Nonparametric Approaches

As we discussed earlier, nonparametric duration models do not make assumptions about (a) the distribution of failure times or (b) how the independent variables change survival experiences. The applicability of nonparametric methods to duration data is somewhat limited because nonparametric methods cannot deal adequately with censoring and other issues. However, they can be potentially useful when there are no independent variables or when the independent variables are qualitative in nature.[1]

1 The Kaplan-Meier Estimator – $S(t)$

The Kaplan-Meier (KM) estimator is a nonparametric estimate of the survivor function $S(t)$, which indicates the probability of surviving past time t. If we had a data set with observed failure times, $t_1, \ldots, t_k$, where k is the number of distinct failure times observed in the data, then the KM estimate at any t is given as:

$$\widehat{S(t)} = \prod_{j|t_j \leq t} \left(\frac{n_j - d_j}{n_j} \right) \tag{1}$$

where n_j = the number of cases 'at risk' for the event at time t_j and d_j = is the number of cases which experience the event at time t_j. The product is over all observed failure times less than or equal to t.

So, how exactly does this work? Look at Table 1.

Table 1: Kaplan-Meier Estimator

At time	# at Risk	# Failed	# Censored	p	$\widehat{S(t)}$
2	6	1	0	5/6	5/6
4	5	2	0	3/5	1/2
5	3	0	1	1	1/2
7	2	1	0	1/2	1/4
8	1	0	1	1	1/4

At time $t = 2$, we have 6 observations at risk. At $t = 4$, 5 subjects are at risk because 1 failed, and so on. So, we start by asking what the probability of survival beyond time $t = 2$ is. Since 5 subjects out of six survived beyond this point, $p = 5/6$. What is the probability of surviving beyond $t = 4$ given survival up to $t = 4$? We had 5 subjects at risk at $t = 4$ but two then failed. Thus, the probability $p = 3/5$. The probability of surviving beyond $t = 5$ given survival to $t = 5$ is $3/3 = 1$ since there were 3 at risk in period 5 and all survived (I'll come back to the censored observation in a moment). This is where the p's come from in the table. Thus, we can use these conditional probabilities to calculate unconditional probabilities. For example,

[1] Much of these notes is based on Cleves, Gould, and Gutierrez (2008).

the unconditional probability of surviving beyond $t = 2$ is just 5/6 since $t = 2$ is the first time period. However, the unconditional probability of surviving beyond $t = 4$ is $(5/6)(3/5) = 1/2$. The unconditional probability of surviving beyond $t = 5$ is $(5/6)(3/5)(1) = 1/2$, and so on. These unconditional probabilities are $\widehat{S(t)}$ in the table and represent the Kaplan-Meier estimate of the survivor function. Note that the Kaplan-Meier estimate in Eq. (1) only operates on the observed failure times (not at censored times), the net effect is to ignore cases where $p = 1$ in calculating $\widehat{S(t)}$. Thus, the occurrence of events (deaths) affects the Kaplan-Meier estimate but not the number of censored observations. So, how are censored observations taken into account? Well, when an observation is censored it simply reduces the 'at risk' category i.e. n_j.[2]

We can look at the Kaplan-Meier estimates in tabular form. The NET LOST column is related to our censored observations. The table basically tells us the probability of units surviving past time t.

```
. sts list

         failure _d:  event
   analysis time _t:  duration
```

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf.	Int.]
7	424	1	0	0.9976	0.0024	0.9834	0.9997
9	423	1	0	0.9953	0.0033	0.9813	0.9988
11	422	2	0	0.9906	0.0047	0.9751	0.9964
12	420	2	0	0.9858	0.0057	0.9688	0.9936
14	418	1	0	0.9835	0.0062	0.9657	0.9921
(output omitted)							
604	201	2	0	0.4837	0.0245	0.4349	0.5307
605	199	1	0	0.4813	0.0245	0.4325	0.5283
608	198	0	1	0.4813	0.0245	0.4325	0.5283
611	197	2	0	0.4764	0.0245	0.4277	0.5234
(output omitted)							
1799	5	1	0	0.0108	0.0054	0.0037	0.0259
1822	4	1	0	0.0081	0.0046	0.0023	0.0221
1831	3	1	0	0.0054	0.0038	0.0011	0.0182
1925	2	1	0	0.0027	0.0027	0.0003	0.0143
1936	1	1	0	0.0000	.	.	.

