The Stata Journal (2002)
2, Number 3, pp. 301–313

From the help desk: Comparing areas under receiver operating characteristic curves from two or more probit or logit models

Mario A. Cleves, Ph.D.
Department of Pediatrics
University of Arkansas for Medical Sciences
Little Rock, Arkansas

Abstract. Occasionally, there is a need to compare the predictive accuracy of several fitted logit (logistic) or probit models by comparing the areas under the corresponding receiver operating characteristic (ROC) curves. Although Stata currently does not have a ready routine for comparing two or more ROC areas generated from these models, this article describes how these comparisons can be performed using Stata's `roccomp` command.

Keywords: st0023, Receiving Operating Characteristic (ROC) curve

1 Background

Stata's `roccomp` command is one of Stata's general-purpose programs for computing, analyzing, and comparing areas under the ROC curve. See [R] **roc** for more details on this and other ROC commands. The `roccomp` command 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. That is, there are two "flavors" of `roccomp`. The first is used to analyze correlated data, where several test modalities are applied to the same set of observations. The second is used to analyze independent data, where each test modality is applied to a different (disjointed) set of observations. In order to compare areas under ROC curves from different models, we must first determine if each model to be compared was estimated on the same set of observations or on different disjointed sets. This article describes methods for comparing areas from several receiver operating curves produced by logit and probit models under these two data scenarios. Although in most of the examples in this article, we estimate logistic (logit) models, the procedures described are identically applied to probit models.

Before describing the procedure for comparing areas under two or more ROC curves, let's examine the similarity between Stata's `lroc` command, used to produce ROC curves after logistic regression, and the `roctab` command. We illustrate this using the auto data distributed with Stata 7.0. We begin by fitting a logistic model with `foreign` as the dependent variable and `price` as the only covariate:

```
. use http://www.stata-press.com/data/r7/auto, clear
(1978 Automobile Data)

. logistic foreign price

Logit estimates                                   Number of obs   =         74
                                                  LR chi2(1)      =       0.17
                                                  Prob > chi2     =     0.6784
Log likelihood =  -44.94724                       Pseudo R2       =     0.0019

------------------------------------------------------------------------------
     foreign | Odds Ratio   Std. Err.      z    P>|z|      [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |   1.000035   .0000844     0.42   0.676      .9998699    1.000201
------------------------------------------------------------------------------

. lroc, nograph

Logistic model for foreign

number of observations =       74
area under ROC curve   =   0.5769
```

After fitting the logistic model, we use `lroc` to compute the area under the ROC curve (0.5769). We now use `predict` to obtain the predicted probability of a positive outcome.

```
. predict p
(option p assumed; Pr(foreign))
```

The new variable, `p`, containing the model-predicted probability of a positive outcome, has been added to our data in memory. Note that we did not specify any options for `predict` because the predicted probability is the default after logistic regression. A safer way would have been to type

```
. predict p, p
```

or, even better,

```
. predict p if e(sample), p
```

Although specifying `if e(sample)` is not needed in this case because all 74 observations in the data were used during estimation, as we will see later, this is not always the case. Thus, we recommend that `if e(sample)` always be specified when predicting probabilities for ROC comparison.

Returning to our example, we can now use the `roctab` command to generate a ROC curve. `roctab` is used to perform nonparametric ROC analyses. It calculates the area under a single ROC curve, and optionally, it can plot the ROC curve. The simplest syntax for `roctab` is

`roctab` *refvar classvar*

where *refvar*, the reference variable, is a dichotomous variable indicating the true state of each observation, such as diseased and non-diseased or normal and abnormal, and variable *classvar* contains the outcome of the classification test. See [R] **roc** or type `help roctab` for more details and for additional options.

Using the original outcome variable `foreign` as the reference variable and the predicted probabilities from `predict` as the classification variable, we obtain

```
. roctab foreign p
                      ROC                    —Asymptotic Normal——
           Obs       Area     Std. Err.      [95% Conf. Interval]
         ————————————————————————————————————————————————————————
            74     0.5769       0.0747        0.43053     0.72331
```

Note that the area under the ROC curve computed by `roctab` is the same as that previously reported by `lroc`; thus, the two commands are equivalent, and in fact, they are identical. Note, however, that unlike `lroc`, `roctab` also reports the standard error and 95% confidence interval for the area under the curve.

