Title

> **lroc —** Compute area under ROC curve and graph the curve

Description

lroc graphs the ROC curve and calculates the area under the curve.

lroc requires that the current estimation results be from logistic, logit, probit, or ivprobit; see [R] **logistic**, [R] **logit**, [R] **probit**, or [R] **ivprobit**.

Quick start

Graph and compute area under ROC curve for current estimation results

    lroc

Add "My Title" as title of graph

    lroc, title(My Title)

Suppress graph

    lroc, nograph

Menu

Statistics > Binary outcomes > Postestimation > ROC curve after logistic/logit/probit/ivprobit

1

Syntax

> lroc [*depvar*] [*if*] [*in*] [*weight*] [, *options*]

options	Description
[Main]	
all	compute area under ROC curve and graph curve for all observations
nograph	suppress graph
Advanced	
beta(*matname*)	row vector containing model coefficients
Plot	
cline_options	change look of the line
marker_options	change look of markers (color, size, etc.)
marker_label_options	add marker labels; change look or position
Reference line	
rlopts(*cline_options*)	affect rendition of the reference line
Add plots	
addplot(*plot*)	add other plots to the generated graph
Y axis, X axis, Titles, Legend, Overall	
twoway_options	any options other than by() documented in [G-3] *twoway_options*

collect is allowed; see [U] **11.1.10 Prefix commands**.

fweights are allowed; see [U] **11.1.6 weight**.

lroc is not appropriate with svy estimation results.

Options

[Main]

all requests that the statistic be computed for all observations in the data, ignoring any if or in restrictions specified by the estimation command.

nograph suppresses graphical output.

[Advanced]

beta(*matname*) specifies a row vector containing model coefficients. The columns of the row vector must be labeled with the corresponding names of the independent variables in the data. The dependent variable *depvar* must be specified immediately after the command name. See *Models other than the last fitted model* later in this entry.

[Plot]

cline_options, *marker_options*, and *marker_label_options* affect the rendition of the ROC curve—the plotted points connected by lines. These options affect the size and color of markers, whether and how the markers are labeled, and whether and how the points are connected; see [G-3] *cline_options*, [G-3] *marker_options*, and [G-3] *marker_label_options*.

⌐ Reference line ⌐

rlopts(*cline_options*) affects the rendition of the reference line; see [G-3] *cline_options*.

⌐ Add plots ⌐

addplot(*plot*) provides a way to add other plots to the generated graph; see [G-3] *addplot_option*.

⌐ Y axis, X axis, Titles, Legend, Overall ⌐

twoway_options are any of the options documented in [G-3] *twoway_options*, excluding by(). These include options for titling the graph (see [G-3] *title_options*) and for saving the graph to disk (see [G-3] *saving_option*).

Remarks and examples

stata.com

Remarks are presented under the following headings:

> Introduction
> Samples other than the estimation sample
> Models other than the last fitted model

Introduction

Stata also has a suite of commands for performing both parametric and nonparametric receiver operating characteristic (ROC) analysis. See [R] **roc** for an overview of these commands.

lroc graphs the ROC curve—a graph of sensitivity versus one minus specificity as the cutoff c is varied—and calculates the area under it. Sensitivity is the fraction of observed positive-outcome cases that are correctly classified; specificity is the fraction of observed negative-outcome cases that are correctly classified. When the purpose of the analysis is classification, you must choose a cutoff.

The curve starts at $(0, 0)$, corresponding to $c = 1$, and continues to $(1, 1)$, corresponding to $c = 0$. A model with no predictive power would be a $45°$ line. The greater the predictive power, the more bowed the curve, and hence the area beneath the curve is often used as a measure of the predictive power. A model with no predictive power has area 0.5; a perfect model has area 1.

The ROC curve was first discussed in signal detection theory (Peterson, Birdsall, and Fox 1954) and then was quickly introduced into psychology (Tanner and Swets 1954). It has since been applied in other fields, particularly medicine (for instance, Metz [1978]). For a classic text on ROC techniques, see Green and Swets (1966).

lsens also plots sensitivity and specificity; see [R] **lsens**.

▷ Example 1

Hardin and Hilbe (2018) examine data from the National Canadian Registry of Cardiovascular Disease (FASTRAK), sponsored by Hoffman-La Roche Canada. They model death within 48 hours based on whether a patient suffers an anterior infarct (heart attack) rather than an inferior infarct using a logistic regression and evaluate the model using an ROC curve. We replicate their analysis here.

