sttocc — Convert survival-time data to case–control data

Description

`sttocc` generates a nested case–control study dataset from a cohort-study dataset by sampling controls from the risk sets. For each case, the controls are chosen randomly from those members of the cohort who are at risk at the failure time of the case. That is, the resulting case–control sample is matched with respect to analysis time—the time scale used to compute risk sets. The following variables are added to the dataset:

`_case`	Coded 0 for controls, 1 for cases
`_set`	Case–control ID; matches cases and controls that belong together
`_time`	Analysis time of the case's failure

The names of these three variables can be changed by specifying the `generate()` option. *varlist* defines variables that, in addition to those used in the creation of the case–control study, will be retained in the final dataset. If *varlist* is not specified, all variables are carried over into the resulting dataset.

When the resulting dataset is analyzed as a matched case–control study, odds ratios will estimate corresponding rate-ratio parameters in the proportional hazards model for the cohort study.

Randomness in the matching is obtained using Stata's `runiform()` function. To ensure that the sample truly is random, you should set the random-number seed; see [R] **set seed**.

Quick start

Create a nested case–control dataset from a cohort dataset that has been `stset`, matching cases to controls based on analysis time

```
sttocc
```

Same as above, but match on analysis time and categorical variable `catvar`

```
sttocc, match(catvar)
```

Same as above, but match 3 controls for each case

```
sttocc, match(catvar) number(3)
```

Same as above, and name the case indicator `case`, the matching identifier `mid`, and the case's failure time `ftime`

```
sttocc, match(catvar) number(3) generate(case mid ftime)
```

Menu

Statistics > Survival analysis > Setup and utilities > Convert survival-time data to case–control data

Syntax

```
sttocc [varlist] [ , options]
```

options	Description
Main	
match(*matchvarlist*)	match cases and controls on analysis time and specified categorical variables; default is to match on analysis time only
number(#)	use # controls for each case; default is number(1)
nodots	suppress displaying dots during calculation
Options	
generate(*case set time*)	new variable names; default is _case, _set, and _time

You must stset your data before using sttocc; see [ST] **stset**.
fweights, iweights, and pweights may be specified using stset; see [ST] **stset**.

Options

Main

match(*matchvarlist*) specifies more categorical variables for matching controls to cases. When match() is not specified, cases and controls are matched with respect to time only. If match(*matchvarlist*) is specified, the cases will also be matched by *matchvarlist*.

number(#) specifies the number of controls to draw for each case. The default is 1, even though this is not a sensible choice.

nodots requests that dots not be placed on the screen at the beginning of each case–control group selection. By default, dots are displayed to show progress.

Options

generate(*case set time*) specifies variable names for the three new variables; the default is _case, _set, and _time.

Remarks and examples

Nested case–control studies are an attractive alternative to full Cox regression analysis, particularly when time-varying explanatory variables are involved. They are also attractive when some explanatory variables involve laborious coding. For example, you can create a file with a subset of variables for all subjects in the cohort, generate a nested case–control study, and go on to code the remaining data only for those subjects selected.

In the same way as with Cox regression, the results of the analysis are critically dependent on the choice of analysis time (time scale). The choice of analysis time may be calendar time—so that controls would be chosen from subjects still being monitored on the date that the case fails—but other time scales, such as age or time in study, may be more appropriate in some studies. Remember that the analysis time set in selecting controls is implicitly included in the model in subsequent analysis.

match() requires that controls also be matched to the case with respect to other categorical variables, such as sex. This produces an analysis closely mirroring stratified Cox regression. If we wanted to match on calendar time and 5-year age bands, we could first type stsplit ageband ... to create the age bands

and then specify `match(ageband)` on the `sttocc` command. Analyzing the resulting data as a matched case–control study would estimate rate ratios in the underlying cohort that are controlled for calendar time (very finely) and age (less finely). Such analysis could be carried out by Mantel–Haenszel (odds ratio) calculations, for example, using `mhodds`, or by conditional logistic regression using `clogit`.

When ties occur between entry times, censoring times, and failure times, the following convention is adopted:

$$\text{Entry time} < \text{Failure time} < \text{Censoring time}$$

Thus censored subjects and subjects entering at the failure time of the case are included in the risk set and are available for selection as controls. Tied failure times are broken at random.

