Fitting Cox Proportional-Hazards Model for Interval-Censored Event-Time Data

Cox 比例风险模型在区间删失数据中的应用

杨筱

StataCorp LLC

2021 年 8 月 20 日

演讲大纲

- 什么是生存分析？
- 什么是区间删失数据？
- 半参数 Cox 比例风险模型在区间删失数据中的应用
- `stintcox` 命令简介
- `stintcox` 的后估命令功能简介
- Cox 比例风险假定的图形检验
- 结语

什么是生存分析？

- 生存分析（**Survival analysis**）又被称为久期分析（**duration analysis**），风险分析（**hazard analysis**），信度分析（**reliability analysis**），失效时间分析（**failure-time analysis**）等，是描述，测量和分析事件的特征，寻找其发生的原因，并对生存以及事件发生的时间进行预测的分析方法。
  - 一个刚确诊的癌症患者还有多少时间存活？
  - 癌症复发时间和哪些因素有关？
  - 一个上市公司退市的风险有多大？
  - 一个企业存活时间长短和哪些因素有关？
  - 婚姻的维系时间长短和哪些因素有关？
- 生存分析探索的因变量是一个包含了删失或者截除数据的事件时间变量。我们研究的是哪些因素影响了事件的发生速度及生存时间的长短。

生存分析的基本概念

- 生存时间 (**survival time**)：又叫关注事件时间 (**event time**), 泛指所关心的某现象的持续时间。生存时间分为两类：
  - 完全数据 (**complete data**)：指从观察起点到发生“死亡”事件的时间。
  - 删失数据 (**censored data**)：指从观察起点到发生非“死亡”事件的时间。
- 删失 (**censoring**)：又叫截尾，删截，是指研究者只能观察调查期的两个时间节点之间的相关信息，造成了观察对象在时间窗口外的信息缺失，被称为删失。
- 预测变量 (**covariates**)：我们敢兴趣的影响生存时间的其他变量，包括时间无关变量 (**time-independent covariates**) 和时间相关变量 (**time-dependent covariates**)。
- 生存时间，删失，和预测变量是绝大多数生存分析数据的要素。

生存分析的基本概念

- 生存函数 (**survival function**) $S(t)$：生存超过某个时间 t 的概率。

$$S(t) = Pr(T > t) = \int_t^{\infty} f(u)du = 1 - F(t)$$

- 风险函数（**hazard function**）$\lambda(t)$：关注事件在 t 时刻发生的概率。

$$\lambda(t) = \lim_{dt \to 0} \frac{Pr(t \leq T < t + dt | T \geq t)}{dt} = \frac{f(t)}{S(t)}$$

- 累积风险函数（**cumulative hazard function**）$\Lambda(t)$：关注事件到 t 时刻为止发生的概率。

$$\Lambda(t) = \int_0^t \lambda(u)du = -lnS(t)$$

- 生存曲线（**survival curve**）：以生存时间为横轴，生存率为纵轴，将各个时间点对应的生存率连接在一起的曲线图

删失的类型
Types of censoring

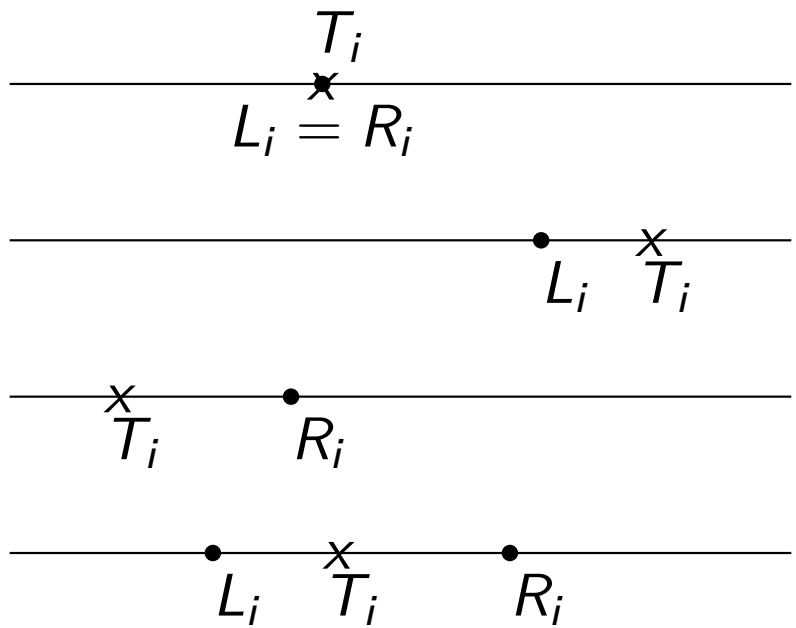
Event time T_i is not always exactly observed. $(L_i, R_i]$ denotes the interval（观察窗口期）in which T_i is observed.

