Duration Data

1 Recording Duration Data

Given the possibility of censoring and truncation, Table 1 illustrates a good way to record duration data.[1] By recording the analysis times during which cases are observed (t_0 and t_1) and whether

Table 1: Sample Duration Data

id	t_0	t_1	event	x
1	0	1	1	3
2	0	5	1	2
3	3	9	1	9
4	0	20	1	4
5	0	26	0	10

an event took place or not, it is possible to to take account of both censoring and truncation. For example, we know that observation 5 is right-censored since EVENT= 0. And we know that observation 3 is left-truncated since it enters our data set only in observation period 3.

It is important to recognize that we can record the same information about the duration of a case in multiple different, but equivalent, ways. For example, consider case 2 from Table 1. The duration data for case 2 can be recorded as in Table 1 or in any of the following ways:

Table 2: Sample Duration Recording II

id	t_0	t_1	event	x
2	0	2	0	2
2	2	5	1	2

or

Table 3: Sample Duration Recording III

id	t_0	t_1	event	x
2	0	2	0	2
2	2	4	0	2
2	4	5	1	2

All three of these different ways of recording the duration data for case 2 contain the same in-

[1] Much of the information in this section draws on Cleves, Gould, and Gutierrez (2008).

formation – case 2 was observed between analysis time 0 and analysis time 5, and at $t = 5$, it failed.

Why split the information into pieces? Well, it can help you incorporate (a) multiple failures and (b) time varying covariates. Consider the following information in Table 4. As you can see, we observed case 14 between analysis time 0 and analysis time 9. Case 14 experienced an event at analysis time 4 and at analysis time 7. In other words, this way of recording the data allows one to deal with multiple failures or events. Note also the independent variable x_2. This variable changes value across time – from analysis time 0 to 2, its value is 0; from analysis time 4 to 5, its value is 1; and from analysis time 5 to 9, its value is 0 again. In other words, this way of recording the data also allows one to deal with time varying covariates.

Table 4: Sample Duration Recording: Multiple Failures and TVCs

id	t_0	t_1	event	x_1	x_2
14	0	2	0	2	0
14	2	4	1	2	1
14	4	5	0	2	1
14	5	7	1	2	0
14	7	9	0	2	0

Note that all of this raises issues about the 'size' of the data set. In most statistical analyses, the number of observations or data records is taken as the size of the data set. However, this is not the case in duration analysis because, as you have seen, we can break the data records up into more and more time units. One way to measure the number of observations is to count how many cases are at risk of failure at any particular point in time – the sample size will obviously change over time as cases fail. The bottom line is that the notion of sample size is difficult to think about in duration analysis.

2 Preparing to Use Duration Data in STATA: Using stset

Before you can conduct duration analyses in STATA, you need to tell STATA what format your duration data is in. This requires using the STSET command. Let's go back to our government duration data from earlier. I have data from the Constitutional Change and Parliamentary Democracy (CCPD) project on the duration of governments in 17 European countries from 1946 to 1998/1999. Figure 1 shows what some of the data look like for Luxembourg and the Netherlands. In this data set, each observation records a failure time. There are some issues with the data set as you will see. First, there is no fail date for the last government in each country; that is, DATE_OUT is missing. The reason for this is that the CCPD stopped observing government duration on December 31, 1998. As a result, any government that was still in power at that date was simply coded as having no end date. In effect, the last case for each country is right-censored. At the moment, there is no variable in the data set indicating whether a case is right-censored or not. So, we should create one.

