The Stata Journal (2017)
17, Number 4, pp. 1015–1023

Estimating receiver operative characteristic curves for time-dependent outcomes: The stroccurve package

Mattia Cattaneo
Department of Management
Information and Production Engineering
University of Bergamo
Bergamo, Italy
mattia.cattaneo@unibg.it

Paolo Malighetti
Department of Management
Information and Production Engineering
University of Bergamo
Bergamo, Italy
paolo.malighetti@unibg.it

Daniele Spinelli
Department of Management
Information and Production Engineering
University of Bergamo
Bergamo, Italy
daniele.spinelli@unibg.it

Abstract. Receiver operating characteristic (ROC) curves are an established method for assessing the predictive capacity of a continuous biomarker for a binary outcome. However, in some cases, outcomes are time dependent. Although the literature has proposed packages for performing ROC analysis of time-independent outcomes, a package is not yet available for analyzing the predictive capacity of continuous biomarkers when the binary outcome is time dependent. In this article, we present `stroccurve`, a new command for performing ROC analysis within a survival framework.

Keywords: st0509, stroccurve, receiver operating characteristic curves, time-dependent outcome, survival ROC

1 Introduction

Receiver operative characteristic (ROC) curves are a well-known method for assessing the predictive accuracy of a continuous biomarker. They provide estimates of sensitivity and one minus specificity of every possible cutoff in the biomarker distribution for a binary outcome. For an introduction to ROC curves, see Pepe, Longton, and Janes (2009).

While packages estimating the ROC curve for time-independent outcomes have already been developed for Stata users—that is, the `roccurve` package by Pepe, Longton, and Janes (2009)—a command that allows users to obtain sensitivity and specificity measure within a failure-time (survival) data framework is not yet available.

To this end, we present a new command for Stata users, `stroccurve`, that can assess the classification accuracy of a continuous biomarker when failures occur at different

points in time and when observations are subject to censoring. `stroccurve` is mainly based on the contribution of Heagerty, Lumley, and Pepe (2000).

This article is organized as follows: Section 2 illustrates the methods for assessing the predictive accuracy of a continuous biomarker within a survival framework, section 3 describes the syntax of the new command, and section 4 provides examples to outline the use of `stroccurve`.

2 Estimation of time-dependent ROC curves

Heagerty, Lumley, and Pepe (2000) defined two approaches to estimate ROC curves for failure-time data. The first applies Bayes' theorem and the Kaplan–Meier (KM) survival estimator (Kaplan and Meier 1958), while the second smooths the conditional survival function through the method provided by Akritas (1994).

In this section, the notation is the following:

- n: number of individuals
- T_i: failure time for individual i
- X_i: marker value for individual i
- C_i: censoring time for individual i
- $Z_i = \min(T_i, C_i)$: follow-up time for individual i
- d_i: censoring indicator for individual i. $d_i = 1$ if $T_i \leq C_i$ and $d_i = 0$ if $T_i > C$
- $D_i(t)$: failure status prior to time t. $D_i(t) = 1$ if $T_i \leq t$ and $D_i(t) = 0$ if $T_i > t$
- $T_n(t)$: unique levels of Z_i for observed events $D_i = 1$ in $Z_i \leq t$
- $S(t) = P(T > t)$: the survival function

2.1 KM method

The estimates of sensitivity (true positive rate, TP) and specificity (true negative rate, TN) for each possible value c of the biomarker X can be derived through Bayes' theorem,

$$\text{TP}(c,t) = P\{X > c | D(t) = 1\} = \frac{P(X > c)\{1 - S(t|X > c)\}}{1 - S(t)}$$

$$\text{TN}(c,t) = P\{X \leq c | D(t) = 0\} = \frac{P(X \leq c)S(t|X \leq c)}{S(t)}$$

where $P(X \leq c)$ is estimated by the biomarker empirical cumulative distribution $F_X(X_i) = 1/n \sum_i \mathbf{1}(X_i \leq c)$, while the estimates of the survival function $S(t)$ and of

the conditional survival function $S(t|X > c)$ are produced relying on the KM estimator $S_{\text{KM}(t)}$. The KM estimator is defined as

$$\widehat{S}_{\text{KM}}(t) = \sum_{s \in T_n(t)} \left\{ 1 - \frac{\mathbf{1}(Z_j = s)\delta_j}{\mathbf{1}(Z_j \geq s)} \right\}$$

