

SC706: Longitudinal Data Analysis
Instructor: Natasha Sarkisian

Event History Analysis: Nonparametric Models

Continuous-time models

We usually use these when time is measured more precisely and when there are more observation points. Since we assume that time is continuous, we cannot calculate hazard rate as the probability that respondent experiences an event at time t given that this respondent is at risk at time t . That's because probability that the event will happen at exactly that time is very tiny for every t . So instead we focus on intervals and examine probability that an individual experiences an event between time t and time $t+s$, given that this individual is in the risk set. Formally, the continuous-time hazard becomes:

$$h(t) = \lim_{s \rightarrow 0} P(t, t+s)/s$$

It is an unobserved rate at which events (failures) occur (number of events per unit of time). Can compare to the notion of risk: if one person's hazard is 1 and another one's is 2, it is appropriate to say that the second person's risk is twice as high.

What we model in continuous-time models is the log of this hazard rate. There are three types of ways to model it: non-parametric, semi-parametric, and parametric. They differ in how we treat the baseline hazard and the independent variables.

Example: dropout.dta

Hypothetical data about time-to-dropout for a sample of 254 college entrants (copied from Yamaguchi 1991, Table 6.6).

Let's look at the data:

```
. tab1 dur- prt
```

```
-> tabulation of dur
```

Time until school dropout (#months since entry)	Freq.	Percent	Cum.
0	1	0.38	0.38
1	2	0.75	1.13
2	10	3.77	4.91
3	8	3.02	7.92
4	7	2.64	10.57
5	3	1.13	11.70
6	3	1.13	12.83
7	3	1.13	13.96
8	8	3.02	16.98
9	10	3.77	20.75

10	2	0.75	21.51
11	2	0.75	22.26
12	1	0.38	22.64
14	1	0.38	23.02
15	3	1.13	24.15
16	1	0.38	24.53
17	1	0.38	24.91
18	1	0.38	25.28
20	4	1.51	26.79
21	10	3.77	30.57
23	1	0.38	30.94
24	1	0.38	31.32
26	1	0.38	31.70
27	3	1.13	32.83
29	2	0.75	33.58
30	2	0.75	34.34
31	1	0.38	34.72
32	4	1.51	36.23
33	6	2.26	38.49
34	1	0.38	38.87
36	2	0.75	39.62
39	3	1.13	40.75
40	14	5.28	46.04
41	113	42.64	88.68
42	10	3.77	92.45
43	1	0.38	92.83
44	15	5.66	98.49
45	3	1.13	99.62
47	1	0.38	100.00

Total	265	100.00	

-> tabulation of evt

1=drop-out, 0=censored	Freq.	Percent	Cum.
0	158	59.62	59.62
1	107	40.38	100.00

Total	265	100.00	

-> tabulation of sex

0=male,1=fe male	Freq.	Percent	Cum.
0	114	43.02	43.02
1	151	56.98	100.00

Total	265	100.00	

-> tabulation of grd

High-school grades (self reported)	Freq.	Percent	Cum.
1	88	33.21	33.21
2	83	31.32	64.53
3	55	20.75	85.28
4	31	11.70	96.98

5	8	3.02	100.00

Total	265	100.00	

-> tabulation of prt

1=part-time student,0=f ull-time	Freq.	Percent	Cum.
0	240	90.57	90.57
1	25	9.43	100.00

Total	265	100.00	

When using these data for continuous-time modeling, we don't need to transform them, but need to inform Stata what are our time and event variables.

```
. stset dur, failure(evt)
```

```
failure event:  evt != 0 & evt < .
obs. time interval:  (0, dur]
exit on or before:  failure
```

```
-----
265 total obs.
1 obs. end on or before enter()
-----
264 obs. remaining, representing
106 failures in single record/single failure data
8086 total analysis time at risk, at risk from t = 0
earliest observed entry t = 0
last observed exit t = 47
```

To check the dataset, we need to examine the output of stset, look at the actual data (use data browser or list command), and use stdes:

```
. stdes
```

```
failure _d:  evt
analysis time _t:  dur
```

Category	total	per subject			
		mean	min	median	max
no. of subjects	264				
no. of records	264	1	1	1	1
(first) entry time		0	0	0	0
(final) exit time		30.62879	1	41	47
subjects with gap	0				
time on gap if gap	0				
time at risk	8086	30.62879	1	41	47
failures	106	.4015152	0	0	1

Nonparametric models

Nonparametric analysis follows the philosophy of letting the dataset speak for itself. It makes no assumptions about the functional form of hazard and does not model the effect of independent variables. Here, we compare survival experience and cumulative hazards at a qualitative level across the values of independent variables.