[2]STATA follows the convention that if censoring and failure are recorded as occurring at the same time, then we assume that failure occurs before censoring when it comes to figuring out the 'at risk' category. This convention is used elsewhere as well – failures occur before censoring.

You can also look at the Kaplan-Meier estimates across some qualitative variable such as the minimal winning coalition. In the output below, 0 refers to the survivor function for governments that were not minimal winning coalitions and 1 refers to the survivor function for governments that were minimal winning coalitions.

```
. sts list, by(mwc) compare

         failure _d:  event
   analysis time _t:  duration

                 Survivor Function
mwc                   0          1
---------------------------------
time       7     0.9965     1.0000
         248     0.7133     0.8789
         489     0.5259     0.6999
         730     0.3610     0.5767
         971     0.2448     0.4680
        1212     0.1442     0.3176
        1453     0.0477     0.1629
        1694     0.0040     0.0543
        1935          .     0.0090
        2176          .          .
---------------------------------
```

From before, we know that the cumulative probability of failure, $F(t) = 1 - S(t)$. You can obtain this by typing

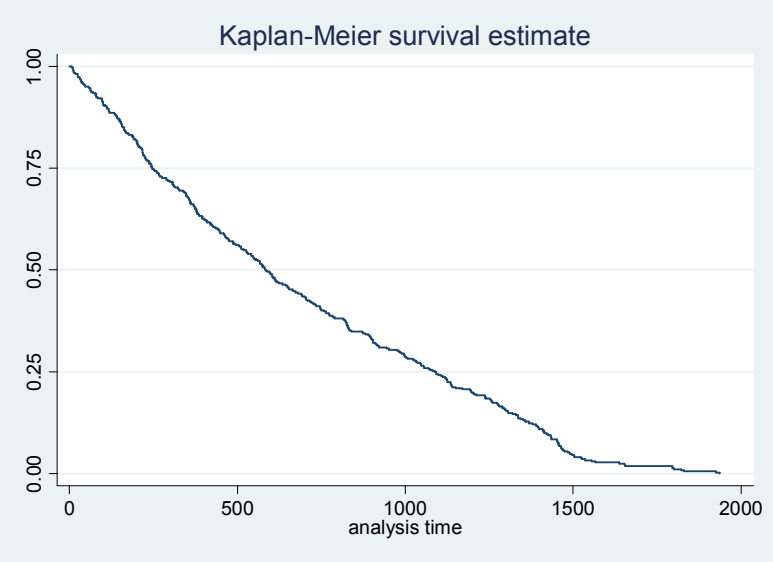
```
 sts list, failure

         failure _d:  event
   analysis time _t:  duration

           Beg.          Net            Failure       Std.
  Time    Total   Fail   Lost           Function      Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
     7      424      1      0             0.0024     0.0024     0.0003    0.0166
     9      423      1      0             0.0047     0.0033     0.0012    0.0187
    11      422      2      0             0.0094     0.0047     0.0036    0.0249
    12      420      2      0             0.0142     0.0057     0.0064    0.0312
    14      418      1      0             0.0165     0.0062     0.0079    0.0343
    17      417      1      0             0.0189     0.0066     0.0095    0.0374
    24      416      3      0             0.0259     0.0077     0.0145    0.0464
```

(output omitted)

Rather than look at the K-M estimate in tabular form, you can look at it graphically by typing:

```
sts graph
```

Figure 1: Kaplan-Meier Plot of Government Survival



So how do we interpret these plots? It is important to remember that the plot does NOT indicate the proportion of observations surviving to time t.[3] Instead, the plot indicates the unconditional probability that an observation will survive beyond time t (no covariates). Naturally, the survivor function starts at 1 since all observations are alive at $t = 0$. However, it declines as units 'fail' over time. The survival probability of an observation lasting beyond time period 500 is about 0.56 and the survival probability of an observation lasting beyond time period 1,000 is about 0.28. Since we know that all observations fail eventually by assumption, the survivor plot will asymptote to 0 if there are no censored observations. In this particular example, all observations are censored at time period 1936 and so the survival function never quite gets to zero.