Why does this work? Each logistic predicted probability is a possible cut-point for classifying subjects. For example, if in the above model we use `p` = 0.5 as a classification cut-point, then we could classify automobiles with `p` >= 0.5 as domestic and those with `p` < 0.5 as foreign, and then construct the following table:

Logistic	Actual Origin	
Classification	Foreign	Domestic
Foreign	A	B
Domestic	C	D

where A is the number of foreign cars correctly classified, and similarly, D is the number of domestic cars correctly classified. From the above table, we can compute the sensitivity, A/(A + C), and specificity, D/(B + D), of our classification cut-point. A perfectly discriminate cut-point would classify every automobile correctly; that is, both sensitivity and specificity would equal one.

If instead of selecting `p` = 0.5 we select `p` = 0.3, we would obtain different counts for A, B, C, and D, and consequently, different values for sensitivity and specificity. If we use each predicted probability value obtained from our model as a possible cut-point, we would obtain for each probability value an associated sensitivity and specificity. By plotting these sensitivity and specificity values, we generate a ROC curve. This is how both `lroc` and `roctab` construct a ROC curve.

The same approach for computing the area under the ROC is followed for a probit model. That is, estimate the model, predict the predicted probabilities, and then use these probabilities in `roctab` to produce the ROC. We illustrate with the same setup as before:

```
. probit foreign price
Iteration 0:   log likelihood =  -45.03321
Iteration 1:   log likelihood =  -44.94401
Iteration 2:   log likelihood = -44.943972
Probit estimates                                  Number of obs   =         74
                                                  LR chi2(1)      =       0.18
                                                  Prob > chi2     =     0.6727
Log likelihood = -44.943972                       Pseudo R2       =     0.0020

------------------------------------------------------------------------------
     foreign |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |   .0000222   .0000522     0.42   0.671    -.0000802    .0001245
       _cons |  -.6701415   .3605534    -1.86   0.063    -1.376813    .0365302
------------------------------------------------------------------------------

. predict lp if e(sample), p
. roctab foreign lp
                      ROC                    -Asymptotic Normal--
           Obs       Area     Std. Err.      [95% Conf. Interval]
         --------------------------------------------------------
            74     0.5769       0.0747        0.43053     0.72331
```

Note that, not surprisingly, given the similarity between the probit and logit models, the area under this curve is the same as that previously obtained, at least out to the reported precision.

2 Comparing models estimated using the same set of observations

This entry refers to situations where all models to be compared are being estimated on the same set of observations. In this situation, difficulties can arise if there are missing values in covariates included in some models and not in others. For now, let's put this issue aside and look at the simple case where there are no missing covariate values and all models to be compared use the same observations.

In the previous section, we saw how to obtain the ROC from logit and probit models using `roctab`. We did this because `roccomp` computes ROC areas in the same way, except that it repeats the process for each curve to be compared.

We now illustrate how to compute ROC curves from two nested logistic models and compare their areas.

We begin with the same logistic model that we estimated before and save the predicted probabilities in the `p1` variable:

(*Continued on next page*)

```
. use http://www.stata-press.com/data/r7/auto, clear
(1978 Automobile Data)

. logistic foreign price

Logit estimates                                   Number of obs   =         74
                                                  LR chi2(1)      =       0.17
                                                  Prob > chi2     =     0.6784
Log likelihood =  -44.94724                       Pseudo R2       =     0.0019

------------------------------------------------------------------------------
     foreign | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |   1.000035   .0000844     0.42   0.676     .9998699    1.000201
------------------------------------------------------------------------------

. predict p1 if e(sample), p

. lroc, nograph

Logistic model for foreign

number of observations =       74
area under ROC curve   =   0.5769
```

We now add the variable `mpg` (miles per gallon) as an independent variable in the model and save the predicted probabilities from this second model in variable `p2`:

```
. logistic foreign price mpg

Logit estimates                                   Number of obs   =         74
                                                  LR chi2(2)      =      17.14
                                                  Prob > chi2     =     0.0002
Log likelihood = -36.462189                       Pseudo R2       =     0.1903

------------------------------------------------------------------------------
     foreign | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |   1.000266   .0001166     2.28   0.022     1.000038    1.000495
         mpg |   1.263436   .0848332     3.48   0.000     1.107642    1.441143
------------------------------------------------------------------------------

. predict p2 if e(sample), p

. lroc, nograph

Logistic model for foreign

number of observations =       74
area under ROC curve   =   0.8112
```

Note that in both models, the complete data and the same 74 observations were used. We asked Stata to compute the area under the ROC curve after estimating each model. Although this is not necessary, we did it so that we can compare these areas with those reported by `roccomp`. In variable `p1`, we have the predicted probabilities from the first model, and in `p2`, the predicted probabilities from the second model. To compare the areas under the two corresponding ROCs, we use `roccomp`.