We can examine the nonparametric estimate of survival function (the Kaplan-Meier estimate) and the nonparametric estimate of cumulative hazard function (the Nelson-Aalen estimate). We can also examine the hazard function itself. We start with Kaplan-Meier:

```
. sts list
```

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf. Int.]	
1	264	2	0	0.9924	0.0053	0.9701	0.9981
2	262	10	0	0.9545	0.0128	0.9213	0.9739
3	252	8	0	0.9242	0.0163	0.8850	0.9504
4	244	7	0	0.8977	0.0186	0.8544	0.9287
5	237	3	0	0.8864	0.0195	0.8415	0.9191
6	234	3	0	0.8750	0.0204	0.8287	0.9095
7	231	3	0	0.8636	0.0211	0.8160	0.8997
8	228	8	0	0.8333	0.0229	0.7826	0.8732
9	220	10	0	0.7955	0.0248	0.7416	0.8393
10	210	2	0	0.7879	0.0252	0.7335	0.8325
11	208	2	0	0.7803	0.0255	0.7254	0.8256
12	206	1	0	0.7765	0.0256	0.7213	0.8221
14	205	1	0	0.7727	0.0258	0.7173	0.8187
15	204	3	0	0.7614	0.0262	0.7052	0.8083
16	201	1	0	0.7576	0.0264	0.7012	0.8048
17	200	1	0	0.7538	0.0265	0.6972	0.8013
18	199	1	0	0.7500	0.0267	0.6932	0.7979
20	198	4	0	0.7348	0.0272	0.6772	0.7839
21	194	10	0	0.6970	0.0283	0.6376	0.7485
23	184	1	0	0.6932	0.0284	0.6337	0.7450
24	183	1	0	0.6894	0.0285	0.6298	0.7414
26	182	1	0	0.6856	0.0286	0.6258	0.7379
27	181	3	0	0.6742	0.0288	0.6141	0.7271
29	178	1	1	0.6705	0.0289	0.6102	0.7236
30	176	2	0	0.6628	0.0291	0.6023	0.7163
31	174	1	0	0.6590	0.0292	0.5984	0.7127
32	173	3	1	0.6476	0.0294	0.5867	0.7019
33	169	6	0	0.6246	0.0298	0.5632	0.6799
34	163	0	1	0.6246	0.0298	0.5632	0.6799
36	162	1	1	0.6208	0.0299	0.5592	0.6762
39	160	2	1	0.6130	0.0300	0.5513	0.6688
40	157	1	13	0.6091	0.0301	0.5473	0.6651
41	143	1	112	0.6048	0.0302	0.5430	0.6610
42	30	0	10	0.6048	0.0302	0.5430	0.6610
43	20	0	1	0.6048	0.0302	0.5430	0.6610
44	19	1	14	0.5730	0.0422	0.4860	0.6506
45	4	1	2	0.4297	0.1280	0.1852	0.6550
47	1	0	1	0.4297	0.1280	0.1852	0.6550

We can also use “by” option to compare the functions for different levels of independent variable:

```
. sts list, by(prt)
      failure _d: evt
      analysis time _t: dur
```

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf. Int.]	

