Modeling interval-censored event-time data with Stata

Xiao Yang
Principal Statistician and Software Developer

StataCorp LLC

Northern European Stata Conference
August 29, 2025

Outline

Univariate interval-censored event-time data

- What are interval-censored event-time data?
- Methodology for interval-censored Cox model
- Applications using the `stintcox` command
- Postestimation features

Multivariate interval-censored event-time data

- What are multivariate interval-censored event-time data?
- Marginal Cox PH model for multivariate interval-censored data
- Applications using the `stmgintcox` command
- Postestimation features

Univariate interval-censored event-time data

Interval-censored event-time data

- The event of interest is not always observed exactly but is known to occur within some time interval. For example, cancer recurrence, time of COVID infection, etc.
- Interval-censored event-time data arise in many areas, including medical, epidemiological, economic, financial, and sociological studies.
- Interval-censored data might contain four types of censoring: left-censoring, right-censoring, interval-censoring, and no censoring.
- These data are usually stored in one of two formats: single record per subject or multiple records per subject.
- Ignoring interval-censoring may lead to biased estimates.

Four types of censoring

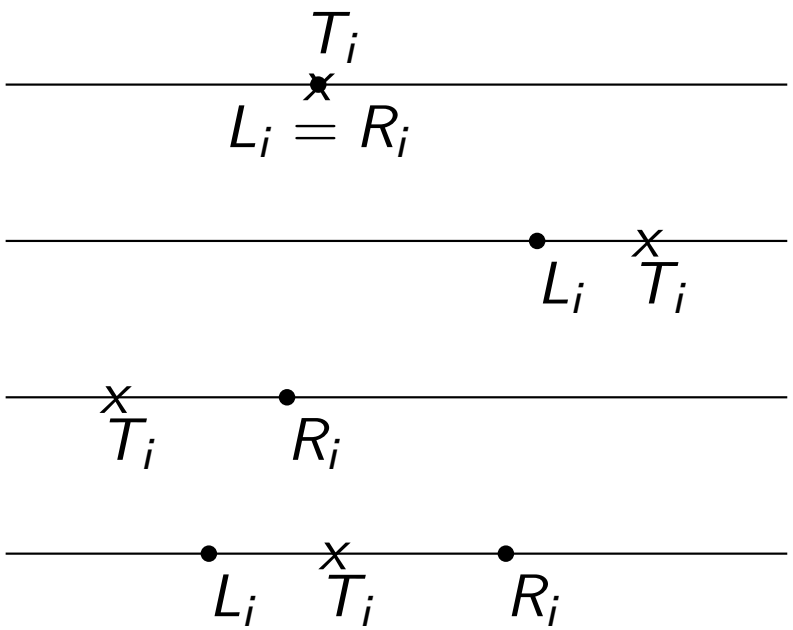
For each subject i, event time T_i is not always exactly observed. Instead, $(L_i, R_i]$ denotes the event-time interval in which T_i is observed.

No censoring
$L_i = R_i = T_i$

Right-censoring
$(L_i, R_i = +\infty)$

Left-censoring
$(L_i = 0, R_i]$

Interval-censoring
$(L_i, R_i]$



Storage formats for interval-censored data

Single-record-per-subject (single-record) format:

- contains one record per subject
- contains lower and upper endpoints of the event-time interval
- censoring type is determined by the event-time interval
- covariates are time-independent (baseline covariates)

	id	ltime	rtime	x1	x2	x3
1.	101	0	6	17	22	0
2.	102	4	9	12	22	1
3.	103	13	.	13	22	0

Multiple-record-per-subject (multiple-record) format:

- contains multiple records for some subjects
- contains an examination time and an event-status indicator for each record
- censoring type and the event-time interval are determined by the examination time and event status
- allows you to record time-varying covariates (TVCs) – variable x3 below

	id	time	status	x1	x2	x3
1.	101	6	1	17	22	0
2.	102	4	0	12	22	1
3.	102	6	0	12	22	0
4.	102	9	1	12	22	1
5.	103	13	0	13	22	0

Methods for analyzing interval-censored data

- Simple imputation methods (often biased)
- Nonparametric maximum-likelihood estimation (NPMLE) – Stata command `stintnp` (forthcoming), R function `ic_np`, SAS PROC ICLIFETEST
- Parametric regression models – Stata command `stintreg`, R function `ic_par`, SAS PROC LIFEREG
- Semiparametric Cox proportional hazards (PH) model – Stata command `stintcox`, R function `ic_sp`, SAS PROC ICPHREG
- Marginal Cox PH model for multivariate data – Stata command `stmgintcox`
- In what follows, we'll focus on the semiparametric Cox PH model with univariate and multivariate interval-censored event-time data in Stata

Cox PH model

- The Cox PH model was first introduced by Sir David Cox in 1972 and was used routinely to analyze uncensored and right-censored event-time data:

$$h(t;\boldsymbol{x}) = h_0(t)\exp(\boldsymbol{x}'\boldsymbol{\beta})$$

  where $h(t;\boldsymbol{x})$ is a hazard function at time t for a subject with a $p \times 1$ vector of covariate values $\boldsymbol{x}$, $h_0(t)$ is a baseline (with $\boldsymbol{x} = 0$) hazard function at time t, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients.

- It does not require parameterization of the baseline hazard function.
- Under the PH assumption, the hazard ratios are constant over time.

$$\frac{h(t;\boldsymbol{x}_i)}{h(t;\boldsymbol{x}_j)} = \frac{h_0(t)\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{h_0(t)\exp(\boldsymbol{x}_j'\boldsymbol{\beta})} = \exp(\boldsymbol{x}_i - \boldsymbol{x}_j)'\boldsymbol{\beta}$$

Challenges for interval-censored data

- The Cox model is challenging for interval-censored event-time data because none of the event times are observed exactly. In particular, the traditional partial-likelihood approach is not applicable.
- Several authors have proposed spline methods to fit the Cox model to interval-censored data and those methods have their limitations.
- The direct maximum-likelihood optimization using the Newton-Raphson algorithm is highly unstable.
- Zeng et al. (2016) developed a genuine EM algorithm for efficient nonparametric maximum-likelihood estimation (NPMLE) method to fit the Cox model to interval-censored data.

