Professor Brad Jones
Department of Political Science
University of Arizona
Software Tutorial: Companion to Lecture Notes

1. Illustrating the Kaplan-Meier Estimate with Stata and R

Stata:

Stata's "sts list" invokes the KM estimator (results are also in lecture notes):

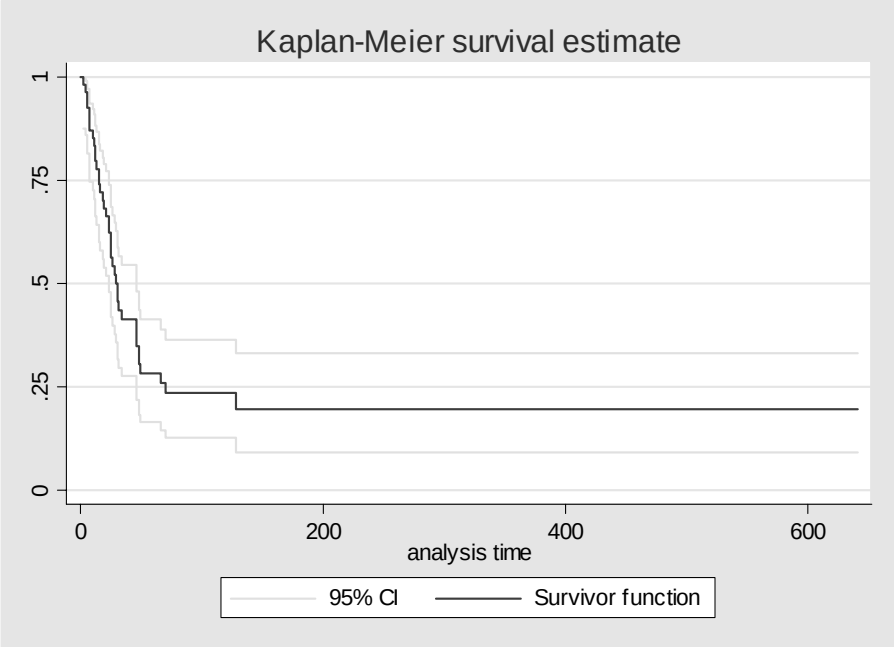
```
. sts list

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

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf.	Int.]
2	54	1	0	0.9815	0.0183	0.8757	0.9974
4	53	1	0	0.9630	0.0257	0.8599	0.9906
5	52	2	0	0.9259	0.0356	0.8146	0.9715
7	50	3	0	0.8704	0.0457	0.7472	0.9360
10	47	1	0	0.8519	0.0483	0.7255	0.9230
11	46	1	0	0.8333	0.0507	0.7042	0.9096
12	45	2	0	0.7963	0.0548	0.6624	0.8816
13	43	1	1	0.7778	0.0566	0.6420	0.8672
15	41	2	0	0.7398	0.0598	0.6005	0.8369
16	39	1	1	0.7209	0.0612	0.5802	0.8214
18	37	1	0	0.7014	0.0626	0.5594	0.8053
19	36	1	0	0.6819	0.0638	0.5389	0.7889
21	35	1	0	0.6624	0.0649	0.5186	0.7723
23	34	2	1	0.6235	0.0667	0.4789	0.7385
25	31	3	1	0.5631	0.0687	0.4185	0.6848
26	27	1	0	0.5423	0.0693	0.3980	0.6660
28	26	1	0	0.5214	0.0697	0.3777	0.6469
29	25	1	1	0.5005	0.0699	0.3577	0.6276
30	23	2	0	0.4570	0.0703	0.3164	0.5870
31	21	1	0	0.4353	0.0702	0.2962	0.5663
34	20	1	0	0.4135	0.0700	0.2764	0.5453
46	19	3	0	0.3482	0.0684	0.2189	0.4807
48	16	2	0	0.3047	0.0664	0.1823	0.4362
49	14	1	0	0.2829	0.0651	0.1645	0.4134
59	13	0	1	0.2829	0.0651	0.1645	0.4134
66	12	1	0	0.2593	0.0638	0.1453	0.3890
70	11	1	0	0.2358	0.0622	0.1266	0.3642
71	10	0	1	0.2358	0.0622	0.1266	0.3642
99	9	0	1	0.2358	0.0622	0.1266	0.3642
127	8	0	2	0.2358	0.0622	0.1266	0.3642
128	6	1	0	0.1965	0.0630	0.0912	0.3310
284	5	0	1	0.1965	0.0630	0.0912	0.3310
319	4	0	1	0.1965	0.0630	0.0912	0.3310
452	3	0	1	0.1965	0.0630	0.0912	0.3310
634	2	0	1	0.1965	0.0630	0.0912	0.3310