Figure 1: Government Duration Data

	country	cabinetcode	countryname	cabinet_name	cabinet_parties	date_in	date_out	duration	mwc
290	11	1101	luxembourg	Dupong I	CSV, LSAP, DP, KPL	451114	470301	472	0
291	11	1102	luxembourg	Dupong II	CSV, DP	470301	480607	464	1
292	11	1103	luxembourg	Dupong III	CSV, DP	480714	510604	1055	1
293	11	1104	luxembourg	Dupong IV	CSV, LSAP	510703	531223	904	1
294	11	1105	luxembourg	Bech I	CSV, LSAP	531229	540531	153	1
295	11	1106	luxembourg	Bech II	CSV, LSAP	540629	580329	1369	1
296	11	1107	luxembourg	Frieden	CSV, LSAP	580329	590202	310	1
297	11	1108	luxembourg	Werner I	CSV, DP	590302	640608	1925	1
298	11	1109	luxembourg	Werner II	CSV, LSAP	640715	681029	1567	1
299	11	1110	luxembourg	Werner III	CSV, DP	690206	740527	1936	1
300	11	1111	luxembourg	Thorn	DP, LSAP	740615	790611	1822	1
301	11	1112	luxembourg	Werner IV	CSV, DP	790716	840618	1799	1
302	11	1113	luxembourg	Santer I	CSV, LSAP	840720	890619	1795	1
303	11	1114	luxembourg	Santer II	CSV, LSAP	890714	940613	1795	1
304	11	1115	luxembourg	Santer III	CSV, LSAP	940713	950120	191	1
305	11	1116	luxembourg	Juncker	CSV, LSAP	950126	.	.	1
306	12	1201	netherlands	Schermerhorn	KVP, PvdA, ARP	450624	460517	327	0
307	12	1202	netherlands	Beel I	KVP, PvdA	460703	480707	735	1
308	12	1203	netherlands	Drees I	PvdA, KVP, CHU, VVD	480807	510124	900	0
309	12	1204	netherlands	Drees II	PvdA, KVP, CHU, VVD	510315	520625	468	0
310	12	1205	netherlands	Drees III	PvdA, KVP, CHU, ARP	520902	560613	1380	0
311	12	1206	netherlands	Drees IV	PvdA, KVP, CHU, ARP	561013	581212	790	0
312	12	1207	netherlands	Beel II	KVP, CHU, ARP	581222	590312	80	1
313	12	1208	netherlands	De Quay	KVP, CHU, ARP, VVD	590519	630515	1457	0
314	12	1209	netherlands	Marijnen	KVP, CHU, ARP, VVD	630724	650227	584	0
315	12	1210	netherlands	Cals	KVP, PvdA, ARP	650414	661015	549	0
316	12	1211	netherlands	Zijlstra	ARP, KVP	661122	670215	85	0
317	12	1212	netherlands	De Jong	KVP, ARP, CHU, VVD	670405	710428	1484	1
318	12	1213	netherlands	Biesheuvel I	ARP, KVP, CHU, VVD, DS70	710706	720720	358	1
319	12	1214	netherlands	Biesheuvel II	ARP, KVP, CHU, VVD	720809	721128	132	0
320	12	1215	netherlands	Den Uyl	PvdA, PPR, D66, KVP, ARP	730511	770322	1411	0
321	12	1216	netherlands	Van Agt I	CDA, VVD	771219	810527	1254	1
322	12	1217	netherlands	Van Agt II	CDA, PvdA, D66	810911	820512	243	0
323	12	1218	netherlands	Van Agt III	CDA, D66	820529	820909	102	0
324	12	1219	netherlands	Lubbers I	CDA, VVD	821104	860521	1296	1
325	12	1220	netherlands	Lubbers II	CDA, VVD	860714	890502	1151	1
326	12	1221	netherlands	Lubbers III	CDA, PvdA	891107	940503	1638	1
327	12	1222	netherlands	Kok I	PvdA, D66, VVD	940822	980505	1355	1
328	12	1223	netherlands	Kok II	PvdA, D66, VVD	980803	.	.	0
329	13	1301	norway	Gerhardsen II	A	451105	491010	1435	0

```
generate event=.;

replace event=1 if  date_out~=.;

replace event=0 if date_out==.;

label var event "1 = government collapse, 0 = right censored";
```

