Title

stata.com

rocfit — Parametric ROC models

Description

`rocfit` fits maximum-likelihood ROC models assuming a binormal distribution of the latent variable.

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.

See [R] **roc** for other commands designed to perform receiver operating characteristic (ROC) analyses with rating and discrete classification data.

Quick start

Binary true state, `true`, as a function of classification variable `class`

```
rocfit true class
```

As above, but with frequency weights `wvar`

```
rocfit true class [fweight = wvar]
```

Specify that `class` is continuous and generate `v1` containing classification groups

```
rocfit true class, continuous(3) generate(v1)
```

Menu

Statistics > Epidemiology and related > ROC analysis > Parametric ROC analysis without covariates

Syntax

```
rocfit refvar classvar [if] [in] [weight] [, rocfit_options]
```

rocfit_options	Description
Model	
`continuous(#)`	divide *classvar* into # groups of approximately equal length
`generate(`*newvar*`)`	create *newvar* containing classification groups
SE	
`vce(`*vcetype*`)`	*vcetype* may be `oim` or `opg`
Reporting	
`level(#)`	set confidence level; default is `level(95)`
Maximization	
maximize_options	control the maximization process; seldom used

`fp`, is allowed; see [U] **11.1.10 Prefix commands**.

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

See [U] **20 Estimation and postestimation commands** for more capabilities of estimation commands.

Options

Model

`continuous(#)` specifies that the continuous *classvar* be divided into # groups of approximately equal length. This option is required when *classvar* takes on more than 20 distinct values.

`continuous(.)` may be specified to indicate that *classvar* be used as it is, even though it could have more than 20 distinct values.

`generate(`*newvar*`)` specifies the new variable that is to contain values indicating the groups produced by `continuous(#)`. `generate()` may be specified only with `continuous()`.

SE

`vce(`*vcetype*`)` specifies the type of standard error reported. *vcetype* may be either `oim` or `opg`; see [R] ***vce_option***.

Reporting

`level(#)`; see [R] **estimation options**.

Maximization

maximize_options: `difficult`, `technique(`*algorithm_spec*`)`, `iterate(#)`, `[no]log`, `trace`, `gradient`, `showstep`, `hessian`, `showtolerance`, `tolerance(#)`, `ltolerance(#)`, `nrtolerance(#)`, `nonrtolerance`, and `from(`*init_specs*`)`; see [R] **maximize**. These options are seldom used.

Setting the optimization type to `technique(bhhh)` resets the default *vcetype* to `vce(opg)`.

Remarks and examples

stata.com

Dorfman and Alf (1969) developed a generalized approach for obtaining maximum likelihood estimates of the parameters for a smooth fitting ROC curve. The most commonly used method for ordinal data, and the one implemented here, is based upon the binormal model; see Pepe (2003), Pepe, Longton, and Janes (2009), and Janes, Longton, and Pepe (2009) for methods of ROC analysis for continuous data, including methods for adjusting for covariates.

The model assumes the existence of an unobserved, continuous, latent variable that is normally distributed (perhaps after a monotonic transformation) in both the normal and abnormal populations with means μ_n and μ_a and variances σ_n^2 and σ_a^2, respectively. The model further assumes that the K categories of the rating variable result from partitioning the unobserved latent variable by $K-1$ fixed boundaries. The method fits a straight line to the empirical ROC points plotted using normal probability scales on both axes. Maximum likelihood estimates of the line's slope and intercept and the $K-1$ boundaries are obtained simultaneously. See *Methods and formulas* for details.

The intercept from the fitted line is a measurement of $(\mu_a - \mu_n)/\sigma_a$, and the slope measures σ_n/σ_a.

Thus the intercept is the standardized difference between the two latent population means, and the slope is the ratio of the two standard deviations. The null hypothesis that there is no difference between the two population means is evaluated by testing that the intercept $= 0$, and the null hypothesis that the variances in the two populations are equal is evaluated by testing that the slope $= 1$.