No censoring
$L_i = R_i = T_i$

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

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

Interval-censoring
$(L_i, R_i]$



什么是区间删失数据？
What is interval-censored data?

- 关注事件的发生点不是总能被精确地观察到的，但是我们知道事件的发生是在一个时间区间里面。比如，癌症复发，何时感染上新冠病毒。
- 区间删失数据被越来越多的运用到各个领域，包括医学，流行病学，经济学，金融学和社会学的研究。
- 忽略区间删失会导致有偏估计。
- 区间删失数据中的每个观察对象可能包括各种不同删失：左删失，右删失，区间删失和没有删失。

区间删失数据集的类型

- Case I interval-censored data (**current status data**):
  每个观察对象只能被观察一次，我们只知道关注事件有没有在观察点前发生。所以每个观察对象的删失类型只可能是左删失或者右删失。
- Case II (**general**) interval-censored data:
  每个观察对象有两个或者多个观察点，包括关注事件的观察窗口的左右时间节点被记录在数据中。每个观察对象的删失类型是左删失，右删失，或者区间删失中的一种。

分析区间删失数据的方法

- 简单的插补方法（Simple imputation methods）
- 非参数极大似然估计（Nonparametric maximum-likelihood estimation）
- 参数回归模型（Parametric regression models）– `stintreg`
- 半参数 Cox 比例风险模型（Semiparametric Cox proportional hazards model）– `stintcox`
- 贝叶斯模型（Bayesian analysis）

什么是 Cox 比例风险模型？

- Cox 比例风险模型是由英国统计学家 D.R.Cox 于 1972 年首先提出的一种半参数回归模型。该模型以生存结局和生存时间为因变量，可同时分析众多因素对生存期的影响，能分析带有删失生存时间的资料，且不要求估计资料的生存分布类型。

$$\lambda(t;\mathbf{x}) = \lambda_0(t)\exp(\mathbf{x}'\boldsymbol{\beta})$$

- Cox PH 模型不需要假设基准风险 $\lambda_0(t)$ 的具体形式。

$$\frac{\lambda(t;\mathbf{x_i})}{\lambda(t;\mathbf{x_j})} = \frac{\lambda_0(t)\exp(\mathbf{x_i}'\boldsymbol{\beta})}{\lambda_0(t)\exp(\mathbf{x_j}'\boldsymbol{\beta})} = e^{(\mathbf{x_i}-\mathbf{x_j})'\boldsymbol{\beta}}$$

- 当比例风险的假设成立时，风险比率不随时间变化而变化。
- 基于 Cox 比例风险模型的以上特点，它被大量广泛的运用在右删失数据中。

Cox 模型在区间删失数据中遇到的挑战

- 在区间删失数据中，因为没有任何关注事件的发生点是确切被观察到的，所以 Cox 模型中传统用的部分似然函数 (partial-likelihood) 方法不能被直接运用。换一句话说，我们需要同时估计基准风险和预测变量的参数值。
- 一些研究提出使用基于 Newton-Raphson 算法的直接最大似然法，但是这种方法非常不稳定。
- 另外还有一些研究提出用 spline methods 估算基准风险的参数值，这些方法也都有他们自己的限制性，比如不同的样条函数会给出不同的结果。
- Zeng, Mao, and Lin (2016) 针对区间删失数据中 Cox 模型遇到的挑战，研发了一种全新的基于期望最大化算法（EM algorithm）的高效非参数最大似然估计方法（Nonparametric maximun-likelihood estimation，NPMLE)。这个全新的模型就是 Stata 17 推出的 `stintcox` 命令。

stintcox 命令的模型分析

- Suppose that the observed data consist of $(t_{li}, t_{ui}, \mathbf{x}_i)$ for $i = 1, \ldots, n$, where t_{li} and t_{ui} define the observed time interval and $\mathbf{x}_i$ records covariate values for a subject i.
- Under the NPMLE approach, the baseline cumulative hazard function Λ_0 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

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

stintcox 命令的模型分析

- Let W_{ik} $(i = 1, \ldots, n; k = 1, \ldots, m)$ be independent latent Poisson random variables with means $h_k \exp(\mathbf{x}_i \boldsymbol{\beta})$. 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 $(t_{li}, t_{ui}, \mathbf{x}_i, A_i = 0, B_i > 0)$ is

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

- (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.

stintcox 命令的模型分析

- We maximize (2) through an EM algorithm treating W_{ik} as missing data.
  - In the E-step, we evaluate the posterior means of W_{ik}.
  - In the M-step, we update β 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.

stintcox 命令概述

stintcox 命令是用于建立适用于区间删失数据的半参数 Cox 比例模型。

- 适用于 current-status data 和 general interval-censored data.
- 提供了四种计算标准误差估计的方法，并且能够在重播时（on replay）计算。
- 提供不同的选项以平衡命令执行的速度和准确度。
- 支持两种不同的选择基准风险函数时间间隔的方法。
- 支持分层模型（stratification）。

stintcox 命令基本语法

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

- Option `interval()` is required and is used to specify two time variables that contain the endpoints of the event-time interval.
- *indepvars* is optional. You can fit a Cox model without any covariates.
- `st` setting the data is not necessary and will be ignored.