Both anterior and inferior refer to sites on the heart where damage occurs. The model is also adjusted for hcabg, whether the subject has had a cardiac bypass surgery (CABG); age, a four-category age-group indicator; and killip, a four-level risk indicator.

We load the data and then estimate the parameters of the logistic regression with `logistic`. Factor-variable notation is used for each predictor, because they are categorical; see [U] **11.4.3 Factor variables**.

```
. use https://www.stata-press.com/data/r18/heart
(Heart attacks)
. logistic death i.site i.hcabg i.killip i.age
```

Logistic regression					Number of obs	= 4,483
					LR chi2(8)	= 211.37
					Prob > chi2	= 0.0000
Log likelihood = -636.62553					Pseudo R2	= 0.1424

death	Odds ratio	Std. err.	z	P>|z|	[95% conf. interval]	
site						
Anterior	1.901333	.3185757	3.83	0.000	1.369103	2.640464
1.hcabg	2.105275	.7430694	2.11	0.035	1.054076	4.204801
killip						
2	2.251732	.4064423	4.50	0.000	1.580786	3.207453
3	2.172105	.584427	2.88	0.004	1.281907	3.680487
4	14.29137	5.087654	7.47	0.000	7.112964	28.71423
age						
60–69	1.63726	.5078582	1.59	0.112	.8914261	3.007115
70–79	4.532029	1.206534	5.68	0.000	2.689568	7.636647
>=80	8.893222	2.41752	8.04	0.000	5.219991	15.15125
_cons	.0063961	.0016541	-19.54	0.000	.0038529	.010618

Note: _cons estimates baseline odds.

The odds ratios for a unit change in each covariate are reported by `logistic`. At fixed values of the other covariates, patients who enter Canadian hospitals with an anterior infarct have nearly twice the odds of death within 48 hours than those with an inferior infarct. Those who have had a previous CABG have approximately twice the risk of death of those who have not. Those with higher Killip risks and those who are older are also at greater risk of death.

We use `lroc` to draw the ROC curve for the model. The area under the curve of approximately 0.8 indicates acceptable discrimination for the model.

```
. lroc

Logistic model for death

Number of observations =      4483
Area under ROC curve   =    0.7965
```


Area under ROC curve = 0.7965

◁
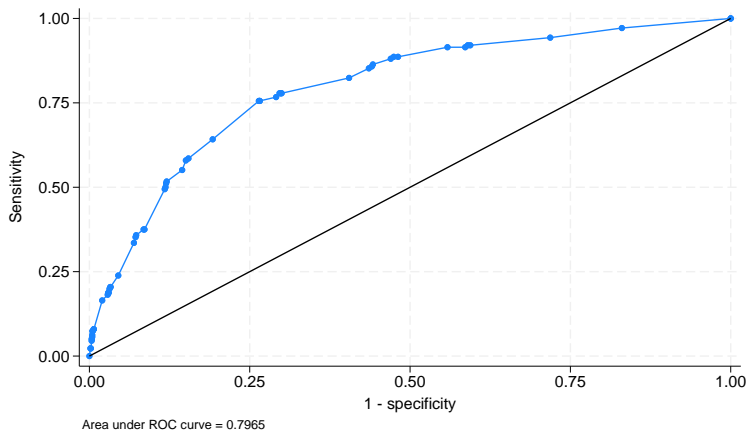
Samples other than the estimation sample

lroc can be used with samples other than the estimation sample. By default, lroc remembers the estimation sample used with the last logistic, logit, probit, or ivprobit command. To override this, simply use an if or in restriction to select another set of observations, or specify the all option to force the command to use all the observations in the dataset.

See example 3 in [R] estat gof for an example of using lroc with a sample other than the estimation sample.

Models other than the last fitted model

By default, lroc uses the last model fit by logistic, logit, probit, or ivprobit. You may also directly specify the model to lroc by inputting a vector of coefficients with the beta() option and passing the name of the dependent variable *depvar* to lroc.

