ROC: a gretl function package for receiver operating characteristic curves

Peter M. Summers
High Point University*

July 7, 2015

This document describes the gretl function package roc.gfn, which can be used to compute and plot a model's receiver operating characteristic (ROC) curve and related statistics. It is intended primarily as a complement to estimation of logit or probit models, and provides additional information about how well such models fit the data.

1 Analysis of ROC curves

Consider the following logit model based on Mroz (1987), estimated using the native gretl data set mroz87.gdt:

```
logit LFP const WA WE KL6
```

In this model, the dependent variable LFP is the labor force participation of a woman in 1975, equal to 1 if she worked outside the home and zero otherwise. The explanatory variables (apart from the constant) are her age (WA), her educational attainment (WE), and the number of children in the household under six years old. Estimating this model in gretl gives the results in table 1:

Part of the standard gretl output is the number of cases correctly predicted, which in this case is 66.1% of the sample. In addition, estimating the model in a script generates the following table:

		Predicted	
		0	1
Actual	0	156	169
	1	86	342

*Email: psummers@highpoint.edu. Thanks to Artur Bala for many helpful suggestions.

Table 1: Logit estimation results

Model 1: Logit, using observations 1–753
Dependent variable: LFP
Standard errors based on Hessian

	Coefficient	Std. Error	z	Slope*
const	0.523316	0.675893	0.7743	
WA	−0.0558658	0.0111462	−5.0121	−0.0136657
WE	0.202588	0.0372229	5.4426	0.0495564
KL6	−1.43215	0.191385	−7.4831	−0.350328

Mean dependent var	0.568393	S.D. dependent var	0.495630
McFadden R^2	0.094843	Adjusted R^2	0.087074
Log-likelihood	−466.0412	Akaike criterion	940.0823
Schwarz criterion	958.5786	Hannan–Quinn	947.2080

*Evaluated at the mean

Number of cases 'correctly predicted' = 498 (66.1 percent)
Likelihood ratio test: $\chi^2(3)$ = 97.664 [0.0000]

To generalize from this example, let the actual and fitted values from the model be y and $\hat{y}$, respectively. Both the fraction correctly predicted and the "predicted/actual" table are based on a classification rule that predicts $y = 1$ if $\hat{y} \geq 0.5$, and zero otherwise. The correctly predicted observations are those in the upper left and lower right cells of the table above (156 and 342), which represent 66.1% of the total number of observations.

The intuition behind the ROC curve is to repeat this sort of classification for each value of $\hat{y}$. Suppose that the value 1 represents a "positive" outcome, and 0 a "negative". Re-write the classification table as

		Predicted	
		0	1
Actual	0	TN	FP
	1	FN	TP

where "TN," "FN," "FP," and "TP" denote the number of "true negative," "false negative," "false positive," and "true positive" outcomes, respectively. The total number of positive outcomes is $P = TP + FN$, and similarly $N = TN + FP$.

The ROC curve is a plot of the true positive rate TP/P (also known as the "sensitivity") against the false positive rate FP/N (also known as one minus the "specificity"), with the rates computed for each value of $\hat{y}$. Figure 1 plots the ROC curve from the model in table 1.

In addition to plotting the curve, the area under the curve (AUROC), its standard error, and a 95% confidence interval are also reported. The 45-degree line represents a completely uninformative classifier; such a model would predict the data no better than a coin flip. The area under this line is obviously 0.5. The ROC curve for a perfect classifier would follow the right-hand and top sides of the figure, and have an AUROC of 1.

The roc function computes AUROC according to the formula in DeLong, DeLong, and Clarke-Pearson (1988). This statistic has an asymptotically Normal distribution (see Mason and Graham (2002) for a discussion). The standard error is computed in three different ways, with the default being that of DeLong et al.[1] Alternatives are the Normal approximation, and the method of Hanley and McNeil (1982).

The syntax for calling the function is[2]

```
? result = roc(y, yhat, verbose)
```

[1] This is also the default standard error computed in Stata's "roctab" function.

[2] The roc function expects y and yhat to be series; a matrix version of the function, named mroc, allows for matrix inputs but is otherwise identical.