研究实例

Modified Bangkok IDU Preparatory Study

- 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 exact time of HIV infection was not observed, but it was known to fall in intervals between blood tests with time variables `ltime` and `rtime`.
- We want to identify the factors that influence HIV infection. The covariates that we are interested in are centered age variable (`age_mean`), and history of drug injection before recruitment (`inject`).

研究实例数据集截图


```
. list in 701/710
```

	ltime	rtime	age_mean	inject
701.	41.049179	.	-1.4617438	Yes
702.	20.09836	.	3.5382562	No
703.	40.918034	.	5.5382562	No
704.	11.934426	16.065575	4.5382562	No
705.	32.327869	.	-10.461744	Yes
706.	40.360657	.	-5.4617438	No
707.	39.901638	.	-9.4617438	No
708.	24.065575	.	7.5382562	Yes
709.	28.163935	32.52459	-7.4617438	No
710.	0	16.196722	3.5382562	Yes

半参数 Cox 回归例子


```
. stintcox age_mean i.inject, interval(ltime rtime)
note: using adaptive step size to compute derivatives.
Performing EM optimization (showing every 100 iterations):
Iteration 0:    log likelihood = -1086.2564
  (output omitted)
Iteration 299:  log likelihood = -601.53336
Computing standard errors: ..............  done
Interval-censored Cox regression                  Number of obs     =  1,124
Baseline hazard: Reduced intervals                      Uncensored  =      0
                                                    Left-censored   =     41
                                                   Right-censored   =    991
                                                   Interval-cens.   =     92
                                                  Wald chi2(2)      =  11.18
Log likelihood = -601.53336                       Prob > chi2       = 0.0037
```

	Haz. ratio	OPG std. err.	z	P>|z|	[95% conf.	interval]
age_mean	.9657816	.0124711	-2.70	0.007	.9416454	.9905365
inject						
Yes	1.590116	.2847623	2.59	0.010	1.11942	2.25873

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

stintcox 的标准误差估计方法

- stintcox estimates VCE for regression coefficients using the profile log-likelihood, which is obtained by maximizing the likelihood by holding the regression coefficients fixed.

Type of VCE	Order of deriv.	Stepsize
vce(opg[,stepsize(adaptive)])	first-order	adaptive
vce(opg, stepsize(fixed [#]))	first-order	fixed
vce(oim[,stepsize(adaptive)])	second-order	adaptive
vce(oim, stepsize(fixed [#]))	second-order	fixed

标准误差估计实例

- 如果分析的数据集太小或者数据集里面的区间截失样本太小，选择不同的标准误差估计方法有时候会给出很不一样的结果。在这种情况下，你可能需要去比较这些不同的方法。
- `stintcox` 提供了在重放（on replay）时重新计算标准误差的功能，不需要再重新计算回归系数。

标准误差估计实例


```
. stintcox, vce(oim)
note: using adaptive step size to compute derivatives.
Computing standard errors: ................. done
Interval-censored Cox regression                  Number of obs     =  1,124
Baseline hazard: Reduced intervals                      Uncensored  =      0
                                                    Left-censored   =     41
                                                   Right-censored   =    991
                                                   Interval-cens.   =     92
                                                  Wald chi2(2)      =  11.18
Log likelihood = -601.53336                       Prob > chi2       = 0.0037
```

	Haz. ratio	OIM std. err.	z	P>|z|	[95% conf.	interval]
age_mean	.9657816	.0121666	-2.76	0.006	.9422274	.9899245
inject						
Yes	1.590116	.3285746	2.24	0.025	1.060572	2.384061

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

平衡命令执行的速度与准确度
favorspeed vs. favoraccuracy

- stintcox 在处理一些很大的数据集时可能会比较费时。非参数最大似然法和 EM 算法都是计算密集型方法。你最后需要的数据精度决定了命令执行的速度。
- 选项 favorspeed 和 favoraccuracy 为你提供了方法来权衡执行速度和准确度。
- stintcox 的默认选项是 favoraccuracy，以确保最后报告的精确度。但是在最初的探索阶段，你可以用 favorspeed 来快速得到结果。
- 当选择 favorspeed 时，stintcox 选用不那么严格的收敛准则。

favorspeed vs favoraccuracy 实例

```
. stintcox age_mean i.inject, 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.
Performing EM optimization (showing every 100 iterations):
Iteration 0:    log likelihood = -1086.2564
Iteration 31:   log likelihood = -602.62237
Computing standard errors: ..... done
Interval-censored Cox regression                  Number of obs     =  1,124
Baseline hazard: Reduced intervals                      Uncensored  =      0
                                                    Left-censored   =     41
                                                   Right-censored   =    991
                                                   Interval-cens.   =     92
                                                  Wald chi2(2)      =  11.19
Log likelihood = -602.62237                       Prob > chi2       = 0.0037
```