As mentioned in `roccomp`'s help file, `roccomp` expects the data to be in wide form when comparing areas estimated from the same sample, and that is exactly how we have our data. Here is a partial list:

```
. list foreign p1 p2 in 1/5
       foreign         p1         p2
  1.   Foreign    .291237   .9669088
  2.   Foreign   .2853879   .8528188
  3.   Foreign   .2797308   .8244022
  4. Domestic   .3731745   .8166367
  5. Domestic   .2842094   .8144851
```

Each observation in the data contains the outcome variable, `foreign`, and the two variables `p1` and `p2` containing the predicted probabilities generated from each of our two logistic models. If we were to compare three models, we would expect each observation to have, in addition to the outcome variable, three new variables containing the predicted probabilities from the three models, and so on.

The syntax for `roccomp`, without options, for comparing ROC areas estimated from the same sample is

`roccomp` *refvar classvar classvar* [*classvars*]

where, as in `roctab`, the reference variable, *refvar*, is a dichotomous variable indicating the true state of each observation, and the *classvar* variables contain the outcome of each of the classification tests applied to the observation. See [R] **roc** or type `help roccomp` for more details and additional options.

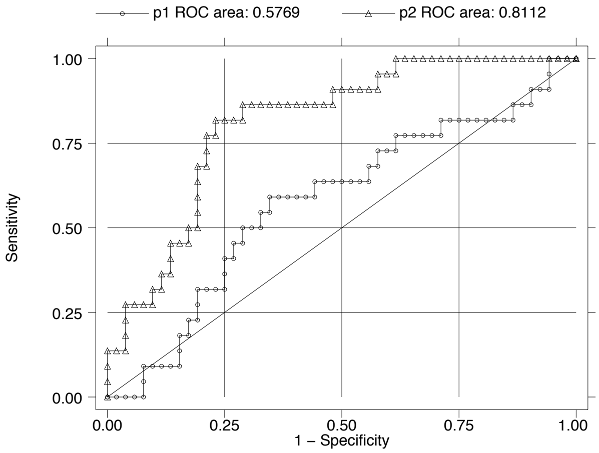
Using the original outcome variable foreign as the reference variable and the predicted probabilities `p1` and `p2` as the classification variables, we obtain

```
. roccomp foreign p1 p2
                               ROC                    —Asymptotic Normal——
                    Obs       Area     Std. Err.      [95% Conf. Interval]
─────────────────────────────────────────────────────────────────────────
p1                   74     0.5769       0.0747        0.43053     0.72331
p2                   74     0.8112       0.0514        0.71040     0.91198
─────────────────────────────────────────────────────────────────────────
Ho: area(p1) = area(p2)
    chi2(1) =     6.03       Prob>chi2 =   0.0141
```

First, note that the areas that `roccomp` reports for the two curves are the same as those computed by `lroc` above. We are therefore confident that we are comparing the correct areas from the two models. `roccomp` computed a significance probability of 0.0141, suggesting that the two models are different in their predictive ability. We can visually examine this difference by specifying the graph option of `roccomp`:

```
. roccomp foreign p1 p2, graph s(oT)
```

(*Continued on next page*)



When comparing areas under ROC curves from models estimated on the same sample, it is important that we be cognizant of the actual estimation sample that each model is using. Stata makes this evaluation easy by identifying the used observations by flagging them with `e(sample)`. We want to make sure that predict only generates predicted probabilities for the sample used in the estimation so that the ROC curves compared are the correct ones based on the estimated models. That is why we previously recommended that `if e(sample)` always be specified when using predict in the current context. Difficulties in comparing ROC curves can arise when there are missing covariate values that drop observations from some models and not from others.

