Appendix: Data Structure and Organization
"Duration Models for Repeated Events" – Box–Steffensmeier and Zorn (2002)

1 Introduction

Data structure and organization plays a key role in the estimation of models for repeated events duration data. Here, we briefly outline the data organization for each of the various models estimated above. The key differences in data organization are between the WLW model and the others, and between the elapsed–time and interevent–time specifications for the PWP model. We begin with the simpler case of data in which covariates do not vary over time that is, data which contain one observation for each event. We go on to illustrate the organization of data with time-varying covariates.

2 Time–Invariant Data

2.1 A–G and PWP Models

Events data in which the covariates do not vary over time are called "time–invariant" data. For such data, each observation corresponds to a single duration, terminated either by the occurrence of the event of interest or through censoring. Repeated events occur when the same unit of analysis experiences two or more events: for example, when the nations which make up a single dyad experience two or more conflicts in the period under study.

Data organization for the AG and PWP models is nearly identical, and is presented in Table 1. In the example data, the first dyad (number 1010) experienced three conflict events, at four, seven and ten years after the start of observation, and the observation remained in the data for a total of ten years. The second dyad (number 1020) experienced no conflicts, and persisted in the data for three years before being censored. The third dyad (number 1030) experienced one conflict two years after entering the data set, and was subsequently censored after a total of seven years, while the fourth dyad (number 1040) had conflicts at three and nine years and was censored in the fifteenth year. Note that each event is recorded as a single observation, containing an identifier for the unit of analysis (here, `DYADID`), an indicator of whether (0) or not (1) the observation was censored (`CENSOR`), and an index for the number of the `EVENT` (first, second, etc.). For illustrative purposes, a variable indicating the `DURATION` until each event is also included.

More critical for the PWP model is the specification of elapsed or interevent time. The counting process formulation of the AG model requires that the data take the form of elapsed time; that is, time since the start of observation for that unit. This formulation of the duration is also used for the elapsed–time variant of the PWP model. In Table 1, data on the elapsed–time durations are represented by `ESTART` and `EFINISH`, which indicate the points at which the durations associated with each event started and ended, respectively. The key difference between the AG and PWP–elapsed time models, then, is that the latter stratifies the estimates by `EVENT`, estimating separate baseline hazards for each event occurrence, while the former does not. By contrast, for the interevent–time variant of the PWP model, the durations are defined by the `ISTART` and `IFINISH` variables, respectively; note that, for the latter two, the start times are all reset to zero following the occurrence of each event. In all instances, robust standard errors, clustered by `DYADID`, are estimated to account for intra-dyad dependence.

2.2 WLW Models

Data organization for the WLW models is significantly different than that for the AG and PWP approaches. Because all observations are potentially "at risk" for every event from the start of the observation period, the WLW data structure requires separate observations for each unit of analysis for every possible event. In our example, suppose that the greatest number of multiple events experienced by any dyad is four. The

Table 1: AG/PWP Data Organization, Time–Invariant Data

DYADID	EVENT	CENSOR	DURATION	ESTART	EFINISH	ISTART	IFINISH
1010	1	1	4	0	4	0	4
1010	2	1	3	4	7	0	3
1010	3	1	3	7	10	0	3
1020	1	0	3	0	3	0	3
1030	1	1	2	0	2	0	2
1030	2	0	5	2	7	0	5
1040	1	1	3	0	3	0	3
1040	2	1	6	3	9	0	6
1040	3	0	6	9	15	0	6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

WLW data then require a separate observation for every dyad for each of the four possible events (e.g. first, second, third and fourth). Dyads which do not experience second, third, etc. events are treated as censored for that observation; this is illustrated in Table 2. Thus, for dyad 1010, we have a total of four observations, three of which end with an event (at times 4, 7 and 10 for the first, second and third events, respectively) and the last of which is censored at $T = 10$. Dyad 1020 also have four observations, all of which are coded as censored at $T = 3$. Dyad 1030 would have four observations, with the first uncensored at $T = 2$ and the remaining three censored at $T = 7$, etc. To estimate the WLW model, the data are stratified by `EVENT`, and robust standard errors, clustered by `DYADID`, are estimated as well.