	Haz. ratio	OPG std. err.	z	P>|z|	[95% conf.	interval]
age_mean	.965774	.012463	-2.70	0.007	.9416534	.9905125
inject						
Yes	1.591654	.2848271	2.60	0.009	1.120794	2.260329

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

stintcox 后估命令功能简介

在进行模型估计后，stintcox 还提供了以下几个功能:

- Predictions of hazard ratios, linear predictions, and standard errors
- Predictions of baseline survivor, baseline cumulative hazard, and baseline hazard contribution functions
- Prediction of martingale-like residuals
- Plots for survivor, hazard, and cumulative hazard function

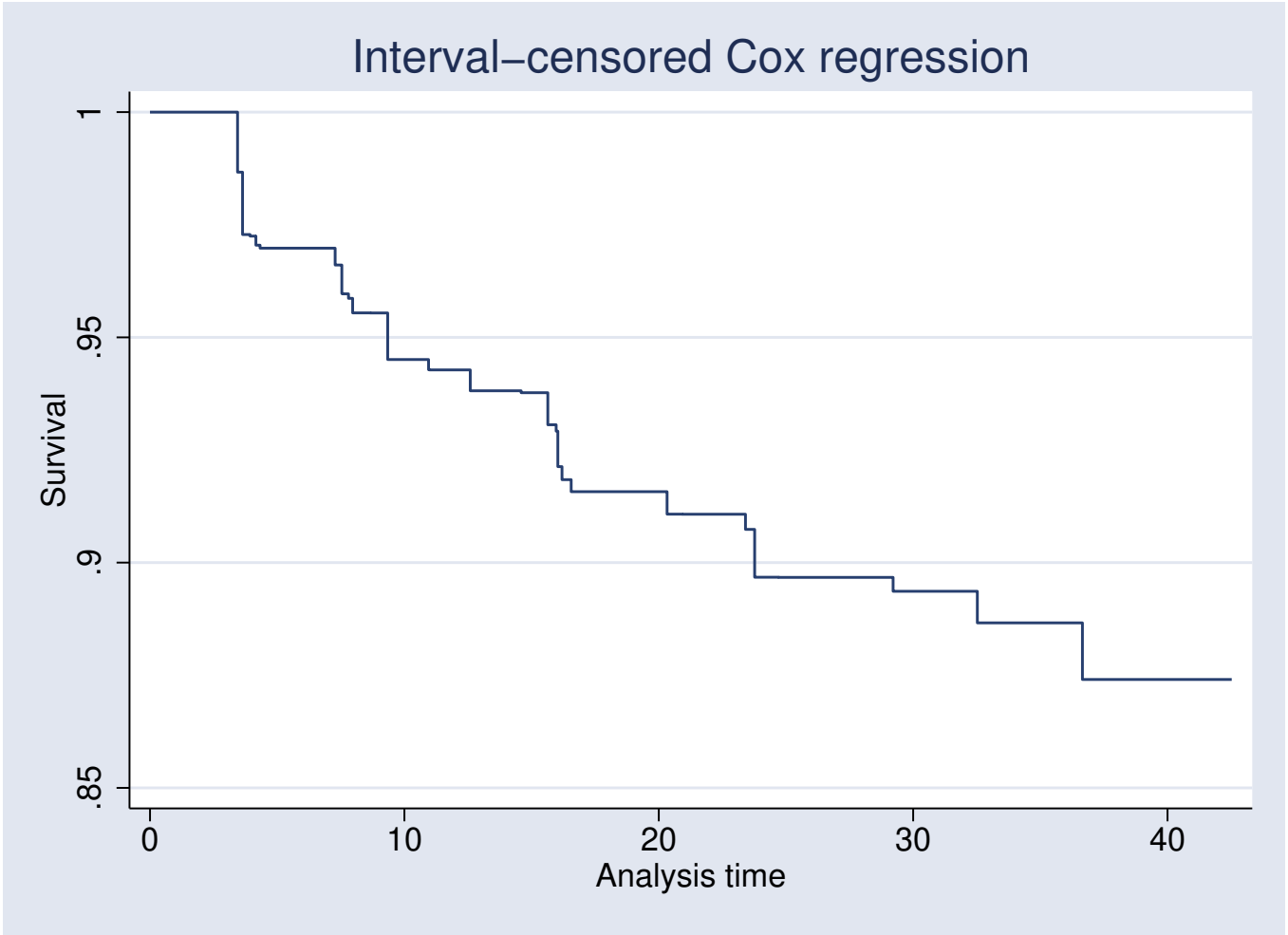
Predict baseline survival functions
预测基准生存函数

```
. stintcox age_mean i.inject, interval(ltime rtime)
  (output omitted)
. predict bs_l bs_u, basesurv
. list bs_l bs_u ltime rtime age_mean inject in 701/710
```