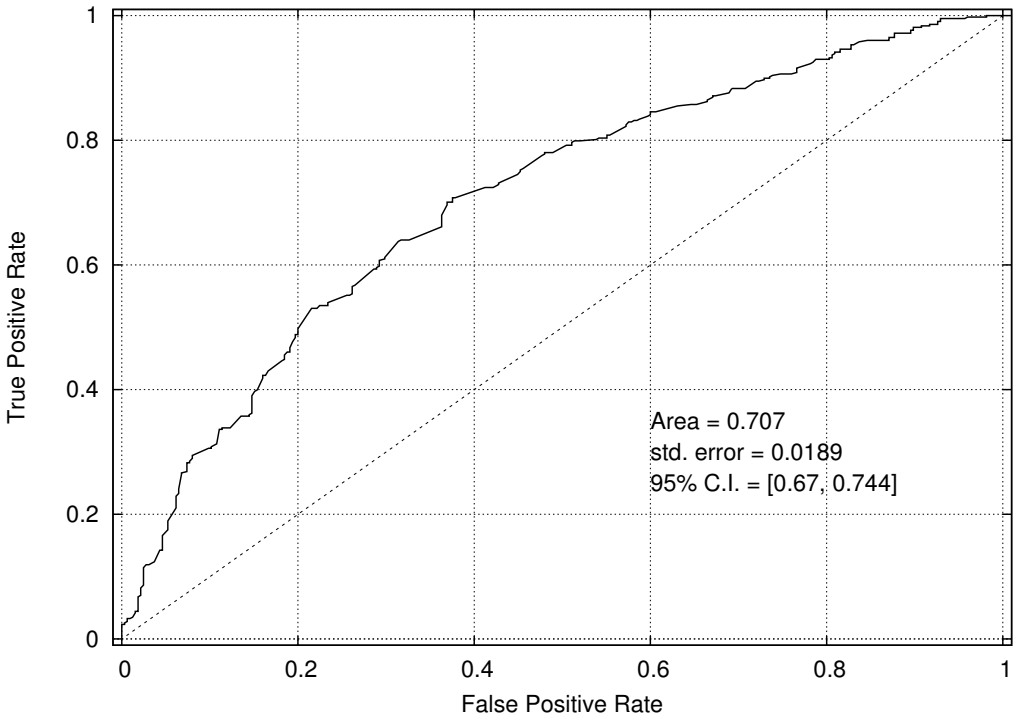

Figure 1: ROC curve from logit model for Mroz data

In addition to the required inputs – the actual and predicted values – the optional third argument controls the amount of output produced. The default value is 2, which produces two plots and a summary table. The first plot is the ROC curve shown in figure 1, and the second plots the fraction correctly predicted versus the value of $\hat{y}$, as shown in figure 2.

The summary table is:

```
--------------------------------------------------
                    ROC Analysis
--------------------------------------------------
Area under curve (std. error) =  0.707 (0.0189)
95% C.I. =   [0.67, 0.744]
Max correctly predicted = 0.672 at threshold 0.557
Youden index  = 0.333
--------------------------------------------------
```

The Youden index is the maximum vertical distance between the ROC curve and the 45-degree line. It is the maximum distance between the true positive rate and the false positive rate, and occurs at the same threshold as the maximum correctly predicted.[3] A verbosity value of 1 suppresses the figures but prints the table, while a value of 0 suppresses all printed output.

The function returns a bundle (named "result" here) with the following elements:

- auroc: the area under the ROC curve
- se: the standard error computed according to DeLong, DeLong, and Clarke-Pearson (1988)
- se_Normal: the standard error using the Normal approximation
- se_Hanley: the standard error using the method of Hanley and McNeil (1982)
- maxfcp: the maximum fraction correctly predicted
- thresh: the "threshold" value of $\hat{y}$ that results in maxfcp% correctly predicted (this may not be unique)
- Youden: the Youden index
- tpr: a $(1 \times n)$ column vector with the true positive rate for each value of $\hat{y}$

[3] See Kumar and Indrayan (2011) for more details.