Table 2: WLW Data Organization, Time–Invariant Data

DYADID	EVENT	CENSOR	DURATION
1010	1	1	4
1010	2	1	7
1010	3	1	10
1010	4	0	10
1020	1	0	3
1020	2	0	3
1020	3	0	3
1020	4	0	3
1030	1	1	2
1030	2	1	7
⋮	⋮	⋮	⋮

3 Time–Varying Data

3.1 A–G and PWP Models

In applications where one's data contains covariates that change their values over time, data organization is somewhat more complex. The key is to maintain the same risk sets as in the time–invariant model. For the AG and PWP models, the data structure requires one observation for every unit, for every time period under study. In the example of international conflict data, this is the familiar dyad–year data common to

such studies (e.g. Oneal and Russett 1997, 1999). Table 3 presents the AG/PWP data organization for the example data discussed above, but now also including a covariate `X` which varies for each dyad from one year to the next. The data contain observations equal to the number of unit–years observed; each observation is now indexed by both `DYADID` and `YEAR`, as well as containing the indicator of the `EVENT` (first, second, etc.) for which they are at risk. Any dyad–year in which no conflict occurred is treated as censored, and the duration variables proceed as year–by–year step functions. Once again, the AG and PWP/elapsed–time models define durations according to the `ESTART` and `EFINISH` variables, while the PWP/interevent–time model analyzes the `ISTART` and `IFINISH` variables. As before, in all cases robust standard errors, clustered by `DYADID`, are estimated, and both PWP models also stratify by event number.

Table 3: AG/PWP Data Organization, Time–Varying Data

DYADID	YEAR	EVENT	X	CENSOR	DURATION	ESTART	EFINISH	ISTART	IFINISH
1010	1950	1	100	0	1	0	1	0	1
1010	1951	1	120	0	2	1	2	1	2
1010	1952	1	155	0	3	2	3	2	3
1010	1953	1	140	1	4	3	4	3	4
1010	1954	2	95	0	1	4	5	0	1
1010	1955	2	80	0	2	5	6	1	2
1010	1956	2	125	1	3	6	7	2	3
1010	1957	3	75	0	1	7	8	0	1
1010	1958	3	110	0	2	8	9	1	2
1010	1959	3	105	1	3	9	10	2	3
1020	1950	1	65	0	1	0	1	0	1
1020	1951	1	70	0	2	1	2	1	2
1020	1952	1	80	0	3	2	3	2	3
1030	1950	1	175	0	1	0	1	0	1
1030	1951	1	190	1	2	1	2	1	2
1030	1952	2	140	0	1	2	3	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

3.2 WLW Models

The time–varying WLW data organization requires that it accurately reflect its assumption that all observations are "at risk" for all possible events at all times prior to experiencing that event. Accordingly, and as in the case of time-invariant data, the data must contain one observation for every dyad-year for every event that dyad has not experienced up to and including that time; this is illustrated in Table 4. For dyad 1010 in our example, the data contain a total of 31 observations: four for the first conflict (which was experienced at $T = 4$), seven for the second conflict (at $T = 7$), ten for the third (at $T = 10$) and ten for the fourth (which never occurred, and is thus treated as censored at $T = 10$). Dyad 1020 consists of twelve observations (three each for the first, second, third and fourth conflicts), all of which are treated as censored. While not shown here, dyad 1030 would have a total of seventeen observations (two for the first conflict, occurring at $T = 2$, and five censored observations each for the second, third and fourth conflicts). As in the model above, model estimation stratifies by `EVENT`, as well as clustering standard error estimates by `DYADID`; the latter is particularly important to avoid drastically underestimating one's standard errors.

Table 4: WLW Data Organization, Time–Varying Data

DYADID	YEAR	EVENT	X	CENSOR	DURATION
1010	1950	1	100	0	1
1010	1951	1	120	0	2
1010	1952	1	155	0	3
1010	1953	1	140	1	4
1010	1950	2	100	0	1
1010	1951	2	120	0	2
1010	1952	2	155	0	3
1010	1953	2	140	0	4
1010	1954	2	95	0	5
1010	1955	2	80	0	6
1010	1956	2	125	1	7
1010	1950	3	100	0	1
1010	1951	3	120	0	2
1010	1952	3	155	0	3
1010	1953	3	140	0	4
1010	1954	3	95	0	5
1010	1955	3	80	0	6
1010	1956	3	125	0	7
1010	1957	3	75	0	8
1010	1958	3	110	0	9
1010	1959	3	105	1	10
1010	1950	4	100	0	1
1010	1951	4	120	0	2
1010	1952	4	155	0	3
1010	1953	4	140	0	4
1010	1954	4	95	0	5
1010	1955	4	80	0	6
1010	1956	4	125	0	7
1010	1957	4	75	0	8
1010	1958	4	110	0	9
1010	1959	4	105	0	10
1020	1950	1	65	0	1
1020	1951	1	70	0	2
1020	1952	1	80	0	3
1020	1950	2	65	0	1
1020	1951	2	70	0	2
1020	1952	2	80	0	3
1020	1950	3	65	0	1
1020	1951	3	70	0	2
1020	1952	3	80	0	3
1020	1950	4	65	0	1
1020	1951	4	70	0	2
1020	1952	4	80	0	3
⋮	⋮	⋮	⋮	⋮	⋮