Example 1: Creating a nested case–control study

Using the `diet` data introduced in example 1 of [ST] **stsplit**, we will illustrate the use of `sttocc`, letting age be analysis time. Controls are chosen from subjects still being monitored at the age at which the case fails.

```
. use https://www.stata-press.com/data/r19/diet
(Diet data with dates)
. stset dox, failure(fail) enter(time doe) id(id) origin(time dob) scale(365.25)
Survival-time data settings
          ID variable: id
        Failure event: fail!=0 & fail<.
Observed time interval: (dox[_n-1], dox]
     Enter on or after: time doe
     Exit on or before: failure
     Time for analysis: (time-origin)/365.25
               Origin: time dob
------------------------------------------------------------------------------
        337  total observations
          0  exclusions
------------------------------------------------------------------------------
        337  observations remaining, representing
        337  subjects
         80  failures in single-failure-per-subject data
  4,603.669  total analysis time at risk and under observation
                                                At risk from t =         0
                                     Earliest observed entry t =  30.07529
                                          Last observed exit t =  69.99863
. set seed 9123456
```

```
. sttocc, match(job) n(5) nodots
note: 2 sets of tied failure times detected; splitting at random.
      Failures are assumed to precede censorings.

Survival-time data settings

         Failure _d: fail
   Analysis time _t: (dox-origin)/365.25
             Origin: time dob
  Enter on or after: time doe
       ID variable: id
      Matching for: job

Converting survival-time data to case-control data:
  Sampling 5 controls for each of 80 cases ...

Data are now case-control data with new variables:
    _case    Case-control indicator
    _set     Case-control ID
    _time    Analysis time of the case's failure
```

The above two commands create a new dataset in which there are five controls per case, matched on `job`, with the age of the subjects when the case failed recorded in the variable `_time`. The case indicator is given in `_case` and the matched set number, in `_set`. Because we did not specify the optional *varlist*, all variables are carried over into the new dataset.

```
. describe

Contains data from https://www.stata-press.com/data/r19/diet.dta
 Observations:           480                  Diet data with dates
    Variables:            14                  1 May 2024 19:01
-------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------
id              int     %9.0g                 Subject identity number
fail            byte    %8.0g                 Outcome (CHD = 1 3 13)
job             byte    %8.0g                 Occupation
month           byte    %8.0g                 Month of survey
energy          float   %9.0g                 Total energy (1000kcals/day)
height          float   %9.0g                 Height (cm)
weight          float   %9.0g                 Weight (kg)
hienergy        byte    %9.0g                 Indicator for high energy
doe             int     %td                   Date of entry
dox             int     %td                   Date of exit
dob             int     %td                   Date of birth
_case           byte    %8.0g                 0 for controls; 1 for cases
_set            long    %12.0g                Case-control ID
_time           double  %10.0g                Analysis time of the case's
                                                failure
-------------------------------------------------------------------------------
Sorted by: _set  _case
     Note: Dataset has changed since last saved.
```

We can verify that the controls were correctly selected:

```
. gen ageentry=(doe-dob)/365.25
. gen ageexit=(dox-dob)/365.25
. sort _set _case id
. list _set id _case _time ageentry ageexit job, sepby(_set)
```

	_set	id	_case	_time	ageentry	ageexit	job
1.	1	65	0	42.57358	40.11225	56.82409	0
2.	1	73	0	42.57358	36.58043	52.70636	0
3.	1	74	0	42.57358	37.09788	53.39083	0
4.	1	75	0	42.57358	31.13484	47.26078	0
5.	1	86	0	42.57358	38.14921	54.10815	0
6.	1	90	1	42.57358	31.4141	42.57358	0
7.	2	203	0	47.8987	41.26215	61.22108	2
8.	2	207	0	47.8987	43.6386	63.51266	2
9.	2	236	0	47.8987	45.30048	57.42368	2
10.	2	281	0	47.8987	44.34223	61.54963	2
11.	2	333	0	47.8987	46.37645	61.8371	2
12.	2	196	1	47.8987	45.46475	47.8987	2
13.	3	37	0	47.964408	35.2115	52.67351	0
14.	3	66	0	47.964408	40.09309	56.9692	0
				(*output omitted*)			
479.	80	180	0	68.596851	61.55784	69.99863	1
480.	80	108	1	68.596851	55.72074	68.59686	1

The controls do indeed belong to the appropriate risk set. The controls in each set enter at an age that is less than the age of the case at failure, and they exit at an age that is greater than the age of the case at failure. To estimate the effect of high energy, use `clogit`, just as you would for any matched case–control study:

```
. clogit _case hienergy, group(_set) or
Iteration 0:  Log likelihood = -143.22071
Iteration 1:  Log likelihood = -143.22071
Conditional (fixed-effects) logistic regression          Number of obs =    480
                                                         LR chi2(1)    =   0.24
                                                         Prob > chi2   = 0.6241
Log likelihood = -143.22071                              Pseudo R2     = 0.0008
```

_case	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
hienergy	.88683	.217505	-0.49	0.624	.54837	1.434191

Acknowledgments

The original version of `sttocc` was written by David Clayton (retired) of the Cambridge Institute for Medical Research and Michael Hills (1934–2021) of the London School of Hygiene and Tropical Medicine.

References

Clayton, D. G., and M. Hills. 1993. *Statistical Models in Epidemiology*. Oxford: Oxford University Press.

Langholz, B., and D. C. Thomas. 1990. Nested case-control and case-cohort methods of sampling from a cohort: A critical comparison. *American Journal of Epidemiology* 131: 169–176. https://doi.org/10.1093/oxfordjournals.aje.a115471.

Also see

[ST] **stbase** — Form baseline dataset

[ST] **stdescribe** — Describe survival-time data

[ST] **stset** — Declare data to be survival-time data

[ST] **stsplit** — Split and join time-span records