Recall that Stata will drop from the estimation any observation in which at least one of the specified model covariate values is missing. Therefore, if we have a dataset in which the variable `age` is never missing and we estimate, for example, a logistic model using `age` as the only covariate, then every observation in the dataset will be included in the estimation. If, on the other hand, the dataset contains the variable `sex` missing in 5% of observations, then a logistic model with `age` and `sex` as covariates would drop 5% of the observations due to the missing value for `sex` in these observations. Additionally, `roccomp` will drop any observations in which at least one of the predicted probability values is missing. If we use `roccomp` to compare the two ROC areas from these two nested models, although `roccomp` will correctly compare the curves based on the nonmissing data, the comparison may not be the one that we think we are making.

To illustrate this using the auto data, assume that `rep78` is a continuous variable that can be included directly in our models. In reality, `rep78` is a categorical variable and would need to be “dummied-up” for model inclusion. In our data, the `rep78` variable is missing in five observations.

```
. use http://www.stata-press.com/data/r7/auto, clear
(1978 Automobile Data)

. logistic foreign price mpg

Logit estimates                                   Number of obs   =         74
                                                  LR chi2(2)      =      17.14
                                                  Prob > chi2     =     0.0002
Log likelihood = -36.462189                       Pseudo R2       =     0.1903

------------------------------------------------------------------------------
     foreign | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |   1.000266   .0001166     2.28   0.022     1.000038    1.000495
         mpg |   1.263436   .0848332     3.48   0.000     1.107642    1.441143
------------------------------------------------------------------------------

. predict p1 if e(sample), p

. lroc, nograph

Logistic model for foreign

number of observations =       74
area under ROC curve   =   0.8112

. logistic foreign price mpg rep78

Logit estimates                                   Number of obs   =         69
                                                  LR chi2(3)      =      34.08
                                                  Prob > chi2     =     0.0000
Log likelihood = -25.362394                       Pseudo R2       =     0.4018

------------------------------------------------------------------------------
     foreign | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |   1.000141   .0001379     1.03   0.305     .9998712    1.000412
         mpg |    1.18063   .0966693     2.03   0.043     1.005583    1.386148
       rep78 |   5.321595   2.656575     3.35   0.001     2.000398    14.15687
------------------------------------------------------------------------------

. predict p2 if e(sample), p
(5 missing values generated)

. lroc, nograph

Logistic model for foreign

number of observations =       69
area under ROC curve   =   0.9147
```

Note that all 74 observations were used in the first logistic model, whereas the second model with `rep78` was estimated using only 69 observations. That is because, as mentioned, `rep78` is missing in five observations. Let's now examine what `roccomp` does with these data.

```
. roccomp foreign p1 p2

                              ROC                    -Asymptotic Normal--
                 Obs       Area     Std. Err.      [95% Conf. Interval]
-------------------------------------------------------------------------
p1                69     0.8264       0.0515        0.72535     0.92742
p2                69     0.9147       0.0352        0.84562     0.98374
-------------------------------------------------------------------------
Ho: area(p1) = area(p2)
    chi2(1) =     3.86       Prob>chi2 =   0.0495
```

First, note that the number of observations for `p1` and `p2` are both 69. Although the first model, on which `p1` was predicted, was estimated using all 74 observations, when comparing the ROC curves, five observations were dropped due to missing `p2` values. However, the number of observations is not the only difference, and more importantly, the ROC area for the first model is not 0.8112 as `lroc` reported but is now 0.8264 based on the 69 observations that remained.

Because `roccomp` performs the correct comparison based on the data that remain after dropping missing values, we may be misled into thinking that all is fine, but it is not. Note that `p1` was predicted based on a model that had 74 observations. Had we dropped the five observations with missing `rep78` before we began, we would obtain different values for `p1` and, consequently, a different ROC area from the one computed by either `lroc` or `roccomp` above.

```
. use http://www.stata-press.com/data/r7/auto, clear
(1978 Automobile Data)
. drop if rep78==.
(5 observations deleted)
. quiet logistic foreign price mpg
. predict p1 if e(sample), p
. lroc, nograph
Logistic model for foreign
number of observations =       69
area under ROC curve   =   0.8284
. quiet logistic foreign price mpg rep78
. predict p2 if e(sample), p
. lroc, nograph
Logistic model for foreign
number of observations =       69
area under ROC curve   =   0.9147
. roccomp foreign p1 p2
                             ROC                    —Asymptotic Normal——
                  Obs       Area     Std. Err.      [95% Conf. Interval]
────────────────────────────────────────────────────────────────────────
p1                 69     0.8284       0.0511        0.72813     0.92862
p2                 69     0.9147       0.0352        0.84562     0.98374
────────────────────────────────────────────────────────────────────────
Ho: area(p1) = area(p2)
    chi2(1) =     3.49       Prob>chi2 =   0.0617
```