There are various extensions of the KM plot that you can do. For example, you can get confidence intervals by typing:

```
sts graph, ci
```

You can compare two different types of government duration by typing

```
sts graph, by(mwc) ci
```

[3]If you want this information, you can get it from the sts list command.

You can also add information about when observations are censored by typing:

```
sts graph, by(cabinet_majority)censored(number) ci
```

You can also put a 'risk table' at the bottom of the figure by typing:

```
sts graph, by(cabinet_majority)censored(number) risktable ci
```

2 Cumulative Hazard Function – $H(t)$

Recall that the cumulative hazard function is defined as:

$$H(t) \equiv \Lambda(t) = \int_0^t h(t)dt \tag{2}$$

This is also referred to as the integrated hazard and measures the total amount of risk that has accumulated up to time t. The integrated hazard can also be written as:[4]

$$H(t) = -\ln[S(t)] \tag{4}$$

As you can see, we could use the KM estimator for $S(t)$ and plug this in to get the cumulative hazard. However, there is another estimate for $H(t)$ that has better small sample properties. This estimator is known as the Nelson-Aalen estimator and is calculated as:

$$\hat{H}(t) = \sum_{j|t_j \leq t} \frac{d_j}{n_j} \tag{5}$$

where n_j is the number at risk at time t_j, d_j is the number of failures at time t_j, and the sum is over all distinct failure times less than or equal to t.

To see exactly how this works, look at Table 2.

Table 2: Kaplan-Meier Plots

At time	n_j	d_j	# Censored	e_j	$\hat{H}(t)$
2	6	1	0	0.1667	0.1667
4	5	2	0	0.4000	0.5667
5	3	0	1	0.0000	0.5667
7	2	1	0	0.5000	1.0667
8	1	0	1	0.0000	1.0667

[4] We can also write $S(t)$ in terms of $H(t)$:

$$S(t) = e^{-H(t)} \tag{3}$$

The expected number of failures at each observed time is just the number of failures at each time period divided by the number at risk i.e. $e_j = \frac{d_j}{n_j}$. The cumulative hazard rate is just the sum of these over time.

To get the cumulative hazard rate with confidence intervals in tabular format, you can type

```
. sts list, cumhaz

         failure _d:  event
   analysis time _t:  duration
```

Time	Beg. Total	Fail	Net Lost	Nelson-Aalen Cum. Haz.	Std. Error	[95% Conf.	Int.]
7	424	1	0	0.0024	0.0024	0.0003	0.0167
9	423	1	0	0.0047	0.0033	0.0012	0.0189
11	422	2	0	0.0095	0.0047	0.0036	0.0252
12	420	2	0	0.0142	0.0058	0.0064	0.0317
14	418	1	0	0.0166	0.0063	0.0079	0.0349

(output omitted)

1795	7	2	0	4.1830	0.4116	3.4493	5.0727
1799	5	1	0	4.3830	0.4576	3.5719	5.3782
1822	4	1	0	4.6330	0.5214	3.7159	5.7765
1831	3	1	0	4.9663	0.6189	3.8901	6.3403
1925	2	1	0	5.4663	0.7956	4.1096	7.2709
1936	1	1	0	6.4663	1.2779	4.3898	9.5252