Genuine semiparametric interval-censored Cox model

- Suppose that the observed data consist of $(t_{li}, t_{ui}, \boldsymbol{x}_i')$ for $i = 1, \ldots, n$, where t_{li} and t_{ui} define the observed time interval and $\boldsymbol{x}_i$ records covariate values for a subject i.
- Under the NPMLE approach, the baseline cumulative hazard function $H_0(t)$ is regarded as a step function with nonnegative jumps $h_1, \ldots, h_m$ at $t_1, \ldots, t_m$, respectively, where $t_1 < \cdots < t_m$ are the distinct time points for all $t_{li} > 0$ and $t_{ui} < \infty$ for $i = 1, \ldots, n$.
- The observed-data likelihood function is

$$\prod_{i=1}^{n} \exp\left\{-\sum_{t_k \le t_{li}} h_k \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right\}\left[1-\exp\left\{-\sum_{t_{li} < t_k \le t_{ui}} h_k \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right\}\right]^{I(t_{ui}<\infty)} \tag{1}$$

- Let W_{ik} $(i = 1, \ldots, n; k = 1, \ldots, m)$ be independent latent Poisson random variables with means $h_k \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$. Define $A_i = \sum_{t_k \le t_{li}} W_{ik}$ and $B_i = I(t_{ui} < \infty) \sum_{t_{li} < t_k \le t_{ui}} W_{ik}$. The likelihood for the observed data $(t_{li}, t_{ui}, \boldsymbol{x}_i', A_i = 0, B_i > 0)$ is

$$\prod_{i=1}^{n} \prod_{t_k \le t_{li}} \Pr(W_{ik} = 0) \Big\{ 1 - \Pr\Big(\sum_{t_{li} < t_k \le t_{ui}} W_{ik} = 0 \Big) \Big\}^{I(t_{ui} < \infty)} \qquad (2)$$

- (1) and (2) are exactly equal. The maximization of a weighted sum of Poisson log-likelihood functions is strictly concave and has a closed-form solution for h_k's.

- We maximize (2) through an EM algorithm treating W_{ik} as missing data.
  1. In the E-step, we evaluate the posterior means of W_{ik}.
  2. In the M-step, we update $\boldsymbol{\beta}$ and h_k for $k = 1, ..., m$.
- This method allows a completely arbitrary baseline hazard function, and the results are consistent, asymptotically normal, and asymptotically efficient.
- This method has been implemented in Stata's `stintcox` command.

Supported features of stintcox

The stintcox command fits semiparametric Cox PH models to interval-censored event-time data. Its features:

- Single-record and multiple-record formats
- Time-varing covariates (TVCs)
- Stratification
- Full and reduced estimation of baseline hazard function
- Several methods for standard error estimation, including robust and cluster-robust standard errors
- Testing and graphical checks for PH assumption
- Predictions of baseline functions, martingale-like residuals, Cox–Snell-like residuals, and time-varying prediction
- Graphs of survivor, cumulative hazard, and hazard functions

Basic syntax

Single-record-per-subject data format

. `stintcox` [*indepvars*], `interval(`*t_l* *t_u*`)` ...

- *indepvars* specifies a list of covariates and is optional. You can fit a Cox model without any covariates.
- For single-record data, the `interval(`*t_l* *t_u*`)` option specifies variables *t_l* and *t_u* containing the respective lower and upper endpoints of the observed time interval.

Multiple-record-per-subject data format

. stintcox [*indepvars*], id(*idvar*) time(*timevar*) status(*statusvar*) ...

- For multiple-record data, the id(*idvar*) option specifies variable *idvar* recording subject identifiers, time(*timevar*) specifies the examination time *timevar*, and status(*statusvar*) specifies the event-status indicator *statusvar*.
- Note for Stata users: Unlike right-censored survival-time data, st setting interval-censored data is not necessary, and any st settings will be ignored by the stintcox command.

Modified Bangkok IDU Preparatory Study

A cohort study of injecting drug users in Thailand.

- 1124 subjects were initially negative for HIV-1 virus.
- They were followed and tested for HIV approximately every four months.
- The event of interest was time to HIV-1 seropositivity.
- The study aim is to identify the factors that influence time to HIV infection.
- Two versions of the data:
  - single-record dataset contains all baseline covariates;
  - multiple-record dataset contains both baseline covariates and TVCs.

Single-record-per-subject data

```
. webuse idu, clear
(Modified Bangkok IDU Preparatory Study)
. generate id = _n
. format ltime rtime age_mean %6.2f
. list id ltime rtime age_mean male needle inject jail ///
> if id >= 271 & id <= 274, noobs
```

id	ltime	rtime	age_mean	male	needle	inject	jail
271	22.00	.	-6.46	Yes	Yes	No	No
272	3.80	9.41	8.54	No	No	No	Yes
273	20.66	.	-11.46	Yes	Yes	No	No
274	0.00	3.87	-4.46	Yes	Yes	Yes	Yes

Fitting interval-censored Cox model with single-record data

```
. stintcox age_mean i.male i.needle i.inject i.jail, interval(ltime rtime)
note: using adaptive step size to compute derivatives.

Performing EM optimization (showing every 100 iterations):
Iteration 0:   Log likelihood = -1086.2564
Iteration 100: Log likelihood = -597.65634
Iteration 200: Log likelihood = -597.57555
Iteration 295: Log likelihood = -597.56443

Computing standard errors: ........................... done

Interval-censored Cox regression                  Number of obs     =  1,124
Baseline hazard: Reduced intervals                       Uncensored =      0
                                                     Left-censored =     41
Event-time interval:                                 Right-censored =    991
  Lower endpoint: ltime                              Interval-cens. =     92
  Upper endpoint: rtime
                                                  Wald chi2(5)      =  17.10
Log likelihood = -597.56443                       Prob > chi2       = 0.0043

-more-
```

Fitting interval-censored Cox model with single-record data (cont.)

	Haz. ratio	OPG std. err.	z	P>|z|	[95% conf.	interval]
age_mean	.9684341	.0126552	-2.45	0.014	.9439452	.9935582
male						
Yes	.6846949	.1855907	-1.40	0.162	.4025073	1.164717
needle						
Yes	1.275912	.2279038	1.36	0.173	.8990401	1.810768
inject						
Yes	1.250154	.2414221	1.16	0.248	.8562184	1.825334
jail						
Yes	1.567244	.3473972	2.03	0.043	1.014982	2.419998

Note: Standard error estimates may be more variable for small datasets and datasets with low proportions of interval-censored observations.

Handling TVCs

- Many datasets include TVCs, covariates that vary over time, such as age or transplantation status.
- TVCs also arise as a result of interacting baseline covariates with functions of time when the effect of a baseline covariate on the outcome is not constant over the follow-up time.
- TVCs are also useful for checking the PH assumption.
- During estimation, it is generally assumed that TVCs are external to the subject and are not directly related to the event status.
- We can incorporate all of these types of TVCs with `stintcox`.

- The `tvc()` option specifies the baseline variables to be included in the model as an interaction with a function of time to form TVCs.
- It is a convenience tool to speed up calculations and avoid splitting the data over many analysis times.
- The `texp()` option is used in conjunction with `tvc()` to specify the function of time that multiplies covariates specified in `tvc()` such as `texp(log(_t))`.
- Observed TVCs are recorded in a multiple-record format and are handled automatically by `stintcox`.