	bs_l	bs_u	ltime	rtime	age_mean	inject
701.	.8740674	0	41.049179	.	-1.4617438	Yes
702.	.9157519	0	20.09836	.	3.5382562	No
703.	.8740674	0	40.918034	.	5.5382562	No
704.	.9427818	.9213125	11.934426	16.065575	4.5382562	No
705.	.8936399	0	32.327869	.	-10.461744	Yes
706.	.8740674	0	40.360657	.	-5.4617438	No
707.	.8740674	0	39.901638	.	-9.4617438	No
708.	.896766	0	24.065575	.	7.5382562	Yes
709.	.8967278	.8866288	28.163935	32.52459	-7.4617438	No
710.	1	.9184227	0	16.196722	3.5382562	Yes

Graph baseline survival functions
绘制基准生存函数

```
. stcurve, survival at(age_mean=0 inject=0)
```



Cox 比例风险假定的图形检验

- stintphplot 用于绘制分组变量的“对数”-“对数”图。若“对数”-“对数”图中的曲线相互平行，则比例风险假定成立。
- stintcoxnp 对于每个分组变量分别绘制非参数生存曲线和 Cox 生存曲线。如果这两种生存曲线基本重合，则比例风险假定成立。
- 以上两个命令是独立于 stintcox 命令的，不需要在用这两个命令之前先执行 stintcox 命令。

stintphplot 命令的基本语法

`stintphplot, interval(`t_l t_u`) by()`

- Computes nonparametric estimates of the survivor function for each level of `by()` variable.

`stintphplot, interval(`t_l t_u`) by() adjustfor()`

- Fits a separate Cox model, which contains all covariates from the `adjustfor()` option, for each level of `by()` variable.

`stintphplot, interval(`t_l t_u`) strata() adjustfor()`

- Fits one stratified Cox model with all covariates from the `adjustfor()` option, then plots the estimated survivor function for each level of `strata()` variable.