prt=0							
1	239	2	0	0.9916	0.0059	0.9670	0.9979
2	237	4	0	0.9749	0.0101	0.9450	0.9886
3	233	6	0	0.9498	0.0141	0.9133	0.9712
4	227	5	0	0.9289	0.0166	0.8881	0.9552
5	222	2	0	0.9205	0.0175	0.8782	0.9485
6	220	3	0	0.9079	0.0187	0.8636	0.9384
7	217	2	0	0.8996	0.0194	0.8539	0.9315
8	215	7	0	0.8703	0.0217	0.8207	0.9069
9	208	8	0	0.8368	0.0239	0.7835	0.8780
10	200	2	0	0.8285	0.0244	0.7744	0.8707
11	198	2	0	0.8201	0.0248	0.7652	0.8633
14	196	1	0	0.8159	0.0251	0.7607	0.8596
15	195	3	0	0.8033	0.0257	0.7471	0.8484
16	192	1	0	0.7992	0.0259	0.7425	0.8446
17	191	1	0	0.7950	0.0261	0.7380	0.8409
18	190	1	0	0.7908	0.0263	0.7335	0.8371
20	189	3	0	0.7782	0.0269	0.7201	0.8258
21	186	10	0	0.7364	0.0285	0.6757	0.7876
23	176	1	0	0.7322	0.0286	0.6713	0.7837
24	175	1	0	0.7280	0.0288	0.6669	0.7798
26	174	1	0	0.7238	0.0289	0.6625	0.7760
27	173	3	0	0.7113	0.0293	0.6493	0.7643
29	170	1	1	0.7071	0.0294	0.6450	0.7604
30	168	2	0	0.6987	0.0297	0.6362	0.7526
31	166	1	0	0.6945	0.0298	0.6318	0.7487
32	165	2	1	0.6861	0.0300	0.6231	0.7408
33	162	6	0	0.6607	0.0307	0.5968	0.7169
34	156	0	1	0.6607	0.0307	0.5968	0.7169
39	155	2	1	0.6521	0.0308	0.5880	0.7088
40	152	1	13	0.6478	0.0309	0.5836	0.7048
41	138	1	109	0.6431	0.0311	0.5787	0.7004
42	28	0	8	0.6431	0.0311	0.5787	0.7004
43	20	0	1	0.6431	0.0311	0.5787	0.7004
44	19	1	14	0.6093	0.0442	0.5168	0.6894
45	4	1	2	0.4570	0.1360	0.1921	0.6895
47	1	0	1	0.4570	0.1360	0.1921	0.6895
prt=1							
2	25	6	0	0.7600	0.0854	0.5420	0.8843
3	19	2	0	0.6800	0.0933	0.4609	0.8253
4	17	2	0	0.6000	0.0980	0.3845	0.7611
5	15	1	0	0.5600	0.0993	0.3479	0.7273
7	14	1	0	0.5200	0.0999	0.3125	0.6924
8	13	1	0	0.4800	0.0999	0.2781	0.6564
9	12	2	0	0.4000	0.0980	0.2128	0.5812
12	10	1	0	0.3600	0.0960	0.1819	0.5420
20	9	1	0	0.3200	0.0933	0.1524	0.5015
32	8	1	0	0.2800	0.0898	0.1242	0.4598
36	7	1	1	0.2400	0.0854	0.0976	0.4167
41	5	0	3	0.2400	0.0854	0.0976	0.4167
42	2	0	2	0.2400	0.0854	0.0976	0.4167

Or we can list side by side for every fifth time point:

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

```
      failure _d:  evt
analysis time _t:  dur
```

prt	Survivor Function		
	0	1	
time			
	1	0.9916	1.0000
	6	0.9079	0.5600
	11	0.8201	0.4000
	16	0.7992	0.3600
	21	0.7364	0.3200
	26	0.7238	0.3200
	31	0.6945	0.3200
	36	0.6607	0.2400
	41	0.6431	0.2400
	46	0.4570	.
	51	.	.

Or you can select points

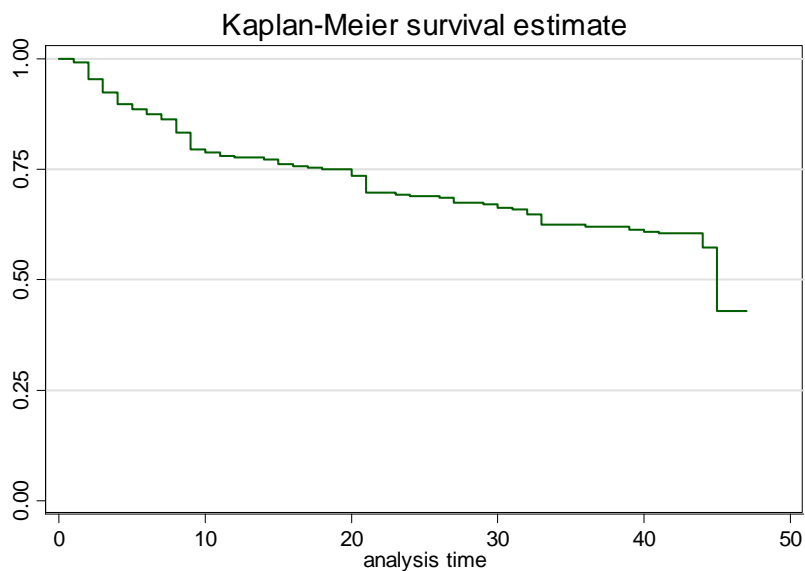
```
. sts list, by(prt) compare at(1 10 20 30 40 50)
```

```
      failure _d:  evt
analysis time _t:  dur
```