Testing the PH assumption

- One way of testing the PH assumption for a covariate (say, x_1) is to test whether the coefficient associated with that covariate is time invariant.
- This can be accomplished by including an interaction between this covariate and a function of time ($g(t)$) in the model and testing whether the corresponding coefficient equals zero, ($H_0 : \gamma_1 = 0$).

$$\begin{aligned} h(t) &= h_0(t)\exp\{\beta_1 x_1 + \gamma_1 g(t) x_1\} \\ &= h_0(t)\exp\left[\{\beta_1 + \gamma_1 g(t)\} x_1\right] \end{aligned}$$

- Let's include all covariates in the tvc() option to test the PH assumption individually and globally.
- We also specify the nohr option to present results as coefficients instead of the default hazard ratios.

```
. stintcox age_mean i.male i.needle i.inject i.jail, interval(ltime rtime) ///
> tvc(age_mean i.male i.needle i.inject i.jail) nohr
note: using adaptive step size to compute derivatives.

(iteration output omitted)
Interval-censored Cox regression                  Number of obs     =      1,124
Baseline hazard: Reduced intervals                       Uncensored =          0
                                                   Left-censored =         41
Event-time interval:                               Right-censored =        991
  Lower endpoint: ltime                            Interval-cens. =         92
  Upper endpoint: rtime
                                                  Wald chi2(10)     =      31.99
Log likelihood = -590.43386                       Prob > chi2       =     0.0004
-more-
```

	OPG Coefficient	std. err.	z	P>|z|	[95% conf.	interval]
main						
age_mean	-.0310177	.0233817	-1.33	0.185	-.076845	.0148097
male						
Yes	-1.271583	.4604788	-2.76	0.006	-2.174105	-.3690615
needle						
Yes	-.1819587	.3297493	-0.55	0.581	-.8282554	.464338
inject						
Yes	.6852961	.3431924	2.00	0.046	.0126513	1.357941
jail						
Yes	-.529615	.4021087	-1.32	0.188	-1.317734	.2585036

-more-

tvc						
age_mean	-.000129	.0017099	-0.08	0.940	-.0034804	.0032224
male						
Yes	.0884102	.042994	2.06	0.040	.0041434	.1726769
needle						
Yes	.0358545	.0238562	1.50	0.133	-.0109027	.0826118
inject						
Yes	-.0361192	.0228754	-1.58	0.114	-.0809541	.0087157
jail						
Yes	.0916036	.0348915	2.63	0.009	.0232176	.1599896

Notes: Standard error estimates may be more variable for small datasets and datasets with low proportions of interval-censored observations.
Variables in tvc equation interacted with _t.

Wald test that [tvc] = 0: chi2(5) = 13.3282 Prob > chi2 = 0.0205

Multiple-record-per-subject data

```
. webuse idu2, clear
(Modified Bangkok IDU Preparatory Study with time-varying variable jail_vary)
. format time age_mean %6.2f
. list id time is_seropos age_mean male needle inject jail_vary ///
> if id >= 271 & id <=274, sepby(id) noobs abbreviate(10) compress
```

id	time	is_seropos	age_mean	male	needle	inject	jail_vary
271	4.89	No	-6.46	Yes	Yes	No	No
271	9.31	No	-6.46	Yes	Yes	No	No
271	13.38	No	-6.46	Yes	Yes	No	Yes
271	17.97	No	-6.46	Yes	Yes	No	Yes
271	22.00	No	-6.46	Yes	Yes	No	No
272	3.80	No	8.54	No	No	No	Yes
272	9.41	Yes	8.54	No	No	No	No
273	3.93	No	-11.46	Yes	Yes	No	No
273	8.00	No	-11.46	Yes	Yes	No	No
273	12.07	No	-11.46	Yes	Yes	No	Yes
273	15.97	No	-11.46	Yes	Yes	No	Yes
273	20.66	No	-11.46	Yes	Yes	No	Yes
274	3.87	Yes	-4.46	Yes	Yes	Yes	Yes

Using stintcox with multiple-record data

Fit a Cox model using multiple-record data, including the time-varying covariate jail_vary

```
. stintcox age_mean i.male i.needle i.inject i.jail_vary, id(id) time(time) ///
> status(is_seropos)
note: time-varying covariates detected in the data; using method nearleft to
      impute their values between examination times.
note: using adaptive step size to compute derivatives.
```

(iteration output omitted)

```
Interval-censored Cox regression                Number of obs      =  6,453
Baseline hazard: Reduced intervals              Number of subjects =  1,124
                                                        Uncensored =      0
ID variable: id                                     Left-censored =     41
Examination time: time                             Right-censored =    991
Status indicator: is_seropos                       Interval-cens. =     92

                                                Wald chi2(5)       =  17.03
Log likelihood = -598.34887                     Prob > chi2        = 0.0044
-more-
```

Using stintcox with multiple-record data (cont.)

time	Haz. ratio	OPG std. err.	z	P>|z|	[95% conf.	interval]
age_mean	.9714605	.012757	-2.20	0.027	.9467762	.9967884
male						
Yes	.6678044	.1816576	-1.48	0.138	.3918353	1.138138
needle						
Yes	1.271409	.2275426	1.34	0.180	.8952546	1.805609
inject						
Yes	1.370672	.2575405	1.68	0.093	.9484142	1.980928
jail_vary						
Yes	1.440966	.2916178	1.81	0.071	.9691488	2.142481

Time varying: jail_vary
Note: Standard error estimates may be more variable for small datasets and datasets with low proportions of interval-censored observations.

Postestimation features

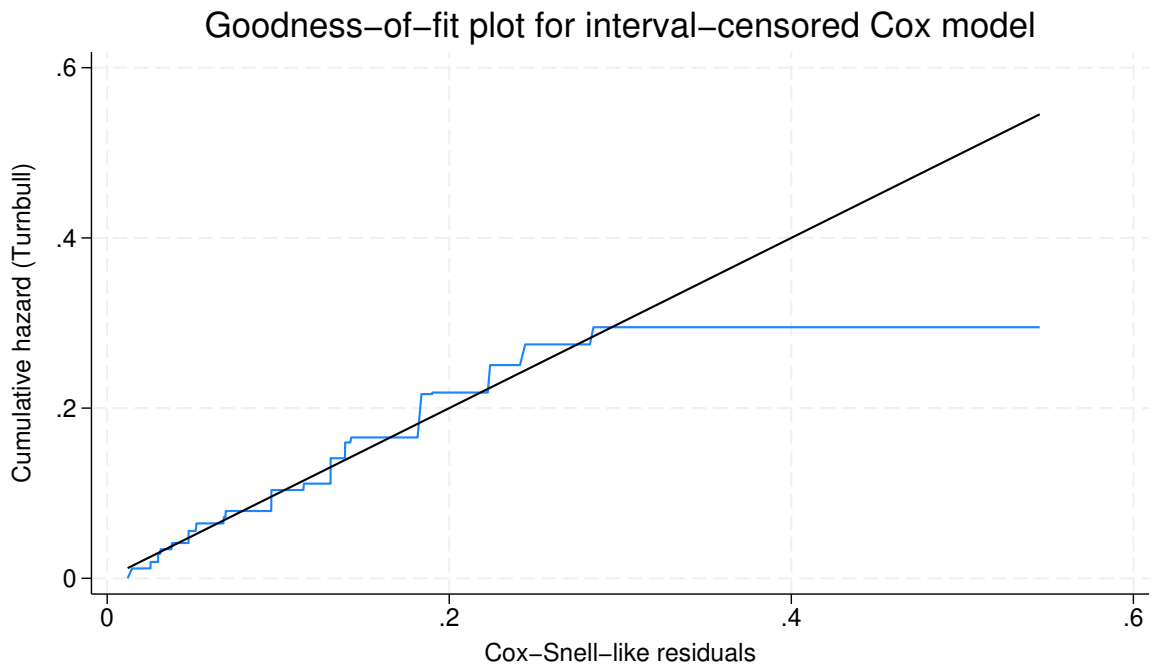
`stintcox` provides several postestimation features:

- Predictions of hazard ratios, linear predictions, and standard errors with support for TVCs
- Predictions of baseline survivor, baseline cumulative hazard, and baseline hazard contribution functions
- Prediction of martingale-like residuals and Cox–Snell-like residuals
- Goodness-of-fit plot
- Plots for survivor, hazard, and cumulative hazard functions

Goodness-of-fit (GOF) plot

- The GOF plot, produced by the `estat gofplot` command after `stintcox`, is used to assess the goodness of fit of the model visually.
- It plots the Cox–Snell-like residuals versus the estimated cumulative hazard function corresponding to these residuals.
- The Cox–Snell-like residuals form the 45° reference line. If the model fits the data well, the plotted estimated cumulative hazards should be close to the reference line.

```
. estat gofplot
```

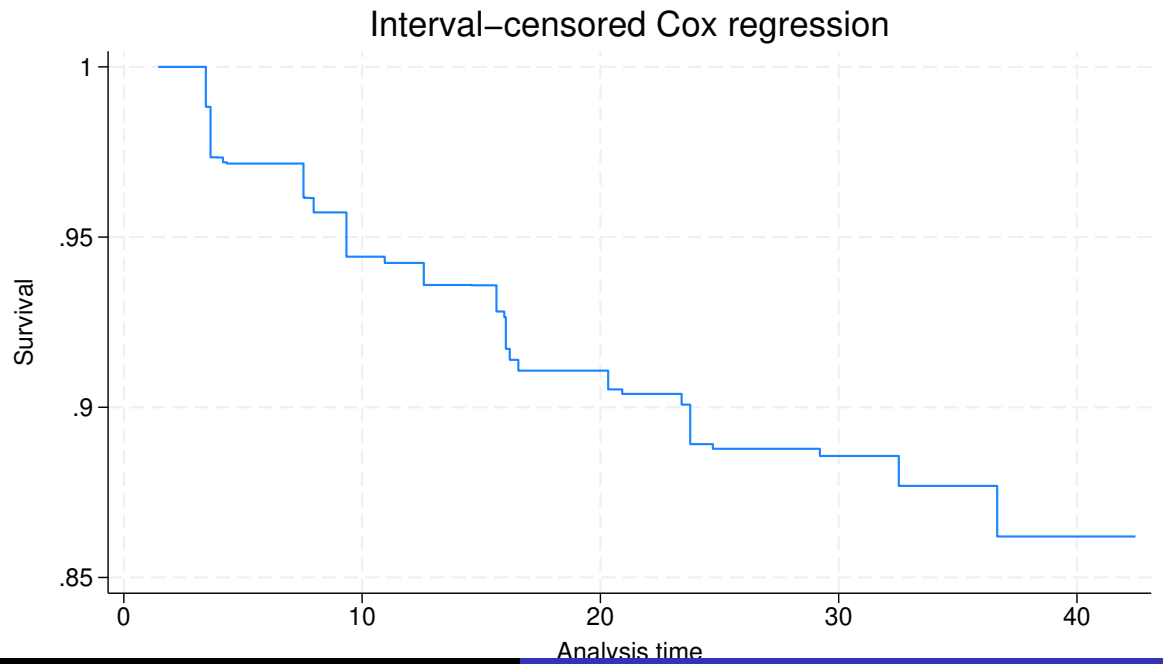


- The model appears to fit the data well, except perhaps in the tails.

Plots of survivor functions

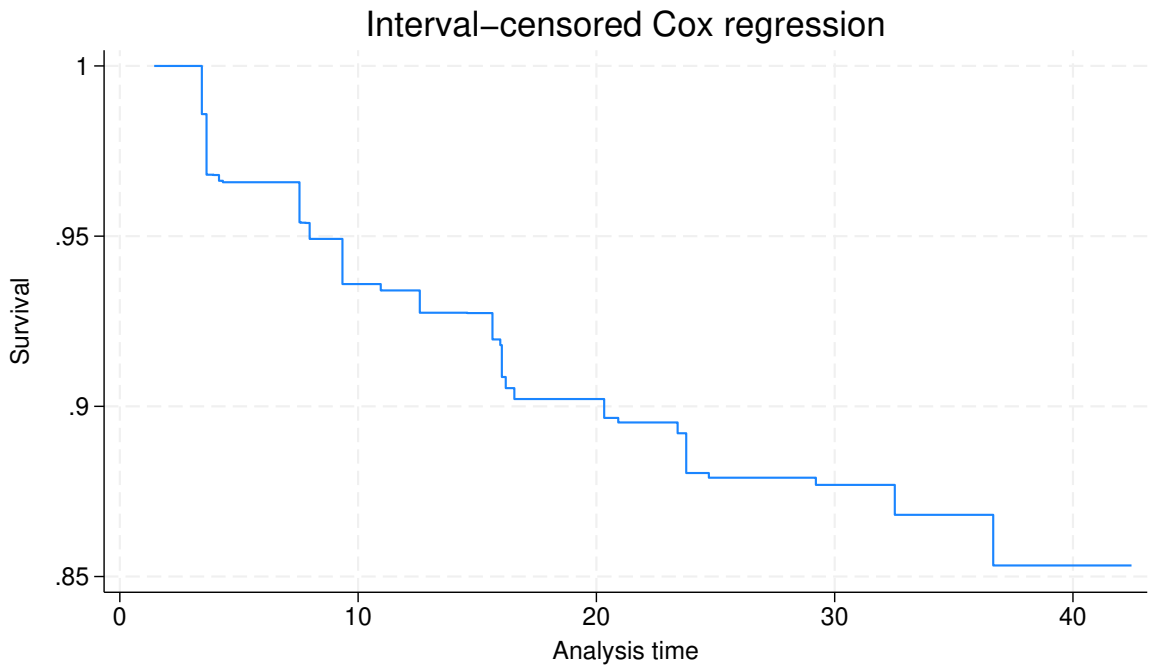
- We can use the `stcurve` command to plot the estimated survivor function.
- By default, `stcurve` evaluates the functions at the overall means of covariates.

```
. stcurve, survival
note: function evaluated at overall means of covariates.
```