Now that we have a variable indicating whether an observation is right censored, we can replace the missing value for the DATE_OUT variable with the last date of observation by the CCPD project; that is, 981231. This is relatively straightforward and I do not show the code to do that here. Now that we have a DATE_IN and DATE_OUT value for each case, we can create a DURATION value for each case. But before we do this, let's have a look at STATA's STSET command. You will notice that there is already a variable in the data set that measures the duration of each government: DURATION. To tell STATA the structure of our duration data, we would type the following:

```
 stset duration;

     failure event:  (assumed to fail at time=duration)
obs. time interval:  (0, duration]
 exit on or before:  failure

--------------------------------------------------------------------------------
      424  total obs.
       19  event time missing (duration>=.)                    PROBABLE ERROR
--------------------------------------------------------------------------------
      405  obs. remaining, representing
      405  failures in single record/single failure data
   268998  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =      1936
```

As you see, when you type STSET, STATA gives you some information in two parts. The first part

```
     failure event:  (assumed to fail at time=duration)
obs. time interval:  (0, duration]
 exit on or before:  failure
```

just tells you what you specified in the STSET command. The second part summarizes the results of applying the definitions in the STSET command to the data set.

```
      424  total obs.
       19  event time missing (duration>=.)                    PROBABLE ERROR
--------------------------------------------------------------------------------
      405  obs. remaining, representing
      405  failures in single record/single failure data
   268998  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =      1936
```

As you can see, it is split into two by a horizontal dashed line. The bit above the dashed line highlights potential problems – this is why it says probable error. The bit below the dashed line tells you about the characteristics of the data set that you have STSET. You should look at the 'probable errors' to see what, if anything, is wrong.

The probable error is occurring because there are 19 cases in the data set for which the value of the DURATION variable is missing. Now that we have a DATE_IN and DATE_OUT value for each case, we can resolve this problem by creating a DURATION value for each case. Let's act as if there is no DURATION variable in the data set and use the DATE_IN and DATE_OUT variables to create a new DURATION variable. To do this, we need to tell STATA that the DATE_IN and DATE_OUT variables are actually date variables. At the moment, the dates in our dataset are recorded as YYMMDD. The problem is that STATA does not know that this indicates a date. There are three basic steps to get STATA to recognize these as dates.

1. Convert to string format

   ```
   tostring date_in , replace;

   tostring date_out , replace;
   ```

2. Convert string variable into STATA date format[2]

   ```
   generate date0=date(date_in,   "YMD", 2000);

   generate date1=date(date_out, "YMD", 2000);
   ```

   The baseline date for STATA is January 1, 1960. In other words, this is 0. January 2, 1960 = 1, December 31, 1959 = -1, and so on.

3. Put a date format on the new date variable

   ```
   format date0 %d;

   format date1 %d;
   ```

We are now in a position to calculate the duration for each government. We will go ahead and replace the old DURATION variable with a new DURATION variable using our new date variables. We do this by typing:

```
replace duration =  date1 - date0;
```

We can now see what our data looks like in Figure 2. As you can see, we now have a value for the DURATION variable for each case. We also have a variable indicating whether a case is right censored or not (EVENT).

[2] '2000' gives STATA a rule for converting two-digit years into four-digit years. Specifying '2000' means that the largest year to be produced is 2000. So year 89 would be interpreted as 1989 instead of 2089, and so on.

