Title

stata.com

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

Syntax | Menu | Description | Options
Remarks and examples | Acknowledgments | References | Also see

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**.

Menu

Statistics > Survival analysis > Setup and utilities > 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**.

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

stata.com

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. See Clayton and Hills (1997) for more information.

▷ 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 http://www.stata-press.com/data/r13/diet
(Diet data with dates)
. stset dox, failure(fail) enter(time doe) id(id) origin(time dob) scale(365.25)

                id:  id
     failure event:  fail != 0 & fail < .
obs. time interval:  (dox[_n-1], dox]
 enter on or after:  time doe
 exit on or before:  failure
    t 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
 4603.669  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

        failure _d:  fail
  analysis time _t:  (dox-origin)/365.25
            origin:  time dob
 enter on or after:  time doe
                id:  id
      matching for:  job
There were 2 tied times involving failure(s)
  - failures assumed to precede censorings,
  - tied failure times split at random
There are 80 cases
Sampling 5 controls for each case
```

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.

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	37	0	42.57358	35.2115	52.67351	0
2.	1	57	0	42.57358	40.04107	56.5859	0
3.	1	86	0	42.57358	38.14921	54.10815	0
4.	1	92	0	42.57358	36.67625	52.38877	0
5.	1	100	0	42.57358	36.15332	46.86653	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	213	0	47.8987	47.23614	67.02532	2
9.	2	219	0	47.8987	41.86721	61.74127	2
10.	2	313	0	47.8987	41.50582	57.05133	2
11.	2	316	0	47.8987	40.68994	56.23545	2
12.	2	196	1	47.8987	45.46475	47.8987	2
13.	3	57	0	47.964408	40.04107	56.5859	0
14.	3	71	0	47.964408	47.52909	64.57221	0
				(*output omitted*)			
479.	80	136	0	68.596851	58.41205	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 = -142.31278
Iteration 1:   log likelihood = -142.31276
Iteration 2:   log likelihood = -142.31276
Conditional (fixed-effects) logistic regression   Number of obs    =        480
                                                  LR chi2(1)       =       2.06
                                                  Prob > chi2      =     0.1516
Log likelihood = -142.31276                       Pseudo R2        =     0.0072
```

_case	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hienergy	.7026801	.1734294	-1.43	0.153	.433183	1.13984

Acknowledgments

The original version of `sttocc` was written by David Clayton of the Cambridge Institute for Medical Research and Michael Hills (retired) 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.

——. 1995. ssa7: Analysis of follow-up studies. *Stata Technical Bulletin* 27: 19–26. Reprinted in *Stata Technical Bulletin Reprints*, vol. 5, pp. 219–227. College Station, TX: Stata Press.

——. 1997. ssa10: Analysis of follow-up studies with Stata 5.0. *Stata Technical Bulletin* 40: 27–39. Reprinted in *Stata Technical Bulletin Reprints*, vol. 7, pp. 253–268. College Station, TX: Stata Press.

Coviello, V. 2001. sbe41: Ordinary case–cohort design and analysis. *Stata Technical Bulletin* 59: 12–18. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 121–129. College Station, TX: Stata 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.

Also see

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

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

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