▷ Example 1

We use Hanley and McNeil's (1982) dataset, described in example 1 of [R] **roctab**, to fit a smooth ROC curve assuming a binormal model.

```
. use http://www.stata-press.com/data/r14/hanley
. rocfit disease rating
Fitting binormal model:
Iteration 0:   log likelihood = -123.68069
Iteration 1:   log likelihood = -123.64867
Iteration 2:   log likelihood = -123.64855
Iteration 3:   log likelihood = -123.64855
Binormal model of disease on rating              Number of obs   =       109
Goodness-of-fit chi2(2) =       0.21
Prob > chi2             =     0.9006
Log likelihood          =  -123.64855
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
intercept	1.656782	0.310456	5.34	0.000	1.048300	2.265265
slope (*)	0.713002	0.215882	-1.33	0.184	0.289881	1.136123
/cut1	0.169768	0.165307	1.03	0.304	-0.154227	0.493764
/cut2	0.463215	0.167235	2.77	0.006	0.135441	0.790990
/cut3	0.766860	0.174808	4.39	0.000	0.424243	1.109477
/cut4	1.797938	0.299581	6.00	0.000	1.210770	2.385106

Index	Indices from binormal fit: Estimate	Std. Err.	[95% Conf.	Interval]
ROC area	0.911331	0.029506	0.853501	0.969161
delta(m)	2.323671	0.502370	1.339044	3.308298
d(e)	1.934361	0.257187	1.430284	2.438438
d(a)	1.907771	0.259822	1.398530	2.417012

(*) z test for slope==1

rocfit outputs the MLE for the intercept and slope of the fitted regression line along with, here, four boundaries (because there are five ratings) labeled /cut1 through /cut4. Also rocfit computes and reports four indices based on the fitted ROC curve: the area under the curve (labeled ROC area), $\delta(m)$ (labeled delta(m)), d_e (labeled d(e)), and d_a (labeled d(a)). More information about these indices can be found in *Methods and formulas* and in Erdreich and Lee (1981).

◁

Stored results

rocfit stores the following in e():

Scalars

e(N)	number of observations
e(k)	number of parameters
e(k_eq)	number of equations in e(b)
e(k_eq_model)	number of equations in overall model test
e(k_dv)	number of dependent variables
e(df_m)	model degrees of freedom
e(ll)	log likelihood
e(chi2_gf)	goodness-of-fit χ^2
e(df_gf)	goodness-of-fit degrees of freedom
e(p_gf)	χ^2 goodness-of-fit significance probability
e(area)	area under the ROC curve
e(se_area)	standard error for the area under the ROC curve
e(deltam)	delta(m)
e(se_delm)	standard area for delta(m)
e(de)	d(e) index
e(se_de)	standard error for d(e) index
e(da)	d(a) index
e(se_da)	standard error for d(a) index
e(rank)	rank of e(V)
e(ic)	number of iterations
e(rc)	return code
e(converged)	1 if converged, 0 otherwise

Macros

e(cmd)	rocfit
e(cmdline)	command as typed
e(depvar)	*refvar* and *classvar*
e(wtype)	weight type
e(wexp)	weight expression
e(title)	title in estimation output
e(chi2type)	GOF; type of model χ^2 test
e(vce)	*vcetype* specified in vce()
e(vcetype)	title used to label Std. Err.
e(opt)	type of optimization
e(which)	max or min; whether optimizer is to perform maximization or minimization
e(ml_method)	type of ml method
e(user)	name of likelihood-evaluator program
e(technique)	maximization technique
e(properties)	b V

Matrices

e(b)	coefficient vector
e(ilog)	iteration log (up to 20 iterations)
e(gradient)	gradient vector
e(V)	variance–covariance matrix of the estimators

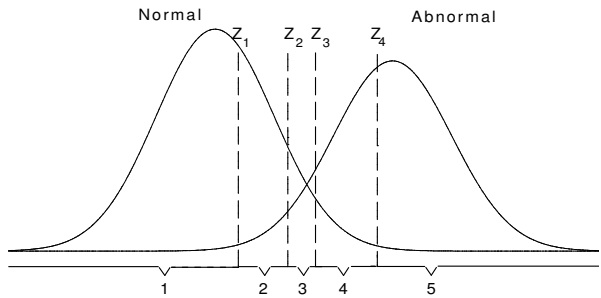
Functions

e(sample)	marks estimation sample

Methods and formulas

Dorfman and Alf (1969) developed a general procedure for obtaining maximum likelihood estimates of the parameters of a smooth-fitting ROC curve. The most common method, and the one implemented in Stata, is based upon the binormal model.

The model assumes that there is an unobserved continuous latent variable that is normally distributed in both the normal and abnormal populations. The idea is better explained with the following illustration:



The latent variable is assumed to be normally distributed for both the normal and abnormal subjects, perhaps after a monotonic transformation, with means μ_n and μ_a and variances σ_n^2 and σ_a^2, respectively.

