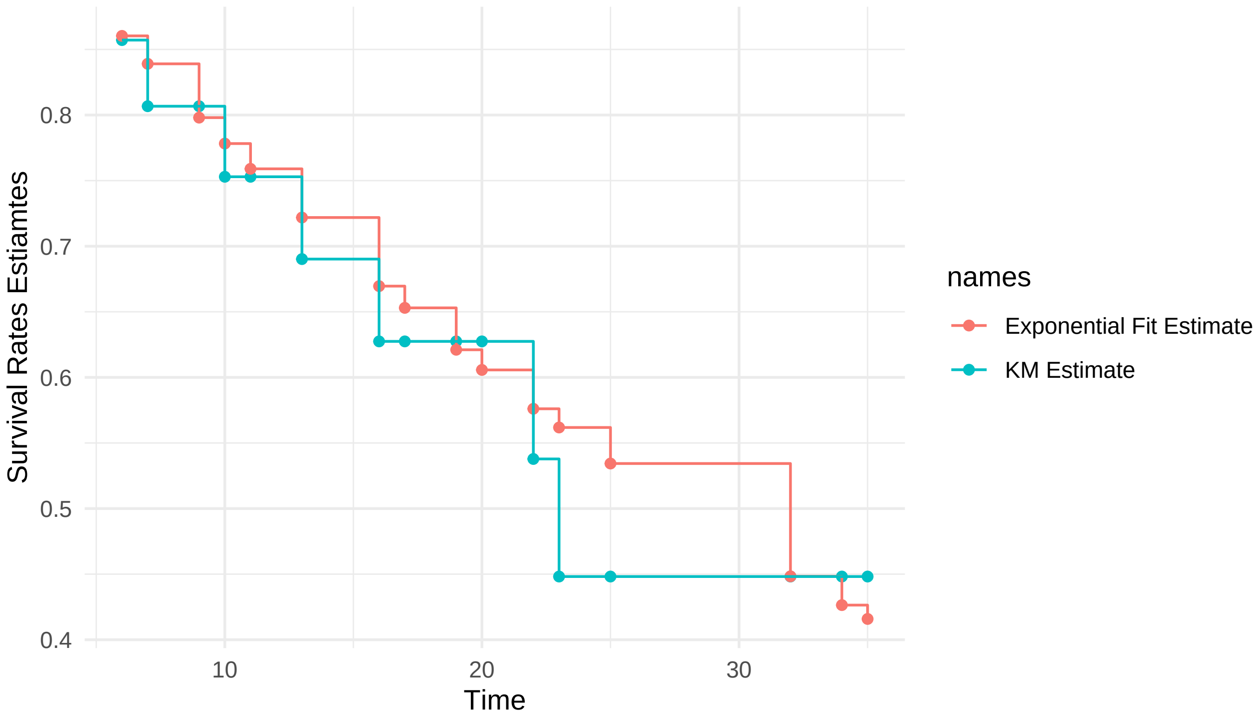
Table 5: KM and Expoenential Model Survival Rates for 6-MP Cases

Time	Number of Events	KM Survival Rate	Exponential Fit Survival Rate
6	3	0.8571429	0.8603484
7	1	0.8067227	0.8390479
10	1	0.7529412	0.7782586
13	1	0.6901961	0.7218736
16	1	0.6274510	0.6695736
22	1	0.5378151	0.5760666
23	1	0.4481793	0.5618043