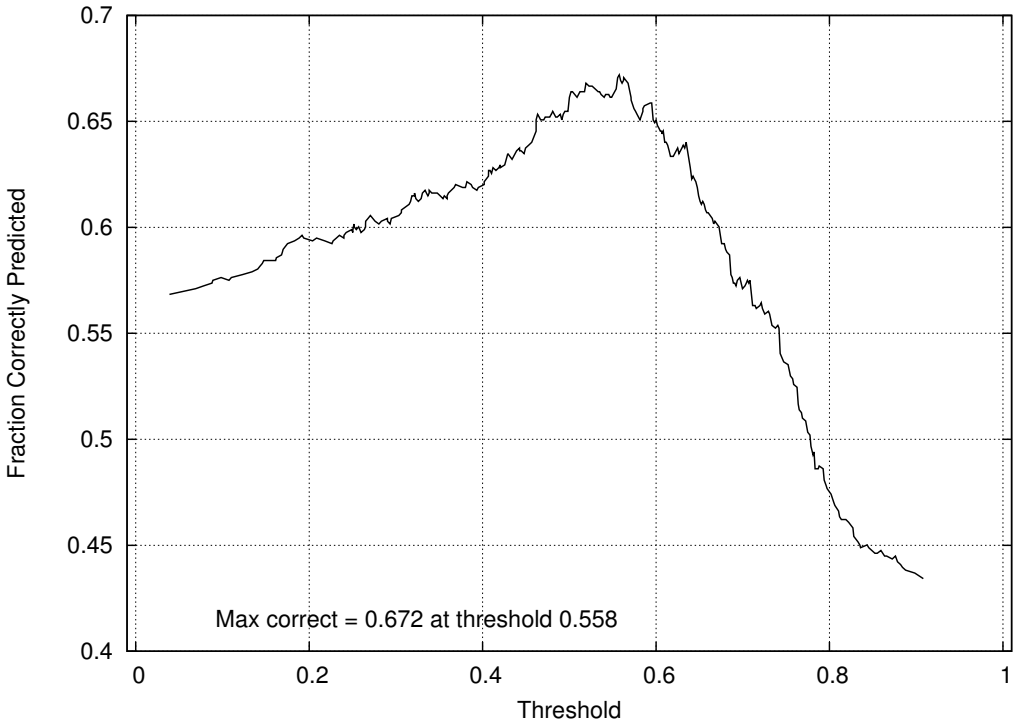

Figure 2: Fraction correctly predicted, logit model for Mroz data

- fpr: a $(1 \times n)$ column vector with the false positive rate for each value of $\hat{y}$
- fcp: a $(1 \times n)$ column vector with the fraction correctly predicted for each value of $\hat{y}$

If desired, the figures can be generated manually from the bundle output using gretl's gnuplot command or the plot environment. The following commands reproduce a version of figure 1:

```
matrix figmat = result.tpr ~ result.fpr ~ result.fpr
gnuplot --matrix=figmat --output=display --with-lines --fit=none \
{set xlabel "False Positive Rate"; set ylabel "True Positive Rate";\
set nokey}
```

For figure 2, use

```
matrix thresh = {yhat1}
thresh = sort(thresh)
matrix figmat = result.fcp ~ thresh
gnuplot --matrix=figmat --output=display --with-lines --fit=none \
{set xlabel "Threshold"; \
set ylabel "Fraction Correctly Predicted"; set nokey}
```

2 Comparing two or more ROC curves

The roc package also includes the function roccomp, which tests whether two or more models have the same area under their ROC curves. The general syntax for calling the function is

```
result = roccomp(y, yhat, names, verbose).
```

The arguments are:

1. y (series): the (binary) dependent variable
2. yhat (list): the predicted values from 2 or more models [4]
3. names (string): optional space-separated string of names used in formatting the output (default = null). If not supplied, the function generates the default labels "yhat1", "yhat2," etc.

[4] As with the roc function, there is a companion function "mroccomp" that takes matrix arguments for y and yhat.