Figure 2: Government Duration Data

	country	cabinetcode	countryname	cabinet_name	cabinet_parties	date_in	date_out	date0	date1	duration	event
1	1	101	austria	Renner	SPÖ, ÖVP, KPÖ	450427	451220	27apr1945	20dec1945	237	1
2	1	102	austria	Figl I	ÖVP, SPÖ, KPÖ	451220	471120	20dec1945	20nov1947	700	1
3	1	103	austria	Figl II	ÖVP, SPÖ	471120	491011	20nov1947	11oct1949	691	1
4	1	104	austria	Figl III	ÖVP, SPÖ	491108	530225	08nov1949	25feb1953	1205	1
5	1	105	austria	Raab I	ÖVP, SPÖ	530402	560514	02apr1953	14may1956	1138	1
6	1	106	austria	Raab II	ÖVP, SPÖ	560629	590512	29jun1956	12may1959	1047	1
7	1	107	austria	Raab III	ÖVP, SPÖ	590716	610411	16jul1959	11apr1961	635	1
8	1	108	austria	Gorbach I	ÖVP, SPÖ	610411	621120	11apr1961	20nov1962	588	1
9	1	109	austria	Gorbach II	ÖVP, SPÖ	630327	640402	27mar1963	02apr1964	372	1
10	1	110	austria	Klaus I	ÖVP, SPÖ	640402	651025	02apr1964	25oct1965	571	1
11	1	111	austria	Klaus II	ÖVP	660419	700303	19apr1966	03mar1970	1414	1
12	1	112	austria	Kreisky I	SPÖ	700421	711019	21apr1970	19oct1971	546	1
13	1	113	austria	Kreisky II	SPÖ	711104	751008	04nov1971	08oct1975	1434	1
14	1	114	austria	Kreisky III	SPÖ	751028	790509	28oct1975	09may1979	1289	1
15	1	115	austria	Kreisky IV	SPÖ	790605	830426	05jun1979	26apr1983	1421	1
16	1	116	austria	Sinowatz	SPÖ, FPÖ	830524	860616	24may1983	16jun1986	1119	1
17	1	117	austria	Vranitzky I	SPÖ, FPÖ	860616	861125	16jun1986	25nov1986	162	1
18	1	118	austria	Vranitzky II	SPÖ, ÖVP	870121	901009	21jan1987	09oct1990	1357	1
19	1	119	austria	Vranitzky III	SPÖ, ÖVP	901217	941011	17dec1990	11oct1994	1394	1
20	1	120	austria	Vranitzky IV	SPÖ, ÖVP	941129	951219	29nov1994	19dec1995	385	1
21	1	121	austria	Vranizky V	SPÖ, ÖVP	960312	970115	12mar1996	15jan1997	309	1
22	1	122	austria	Klima	SPÖ, ÖVP	970115	981231	15jan1997	31dec1998	715	0
23	2	201	belgium	Spaak	PSB/BSP	460313	460320	13mar1946	20mar1946	7	1
24	2	202	belgium	Van Acker III	PSB/BSP, LP/PL, PCB/KPB	460331	460709	31mar1946	09jul1946	100	1
25	2	203	belgium	Huysmans	PSB/BSP, LP/PL, PCB/KPB	460803	470313	03aug1946	13mar1947	222	1
26	2	204	belgium	Spaak II	CVP/PSC, PSB/BSP	470320	490627	20mar1947	27jun1949	830	1

Now we can go back and use the stset command again. As you can see, there are no recorded problems any more.

```
 stset duration

     failure event:  (assumed to fail at time=duration)
obs. time interval:  (0, duration]
 exit on or before:  failure

------------------------------------------------------------------------------
      424  total obs.
        0  exclusions
------------------------------------------------------------------------------
      424  obs. remaining, representing
      424  failures in single record/single failure data
   282534  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =      1936
```

However, it turns out there is still a problem. STATA thinks that all 424 observations failed, when, in fact, we know that some observations were censored. This is because we did not tell STATA that some observations are right censored. And so, we have to type the following to solve this problem:

```
 stset duration, failure(event)

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

--------------------------------------------------------------------------------
      424  total obs.
        0  exclusions
--------------------------------------------------------------------------------
      424  obs. remaining, representing
      407  failures in single record/single failure data
   282534  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =      1936
```

Notice the difference in what is reported. There are now only 407 reported failures instead of 424 reported failures.[3]

If you look at your data after using the STSET command, you will find that STATA has created 4 new variables: $\_t_0$, _t, _d, and _st. $\_t_0$ and _t record the time span in analysis time for each case in your data. Each case will start at $\_t_0$ and end at _t. You will see that _t is the same as our DURATION variable. _d is an indicator variable indicating whether the case ended in failure or not; _d= 1 if there is failure, 0 otherwise. Recall that this is the indicator variable that we used when constructing our likelihood function to take account or right censoring. _st is an indicator

