Title

stata.com

Discrete — Discrete-time survival analysis

Description Acknowledgment References Also see

Description

As of the date that this manual was printed, Stata does not have a suite of built-in commands for discrete-time survival models matching the `st` suite for continuous-time models, but a good case could be made that it should. Instead, these models can be fit easily using other existing estimation commands and data manipulation tools.

Discrete-time survival analysis concerns analysis of time-to-event data whenever survival times are either a) intrinsically discrete (for example, numbers of machine cycles) or b) grouped into discrete intervals of time ("interval-censoring"). If intervals are of equal length, the same methods can be applied to both a) and b); survival times will be positive integers.

You can fit discrete-time survival models with the maximum likelihood method. Data may contain completed or right-censored spells, and late entry (left-truncation) can also be handled, as well as unobserved heterogeneity (also termed "frailty"). Estimation makes use of the property that the sample likelihood can be rewritten in a form identical to the likelihood for a binary dependent variable multiple regression model and applied to a specially organized dataset (Allison 2014, Jenkins 1995). For models without frailty, you can use, for example, `logistic` (or `logit`) to fit the discrete-time logistic hazard model or `cloglog` to fit the discrete-time proportional hazards model (Prentice and Gloeckler 1978). Models incorporating normal frailty may be fit using `xtlogit` and `xtcloglog`. A model with gamma frailty (Meyer 1990) may be fit using `pgmhaz` (Jenkins 1997).

Estimation consists of three steps:

1. *Data organization*: The dataset must be organized so that there is 1 observation for each period when a subject is at risk of experiencing the transition event. For example, if the original dataset contains one row for each subject, i, with information about their spell length, T_i, the new dataset requires T_i rows for each subject, one row for each period at risk. This may be accomplished using `expand` or `stsplit`. (This step is episode splitting at each and every interval.) The result is data of the same form as a discrete panel (`xt`) dataset with repeated observations on each panel (subject).

2. *Variable creation*: You must create at least three types of variables. First, you will need an interval identification variable, which is a sequence of positive integers $t = 1, \ldots, T_i$. For example,

```
. sort subject_id
. by subject_id: generate t = _n
```

   Second, you need a period-specific censoring indicator, d_i. If $d_i = 1$ if subject i's spell is complete and $d_i = 0$ if the spell is right-censored, the new indicator $d^*_{it} = 1$ if $d_i = 1$ and $t = T_i$, and $d^*_{it} = 0$ otherwise.

   Third, you must define variables (as functions of t) to summarize the pattern of duration dependence. These variables are entered as covariates in the regression. For example, for a duration dependence pattern analogous to that in the continuous-time Weibull model, you could define a new variable $x_1 = \log t$. For a quadratic specification, you define variables $x_1 = t$ and $x_2 = t^2$. We can achieve a piecewise constant specification by defining a set of dummy variables, with each group of periods sharing the same hazard rate, or a semiparametric model (analogous to the Cox regression model for continuous survival-time data) using separate dummy variables for each and every duration

interval. No duration variable need be defined if you want to fit a model with a constant hazard rate.

In addition to these three essentials, you may define other time-varying covariates.

3. *Estimation*: You fit a binary dependent variable multiple regression model, with d^*_{it} as the dependent variable and covariates, including the duration variables and any other covariates.

For estimation using spell data with late entry, the stages are the same as those outlined above, with one modification and one warning. To fit models without frailty, you must drop all intervals prior to each subject's entry to the study. For example, if entry is in period e_i, you drop it if $t < e_i$. If you want to fit frailty models on the basis of discrete-time data with late entry, then be aware that the estimation procedure outlined does not lead to correct estimates. (The sample likelihood in the reorganized data does not account for conditioning for late entry here. You will need to write your own likelihood function by using `ml`; see [R] **Maximize**.)

To derive predicted hazard rates, use the `predict` command. For example, after `logistic` or `cloglog`, use `predict, pr`. After `xtlogit` or `xtcloglog`, use `predict, pu0` (which predicts the hazard assuming the individual effect is equal to the mean value). Estimates of the survivor function, S_{it}, can then be derived from the predicted hazard rates, p_{it}, because $S_{it} = (1-p_{i1})(1-p_{i2})(\cdots)(1-p_{it})$.

Acknowledgment

We thank Stephen Jenkins of the London School of Economics and Political Science for drafting this initial entry.

References

Allison, P. D. 2014. *Event History and Survival Analysis*. 2nd ed. Newbury Park, CA: SAGE.

Jenkins, S. P. 1995. Easy estimation methods for discrete-time duration models. *Oxford Bulletin of Economics and Statistics* 57: 129–138. https://doi.org/10.1111/j.1468-0084.1995.tb00031.x.

——. 1997. sbe17: Discrete time proportional hazards regression. *Stata Technical Bulletin* 39: 22–32. Reprinted in *Stata Technical Bulletin Reprints*, vol. 7, pp. 109–121. College Station, TX: Stata Press.

Meyer, B. D. 1990. Unemployment insurance and unemployment spells. *Econometrica* 58: 757–782. https://doi.org/10.2307/2938349.

Prentice, R. L., and L. A. Gloeckler. 1978. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* 34: 57–67. https://doi.org/10.2307/2529588.

Also see

[ST] **stcox** — Cox proportional hazards model

[ST] **stcrreg** — Competing-risks regression

[ST] **streg** — Parametric survival models

[D] **expand** — Duplicate observations

[R] **cloglog** — Complementary log–log regression

[R] **logistic** — Logistic regression, reporting odds ratios

[XT] **xtcloglog** — Random-effects and population-averaged cloglog models

[XT] **xtlogit** — Fixed-effects, random-effects, and population-averaged logit models