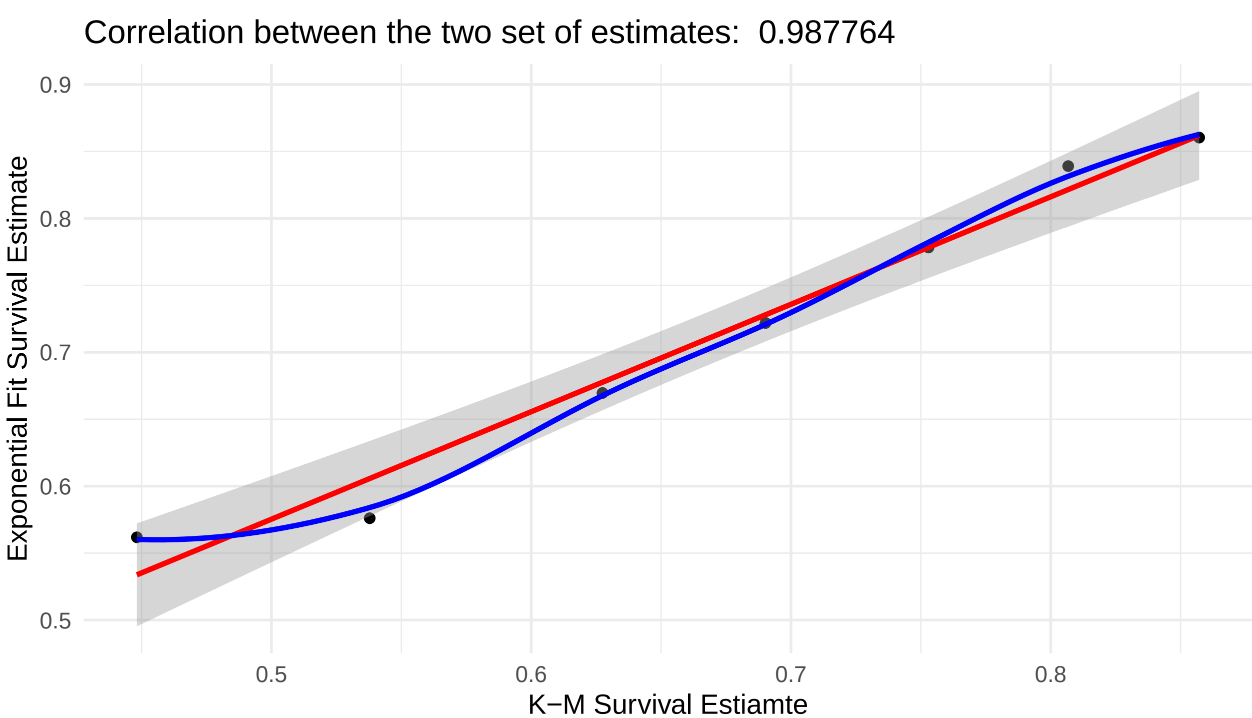
Overall, the two methods produce quite similar estimates, although probabilities from the survival fit appear to be more optimistic. Especially as we get to the higher time records, exponential fit probabilities overstate expected survival chance quite notably.

21.1 - 3

The plot below visualizes the two survival curves on the same plot. We can see that the two curves align quite well up until time $t \approx 22$, where the biggest differences start to occur. The difference occurs because the because KM is fitted to our particular sample, while exponential method is not. Overall, either method might not be well suited for prediction of survial time or extrapolation of any sort.



The plot below visualizes the relationship between the two sets of survival probabilities. As the problem asks in section (2), we remove those time points where no events, and only censoring, occurs. We only have 7 or so data points for this plot, but the two sets of probabilities are almost perfectly correlated. Either can be used for estimation of survival time for this sample up until time point of 25 weeks or so.



21.1 - 4

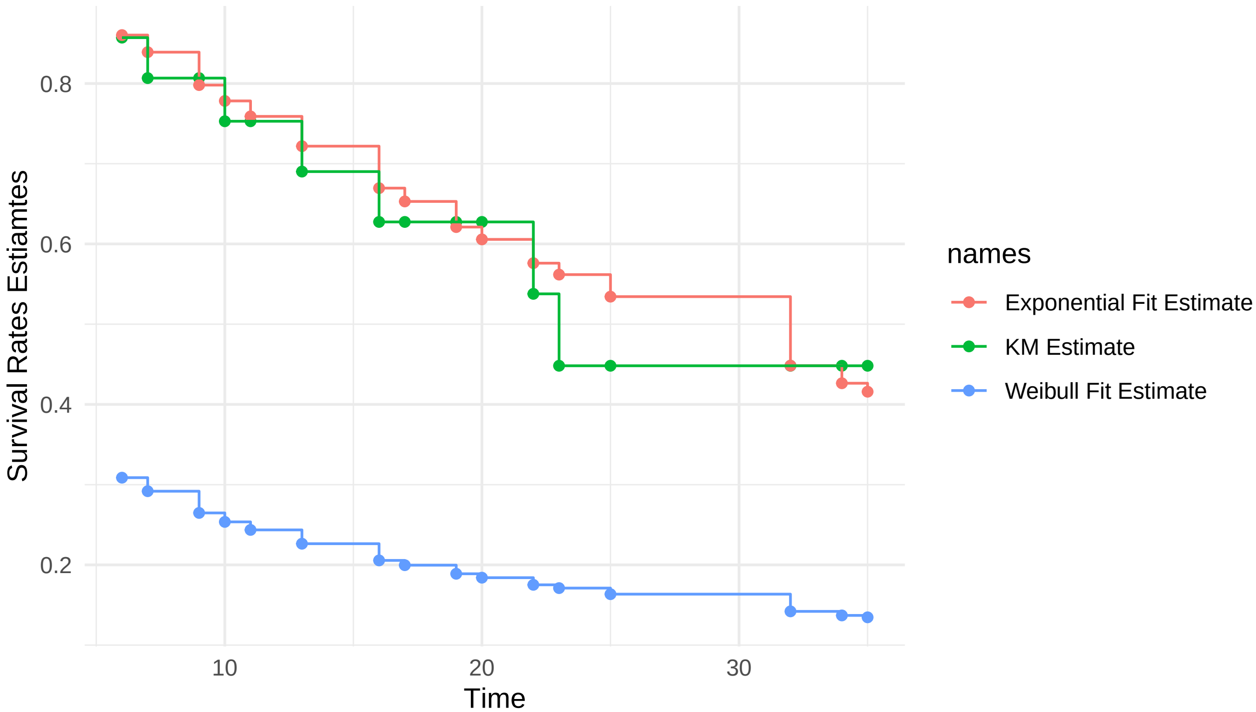
In order to fit the Weibull time-to-event model we will estimate two parameters.