Therefore, the estimates of sensitivity and specificity are calculated as follows:

$$\widehat{\text{TP}}_{\text{KM}}(c,t) = \frac{\{1 - F_X(c)\}\left\{1 - \widehat{S}_{\text{KM}}(t|X > c)\right\}}{1 - \widehat{S}_{\text{KM}}(t)}$$

$$\text{TN}_{\text{KM}}(c,t) = \frac{F_X(c)\widehat{S}_{\text{KM}}(t|X \leq c)}{\widehat{S}_{\text{KM}}(t)}$$

However, this approach produces sensitivity and specificity functions that may not be monotone with respect to the marker values, because the estimator $P(X > c, T > t) = \{1 - F_X(c)\}\{1 - \widehat{S}_{\text{KM}}(t|X > c)\}$ does not provide a valid bivariate distribution. Moreover, the KM-based estimator has a potential problem arising from the assumption that the censoring process is independent from X, which might be violated in practice.

2.2 Nearest-neighbor estimation

To overcome the violation of the monotonicity of the ROC curve with respect to the biomarker and accommodate the possibility that the censoring process is not independent from the marker values, Heagerty, Lumley, and Pepe (2000) provided an alternative method for estimating valid sensitivity and specificity functions using the estimator provided by Akritas (1994),

$$\widehat{S_{\lambda_n}(c,t)} = \frac{1}{n}\sum_i \widehat{S_{\lambda_n}}(t|X = X_i)\mathbf{1}(X_i > c)$$

where $\widehat{S_{\lambda_n}}$ is a smoothed estimator of the conditional survival function depending on parameter λ_n:

$$\widehat{S_{\lambda_n}}(t|X = X_i) = \sum_{s \in T_n(t)} \left\{ 1 - \frac{\sum_j K_{\lambda_n}(X_i, X_j)\mathbf{1}(Z_j = s)d_j}{\sum_j K_{\lambda_n}(X_i, X_j)\mathbf{1}(Z_j \geq s)} \right\}$$

This parameter defines the binary nearest-neighbor kernel $K_{\lambda_n}(X_i, X_j)$, representing the percentage observations included in each neighborhood,

$$K_{\lambda_n}(X_i, X_j) = \mathbf{1}\left\{|F_X(X_i) - F_X(X_j)| < \lambda_n\right\}$$
$$2\lambda_n \in (0, 1)$$

where $\lambda_n = O(n^{-1/3})$ is sufficient to provide a weakly consistent estimator of the bivariate function in a practical situation (Heagerty, Lumley, and Pepe 2000). Hence, estimates of sensitivity and specificity can be obtained using

$$\widehat{\text{TP}}_{\lambda_n}(c,t) = \frac{1 - F_X(c) - \widehat{S}_{\lambda_n}(c,t)}{1 - \widehat{S}_{\lambda_n}(-\infty,t)}$$
$$\widehat{\text{TN}}_{\lambda_n}(c,t) = \frac{\widehat{S}_{\lambda_n}(c,t)}{\widehat{S}_{\lambda_n}(-\infty,t)}$$

where $\widehat{S}_{\lambda_n}(-\infty,t) = 1/n \sum_i \{1(X_i > c)(1 - S_{\lambda_n}(t|X = X_i)\}$.

2.3 Estimation of the optimal cutpoint

Along with ROC curves, it is often necessary to establish a single cutpoint for stratifying individuals into risk categories. For this purpose, the literature provides three decision criteria for defining an optimal cutpoint when the outcome is binary. Such criteria are based on the accuracy measures provided by ROC curves. There are three main strategies for cutpoint estimation based on selecting a biomarker value $c*$ from its distribution, such that

i) the Youden function (Youden 1950), namely, the sum of sensitivity and specificity minus one, is maximized, which is equivalent to optimizing the biomarker's classification accuracy if sensitivity and specificity have the same weight in the decision maker's perspective;

ii) the distance between the selected point and the point representing perfect classification (FP = 0, TP = 1) is minimized (Perkins and Schisterman 2006); and

iii) the concordance probability function (defined as the product of sensitivity and specificity [Liu and Jin 2015]) is maximized.

3 The stroccurve command

3.1 Syntax

The user is required to use `stset` before using `stroccurve`. The syntax of the command is

```
stroccurve markervar [if] [in], timepoint(#) [nne lambda(#) km
   genrocvars replace nograph liu youden nearest]
```

where *markervar* is the continuous biomarker variable for which the time-dependent ROC curve is to be calculated.

3.2 Options

`timepoint(#)` specifies the time point for which the ROC curve is to be calculated. `timepoint()` is required.

`nne` calculates the time-dependent ROC curve with 0/1 nearest-neighbor kernel smoothing of the conditional survival function, the default.