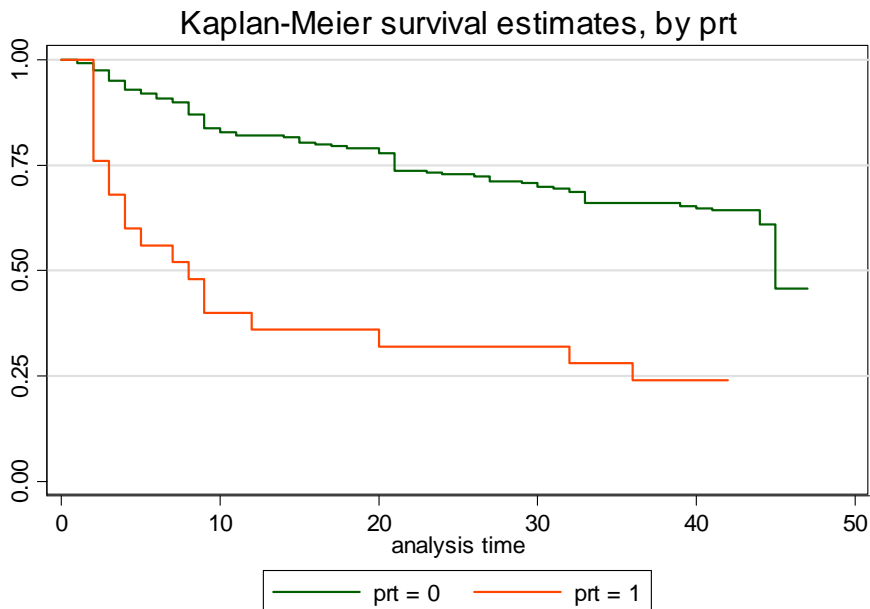
prt	Survivor Function		
	0	1	
time			
	1	0.9916	1.0000
	10	0.8285	0.4000
	20	0.7782	0.3200
	30	0.6987	0.3200
	40	0.6478	0.2400
	50	.	.

It is more helpful, however, to graph the function:

```
. sts graph
```

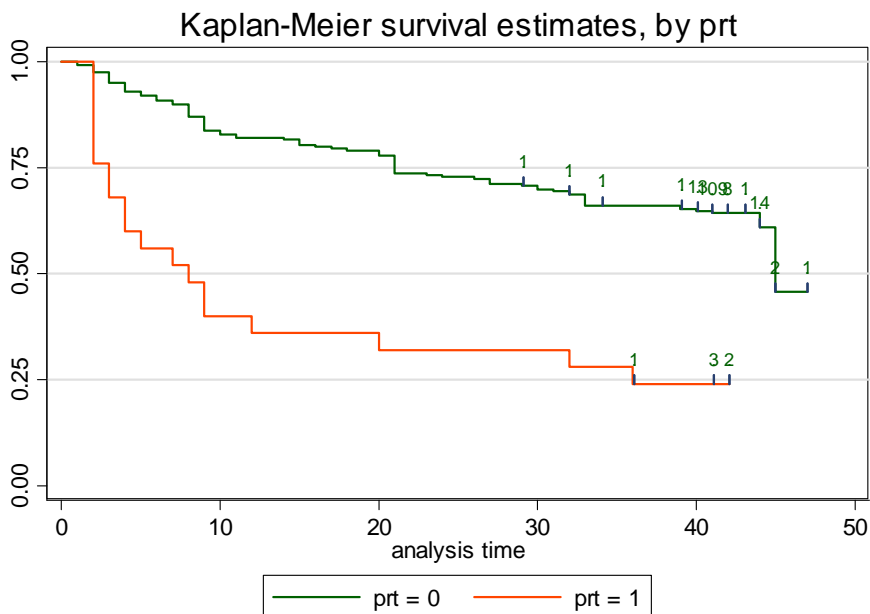


```
. sts graph, by(prt)
```



On this graph, we can also indicate how many observations were censored:

```
. sts graph, by(prt) censored(number)
```