Figure 3: Government Duration Data

	cabinetcode	countryname	cabinet_name	date_in	date_out	date0	date1	duration	event	_st	_d	_t	_t0
1	101	austria	Renner	450427	451220	27apr1945	20dec1945	237	1	1	1	237	0
2	102	austria	Figl I	451220	471120	20dec1945	20nov1947	700	1	1	1	700	0
3	103	austria	Figl II	471120	491011	20nov1947	11oct1949	691	1	1	1	691	0
4	104	austria	Figl III	491108	530225	08nov1949	25feb1953	1205	1	1	1	1205	0
5	105	austria	Raab I	530402	560514	02apr1953	14may1956	1138	1	1	1	1138	0
6	106	austria	Raab II	560629	590512	29jun1956	12may1959	1047	1	1	1	1047	0
7	107	austria	Raab III	590716	610411	16jul1959	11apr1961	635	1	1	1	635	0
8	108	austria	Gorbach I	610411	621120	11apr1961	20nov1962	588	1	1	1	588	0
9	109	austria	Gorbach II	630327	640402	27mar1963	02apr1964	372	1	1	1	372	0
10	110	austria	Klaus I	640402	651025	02apr1964	25oct1965	571	1	1	1	571	0
11	111	austria	Klaus II	660419	700303	19apr1966	03mar1970	1414	1	1	1	1414	0
12	112	austria	Kreisky I	700421	711019	21apr1970	19oct1971	546	1	1	1	546	0
13	113	austria	Kreisky II	711104	751008	04nov1971	08oct1975	1434	1	1	1	1434	0
14	114	austria	Kreisky III	751028	790509	28oct1975	09may1979	1289	1	1	1	1289	0
15	115	austria	Kreisky IV	790605	830426	05jun1979	26apr1983	1421	1	1	1	1421	0
16	116	austria	Sinowatz	830524	860616	24may1983	16jun1986	1119	1	1	1	1119	0
17	117	austria	Vranitzky I	860616	861125	16jun1986	25nov1986	162	1	1	1	162	0
18	118	austria	Vranitzky II	870121	901009	21jan1987	09oct1990	1357	1	1	1	1357	0
19	119	austria	Vranitzky III	901217	941011	17dec1990	11oct1994	1394	1	1	1	1394	0
20	120	austria	Vranitzky IV	941129	951219	29nov1994	19dec1995	385	1	1	1	385	0
21	121	austria	Vranizky V	960312	970115	12mar1996	15jan1997	309	1	1	1	309	0
22	122	austria	Klima	970115	981231	15jan1997	31dec1998	715	0	1	0	715	0
23	201	belgium	Spaak	460313	460320	13mar1946	20mar1946	7	1	1	1	7	0
24	202	belgium	Van Acker III	460331	460709	31mar1946	09jul1946	100	1	1	1	100	0
25	203	belgium	Huysmans	460803	470313	03aug1946	13mar1947	222	1	1	1	222	0
26	204	belgium	Spaak II	470320	490627	20mar1947	27jun1949	830	1	1	1	830	0
27	205	belgium	Eyskens	490811	500318	11aug1949	18mar1950	219	1	1	1	219	0

variable indicating whether the case is to be used in the current analysis or not; _st= 1 if it is to be used, 0 otherwise. In Figure 3, I show the new variables created by STATA.

[3]It turns out that the STSET syntax can be modified in a number of ways to give STATA more information about your duration data. You should look at the STATA manual if you need to do this.

Once you have STSET your data, you should take a look at the data and check for problems before conducting any analyses. One useful command for doing this is STDES. This provides various descriptive information about your data that you might find useful.

```
. stdes

        failure _d:  event
  analysis time _t:  duration

                                |-------------- per subject --------------|
Category                  total        mean         min     median        max
------------------------------------------------------------------------------
no. of subjects             424
no. of records              424           1           1          1          1

(first) entry time                        0           0          0          0
(final) exit time                  666.3538           7      572.5       1936

subjects with gap             0
time on gap if gap            0
time at risk             282534    666.3538           7      572.5       1936

failures                    407    .9599057           0          1          1
------------------------------------------------------------------------------
```

Now we are ready to conduct some analyses.

References

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