- We can instead evaluate the function at time-specific means by specifying the `attmeans` option:

```
. stcurve, survival attmeans
note: function evaluated at time-specific means of covariates.
```



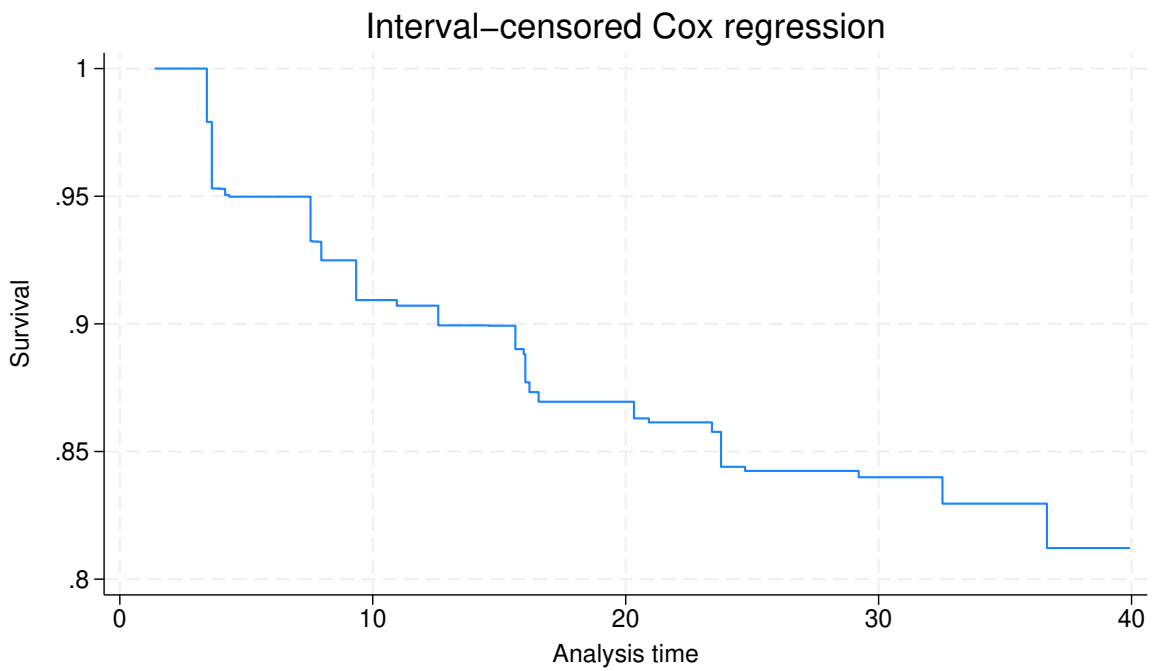
- Alternatively, we can evaluate the survivor function at specific values of covariates that include TVCs via the `atframe()` option.
- Suppose we want to plot the survivor curve for an individual with the same covariate pattern as subject 2.
- We create a new frame called `id2` and use the `frame put` command to copy the relevant information to the new frame.
- We `list` the data in frame `id2`.

```
. frame put time age_mean male needle inject jail_vary if id==2, into(id2)
. frame id2: list
```

	time	age_mean	male	needle	inject	jail_v~y
1.	4.1311475	-6.4617438	Yes	No	Yes	Yes
2.	8.2622951	-6.4617438	Yes	No	Yes	No
3.	12.295082	-6.4617438	Yes	No	Yes	No
4.	16.065574	-6.4617438	Yes	No	Yes	No
5.	20.098361	-6.4617438	Yes	No	Yes	No
6.	24.262295	-6.4617438	Yes	No	Yes	No

- We use the `atframe()` option to graph the survivor curve for this particular profile:

```
. stcurve, survival atframe(id2)
note: function evaluated at specified values of selected covariates and overall
      means of other covariates (if any).
note: covariate values from frame id2 used to evaluate function.
```



Multivariate interval-censored event-time data

Multiple events and clustering

- Multivariate interval-censored event-time data often arise when the study subjects can experience more than one type of event such as the onsets of diabetes and hypertension.
- Clustered data, where there are multiple subjects per cluster, can also be viewed as multivariate event-time data.
- Proper statistical analysis of this type of data requires joint modeling of multiple event times, because the event times may be correlated.

Storage formats for multiple-event interval-censored data

Single-record-per-event (single-record) format:

	id	event	ltime	rtime	x1	x2	x3
1.	101	1	0	6	17	22	0
2.	102	1	4	9	12	22	1
3.	103	1	13	.	13	22	0
4.	101	2	3	9	17	22	1
5.	102	2	0	4	12	22	1
6.	103	2	7	.	13	22	1

Multiple-record-per-event (multiple-record) format:

	id	event	time	status	x1	x2	x3
1.	101	1	6	1	17	22	0
2.	102	1	4	0	0	12	22
3.	102	1	6	0	12	22	0
4.	102	1	9	1	12	22	1
5.	103	1	13	0	13	22	0
6.	101	2	3	0	17	22	1
7.	101	2	9	1	17	22	0
8.	102	2	4	1	12	22	0
9.	103	2	7	0	13	22	1
10.	103	2	13	0	13	22	0

Marginal PH models

- We'll focus on the marginal approach of Xu et al. (2023).
- Their proposed marginal PH models do not require any specification of the dependence structure between multiple event times, and the distribution of event times is estimated nonparametrically, which may lead to more robust inference.
- Marginal models also produce estimates of parameters that can be interpreted as population-average effects.
- Finally, computational algorithms for marginal models are faster and more stable than for random-effects models.

Parameter estimation for marginal PH models

- The model parameters are estimated separately for each event using the EM algorithm described in section *Genuine semiparametric interval-censored Cox model*.
- Then, the joint covariance matrix of all regression coefficients is estimated using a robust or clustered sandwich estimator based on the profile log pseudolikelihood.
  For details, see *Methods and formulas* in **[ST] stmgintcox** (`https://www.stata.com/manuals/ststmgintcox.pdf`).

Supported features of stmgintcox

The stmgintcox command fits marginal PH models to multivariate interval-censored event-time data. Its features:

- Single- and multiple-record-per-event data formats
- Flexible specification of models with event-specific covariates
- Time-varying covariates (TVCs)
- Stratification
- Robust and cluster-robust standard errors
- Powerful test for common effects
- Testing and graphical checks for PH assumption
- Predictions of baseline functions, martingale-like residuals, Cox-Snell-like residuals, and time-varying prediction
- Graphs of survivor, cumulative hazard, and hazard functions

Basic syntax, common covariates

Single-record-per-event data

. stmgintcox *indepvars*, id(*idvar*) event(*eventvar*) interval(*t_l t_u*) ...