4. verbose (boolean): optional argument producing printed output (verbose = 1, default) or suppressing it (verbose = 0)

Regardless of the value of "verbose," the function returns a bundle with elements

- roc_info: The AUROC corresponding to each model in yhat, along with its standard error,
- roc_vcv: The covariance matrix of the AUROCs
- roc_tests: The test statistics and p-values for the pair-wise and joint tests of differences in AUROC between models (see below).

Returning to the mroz87 data set, suppose we estimate the following two models in addition to the one reported above:

```
logit LFP const WA WE KL6 HA HE FAMINC

logit LFP const WA WE KL6 HA HE FAMINC AX MTR
```

The additional regressors are: husband's age (HA) and education (HE), family income (FAMINC), the wife's actual years of experience (AX), and the marginal tax rate facing the wife (MTR).[5] If we label these models "model1," "model2" and "model3" respectively, we can use the roccomp function to compute the area under each model's ROC curve and test for any significant differences between them. The following lines illustrate the procedure:

```
logit LFP const WA WE KL6 --quiet
series yhat1 = $yhat

logit LFP const WA WE KL6 HA HE FAMINC --quiet
series yhat2 = $yhat

logit LFP const WA WE KL6 HA HE FAMINC AX MTR --quiet
series yhat3 = $yhat

list yhat = yhat1 yhat2 yhat3

string names = "model1 model2 model3"
result = roccomp(LFP, yhat, names)
```

[5]The models listed here are for illustrative purposes only, and are not intended as a rigorous analysis of this data set.

The output from the function looks like this:

```
-------------------------------------------------------
                    ROC Comparison
-------------------------------------------------------
Areas under ROC curves and their standard errors:
                Area  std. error
model1        0.7072      0.0189
model2        0.7158      0.0187
model3        0.8014      0.0160

-------------------------------------------------------
Tests of no difference:
                         test stat     p-value
      model1-model2        -1.3231      0.1858
      model2-model3        -5.5933      0.0000
      model1-model3        -5.7565      0.0000
All differences = 0        33.7510      0.0000

Joint test is chi-square with 2 degrees of freedom
```

The top panel of this table reports the AUROC for each model along with its standard error. The bottom panel reports the results of pairwise tests for no significant difference between the AUROC measures, as well as the test that all three AUROC measures are the same. The null hypotheses for the pairwise tests are $H_0 : AUROC_i - AUROC_j = 0, i \neq j$, and the test statistics are Normally distributed. For the joint test, let $\hat{\theta} = (AUROC_1, AUROC_2, AUROC_3)'$. Then the null is $H_0 : L\hat{\theta} = 0$, where L is the difference matrix

$$L = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}.$$

The test statistic is $Q = \hat{\theta}' L' \left[LSL'\right]^{-1} L\hat{\theta}$, where S is the covariance matrix of the AUROCs. Under the null, $Q \sim \chi^2(2)$; more generally, $Q \sim \chi^2(k-1)$ where k is the number of different models being considered.[6] Users can conduct other tests of linear restrictions on the elements of $\hat{\theta}$ by using the bundle elements roc_info and roc_vcv.

An example script reproducing the output shown here is provided in roc_example.inp.

[6] See DeLong, DeLong, and Clarke-Pearson (1988). The joint test is only reported for $k > 2$.

References

DeLong, Elizabeth R., David M. DeLong, and Daniel L. Clarke-Pearson (1988). "Comparing the areas under two or more receiver operating characteristics curves: A nonparametric approach". *Biometrics* 44.3, pp. 837–845.

Hanley, James A. and Barbara J. McNeil (1982). "The meaning and use of the area under a receiver operating characteristic (ROC) curve". *Radiology* 143.1, pp. 29–36.

Kumar, Rajeev and Abhaya Indrayan (2011). "Receiver Operating Charachteristic (ROC) Curve for Medical Researchers". *Indian Pediatrics* 48, pp. 277–287.

Mason, S. J. and N. E. Graham (2002). "Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation". *Quarterly Journal of the Royal Meterological Society* 128, pp. 2145–2166.

Mroz, T. A. (1987). "The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions". *Econometrica* 55, pp. 765–799.