`lambda(#)` specifies the percentage of observations to be included in each neighborhood if the nearest-neighbor estimator of the survival function method is used. It has to be included in the $(0, 0.5)$ interval. The default is `lambda(0.25*`n`^1/3)`, where n is the number of observations.

`km` calculates the time-dependent ROC curve with the KM method.

`genrocvars` generates new specificity and sensitivity variables for *markervar*, `_FP_R`, and `_TP_R` corresponding to their marker values.

`replace` requests `genrocvars` to overwrite the existing `_FP_R` and `_TP_R` variables.

`nograph` suppresses the plot.

`liu` estimates the cutoff by maximizing the concordance probability.

`youden` estimates the cutoff by maximizing the Youden function.

`nearest` estimates the closest cutpoint to $(0, 1)$.

3.3 Stored results

`stroccurve` stores the following in `e()`:

Scalars

`e(AUC)`	returns the area under the ROC curve
`e(youden)`	returns the cutpoint maximizing the Youden criterion if the `youden` option is specified
`e(liu)`	returns the cutpoint maximizing the concordance probability if the `liu` option is specified
`e(nearest)`	returns the nearest cutpoint to $(0, 1)$ if the `nearest` option is specified

Matrices

`e(rocmat)`	returns an $m \times 3$ matrix, where m is the number of unique marker values; the first column includes marker values; the second and third columns report the estimates of sensitivity and one minus specificity for such marker values

Functions

`e(sample)`	marks the estimation sample

4 Examples

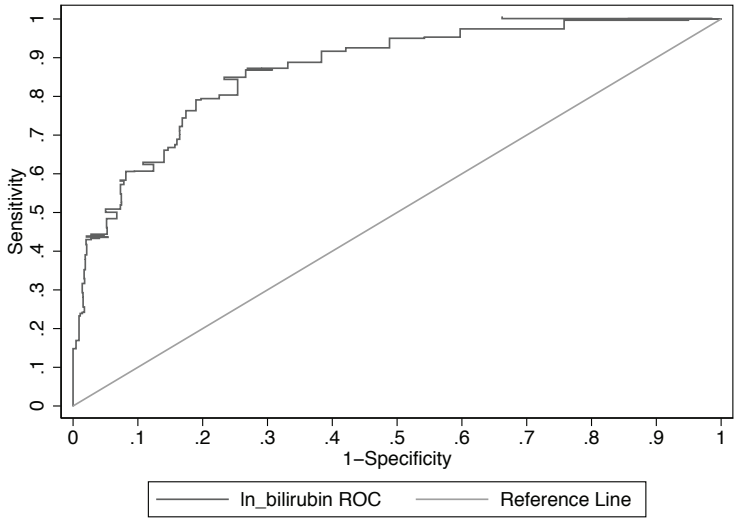
We use primary biliary cirrhosis data from Fleming and Harrington (1991).

```
. use pbc.dta
(PBC data, 3 sources)
. quietly stset survtime, failure(censdead==1)
```

Example 1

We calculate the survival ROC curve for survival at 2,000 days, with the Kaplan–Meier estimator of the survival function.

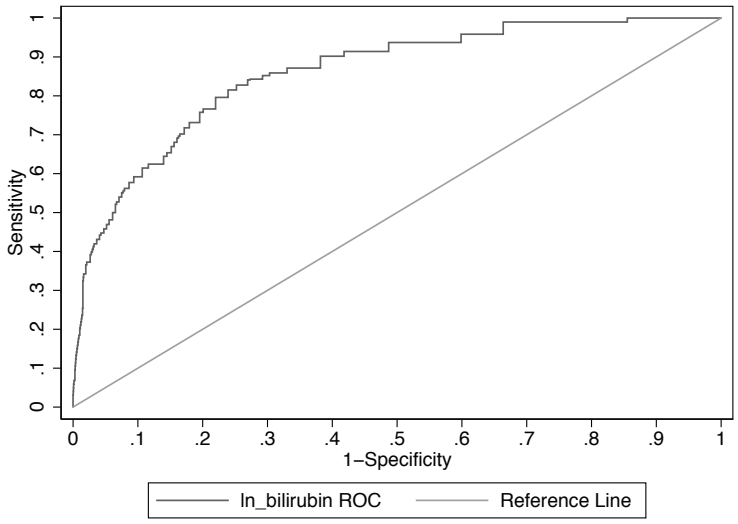
```
. stroccurve ln_bilirubin, timepoint(2000) km
Time dependent ROC curve at time:       2000
Survival Function Estimation Method:    Kaplan Meier
N=                                      312
Area under the ROC Curve:               0.868
```