Although the difference is not large in this example, it can be, and often is, quite substantial for other larger datasets or models.

3 Models estimated using different sets of observations

On occasion, we may want to compare the same model estimated on different sets of similar observations. For example, we may want to compare the ROC curve produced from a model applied to our data to the ROC curve produced by the same model applied to a colleague's data. Or, for example, in a given study, we may want to compare the ROC curve produced from a model using only males to the same model applied to only females. Thus, we may have two or more models estimated on separate datasets, or two or more models estimated on subsets of the same dataset.

We describe the procedure by comparing ROC curves computed from models applied to subsets of data. We again use the auto data distributed with Stata 7.0. We want to compare the area under the ROC curve from a logistic model regressing `price` and `mpg` on `foreign` using only autos with a Repair Record of 3 or less to a similar model fitted to autos with Repair Records of 4 or 5. We begin by creating a dummy or indicator variable, `rep78_dummy`, identifying the two groups that we wish to compare. Note that of the original 74 observations, 5 have a missing repair record and are not included in the analysis.

```
. use http://www.stata-press.com/data/r7/auto, clear
(1978 Automobile Data)

. gen rep78_dummy=1 if rep78<=3
(34 missing values generated)

. replace rep78_dummy=2 if rep78==4 | rep78==5
(29 real changes made)
```

We now fit a logistic model to each subset of data and obtain the corresponding predicted probabilities.

```
. logistic foreign price mpg if rep78_dummy==1

Logit estimates                                   Number of obs   =         40
                                                  LR chi2(2)      =       3.06
                                                  Prob > chi2     =     0.2164
Log likelihood = -9.1246556                       Pseudo R2       =     0.1437

------------------------------------------------------------------------------
     foreign | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |   .9999979   .0002927    -0.01   0.994     .9994243    1.000572
         mpg |   1.307117   .2337931     1.50   0.134     .9205915     1.85593
------------------------------------------------------------------------------

. predict p1 if e(sample), p
(34 missing values generated)

. lroc, nograph

Logistic model for foreign

number of observations =       40
area under ROC curve   =   0.8378

. logistic foreign price mpg if rep78_dummy==2

Logit estimates                                   Number of obs   =         29
                                                  LR chi2(2)      =       8.27
                                                  Prob > chi2     =     0.0160
Log likelihood = -15.112878                       Pseudo R2       =     0.2148

------------------------------------------------------------------------------
     foreign | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |   1.000596   .0003265     1.83   0.068     .9999563    1.001236
         mpg |   1.232908   .1115754     2.31   0.021     1.032521    1.472185
------------------------------------------------------------------------------
```

```
. predict p2 if e(sample), p
(45 missing values generated)
. lroc, nograph
Logistic model for foreign
number of observations =       29
area under ROC curve   =   0.7929
```

The procedure, so far, is similar to that of the previous section, with the exception that we have included a conditional `if` statement in order to estimate the model on the proper data subset. It is very important in this situation to specify `e(sample)` with `predict`. If we fail to specify `e(sample)`, `predict` will generate predicted probabilities for all observations in memory. Let's examine our data in memory by listing a few observations.

```
. list foreign p1 p2 rep78_dummy in 1/5
      foreign         p1          p2  rep78_d~y
  1.  Foreign          .    .9806504          2
  2.  Foreign          .    .9593273          2
  3.  Foreign          .    .9444055          2
  4.  Foreign          .     .899147          2
  5. Domestic          .    .8676886          2
. sort rep78_dummy
. list foreign p1 p2 rep78_dummy in 1/5
      foreign         p1          p2  rep78_d~y
  1. Domestic   .0067503           .          1
  2. Domestic   .0915387           .          1
  3. Domestic   .1472585           .          1
  4.  Foreign   .2277415           .          1
  5. Domestic   .0557233           .          1
```

We see that for observations with `rep78_dummy==2`, the values for `p1` are always missing and the values for `p2` are filled in, and for observations with `rep78_dummy==1`, the opposite is true. This is because we predicted `p1` based on a model fitted for observations with `rep78_dummy==1`, and `p2` was predicted based on a model fitted for observations with `rep78_dummy==2`. This is exactly as it should be. Only those observations included in the estimation sample should contain the predicted probabilities.