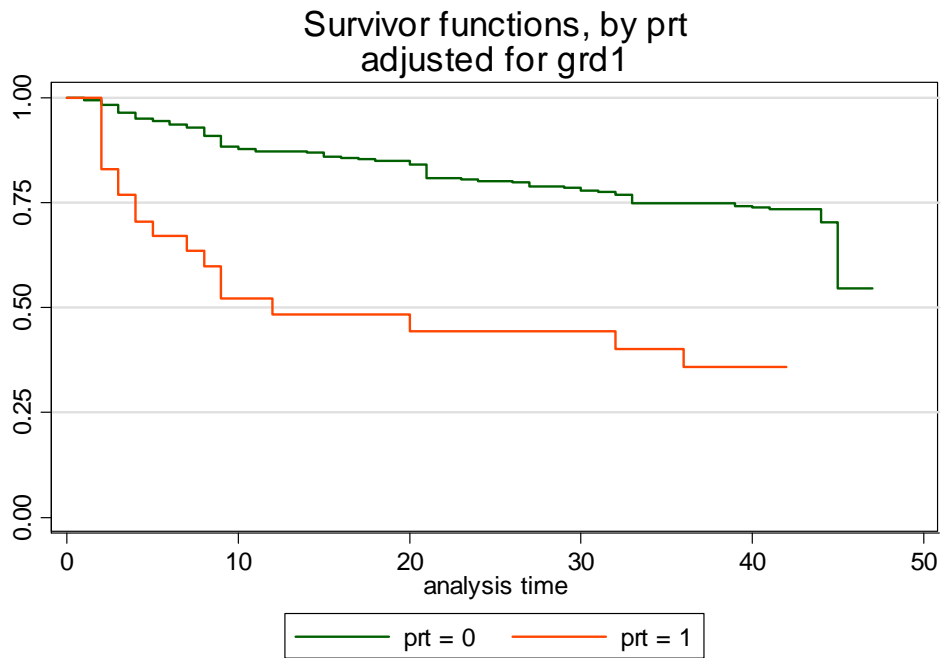


Finally, using either list or graph, we can also generate survival functions adjusted for some other independent variables. It adjusts by setting the value of that variable to zero – so zero has to be a realistic value, and if it's not, we need to generate a new variable where it is. E.g. to adjust for grade:

```
. gen grd1=grd-1
```

```
. sts graph, by(prt) adjustfor(grd1)
```

```
failure _d: evt
analysis time _t: dur
```



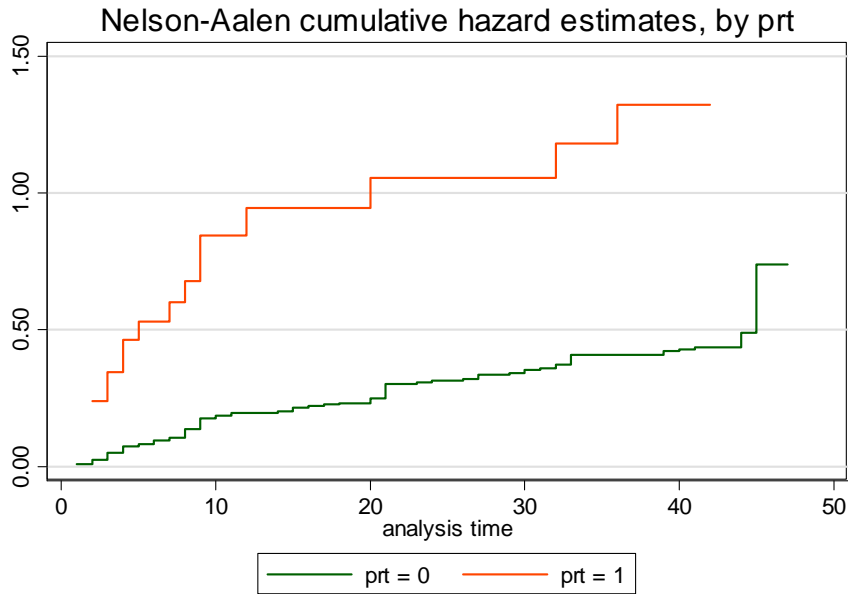
Nelson-Aalen estimate does the same thing the Kaplan-Meier does, but it deals with cumulative hazard. We just have to specify na option. E.g.:

```
. sts list, na by(prt) compare
```

```
failure _d: evt
analysis time _t: dur
```

