Syntax

Test equality of ROC areas

```
roccomp refvar classvar [classvars] [if] [in] [weight] [, roccomp_options]
```

Test equality of ROC area against a standard ROC curve

```
rocgold refvar goldvar classvar [classvars] [if] [in] [weight] [, rocgold_options]
```

**roccomp_options**  

<table>
<thead>
<tr>
<th><strong>Description</strong></th>
<th><strong>Syntax</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main</strong></td>
<td></td>
</tr>
<tr>
<td>by(varname)</td>
<td>split into groups by variable</td>
</tr>
<tr>
<td>test(matname)</td>
<td>use contrast matrix for comparing ROC areas</td>
</tr>
<tr>
<td>graph</td>
<td>graph the ROC curve</td>
</tr>
<tr>
<td>norefline</td>
<td>suppress plotting the 45-degree reference line</td>
</tr>
<tr>
<td>separate</td>
<td>place each ROC curve on its own graph</td>
</tr>
<tr>
<td>summary</td>
<td>report the area under the ROC curve</td>
</tr>
<tr>
<td>binormal</td>
<td>estimate areas by using binormal distribution assumption</td>
</tr>
<tr>
<td>line#opts(cline_options)</td>
<td>affect rendition of the #th binormal fit line</td>
</tr>
<tr>
<td>level(#)</td>
<td>set confidence level; default is level(95)</td>
</tr>
</tbody>
</table>

**Plot**

```
plot#opts(plot_options)          | affect rendition of the #th ROC curve |
```

**Reference line**

```
rlopts(cline_options)           | affect rendition of the reference line |
```

**Y axis, X axis, Titles, Legend, Overall**

```
twoway_options                  | any options other than by() documented in [G-3] twoway_options |
```

1
**rocgold**

**Description**

The above commands are used to perform receiver operating characteristic (ROC) analyses with rating and discrete classification data.

The two variables `refvar` and `classvar` must be numeric. The reference variable indicates the true state of the observation, such as diseased and nondiseased or normal and abnormal, and must be coded as 0 and 1. The rating or outcome of the diagnostic test or test modality is recorded in `classvar`, which must be at least ordinal, with higher values indicating higher risk.
roccomp tests the equality of two or more ROC areas obtained from applying two or more test modalities to the same sample or to independent samples. roccomp expects the data to be in wide form when comparing areas estimated from the same sample and in long form for areas estimated from independent samples.

rocgold independently tests the equality of the ROC area of each of several test modalities, specified by classvar, against a “gold standard” ROC curve, goldvar. For each comparison, rocgold reports the raw and the Bonferroni-adjusted significance probability. Optionally, Šidák’s adjustment for multiple comparisons can be obtained.

See [R] rocfit and [R] rocreg for commands that fit maximum-likelihood ROC models.

Options

Main

by(varname) (roccomp only) is required when comparing independent ROC areas. The by() variable identifies the groups to be compared.

sidak (rocgold only) requests that the significance probability be adjusted for the effect of multiple comparisons by using Šidák’s method. Bonferroni’s adjustment is reported by default.

test(matname) specifies the contrast matrix to be used when comparing ROC areas. By default, the null hypothesis that all areas are equal is tested.

graph produces graphical output of the ROC curve.

norefline suppresses plotting the 45-degree reference line from the graphical output of the ROC curve.

separate is meaningful only with roccomp and specifies that each ROC curve be placed on its own graph rather than one curve on top of the other.

summary reports the area under the ROC curve, its standard error, and its confidence interval. This option is needed only when also specifying graph.

binormal specifies that the areas under the ROC curves to be compared should be estimated using the binormal distribution assumption. By default, areas to be compared are computed using the trapezoidal rule.

line#opts(cline_options) affects the rendition of the line representing the #th ROC curve drawn using the binormal distribution assumption; see [G-3] cline_options. These lines are drawn only if the binormal option is specified.

level(#) specifies the confidence level, as a percentage, for the confidence intervals. The default is level(95) or as set by set level; see [R] level.

Plot

plot#opts(plot_options) affects the rendition of the #th ROC curve—the curve’s plotted points connected by lines. The plot_options can 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] marker_options, [G-3] marker_label_options, and [G-3] cline_options.

For rocgold, plot1opts() are applied to the ROC for the gold standard.