▷ Example 2

Suppose that someone publishes the following logistic model of low birthweight:

$$\Pr(\texttt{low} = 1) = F(-0.02\,\texttt{age} - 0.01\,\texttt{lwt} + 1.3\,\texttt{black} + 1.1\,\texttt{smoke} + 0.5\,\texttt{ptl} + 1.8\,\texttt{ht} + 0.8\,\texttt{ui} + 0.5)$$

where F is the cumulative logistic distribution. These coefficients are not odds ratios; they are the equivalent of what logit produces.

We can see whether this model fits our data. First, we enter the coefficients as a row vector and label its columns with the names of the independent variables plus _cons for the constant (see [P] matrix define and [P] matrix rownames).

```
. use https://www.stata-press.com/data/r18/lbw3, clear
(Hosmer & Lemeshow data)
. matrix input b = (-.02, -.01, 1.3, 1.1, .5, 1.8, .8, .5)
. matrix colnames b = age lwt black smoke ptl ht ui _cons
```

Here we use `lroc` to examine the predictive ability of the model:

```
. lroc low, beta(b) nograph
Logistic model for low

Number of observations =      189
Area under ROC curve   =   0.7275
```

The area under the curve indicates that this model does have some predictive power. We can obtain a graph of sensitivity and specificity as a function of the cutoff probability by typing

```
. lsens low, beta(b)
```


See [R] **lsens**.

◁
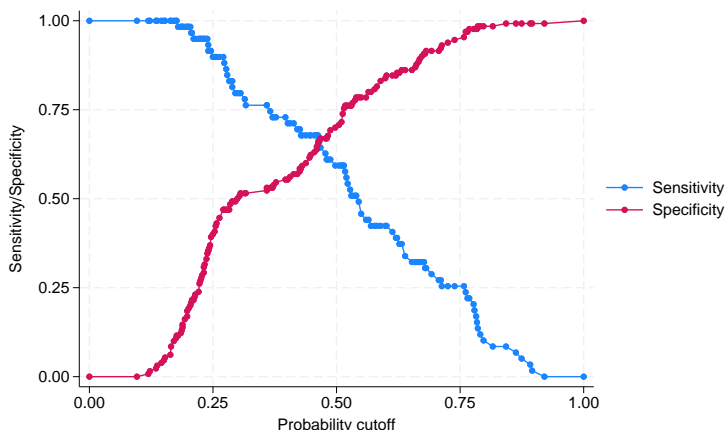
Stored results

`lroc` stores the following in `r()`:

Scalars
    `r(N)`   number of observations
    `r(area)`   area under the ROC curve

Methods and formulas

The ROC curve is a graph of *sensitivity* against $(1 - \textit{specificity})$. This is guaranteed to be a monotone nondecreasing function because the number of correctly predicted successes increases and the number of correctly predicted failures decreases as the classification cutoff c decreases.

The area under the ROC curve is the area on the bottom of this graph and is determined by integrating the curve. The vertices of the curve are determined by sorting the data according to the predicted index, and the integral is computed using the trapezoidal rule.

References

Green, D. M., and J. A. Swets. 1966. *Signal Detection Theory and Psychophysics*. New York: Wiley.

Hardin, J. W., and J. M. Hilbe. 2018. *Generalized Linear Models and Extensions*. 4th ed. College Station, TX: Stata Press.

Hosmer, D. W., Jr., S. A. Lemeshow, and R. X. Sturdivant. 2013. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: Wiley.

Metz, C. E. 1978. Basic principles of ROC analysis. *Seminars in Nuclear Medicine* 8: 283–298. https://doi.org/10.1016/s0001-2998(78)80014-2.

Peterson, W. W., T. G. Birdsall, and W. C. Fox. 1954. The theory of signal detectability. *Transactions IRE Professional Group on Information Theory* PGIT-4: 171–212. https://doi.org/10.1109/TIT.1954.1057460.

Tanner, W. P., Jr., and J. A. Swets. 1954. A decision-making theory of visual detection. *Psychological Review* 61: 401–409. https://doi.org/10.1037/h0058700.

Also see

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

[R] **logit** — Logistic regression, reporting coefficients

[R] **probit** — Probit regression

[R] **ivprobit** — Probit model with continuous endogenous covariates

[R] **lsens** — Graph sensitivity and specificity versus probability cutoff

[R] **estat classification** — Classification statistics and table

[R] **estat gof** — Pearson or Hosmer–Lemeshow goodness-of-fit test

[R] **roc** — Receiver operating characteristic (ROC) analysis

[U] **20 Estimation and postestimation commands**