This latent variable is assumed to be partitioned into the k categories of the rating variable by $k-1$ fixed boundaries. In the above figure, the $k=5$ categories of the rating variable identified on the bottom result from the partition of the four boundaries Z_1 through Z_4.

Let R_j for $j=1,2,\ldots,k$ indicate the categories of the rating variable, let $i=1$ if the subject belongs to the normal group, and let $i=2$ if the subject belongs to the abnormal group.

Then

$$p(R_j|i=1) = F(Z_j) - F(Z_{j-1})$$

where $Z_k = (x_k - \mu_n)/\sigma_n$, F is the cumulative normal distribution, $F(Z_0)=0$, and $F(Z_k)=1$. Also,

$$p(R_j|i=2) = F(bZ_j - a) - F(bZ_{j-1} - a)$$

where $b = \sigma_n/\sigma_a$ and $a = (\mu_a - \mu_n)/\sigma_a$.

The parameters a, b and the $k-1$ fixed boundaries Z_j are simultaneously estimated by maximizing the log-likelihood function

$$\log L = \sum_{i=1}^{2}\sum_{j=1}^{k} r_{ij}\log\{p(R_j|i)\}$$

where r_{ij} is the number of R_js in group i.

The area under the fitted ROC curve is computed as

$$\Phi\left(\frac{a}{\sqrt{1+b^2}}\right)$$

where Φ is the standard normal cumulative distribution function.

Point estimates for the ROC curve indices are as follows:

$$\delta(m) = \frac{a}{b} \qquad d_e = \frac{2a}{b+1} \qquad d_a = \frac{a\sqrt{2}}{\sqrt{1+b^2}}$$

Variances for these indices are computed using the delta method.

The $\delta(m)$ estimates $(\mu_a - \mu_n)/\sigma_n$, d_e estimates $2(\mu_a - \mu_n)/(\sigma_a - \sigma_n)$, and d_a estimates $\sqrt{2}(\mu_a - \mu_n)/(\sigma_a^2 - \sigma_n^2)^2$.

Simultaneous confidence bands for the entire curve are obtained, as suggested by Ma and Hall (1993), by first obtaining Working–Hotelling (1929) confidence bands for the fitted straight line in normal probability coordinates and then transforming them back to ROC coordinates.

References

Bamber, D. 1975. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* 12: 387–415.

Choi, B. C. K. 1998. Slopes of a receiver operating characteristic curve and likelihood ratios for a diagnostic test. *American Journal of Epidemiology* 148: 1127–1132.

Cleves, M. A. 1999. sg120: Receiver operating characteristic (ROC) analysis. *Stata Technical Bulletin* 52: 19–33. Reprinted in *Stata Technical Bulletin Reprints*, vol. 9, pp. 212–229. College Station, TX: Stata Press.

——. 2000. sg120.1: Two new options added to rocfit command. *Stata Technical Bulletin* 53: 18–19. Reprinted in *Stata Technical Bulletin Reprints*, vol. 9, pp. 230–231. College Station, TX: Stata Press.

Dorfman, D. D., and E. Alf, Jr. 1969. Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals–rating-method data. *Journal of Mathematical Psychology* 6: 487–496.

Erdreich, L. S., and E. T. Lee. 1981. Use of relative operating characteristic analysis in epidemiology: A method for dealing with subjective judgment. *American Journal of Epidemiology* 114: 649–662.

Hanley, J. A., and B. J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143: 29–36.

Janes, H., G. M. Longton, and M. S. Pepe. 2009. Accommodating covariates in receiver operating characteristic analysis. *Stata Journal* 9: 17–39.

Ma, G., and W. J. Hall. 1993. Confidence bands for the receiver operating characteristic curves. *Medical Decision Making* 13: 191–197.

Pepe, M. S. 2003. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York: Oxford University Press.

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

Working, H., and H. Hotelling. 1929. Application of the theory of error to the interpretation of trends. *Journal of the American Statistical Association* 24 (Suppl.): 73–85.

Also see

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

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

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

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