```
   641        1      0      1              0.1965    0.0630     0.0912     0.3310
```

This is the same output as in the lecture notes. We can graph the Kaplan-Meier function along with the Greenwood standard errors (which are based on the log-log transformation):

```
. sts graph, gw
```



This graph will look a little different from your lecture notes as it was generated in Stata 8 and in the notes, Stata 7.

R

Because R is an "object-oriented" program, we have to define "survival objects" in order to perform survival analysis. Below is the syntax I used to get my data set into R and produce Kaplan-Meier estimates. Obviously, the location of the data will vary depending upon which subdirectory you use on your own computer. The "library(survival)" command invokes R's survival analysis options.

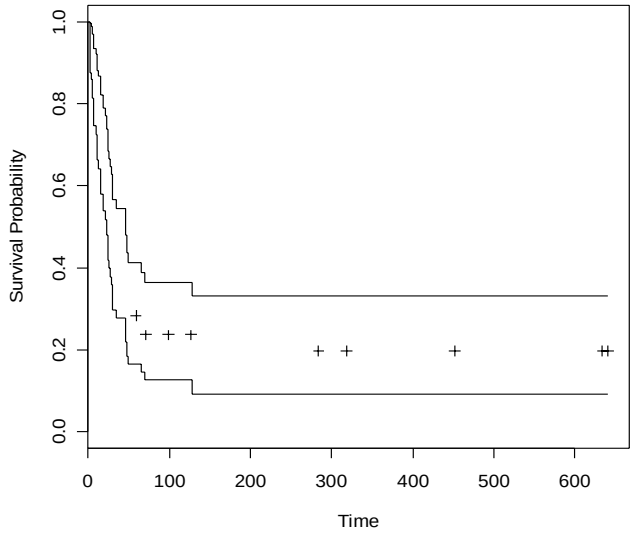
```
> library(survival)

>  unR <- read.dta("c:\\data\\statadata\\unfinal2.dta")

>  unR.surv <- survfit(Surv(duration, failed), unR, conf.type="log-log")

> summary(unR.surv)

Call: survfit(formula = Surv(duration, failed), data = unR, conf.type = "log-
log")

4 observations deleted due to missing
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2     54       1    0.981  0.0183       0.8757        0.997
```

```
  4     53       1    0.963  0.0257       0.8599        0.991
  5     52       2    0.926  0.0356       0.8146        0.972
  7     50       3    0.870  0.0457       0.7472        0.936
 10     47       1    0.852  0.0483       0.7255        0.923
 11     46       1    0.833  0.0507       0.7042        0.910
 12     45       2    0.796  0.0548       0.6624        0.882
 13     43       1    0.778  0.0566       0.6420        0.867
 15     41       2    0.740  0.0598       0.6005        0.837
 16     39       1    0.721  0.0612       0.5802        0.821
 18     37       1    0.701  0.0626       0.5594        0.805
 19     36       1    0.682  0.0638       0.5389        0.789
 21     35       1    0.662  0.0649       0.5186        0.772
 23     34       2    0.623  0.0667       0.4789        0.738
 25     31       3    0.563  0.0687       0.4185        0.685
 26     27       1    0.542  0.0693       0.3980        0.666
 28     26       1    0.521  0.0697       0.3777        0.647
 29     25       1    0.501  0.0699       0.3577        0.628
 30     23       2    0.457  0.0703       0.3164        0.587
 31     21       1    0.435  0.0702       0.2962        0.566
 34     20       1    0.413  0.0700       0.2764        0.545
 46     19       3    0.348  0.0684       0.2189        0.481
 48     16       2    0.305  0.0664       0.1823        0.436
 49     14       1    0.283  0.0651       0.1645        0.413
 66     12       1    0.259  0.0638       0.1453        0.389
 70     11       1    0.236  0.0622       0.1266        0.364
128      6       1    0.196  0.0630       0.0912        0.331
```

The output above, including the log-log Greenwood s.e. are identical (as it should be!) to the Stata output. The only difference is the last four observations, which are censored, are not reported in the R output. This is of no issue as these data points do not contribute any unique information to the function.

Graphically, I can illustrate the KM estimates:

```
> plot(unR.surv, lty = 3, xlab="Time", ylab="Survival Probability")
```



This graph and the Stata version are identical, again, as they should be!

2. Illustrating the Nelson-Aalen Estimate with Stata

Stata:

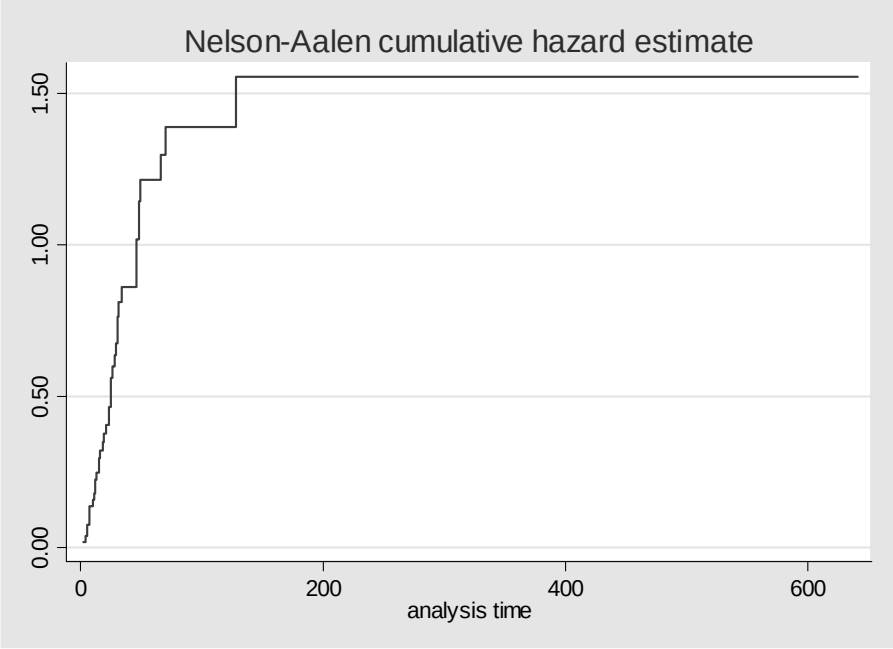
Stata's "sts list" command with the "na" option will give you the Nelson-Aalen estimator of the cumulative hazard. I illustrate it below:

```
.   sts list, na

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

Time	Beg. Total	Fail	Net Lost	Nelson-Aalen Cum. Haz.	Std. Error	[95% Conf.	Int.]
2	54	1	0	0.0185	0.0185	0.0026	0.1315
4	53	1	0	0.0374	0.0264	0.0093	0.1495
5	52	2	0	0.0758	0.0379	0.0285	0.2021
7	50	3	0	0.1358	0.0514	0.0647	0.2850
10	47	1	0	0.1571	0.0556	0.0785	0.3144
11	46	1	0	0.1789	0.0597	0.0930	0.3441
12	45	2	0	0.2233	0.0675	0.1235	0.4037
13	43	1	1	0.2466	0.0714	0.1398	0.4348
15	41	2	0	0.2953	0.0793	0.1745	0.4998
16	39	1	1	0.3210	0.0833	0.1930	0.5338
18	37	1	0	0.3480	0.0876	0.2125	0.5699
19	36	1	0	0.3758	0.0919	0.2327	0.6068
21	35	1	0	0.4044	0.0962	0.2536	0.6446
23	34	2	1	0.4632	0.1048	0.2972	0.7218
25	31	3	1	0.5600	0.1188	0.3695	0.8486
26	27	1	0	0.5970	0.1244	0.3968	0.8982
28	26	1	0	0.6355	0.1302	0.4252	0.9496
29	25	1	1	0.6755	0.1362	0.4549	1.0030
30	23	2	0	0.7624	0.1495	0.5192	1.1196
31	21	1	0	0.8100	0.1569	0.5542	1.1840
34	20	1	0	0.8600	0.1646	0.5910	1.2516
46	19	3	0	1.0179	0.1882	0.7085	1.4625
48	16	2	0	1.1429	0.2079	0.8001	1.6326
49	14	1	0	1.2144	0.2198	0.8516	1.7316
59	13	0	1	1.2144	0.2198	0.8516	1.7316
66	12	1	0	1.2977	0.2351	0.9098	1.8509
70	11	1	0	1.3886	0.2521	0.9729	1.9820
71	10	0	1	1.3886	0.2521	0.9729	1.9820
99	9	0	1	1.3886	0.2521	0.9729	1.9820
127	8	0	2	1.3886	0.2521	0.9729	1.9820
128	6	1	0	1.5553	0.3022	1.0627	2.2761
284	5	0	1	1.5553	0.3022	1.0627	2.2761
319	4	0	1	1.5553	0.3022	1.0627	2.2761
452	3	0	1	1.5553	0.3022	1.0627	2.2761
634	2	0	1	1.5553	0.3022	1.0627	2.2761
641	1	0	1	1.5553	0.3022	1.0627	2.2761

As with KM estimates, it is often useful to graph Nelson-Aalen estimates. I do this in the following way:

```
. sts graph, na
```



The Nelson-Aalen estimator is useful, particularly when it comes to diagnostic methods in the Cox model.

---

3. Illustrating stratified Kaplan-Meier Estimates

STATA

Stata's "sts list" command along with a "by" option can be used to generate stratified KM estimates:

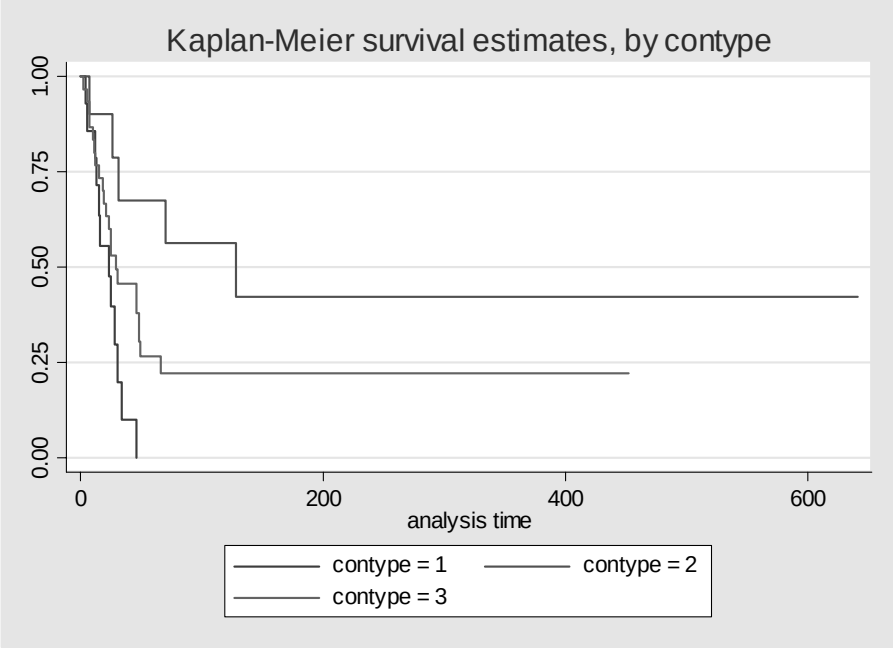
```
. sts list, by(contype)

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

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf.	Int.]
contype=1							
4	14	1	0	0.9286	0.0688	0.5908	0.9896
5	13	1	0	0.8571	0.0935	0.5394	0.9622
12	12	1	0	0.7857	0.1097	0.4725	0.9254
13	11	1	1	0.7143	0.1207	0.4063	0.8819
15	9	1	0	0.6349	0.1308	0.3312	0.8297
16	8	1	0	0.5556	0.1364	0.2636	0.7717
23	7	1	0	0.4762	0.1381	0.2026	0.7083
25	6	1	1	0.3968	0.1360	0.1478	0.6396
28	4	1	0	0.2976	0.1334	0.0820	0.5559
30	3	1	0	0.1984	0.1203	0.0343	0.4603
34	2	1	0	0.0992	0.0924	0.0061	0.3504
46	1	1	0	0.0000	.	.	.
contype=2							
7	10	1	0	0.9000	0.0949	0.4730	0.9853
16	9	0	1	0.9000	0.0949	0.4730	0.9853
26	8	1	0	0.7875	0.1340	0.3809	0.9426