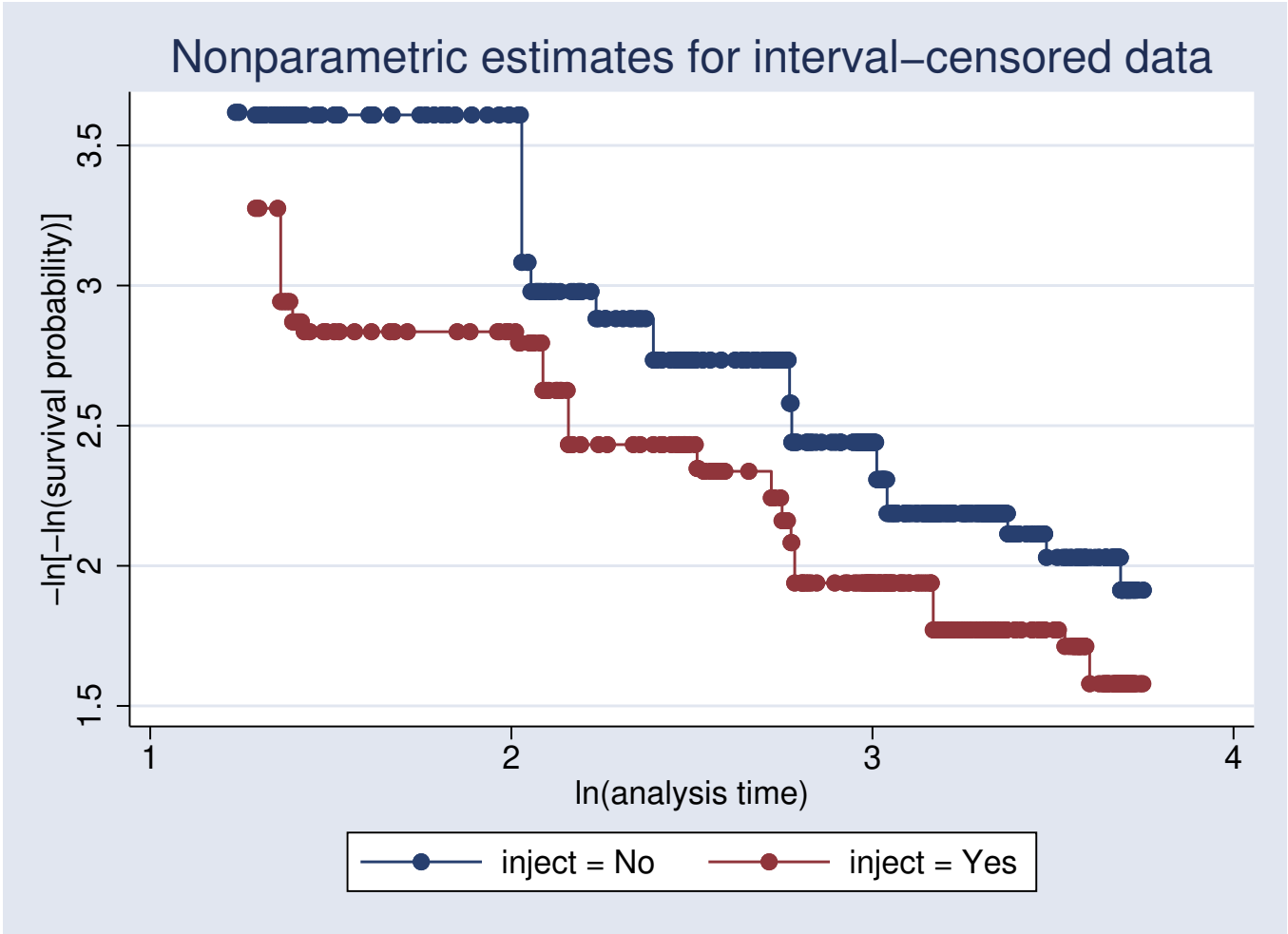
stintcoxnp 命令的基本语法

`stintcoxnp`, `interval`(t_l t_u) `by()` [`separate`]

- The nonparametric and Cox predicted survivor functions are plotted for each level of `by()` variable.
- Option `separate` produces separate plots of nonparametric and Cox predicted survivor functions for each level of `by()` variable.

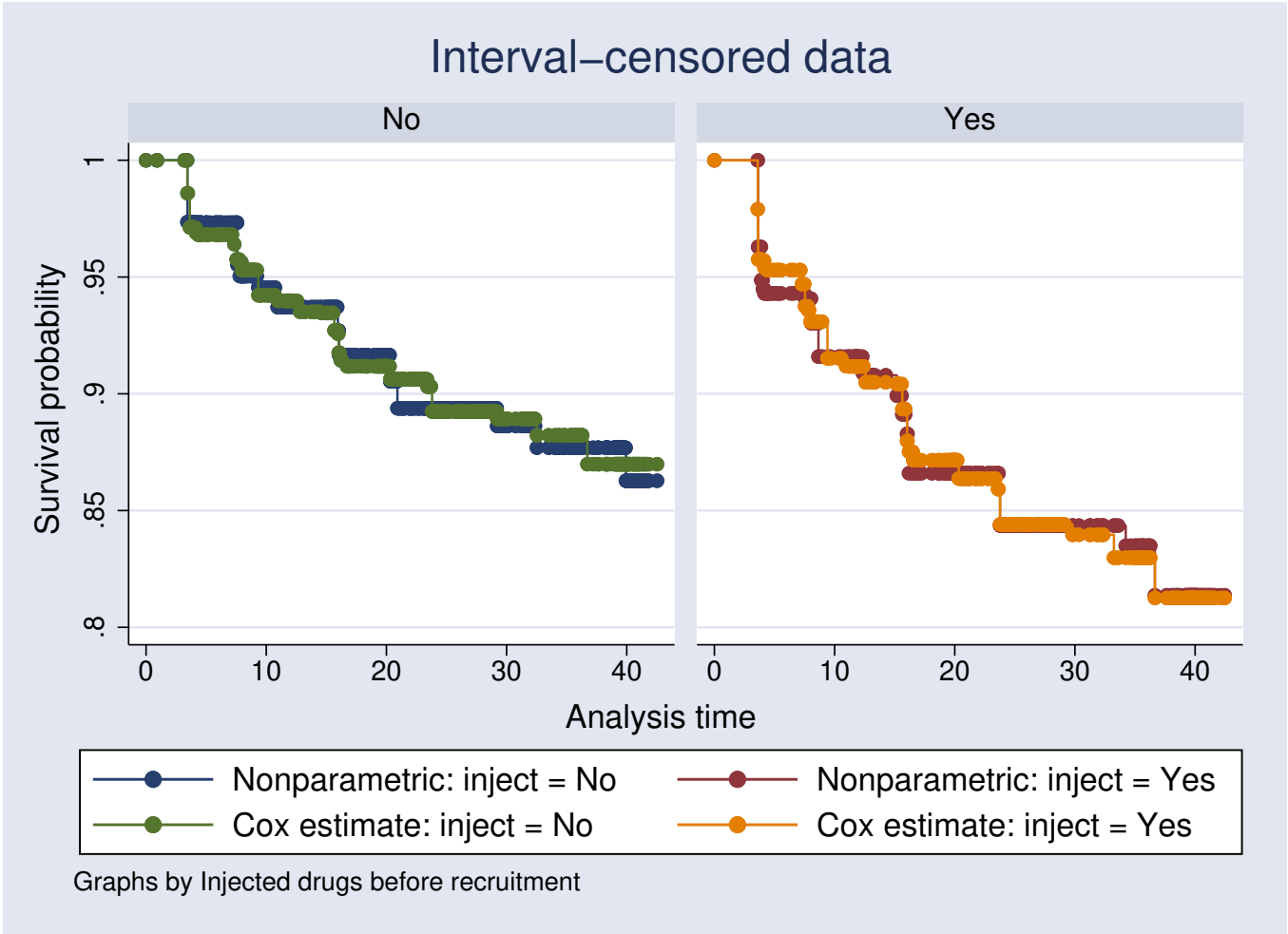
检验 Cox 比例风险假设（模型里面只有一个自变量）

```
. stintphplot, interval(ltime rtime) by(inject)
Computing nonparametric estimates for inject = No ...
Computing nonparametric estimates for inject = Yes ...
```