Multiple-record-per-event data

. stmgintcox *indepvars*, id(*idvar*) event(*eventvar*) time(*timevar*) status(*statusvar*) ...

- The basic syntax is similar to the univariate case, except that the id() and event() options are required.
- The id(*idvar*) option specifies a subject identifier *idvar*.
- The event(*eventvar*) option specifies an event identifier *eventvar*.

Basic syntax, event-specific covariates

Single-record-per-event data

. stmgintcox [*indepvars*] ([*event1*:] [*indepvars1*]) ([*event2*:] [*indepvars2*]) ...,
id(*idvar*) event(*eventvar*) interval(*t_l t_u*) ...

Multiple-record-per-event data

. stmgintcox [*indepvars*] ([*event1*:] [*indepvars1*]) ([*event2*:] [*indepvars2*]) ...,
id(*idvar*) event(*eventvar*) time(*timevar*) status(*statusvar*) ...

- *indepvars* specifies a list of common covariates across all events.
- *indepvar1* specifies a list of covariates for event *event1*.
- *indepvar2* specifies a list of covariates for event *event2*.

Flexible ways to specify models

- All covariates x1, x2, and x3 are common across all events:

```
. stmgintcox x1 x2 x3
```

- Covariate x1 is common across all events, but x2 is included only for "event2" and x3 is only for "event3":

```
. stmgintcox x1 ("event2": x2) ("event3": x3)
```

or, equivalently,

```
. stmgintcox ("event1": x1) ("event2": x1 x2) ("event3": x1 x3)
```

- Event-specific TVCs via the tvc() and texp() options:

```
. stmgintcox ("event1": x1, tvc(x1)) ///
             ("event2": x2, tvc(x2) texp(log(_t))) ///
             ("event3": x3)
```

Atherosclerosis Risk in Communities study

A cohort study of 14751 Caucasian and African-American individuals from four US communities.

- The participants were followed over time and assessed for both diabetes and hypertension during several follow-up exams.
- The exact onset times of these diseases were not observed, but they are known to fall in intervals between doctor visits.
- The goal is to identify the factors that influence time to onset of diabetes and hypertension.
- The factors of interest include three demographic variables – `race`, `male`, and `community`– and five risk factors: `age`, `bmi`, `glucose`, `sysbp`, and `diabp`.
- Two versions of the data: single- and multiple-record-per-event datasets. The dataset used for demonstration is simulated based on the above study.

```
. clear
. webuse aric
(Simulated ARIC data)
. format bmi glucose %6.2f
. list if id==180
```

359.

id	event	ltime	rtime	age	commun~y	male	race
180	Diabetes	1532	.	49	Forsyth	No	Black

bmi	glucose	sysbp	diabp
28.32	95.47	136	79

360.

id	event	ltime	rtime	age	commun~y	male	race
180	Hypertension	319	1532	49	Forsyth	No	Black

bmi	glucose	sysbp	diabp
28.32	95.47	136	79

Using common covariates for both events

```
. stmgintcox age i.male i.community i.race bmi glucose sysbp diabp,
> id(id) event(event) interval(ltime rtime) favorspeed
note: using fixed step size with a multiplier of 5 to compute derivatives.
note: using EM and VCE tolerances of 0.0001.
note: option noemhsgtolerance assumed.
```

(iteration output omitted)

```
Marginal interval-censored Cox regression          Number of events   =      2
Baseline hazard: Reduced intervals                 Number of subjects =    200
                                                   Number of obs      =    400
ID variable: id                                           Uncensored  =      0
Event variable: event                                Left-censored    =     47
Event-time interval:                                 Right-censored   =    240
  Lower endpoint: ltime                              Interval-cens.   =    113
  Upper endpoint: rtime
                                                   Wald chi2(20)      =  84.36
Log pseudolikelihood = -270.83984                  Prob > chi2        = 0.0000
-more-
```

Using common covariates for both events (cont.)

	Haz. ratio	Robust std. err.	z	P>|z|	[95% conf.	interval]
Diabetes						
age	.9552606	.0295589	-1.48	0.139	.8990481	1.014988
male						
Yes	.8084224	.2400335	-0.72	0.474	.451755	1.446684
community						
Jackson	1.597828	.6069935	1.23	0.217	.7588748	3.364265
Minneapolis	1.028054	.342976	0.08	0.934	.5346148	1.976929
Washington	1.407869	.5192024	0.93	0.354	.6833627	2.900504
race						
White	.4289702	.1273669	-2.85	0.004	.2397145	.7676444
bmi	1.116579	.034187	3.60	0.000	1.051545	1.185636
glucose	1.139753	.0303702	4.91	0.000	1.081756	1.200859
sysbp	1.020295	.0122308	1.68	0.094	.9966021	1.04455
diabp	.9928634	.0127512	-0.56	0.577	.9681835	1.018172

Using common covariates for both events (cont.)

Hypertension						
age	.9950085	.0225503	-0.22	0.825	.9517779	1.040203
male						
Yes	.6671401	.1599892	-1.69	0.091	.4169533	1.067448
community						
Jackson	.6085406	.1953944	-1.55	0.122	.3243246	1.141824
Minneapolis	.9040647	.2719638	-0.34	0.737	.5013468	1.630275
Washington	.674088	.2085739	-1.27	0.202	.3675707	1.23621
race						
White	1.261355	.425064	0.69	0.491	.6516152	2.441652
bmi	1.012196	.0195117	0.63	0.529	.9746672	1.05117
glucose	.989899	.0101396	-0.99	0.322	.9702238	1.009973
sysbp	1.075011	.0162901	4.77	0.000	1.043553	1.107418
diabp	1.025533	.0134835	1.92	0.055	.9994433	1.052303

Note: Standard error estimates may be more variable for small datasets and datasets with low proportions of interval-censored observations.