```
     31        7      1      0            0.6750    0.1551     0.2906    0.8825
     70        6      1      0            0.5625    0.1651     0.2094    0.8092
    127        5      0      1            0.5625    0.1651     0.2094    0.8092
    128        4      1      0            0.4219    0.1737     0.1110    0.7126
    319        3      0      1            0.4219    0.1737     0.1110    0.7126
    634        2      0      1            0.4219    0.1737     0.1110    0.7126
    641        1      0      1            0.4219    0.1737     0.1110    0.7126
contype=3
      2       30      1      0            0.9667    0.0328     0.7861    0.9952
      5       29      1      0            0.9333    0.0455     0.7589    0.9829
      7       28      2      0            0.8667    0.0621     0.6828    0.9478
     10       26      1      0            0.8333    0.0680     0.6450    0.9270
     11       25      1      0            0.8000    0.0730     0.6080    0.9048
     12       24      1      0            0.7667    0.0772     0.5720    0.8813
     15       23      1      0            0.7333    0.0807     0.5369    0.8567
     18       22      1      0            0.7000    0.0837     0.5026    0.8312
     19       21      1      0            0.6667    0.0861     0.4692    0.8047
     21       20      1      0            0.6333    0.0880     0.4365    0.7775
     23       19      1      1            0.6000    0.0894     0.4045    0.7495
     25       17      2      0            0.5294    0.0918     0.3378    0.6889
     29       15      1      1            0.4941    0.0922     0.3059    0.6573
     30       13      1      0            0.4561    0.0926     0.2716    0.6232
     46       12      2      0            0.3801    0.0915     0.2070    0.5521
     48       10      2      0            0.3041    0.0876     0.1477    0.4766
     49        8      1      0            0.2661    0.0845     0.1202    0.4371
     59        7      0      1            0.2661    0.0845     0.1202    0.4371
     66        6      1      0            0.2217    0.0812     0.0884    0.3924
     71        5      0      1            0.2217    0.0812     0.0884    0.3924
     99        4      0      1            0.2217    0.0812     0.0884    0.3924
    127        3      0      1            0.2217    0.0812     0.0884    0.3924
    284        2      0      1            0.2217    0.0812     0.0884    0.3924
    452        1      0      1            0.2217    0.0812     0.0884    0.3924
-------------------------------------------------------------------------------
```

Graphing the KM estimates can be done by:

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



I haven't altered this graph by changing legend labels, but you get the point as to what it is you're looking at I hope! (Interstate Conflicts are in the top function; Internationalized Civil Wars are in the middle function; Civil Wars are in the bottom function).

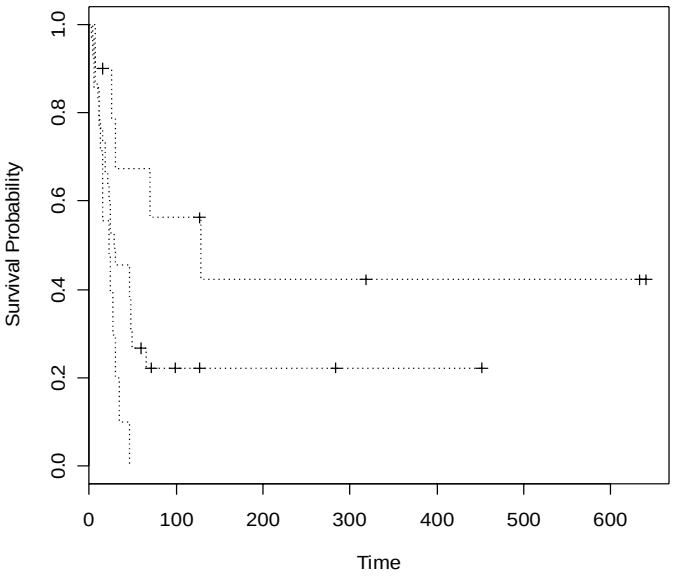
R

I can replicate the Stata results in R in the following way:

```
> unR.surv <- survfit(Surv(duration, failed)~ strata(contype), unR, conf.type="log-
log")

> summary(unR.surv)

Call: survfit(formula = Surv(duration, failed) ~ strata(contype), data =
unR,
    conf.type = "log-log")

4 observations deleted due to missing
                strata(contype)=contype=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    4     14       1   0.9286  0.0688      0.59077        0.990
    5     13       1   0.8571  0.0935      0.53945        0.962
   12     12       1   0.7857  0.1097      0.47246        0.925
   13     11       1   0.7143  0.1207      0.40630        0.882
   15      9       1   0.6349  0.1308      0.33116        0.830
   16      8       1   0.5556  0.1364      0.26365        0.772
   23      7       1   0.4762  0.1381      0.20263        0.708
   25      6       1   0.3968  0.1360      0.14782        0.640
   28      4       1   0.2976  0.1334      0.08196        0.556
   30      3       1   0.1984  0.1203      0.03433        0.460
   34      2       1   0.0992  0.0924      0.00615        0.350
   46      1       1   0.0000      NA           NA           NA

                strata(contype)=contype=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    7     10       1    0.900  0.0949        0.473        0.985
   26      8       1    0.787  0.1340        0.381        0.943
   31      7       1    0.675  0.1551        0.291        0.882
   70      6       1    0.563  0.1651        0.209        0.809
  128      4       1    0.422  0.1737        0.111        0.713

                strata(contype)=contype=3
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2     30       1    0.967  0.0328       0.7861        0.995
    5     29       1    0.933  0.0455       0.7589        0.983
    7     28       2    0.867  0.0621       0.6828        0.948
   10     26       1    0.833  0.0680       0.6450        0.927
   11     25       1    0.800  0.0730       0.6080        0.905
   12     24       1    0.767  0.0772       0.5720        0.881
   15     23       1    0.733  0.0807       0.5369        0.857
   18     22       1    0.700  0.0837       0.5026        0.831
   19     21       1    0.667  0.0861       0.4692        0.805
   21     20       1    0.633  0.0880       0.4365        0.778
   23     19       1    0.600  0.0894       0.4045        0.750
   25     17       2    0.529  0.0918       0.3378        0.689
   29     15       1    0.494  0.0922       0.3059        0.657
   30     13       1    0.456  0.0926       0.2716        0.623
   46     12       2    0.380  0.0915       0.2070        0.552
   48     10       2    0.304  0.0876       0.1477        0.477
   49      8       1    0.266  0.0845       0.1202        0.437
   66      6       1    0.222  0.0812       0.0884        0.392
```

Graphically, R gives me:

```
> plot(unR.surv, lty=c(3), xlab="Time", ylab="Survival Probability")
```



Again, I could improve the appearance of the graph, but the point here is to demonstrate the equivalence to Stata.