Reference line

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

Remarks are presented under the following headings:

Introduction
Comparing areas under the ROC curve
Correlated data
Independent data
Comparing areas with a gold standard

Introduction

roccomp provides comparison of the ROC curves of multiple classifiers. rocgold compares the ROC curves of multiple classifiers with a single “gold standard” classifier. Adjustment of inference for multiple comparisons is also provided by rocgold.


Comparing areas under the ROC curve

The area under multiple ROC curves can be compared by using roccomp. The command syntax is slightly different if the ROC curves are correlated (that is, different diagnostic tests are applied to the same sample) or independent (that is, diagnostic tests are applied to different samples).

Correlated data

Example 1

Hanley and McNeil (1983) presented data from an evaluation of two computer algorithms designed to reconstruct CT images from phantoms. We will call these two algorithms’ modalities 1 and 2. A sample of 112 phantoms was selected; 58 phantoms were considered normal, and the remaining 54 were abnormal. Each of the two modalities was applied to each phantom, and the resulting images were rated by a reviewer using a six-point scale: 1 = definitely normal, 2 = probably normal, 3 = possibly normal, 4 = possibly abnormal, 5 = probably abnormal, and 6 = definitely abnormal. Because each modality was applied to the same sample of phantoms, the two sets of outcomes are correlated.
We list the first 7 observations:

```
. use http://www.stata-press.com/data/r13/ct
. list in 1/7, sep(0)
```

<table>
<thead>
<tr>
<th>mod1</th>
<th>mod2</th>
<th>status</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

The data are in wide form, which is required when dealing with correlated data. Each observation corresponds to one phantom. The variable `mod1` identifies the rating assigned for the first modality, and `mod2` identifies the rating assigned for the second modality. The true status of the phantoms is given by `status=0` if they are normal and `status=1` if they are abnormal. The observations with at least one missing rating were dropped from the analysis.

We plot the two ROC curves and compare their areas.

```
. roccomp status mod1 mod2, graph summary
```

<table>
<thead>
<tr>
<th></th>
<th>Obs</th>
<th>ROC Area</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>mod1</td>
<td>112</td>
<td>0.8828</td>
<td>0.0317</td>
<td>0.82067 0.94498</td>
</tr>
<tr>
<td>mod2</td>
<td>112</td>
<td>0.9302</td>
<td>0.0256</td>
<td>0.88005 0.98042</td>
</tr>
</tbody>
</table>

Ho: area(mod1) = area(mod2)

\[ \text{chi2}(1) = 2.31 \quad \text{Prob}>\text{chi2} = 0.1282 \]

By default, `roccomp`, with the `graph` option specified, plots the ROC curves on the same graph. Optionally the curves can be plotted side by side, each on its own graph, by also specifying `separate`.

For each curve, `roccomp` reports summary statistics and provides a test for the equality of the area under the curves, using an algorithm suggested by DeLong, DeLong, and Clarke-Pearson (1988).
Although the area under the ROC curve for modality 2 is larger than that of modality 1, the chi-squared test yielded a significance probability of 0.1282, suggesting that there is no significant difference between these two areas.

The `roccomp` command can also be used to compare more than two ROC areas. To illustrate this, we modified the previous dataset by including a fictitious third modality.

```
. use http://www.stata-press.com/data/r13/ct2
. roccomp status mod1 mod2 mod3, graph summary

<table>
<thead>
<tr>
<th></th>
<th>ROC Area</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>mod1</td>
<td>0.8828</td>
<td>0.0317</td>
<td>0.82067 0.94498</td>
</tr>
<tr>
<td>mod2</td>
<td>0.9302</td>
<td>0.0256</td>
<td>0.88005 0.98042</td>
</tr>
<tr>
<td>mod3</td>
<td>0.9240</td>
<td>0.0241</td>
<td>0.87670 0.97132</td>
</tr>
</tbody>
</table>
```

Ho: area(mod1) = area(mod2) = area(mod3)

\[ \chi_2(2) = 6.54 \quad \text{Prob}>\chi_2 = 0.0381 \]

By default, `roccomp` tests whether the areas under the ROC curves are all equal. Other comparisons can be tested by creating a contrast matrix and specifying `test(matname)`, where `matname` is the name of the contrast matrix.

For example, assume that we are interested in testing whether the area under the ROC for mod1 is equal to that of mod3. To do this, we can first create an appropriate contrast matrix and then specify its name with the `test()` option.

Of course, this is a trivial example because we could have just specified

```
. roccomp status mod1 mod3
```

without including mod2 to obtain the same test results. However, for illustration, we will continue with this example.