Using event-specific covariates

- From the model above, we can see that body mass index and glucose level are important risk factors for diabetes but not for hypertension.
- Conversely, systolic and diastolic blood pressure may play a role in the risk of hypertension but not the risk of diabetes.
- We can use different sets of covariates to model the two events. The following two specifications will yield the same results.

```
. stmgintcox age i.male i.community i.race ///
> ("Diabetes": bmi glucose) ///
> ("Hypertension": sysbp diabp), ///
> id(id) event(event) interval(ltime rtime) favorspeed
  (output omitted)
==============================================================================
. stmgintcox ("Diabetes": age i.male i.community i.race bmi glucose) ///
>            ("Hypertension": age i.male i.community i.race sysbp diabp), ///
>            id(id) event(event) interval(ltime rtime) favorspeed
note: using fixed step size with a multiplier of 5 to compute derivatives.
note: using EM and VCE tolerances of 0.0001.
note: option noemhsgtolerance assumed.

(iteration output omitted)

Marginal interval-censored Cox regression          Number of events   =      2
Baseline hazard: Reduced intervals                 Number of subjects =    200
                                                   Number of obs      =    400
ID variable: id                                           Uncensored =      0
Event variable: event                                Left-censored =     47
Event-time interval:                                Right-censored =    240
  Lower endpoint: ltime                             Interval-cens. =    113
  Upper endpoint: rtime
                                                   Wald chi2(16)      =  77.01
Log pseudolikelihood = -272.76543                  Prob > chi2        = 0.0000

-more-
```

	Haz. ratio	Robust std. err.	z	P>|z|	[95% conf.	interval]
Diabetes						
age	.9693495	.0293552	-1.03	0.304	.9134885	1.028626
male						
Yes	.8021755	.2273265	-0.78	0.437	.4603091	1.397942
community						
Jackson	1.549902	.6274179	1.08	0.279	.7010166	3.426733
Minneapolis	.9649113	.3361108	-0.10	0.918	.4875122	1.909806
Washington	1.36829	.5112313	0.84	0.401	.6578786	2.845842
race						
White	.4412767	.135994	-2.65	0.008	.2412044	.8073037
bmi	1.112781	.0314166	3.79	0.000	1.052878	1.176092
glucose	1.141379	.0304922	4.95	0.000	1.083153	1.202735

-more-

Hypertension						
age	.9945906	.0220662	-0.24	0.807	.9522686	1.038794
male						
Yes	.6229044	.1403048	-2.10	0.036	.4005846	.9686091
community						
Jackson	.606375	.1824113	-1.66	0.096	.3362643	1.093457
Minneapolis	.8873364	.2642854	-0.40	0.688	.4949546	1.590784
Washington	.6548935	.1999546	-1.39	0.166	.3599802	1.191414
race						
White	1.26674	.4058107	0.74	0.460	.6760798	2.373433
sysbp	1.072573	.0149785	5.02	0.000	1.043614	1.102336
diabp	1.025091	.0138294	1.84	0.066	.9983414	1.052558

Note: Standard error estimates may be more variable for small datasets and datasets with low proportions of interval-censored observations.

Postestimation features

`stmgintcox` provides several postestimation features:

- Powerful test for and estimation of a common covariate effect across all events
- Predictions of hazard ratios, linear predictions, and standard errors of linear predictions with support for TVCs
- Predictions of baseline survivor, baseline cumulative hazard, and baseline hazard contribution functions
- Predictions of martingale-like residuals and Cox–Snell-like residuals
- Goodness-of-fit plots for all events or specific ones
- Event-specific plots of survivor, hazard, and cumulative hazard functions

Powerful test for common covariate effects

- The `estat common` command estimates an optimal weighted average effect of a covariate across all events and conducts a test to determine whether this average effect is zero.

```
. estat common age i.male
    _avg_age: .294*[Diabetes]age + .706*[Hypertension]age
 _avg_1_male: .36*[Diabetes]1.male + .64*[Hypertension]1.male
```

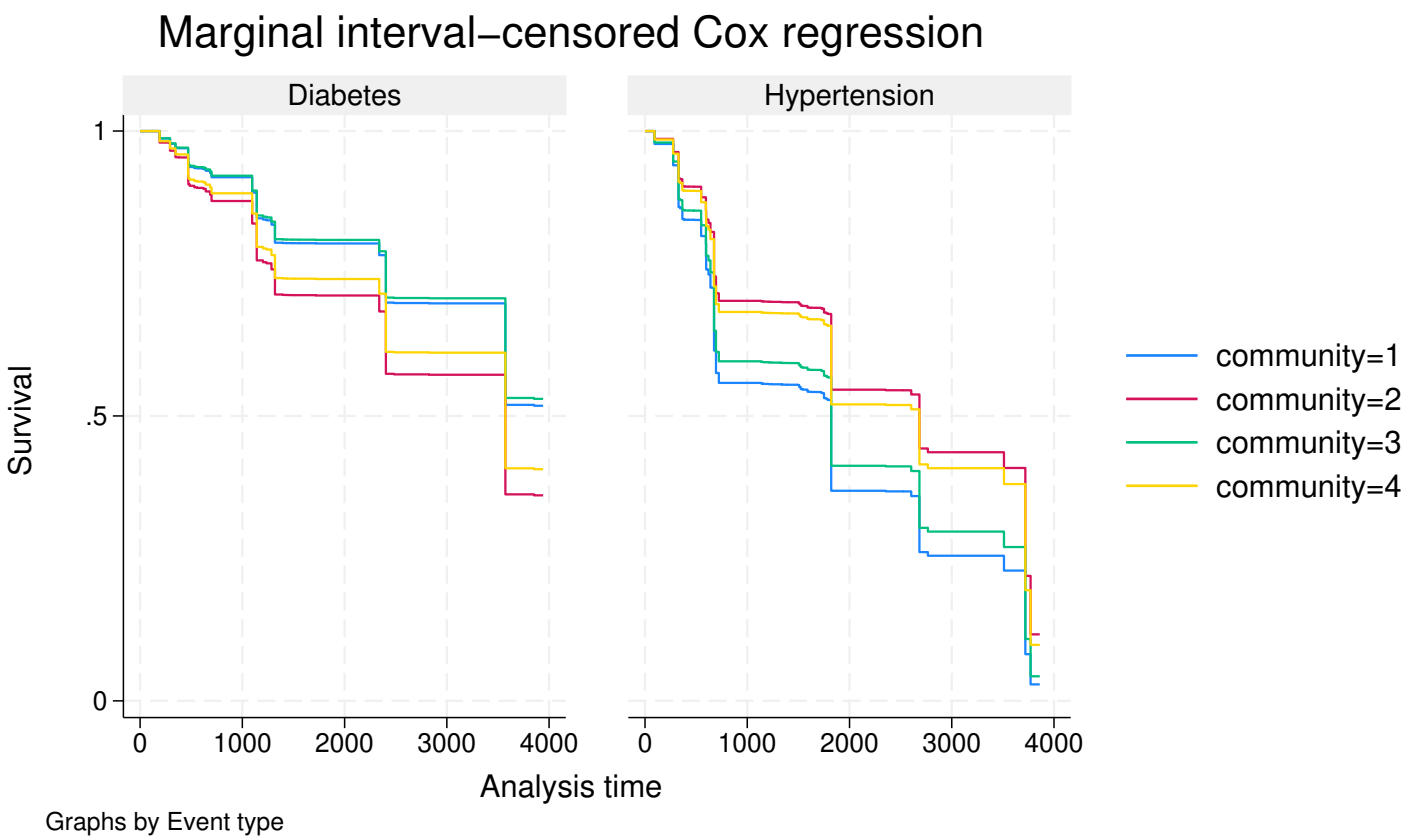
	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
_avg_age	-.0129749	.0200839	-0.65	0.518	-.0523386	.0263887
_avg_1_male	-.3823701	.1925874	-1.99	0.047	-.7598346	-.0049057

- When the effect of a covariate is similar across different events, the reported test is more powerful than the classic multivariate Wald test.