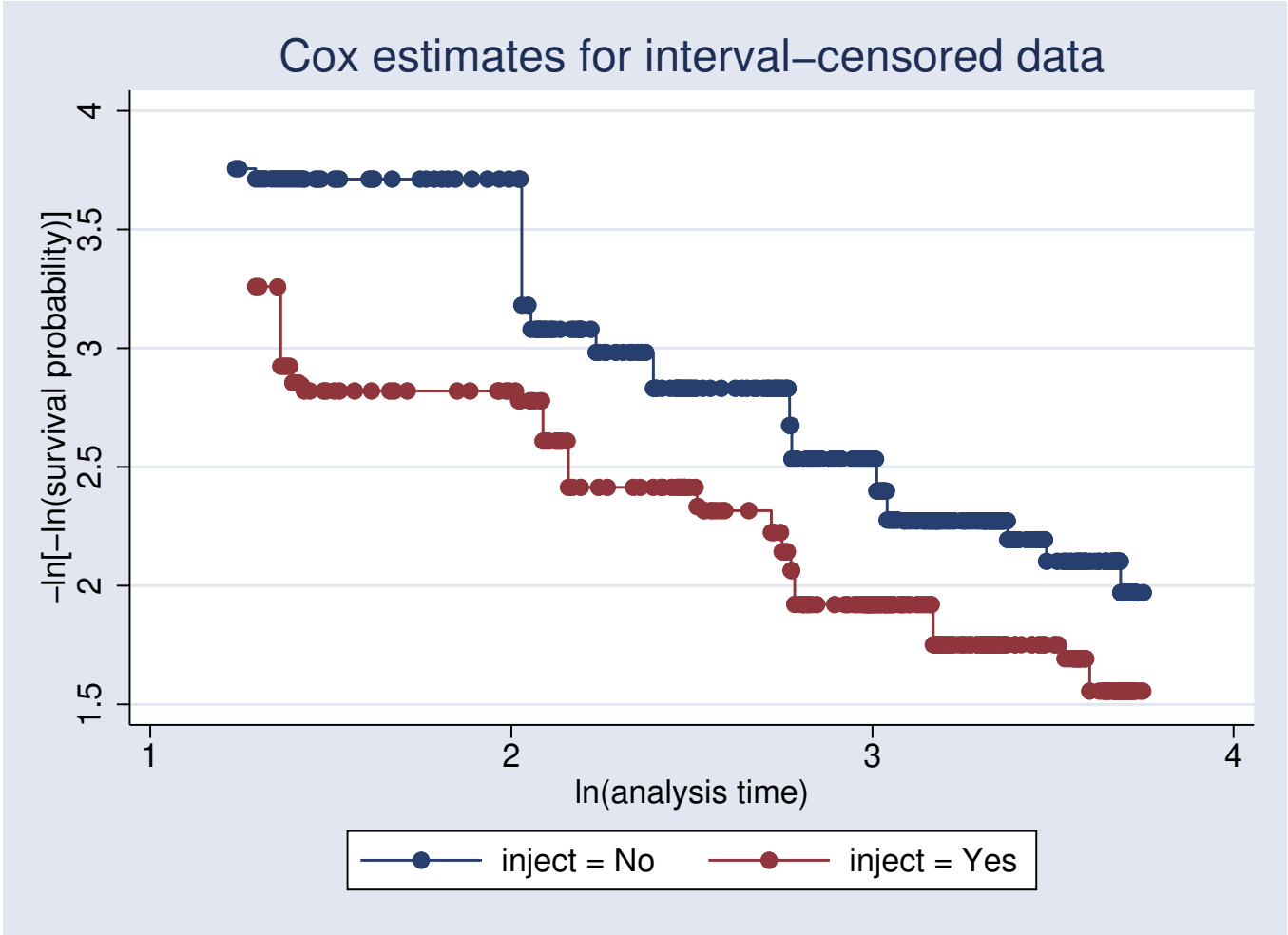
检验 Cox 比例风险假设（模型里面只有一个自变量）

```
. stintcoxnp, interval(ltime rtime) by(inject) separate
Computing nonparametric estimates ...
Computing Cox estimates ...
```