The contrast matrix must have its number of columns equal to the number of `classvars` (that is, the total number of ROC curves) and a number of rows less than or equal to the number of `classvars`, and the elements of each row must add to zero.
. matrix C=(1,0,-1)
. roccomp status mod1 mod2 mod3, test(C)

ROC

<table>
<thead>
<tr>
<th></th>
<th>Area</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>mod1</td>
<td>112</td>
<td>0.8828</td>
<td>0.0317</td>
</tr>
<tr>
<td>mod2</td>
<td>112</td>
<td>0.9302</td>
<td>0.0256</td>
</tr>
<tr>
<td>mod3</td>
<td>112</td>
<td>0.9240</td>
<td>0.0241</td>
</tr>
</tbody>
</table>

Ho: Comparison as defined by contrast matrix: C
chi2(1) = 5.25 Prob>chi2 = 0.0220

Although all three areas are reported, the comparison is made using the specified contrast matrix.

Perhaps more interesting would be a comparison of the area from mod1 and the average area of mod2 and mod3.

. matrix C=(1,-.5,-.5)
. roccomp status mod1 mod2 mod3, test(C)

ROC

<table>
<thead>
<tr>
<th></th>
<th>Area</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>mod1</td>
<td>112</td>
<td>0.8828</td>
<td>0.0317</td>
</tr>
<tr>
<td>mod2</td>
<td>112</td>
<td>0.9302</td>
<td>0.0256</td>
</tr>
<tr>
<td>mod3</td>
<td>112</td>
<td>0.9240</td>
<td>0.0241</td>
</tr>
</tbody>
</table>

Ho: Comparison as defined by contrast matrix: C
chi2(1) = 3.43 Prob>chi2 = 0.0642

Other contrasts could be made. For example, we could test if mod3 is different from at least one of the other two by first creating the following contrast matrix:

. matrix C=(-1,0,1 \ 0,-1,1)
. mat list C

C[2,3]
c1  c2  c3
r1 -1  0  1
r2  0 -1  1

Independent data

Example 2

In example 1, we noted that because each test modality was applied to the same sample of phantoms, the classification outcomes were correlated. Now assume that we have collected the same data presented by Hanley and McNeil (1983), except that we applied the first test modality to one sample of phantoms and the second test modality to a different sample of phantoms. The resulting measurements are now considered independent.
Here are a few of the observations.

. list in 1/7, sep(0)

<table>
<thead>
<tr>
<th>pop</th>
<th>status</th>
<th>rating</th>
<th>mod</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>28</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>19</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

The data are in long form, which is required when dealing with independent data. The data consist of 24 observations: 6 observations corresponding to abnormal phantoms and 6 to normal phantoms evaluated using the first modality, and similarly 6 observations corresponding to abnormal phantoms and 6 to normal phantoms evaluated using the second modality. The number of phantoms corresponding to each observation is given by the `pop` variable. Once again we have frequency-weighted data. The variable `mod` identifies the modality, and `rating` is the assigned classification.

We can better view our data by using the `table` command.

. table status rating [fw=pop], by(mod) row col

<table>
<thead>
<tr>
<th>mod and status</th>
<th>rating</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>12</td>
<td>28</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>13</td>
<td>22</td>
<td>9</td>
<td>54</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>13</td>
<td>31</td>
<td>14</td>
<td>19</td>
<td>26</td>
<td>9</td>
<td>112</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>31</td>
<td>19</td>
<td>5</td>
<td>3</td>
<td></td>
<td></td>
<td>58</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>19</td>
<td>15</td>
<td>10</td>
<td>54</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>34</td>
<td>21</td>
<td>10</td>
<td>22</td>
<td>15</td>
<td>10</td>
<td>112</td>
</tr>
</tbody>
</table>

The `status` variable indicates the true status of the phantoms: `status = 0` if they are normal and `status = 1` if they are abnormal.

We now compare the areas under the two ROC curves.
Comparing areas with a gold standard

The area under multiple ROC curves can be compared with a gold standard using `rocgold`. The command syntax is similar to that of `roccomp`. The tests are corrected for the effect of multiple comparisons.

Example 3

We will use the same data (presented by Hanley and McNeil [1983]) as in the `roccomp` examples. Let’s assume that the first modality is considered to be the standard against which both the second and third modalities are compared.