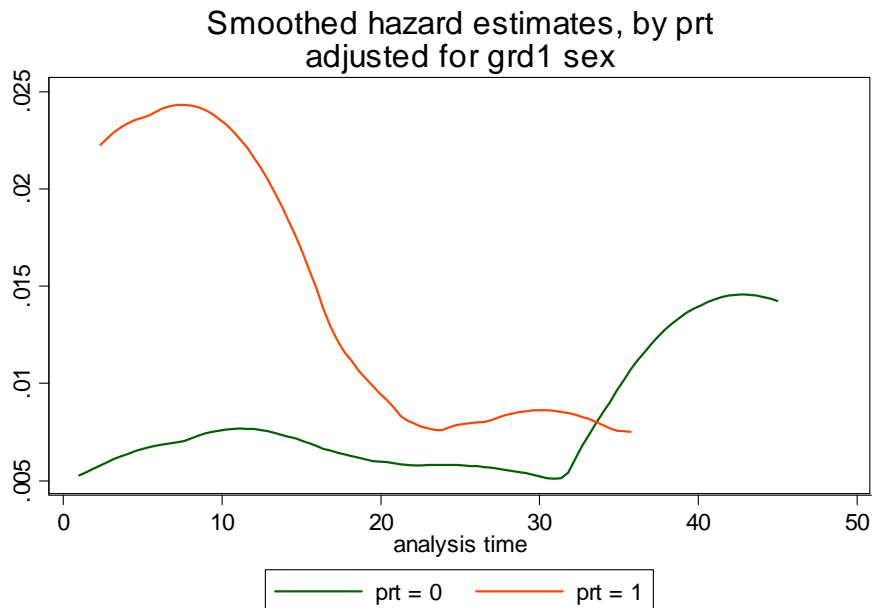
prt	Nelson-Aalen Cum. Haz.	
	0	1
time		
1	0.0084	0.0000
6	0.0957	0.5296
11	0.1960	0.8446
16	0.2217	0.9446
21	0.3018	1.0557
26	0.3190	1.0557
31	0.3601	1.0557
36	0.4093	1.3236
41	0.4360	1.3236
46	0.7387	.
51	.	.

```
. sts graph, by(prt) na
```



Note that adjustfor doesn't work for cumulative hazard. But, we can also obtain the hazard function, and that one we can adjust. Here, we adjust it for grade and sex:

```
. sts graph, hazard by(prt) adjustfor(grd1 sex)
```



These kinds of analyses are very helpful as exploratory analyses before you start constructing semi-parametric or parametric models.

It might also be useful to conduct tests of equality of survivor functions across groups to assess whether or not to subsequently include the predictor in our model. We can only conduct such a test for categorical variables; for continuous variables, we would have to use a bivariate Cox proportional hazard regression model which is a semi-parametric model, or we can use the trend option.

```
. sts test prt
      failure _d: evt
      analysis time _t: dur

Log-rank test for equality of survivor functions
```

prt	Events observed	Events expected
0	87	100.06
1	19	5.94
Total	106	106.00

chi2(1) = 31.38
Pr>chi2 = 0.0000

The default is logrank; it gives equal weight to all observations, but in effect the observations with later failure times might have undue influence because there are fewer cases remaining in the sample, so we can get small p-value even when the two survival curves appear to be very close. Another type of test that gives heavier weights to earlier failure times is Wilcoxon test:

```
. sts test prt, wilcoxon
      failure _d: evt
      analysis time _t: dur

Wilcoxon (Breslow) test for equality of survivor functions
```

prt	Events observed	Events expected	Sum of ranks
0	87	100.06	-3113
1	19	5.94	3113
Total	106	106.00	0

chi2(1) = 35.27
Pr>chi2 = 0.0000

If the chi-squared value associated with the test is sufficiently large and the associated p-value sufficiently small, then we reject the null hypothesis of no group differences in survivor functions. Note that in general, these tests (especially log-rank) place more emphasis on differences in the curves at larger time values.

You can also use trend option to evaluate if there is a linear trend among categories if they are ordered:

```
. sts test grd, trend
```

```
failure _d: evt
analysis time _t: dur
```

Log-rank test for equality of survivor functions

grd	Events observed	Events expected
1	27	38.51
2	32	34.24
3	23	20.94
4	17	10.45
5	7	1.85
Total	106	106.00

chi2(4) = 22.87
Pr>chi2 = 0.0001

Test for trend of survivor functions

chi2(1) = 15.60
Pr>chi2 = 0.0001

Here, the null hypothesis of no difference (compared to the alternative hypothesis of linear trend) is rejected.