Two parameters: scale - location parameter - $\lambda = 3.5194$

shape - $\rho = 0.3029$

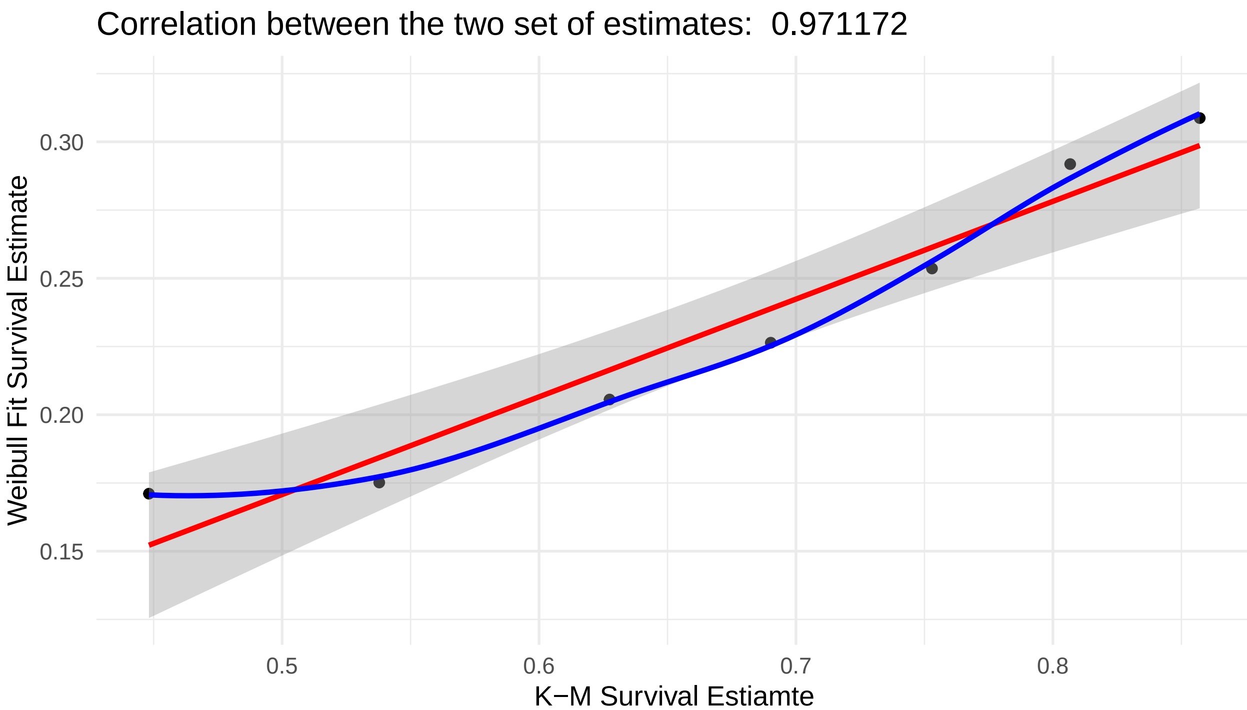
Survival rate $= S(t) = exp(-(t/\lambda)^{\rho})$

We can now use weibull survival model to estimate survival probabilities. As we can see, weibull model is a poor method for our sample.



21.1 - 5

Again, for correlation plot, we only keep those time point where no events and only censoring occurred. The two sets of probabilities are highly correlated, however, Weibull tends to severely underestimate the survival probability at any given point in time.


Correlation between the two set of estimates: 0.971172
0.30
0.25
0.20
0.15
Weibull Fit Survival Estimate
0.5
0.6
0.7
0.8
K−M Survival Estiamte