We want to plot and compare both the areas of the ROC curves of `mod2` and `mod3` with `mod1`. Because we consider `mod1` to be the gold standard, it is listed first after the reference variable in the `rocgold` command line.
. use http://www.stata-press.com/data/r13/ct2
. rocgold status mod1 mod2 mod3, graph summary

<table>
<thead>
<tr>
<th>ROC Area</th>
<th>Std. Err.</th>
<th>chi2</th>
<th>df</th>
<th>Pr&gt;chi2</th>
<th>Pr&gt;chi2</th>
</tr>
</thead>
<tbody>
<tr>
<td>mod1 (standard)</td>
<td>0.8828</td>
<td>0.0317</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mod2</td>
<td>0.9302</td>
<td>0.0256</td>
<td>2.3146</td>
<td>1</td>
<td>0.1282</td>
</tr>
<tr>
<td>mod3</td>
<td>0.9240</td>
<td>0.0241</td>
<td>5.2480</td>
<td>1</td>
<td>0.0220</td>
</tr>
</tbody>
</table>

Equivalently, we could have done this in two steps by using the roccomp command.

. roccomp status mod1 mod2, graph summary
. roccomp status mod1 mod3, graph summary

Stored results

roccomp stores the following in \( r() \):

Scalars
- \( r(N_g) \): number of groups
- \( r(p) \): significance probability

Matrices
- \( r(V) \): variance–covariance matrix

rocgold stores the following in \( r() \):

Scalars
- \( r(N_g) \): number of groups

Matrices
- \( r(V) \): variance–covariance matrix
- \( r(chi2) \): \( \chi^2 \) vector
- \( r(df) \): \( \chi^2 \) degrees-of-freedom vector
- \( r(p) \): significance-probability vector
- \( r(p_adj) \): adjusted significance-probability vector
Assume that we applied a diagnostic test to each of \( N_n \) normal and \( N_a \) abnormal subjects. Further assume that the higher the outcome value of the diagnostic test, the higher the risk of the subject being abnormal. Let \( \hat{\theta} \) be the estimated area under the curve, and let \( X_i, i = 1, 2, \ldots, N_a \) and \( Y_j, j = 1, 2, \ldots, N_n \) be the values of the diagnostic test for the abnormal and normal subjects, respectively.

Areas under ROC curves are compared using an algorithm suggested by DeLong, DeLong, and Clarke-Pearson (1988). Let \( \hat{\theta} = (\hat{\theta}^1, \hat{\theta}^2, \ldots, \hat{\theta}^k) \) be a vector representing the areas under \( k \) ROC curves. See Methods and formulas in [R] roctab for the definition of these area estimates.

For the \( r \)th area, define

\[
V_{10}^r(X_i) = \frac{1}{N_n} \sum_{j=1}^{N_n} \psi(X_i, Y_j)
\]

and for each normal subject, \( j \), define

\[
V_{01}^r(Y_j) = \frac{1}{N_a} \sum_{i=1}^{N_a} \psi(X_i, Y_j)
\]

where

\[
\psi(X, Y) = \begin{cases} 
1 & Y < X \\
\frac{1}{2} & Y = X \\
0 & Y > X 
\end{cases}
\]

Define the \( k \times k \) matrix \( S_{10} \) such that the \((r, s)\)th element is

\[
S_{10}^{r,s} = \frac{1}{N_a - 1} \sum_{i=1}^{N_a} \{V_{10}^r(X_i) - \hat{\theta}^r\} \{V_{10}^s(X_i) - \hat{\theta}^s\}
\]

and \( S_{01} \) such that the \((r, s)\)th element is

\[
S_{01}^{r,s} = \frac{1}{N_n - 1} \sum_{j=1}^{N_n} \{V_{01}^r(Y_i) - \hat{\theta}^r\} \{V_{01}^s(Y_i) - \hat{\theta}^s\}
\]

Then the covariance matrix is

\[
S = \frac{1}{N_a} S_{10} + \frac{1}{N_n} S_{01}
\]

Let \( L \) be a contrast matrix defining the comparison, so that

\[
(\hat{\theta} - \theta)' L (L L')^{-1} L (\hat{\theta} - \theta)
\]

has a chi-squared distribution with degrees of freedom equal to the rank of \( L L' \).
References


——. 2002b. From the help desk: Comparing areas under receiver operating characteristic curves from two or more probit or logit models. *Stata Journal* 2: 301–313.


Also see

[R] **logistic postestimation** — Postestimation tools for logistic

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

[R] **rocfit** — Parametric ROC models

[R] **rocreg** — Receiver operating characteristic (ROC) regression

[R] **roctab** — Nonparametric ROC analysis