Plots of survivor functions

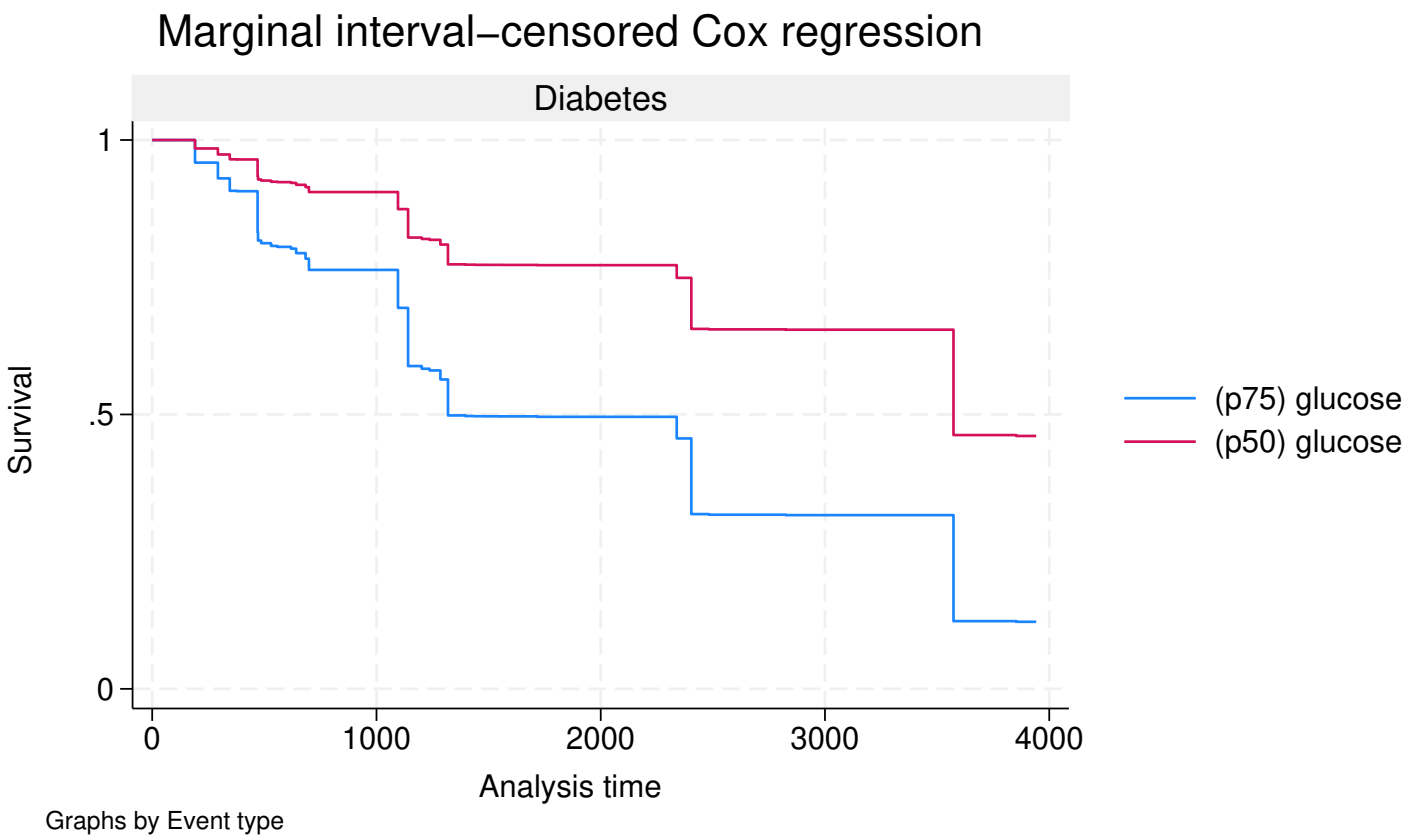
Plots of survivor functions for both events for different communities:

```
. stcurve, survival at(community=(1 2 3 4))
note: function evaluated at specified values of selected covariates and
      overall means of other covariates (if any) for specified events.
```



Plots of survivor functions for diabetes at two different glucose levels:

```
. stcurve, survival at((p75) glucose) at((p50) glucose) event("Diabetes")
note: function evaluated at specified values of selected covariates and
      overall means of other covariates (if any) for specified event.
```



Summary

With `stintcox`, you can

- fit genuine semiparametric Cox PH models to univariate interval-censored event-time data in one of two storage formats (single- or multiple-record-per-subject);
- incorporate TVCs in the model and use them to test the PH assumption;
- choose from several methods for standard error computation;
- obtain diagnostic measures, predictions, and much more after fitting the model;
- access convenient graphical tools for assessing the goodness of fit of the model and for plotting the survivor, cumulative hazard, and hazard functions; and
- incorporate TVCs in predictions, including survivor and other functions.

With `stmgintcox`, you can

- fit marginal Cox PH models to multivariate interval-censored event-time data, including multiple events and clustering;
- specify flexible models with event-specific covariates;
- incorporate TVCs in the model and use them to test the PH assumption;
- perform a more powerful test than the classic multivariate Wald test for testing and estimating the average effect of a covariate across all events;
- obtain diagnostic measures, predictions, and much more after fitting the model;
- access convenient graphical tools for assessing the goodness of fit of the model and for plotting the survivor, cumulative hazards, and hazard functions; and
- incorporate TVCs in predictions, including survivor and other functions.

References

Xu, Y., D. Zeng, and D. Y. Lin (2023). Marginal proportional hazards models for multivariate interval-censored data. *Biometrika 110*, 815–830.

Zeng, D., L. Mao, and D. Lin (2016). Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika 103*, 253–271.

More resources

www.stata.com/features/overview/interval-censored-cox-model/
www.stata.com/new-in-stata/marginal-interval-censored-cox-model/
www.stata.com/manuals/ststintcox.pdf
www.stata.com/manuals/ststintcoxpostestimation.pdf
www.stata.com/manuals/ststmgintcox.pdf
www.stata.com/manuals/ststmgintcoxpostestimation.pdf
www.stata.com/manuals/stestatgofplot.pdf
www.stata.com/manuals/ststcurve.pdf
www.stata.com/flyers/stintcox19.pdf

Questions?

Please contact **tech-support@stata.com** if you have any Stata questions:
https://www.stata.com/support/tech-support/contact/