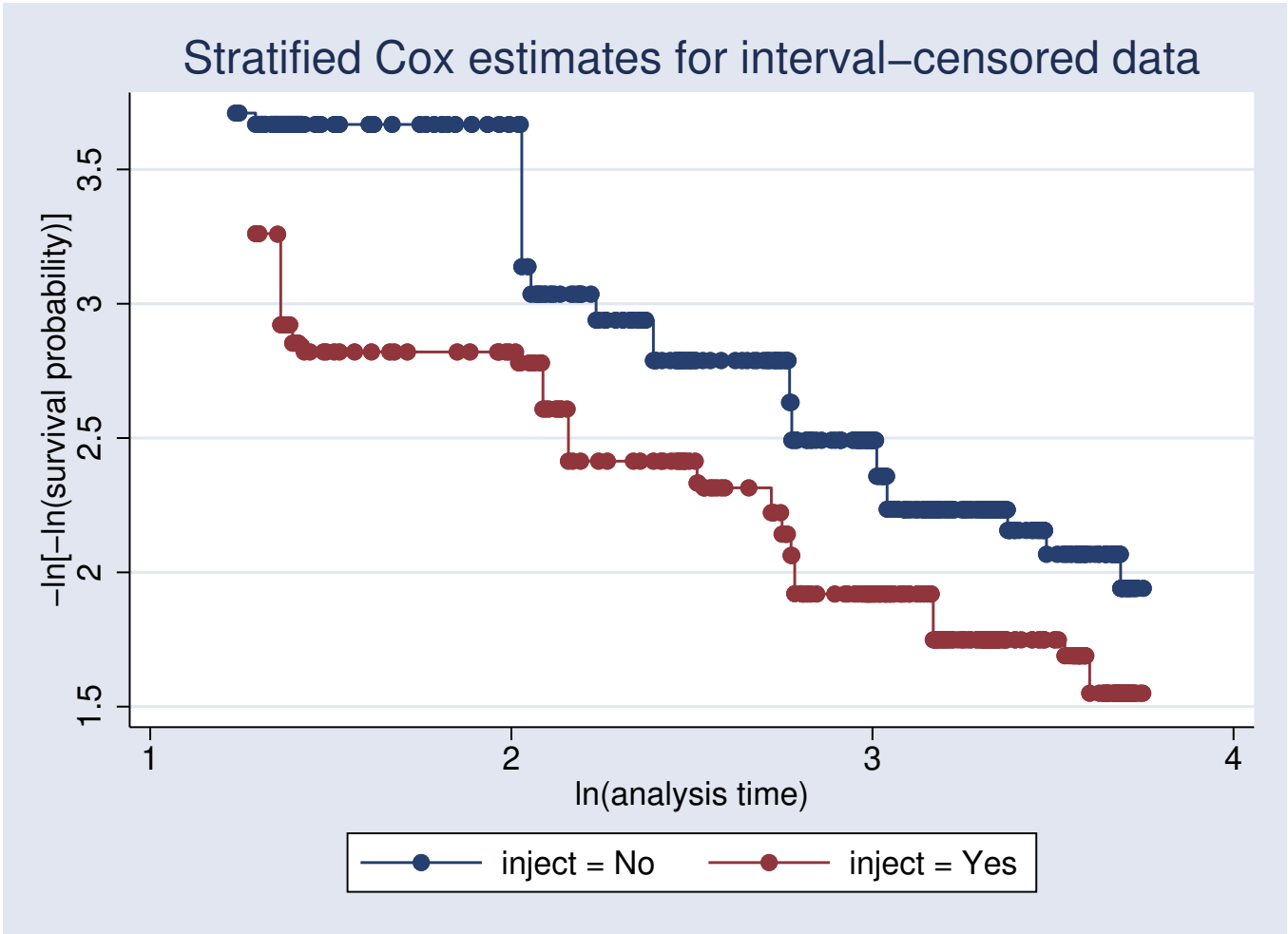
检验 Cox 比例风险假设（模型里面有多个自变量）

```
. stintphplot, interval(ltime rtime) by(inject) adjustfor(age_mean)
Fitting Cox model with covariates from option adjustfor()
for inject = No ...
Fitting Cox model with covariates from option adjustfor()
for inject = Yes ...
```



检验分层 Cox 模型的比例风险假设

```
. stintphplot, interval(ltime rtime) strata(inject) adjustfor(age_mean)
Fitting Cox model stratified on inject with covariates from option adjustfor()
...
```



结语

- Fit a genuine semiparametric Cox proportional-hazards model with time-independent covariates for two types of interval-censored data.
- Support different methods for standard-error computation.
- Support modeling of stratification.
- Support options to control the tradeoff between speed and accuracy.
- Support two ways to choose the time intervals to be estimated for baseline hazard function.
- Provide diagnostic measures, predictions, and much more after fitting the model.
- Provide graphical assessments for proportional-hazard assumption.