You can also graph the cumulative hazard by typing

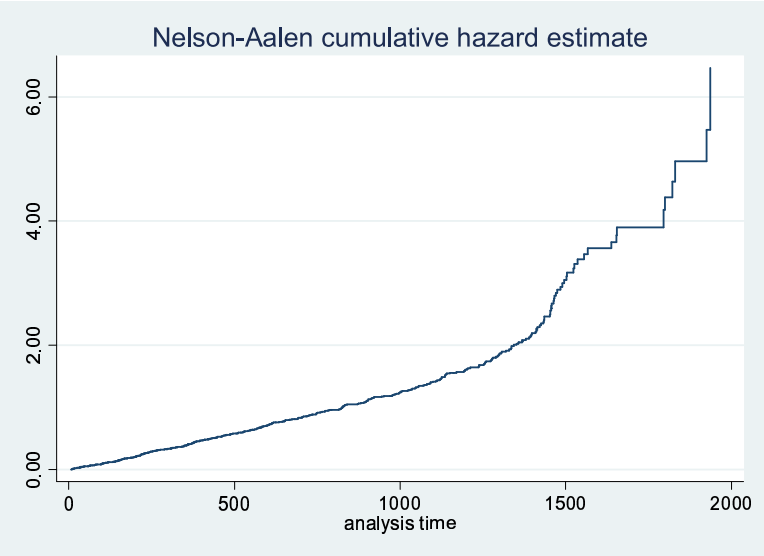
```
sts graph, cumhaz
```

This is shown in Figure 2 on the next page. The cumulative hazard is hard to interpret but can be thought of as the total number of expected failures in $(0, t)$ for a subject, if failure were a repeatable process. Thus, Figure 2 suggests that governments can be expected to fail two times in a period of about 1,300 days if they could fail repeatedly; they can be expected to fail four times in a period of about 1,700 days.

As before, there are various extensions of the NA plot that you can do. For example, you can get confidence intervals by typing:

```
sts graph, cumhaz ci
```

Figure 2: Nelso-Aalen Estimate of Cumulative Hazard of Government Survival



You can also compare two different types of government duration by typing:

```
 sts graph, cumhaz
by(cabinet_majority)
```

Show how KM and NA are related in do-file.

3 Hazard Function – $h(t)$

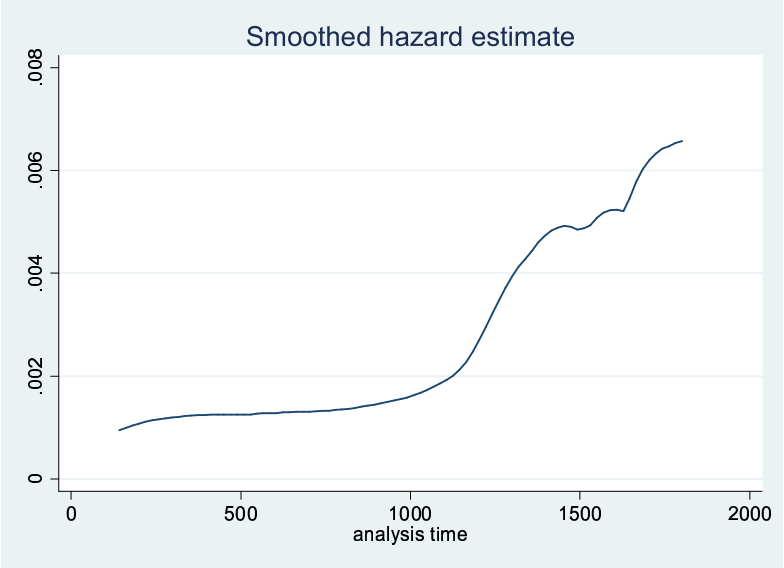
Perhaps a more useful measure would be the hazard rate itself i.e. $h(t)$. This is the probability of having an event at t, given that we have not had an event prior to that point. Put differently, it indicates the failure rate conditional on having survived to time t. This is simply the derivative of the cumulative hazard rate.