Example 2

We calculate the nearest-neighbor ROC curve for survival at 2,000 days, with the smoothing parameter equal to $0.25 \times N^{1/3}$.

```
. stroccurve ln_bilirubin, timepoint(2000)
Smoothing parameter automatically set at 0.25*N^(-1/3)
Time dependent ROC curve at time:       2000
Survival Function Estimation Method:    Nearest Neighbor
Smoothing parameter:                    0.037
N=                                      312
Area under the ROC Curve:               0.858
```



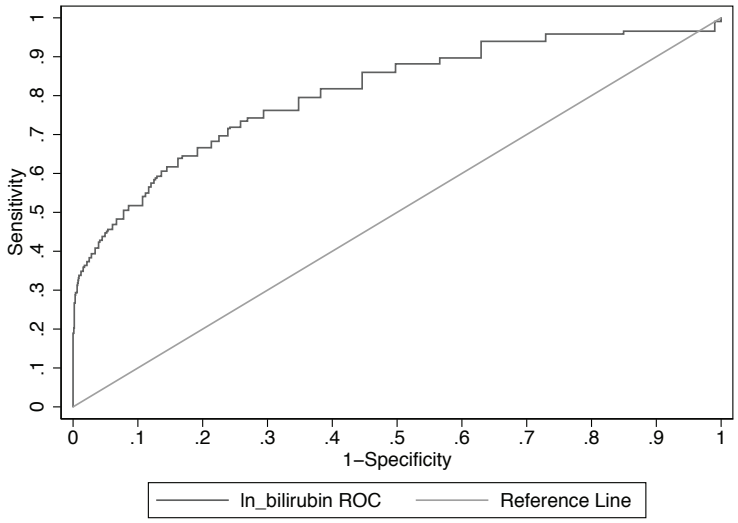
Example 3

We calculate the nearest-neighbor ROC curve for survival at 3,000 days, with the smoothing parameter equal to $0.25 \times N^{1/3}$, and request the optimal cutpoint according to Perkins and Schisterman (2006).

```
. stroccurve ln_bilirubin, timepoint(3000) nearest
Smoothing parameter automatically set at 0.25*N^(-1/3)
Time dependent ROC curve at time:       3000
Survival Function Estimation Method:   Nearest Neighbor
Smoothing parameter:                    0.037
N=                                      312
Area under the ROC Curve:               0.816
------------------------------------------------------------------------------

OPTIMAL CUTPOINTS:
------------------------------------------------------------------------------

Nearest point to (0,1)
Optimal cutpoint:                       3.245
Sensitivity at optimal cutpoint:        0.734
Specificity at optimal cutpoint:        0.742
```



5 Acknowledgments

We thank Anna Falanga, MD, Cinzia Giaccherini, and Cristina Verzeroli (ASST Papa Giovanni XXIII) for their helpful comments.

6 References

Akritas, M. G. 1994. Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics* 22: 1299–1327.

Fleming, T. R., and D. P. Harrington. 1991. *Counting Processes and Survival Analysis*. New York: Wiley.

Heagerty, P. J., T. Lumley, and M. S. Pepe. 2000. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 56: 337–344.

Kaplan, E. L., and P. Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53: 457–481.

Liu, X., and Z. Jin. 2015. Optimal survival time-related cut-point with censored data. *Statistics in Medicine* 34: 515–524.

Pepe, M. S., G. Longton, and H. Janes. 2009. Estimation and comparison of receiver operating characteristic curves. *Stata Journal* 9: 1–16.

Perkins, N. J., and E. F. Schisterman. 2006. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology* 163: 670–675.

Youden, W. J. 1950. Index for rating diagnostic tests. *Cancer* 3: 32–35.

About the authors

Mattia Cattaneo, PhD, is a researcher at the Department of Management, Information and Production Engineering of the University of Bergamo, a scientific coordinator of the International Center for Competitiveness Studies in the Aviation Industry (ICCSAI), and a member of the Higher Education Research Group (HERG). His research interests include healthcare management, transport economics, regional development, and higher education. He is often a reviewer for several journals in the field of economics and higher education.

Paolo Malighetti, PhD, is associate professor at the Department of Management, Information and Production Engineering of the University of Bergamo, where he teaches health technology assessment, business economics, and transport management. His main research interests are related to the pricing and evolution of air transport networks and to healthcare management and economics, specifically, aspects related to the adoption of new technology within clinical practice. Since 2013, he has been director of the Human Factors and Technology in Healthcare (HTH) Center at the University of Bergamo.

Daniele Spinelli is a PhD student in economics and management of technology (DREAMT) at the Department of Management, Information and Production Engineering of the University of Bergamo and at the Department of Economics and Management of the University of Pavia. His main research interests are health economics, econometrics, and biostatistics.