If we attempt to use `roccomp` as before, we will get an error because all observations will be dropped due either to missing `p1` or missing `p2`.

```
. roccomp  foreign p1 p2
Outcome does not vary
r(198);
```

As stated in the help file, `roccomp` expects the data to be in long form for areas estimated from independent samples. In this case, the simplest syntax for `roccomp`, without options, for comparing ROC areas estimated from independent samples is

`roccomp` *refvar classvar*, `by(`*varname*`)`

Note that we can only specify one classification variable and must specify the `by()` option. So, we must create a single classification variable based on our predicted probabilities, and then use `roccomp` specifying `by(rep78_dummy)`.

```
. gen newp=p1 if p1~=.
(34 missing values generated)

. replace newp=p2 if p2~=.
(29 real changes made)

. roccomp foreign newp, by(rep78_dummy)
                              ROC                   -Asymptotic Normal--
rep78_dummy        Obs       Area     Std. Err.      [95% Conf. Interval]
-------------------------------------------------------------------------
1                   40     0.8378       0.0796        0.68184     0.99384
2                   29     0.7929       0.0982        0.60039     0.98547
-------------------------------------------------------------------------
Ho: area(1) = area(2)
    chi2(1) =     0.13       Prob>chi2 =   0.7224
```

We can verify that the ROC areas reported by `roccomp` are the same as those previously obtained by `lroc` for the two models.

Although in the previous examples we have only compared two ROC areas, `roccomp` has no limit on the number of areas that it can compare. For example, we can compare the areas under the ROC curves for `rep78==3`, `rep78==4`, and `rep78==5`.

```
. use http://www.stata-press.com/data/r7/auto, clear
(1978 Automobile Data)

. quiet logistic foreign price mpg if rep78==3

. predict p3 if e(sample), p
(44 missing values generated)

. quiet logistic foreign price mpg if rep78==4

. predict p4 if e(sample), p
(56 missing values generated)

. quiet logistic foreign price mpg if rep78==5

. predict p5 if e(sample), p
(63 missing values generated)

. quiet gen newp=p3 if p3~=.

. quiet replace newp=p4 if p4~=.

. quiet replace newp=p5 if p5~=.

. roccomp foreign newp, by(rep78)
                              ROC                   -Asymptotic Normal--
rep78              Obs       Area     Std. Err.      [95% Conf. Interval]
-------------------------------------------------------------------------
3                   30     0.8642       0.0723        0.72247     1.00000
4                   18     0.8765       0.0905        0.69914     1.00000
5                   11     0.7778       0.1470        0.48969     1.00000
-------------------------------------------------------------------------
Ho: area(3) = area(4) = area(5)
    chi2(2) =     0.35       Prob>chi2 =   0.8407
```

Not a very interesting example, but it illustrates well the procedure for comparing several ROC curves, each computed on a subset of observations.

Finally, a similar procedure works to compare ROC curves from models estimated on different datasets. Simply, estimate the model on each dataset separately, estimate the predicted probabilities and save them in a variable with same name in both datasets, create a dummy variable that identifies the datasets, append the datasets, and then use `roccomp` with the `by()` option. For example, assume that we have two auto datasets, `auto1.dta` and `auto2.dta`. Then, the commands needed to compare the ROC curves are

```
. use auto1.dta, clear
. logistic foreign price mpg
. predict p if e(sample), p
. gen dataset=1
. save temp,replace

. use auto2.dta, clear
. logistic foreign price mpg
. predict p if e(sample), p
. gen dataset=2

. append using temp
. roccomp foreign p, by(dataset)
```

About the Author

Mario Cleves is an Associate Professor in the Department of Pediatrics at the University of Arkansas for Medical Sciences.