You might think that it would be straightforward to get $h(t)$. The problem is that $H(t)$ is a step function and therefore cannot be directly differentiated. What we do to estimate $h(t)$ is to take the steps of the Nelson-Aalen cumulative hazard and smooth them. Without going into the details, we can get the hazard function by typing:

```
sts graph, hazard
```

The hazard function is shown in Figure 3 on the next page. Figure 3 indicates that the probability of failing (conditional on having survived to time t) remains fairly constant below 0.002 for the first 1,100 days. It then climbs rather steeply after this point and continues to climb over time. Figure

Figure 3: Hazard Rate for Coalition Survival



3 indicates that the risk of failing continually increases over time but slowly for the first 1,100 days and then much faster.

As before, there are various extensions of the hazard function plot that you can do. For example, you can get confidence intervals by typing:

```
sts graph, hazard ci
```

You can also compare two different types of government duration by typing:

```
 sts graph, hazard
by(cabinet_majority)
```

You can also change characteristics of the smoothing process too.

4 Median Survival Times

Obtaining mean and median survival times is not obvious in duration data. One reason is that there can be multiple entries for each unit of analysis across time. Another reason is that we have censored observations. Another is that there may be left-truncation. STATA will give you something like a mean or median. For example, STATA will give you the time at which the survival probability is 0.5 if you type:

```
. stci

        failure _d:  1 (meaning all fail)
  analysis time _t:  duration

             |    no. of
             |  subjects         50%     Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------------------
       total |       424         571      34.39676          504         617
```

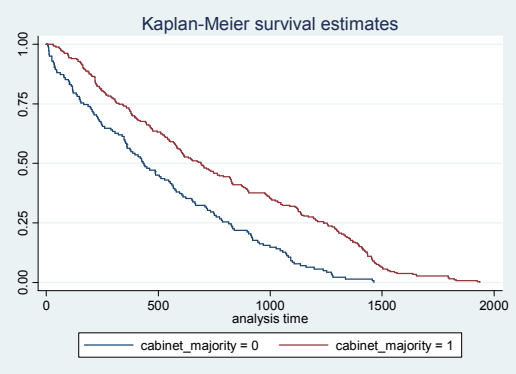
Thus, governments have a 50% chance of surviving beyond 571 days (assuming no covariates). If you look back at the KM plot, from earlier, you will find that 571 days corresponds to the point where the survivor function is equal to 0.5. You can look at the median survival time across different types of government i.e.

```
stci, by(cabinet_majority)
```

5 Log-Rank Test

Above I mentioned that it was possible to compare the survivor functions of different types of government. For example, we could compare the survivor function of those governments that control a legislative majority and those that do not.

Figure 4: Kaplan-Meier Plot of Government Survival



STATA allows you to conduct a formal test of whether two survivor functions are equal across groups. One test is called the log-rank test.[5] Again, you must have a qualitative variable to do this. The null in this test is that the two survivor functions are the same. As the results on the next page indicate, the probability of getting a χ^2 statistic this high (38.58) is tiny ($p = 0.000$). As a result, we reject the null of no difference in the survivor functions. In other words, the survivor function

[5]There are other similar tests that can be done in STATA: Wilcoxon test, Tarone-Ware test and so on. You can look these up to see if they are more appropriate for your problem.

for governments with a majority is different (lower) from the survivor function for governments without a majority.

```
. sts test cabinet_majority, logrank

         failure _d:  1 (meaning all fail)
   analysis time _t:  duration


Log-rank test for equality of survivor functions

             |   Events         Events
cabinet_ma~y |  observed       expected
-------------+-------------------------
0            |       142          91.37
1            |       266         316.63
-------------+-------------------------
Total        |       408         408.00

                  chi2(1) =      38.58
                  Pr>chi2 =     0.0000
```

6 Warning

It is important to note that we should not read too much into these nonparametric plots. This is because virtually all of them are unconditional i.e. there are no covariates. Thus, we should not use the shape of the hazard function, say in Figure 3, to help us decide which type of parametric model to use. We will come back to this point multiple times over the next few days.

References

Cleves, Mario A., William W. Gould & Roberto G. Gutierrez. 2008. *An Introduction to Survival Analysis Using STATA*. Texas: STATA Corporation.