Statistics/Data Analysis

help incroc (version 1.0.2)

Title

incroc — Incremental value of a marker relative to a list of existing predictors. Evaluation is with respect to receiver operating characteristic (ROC) summary statistics. ROC statistics are based on percentile value (PV) calculations

Syntax

incroc *disease_var* [*if*] [*in*], **mod1(***varlist* **)** [**mod2(***varlist* **)** *options*]

options	Description
Models	
* **mod1(***varlist* **)**	model 1 predictor variables
mod2(*varlist* **)**	model 2 predictor variables
Cross-validation	
nsplit(*k***)**	specify number of random partitions for k-fold cross-validation; default is **nsplit(10)**
nocv	omit cross-validation
Summary statistics	
auc	area under ROC curve (AUC)
pauc(*f***)**	partial AUC (pAUC) for false positive rate range FPR < *f*
roc(*f***)**	ROC at specified FPR = *f*
rocinv(*t***)**	ROC^(-1)(t) at specified true positive rate TPR = *t*
Standardization method	
pvcmeth(*method* **)**	specify PV calculation method; **empirical** (default) or **normal**
tiecorr	correction for ties
Covariate adjustment	
adjcov(*varlist* **)**	specify covariates to adjust for
adjmodel(*model* **)**	specify model adjustment; **stratified** (default) or **linear**
Sampling variability	
nsamp(*#***)**	specify number of bootstrap samples; default is **nsamp(1000)**
nobstrap	omit bootstrap sampling
noccsamp	specify cohort rather than case-control bootstrap sampling
nostsamp	draw samples without respect to covariate strata
cluster(*varlist* **)**	specify variables identifying bootstrap resampling clusters
resfile(*filename* **)**	save bootstrap results in *filename*
replace	overwrite specified bootstrap results file if it already exists
level(*#***)**	specify confidence level; default is **level(95)**
New variable	
genxb	generate new variable, **xb***#*, to hold predicted values for model *#*
replace	overwrite existing **xb***#* variables

* **mod1(***varlist* **)** is required.

Description

incroc compares linear predictors from two logit models of disease outcome with respect to one or more ROC statistics: the AUC, the pAUC for FPR < *f*, the ROC at FPR = *f*, and the inverse ROC at TPR = *t* (Pepe Section 4.3, Dodd and Pepe). Model variables are specified with **mod1(***varlist***)** and **mod2(***varlist***)**. Though any two lists of model variables can be specified, **incroc** was motivated by interest in the incremental value of a marker relative to other predictors, implying nested variable lists differing by the inclusion of the marker variable. *disease_var* is the 0/1 disease indicator variable.

Alternatively, a single model can be specified with **mod1(***varlist***)**, in which case the requested ROC statistics are returned without comparison statistics.

The model predicted values used for ROC estimation are obtained by k-fold cross-validation.

All ROC statistics are calculated by using PVs of the disease case measures relative to the corresponding linear predictor distribution among controls (Pepe and Longton, Huang and Pepe).

Optional covariate adjustment can be achieved either by stratification or with a linear regression approach (Janes and Pepe, 2008; Janes and Pepe, in press).

Bootstrap standard errors and confidence intervals (CIs) for the requested statistics and linear predictor differences are calculated. Percentile, normal-based, and bias-corrected CIs are displayed (see estat bootstrap and **[R] bootstrap**).

Wald test results for predictor comparisons are based on the bootstrap standard errors for the difference between predictors.

Many of the standard e-class bootstrap results left behind by **bstat** are available after running **incroc**, in addition to the r-class results listed below.

Options

Models

mod1(*varlist***)** specifies the variables to be included in the linear predictor of the the first logit model. **mod1()** is required.

mod2(*varlist***)** specifies the variables to be included in the linear predictor of the second logit model.

Cross-validation

nsplit(*k***)** specifies the number of random partitions of the data that are to be used to obtain predicted values via k-fold cross-validation. The default is **nsplit(10)**. Specifying **nsplit(***1***)** is equivalent to specifying **nocv**.

nocv specifies that cross-validation is not used. Instead, within sample predicted values are obtained from a single model fit to the full sample. **nocv** overides any specification for **nsplit(***k***)**.

Summary statistics

auc compares predictors with respect to the AUC. This is the default if no summary statistics are specified. +

pauc(*f***)** includes a comparison with respect to the pAUC for FPR < specified *f*. The argument must be between 0 and 1. A tie correction is included in the PV calculation if this option is included among the specified summary statistics options and the empirical PV calculation method is used.

roc(*f***)** compares predictors with respect to the ROC at the specified FPR = *f*. The argument must be between 0 and 1.

rocinv(*t***)** compares predictors with respect to the inverse ROC, ROC^(-1)(t), at the specified TPR = *t*. The argument must be between 0 and 1.

Standardization method

pvcmeth(*method***)** specifies how the PVs are to be calculated. *method* can be one of the following:

empirical, the default, uses the empirical distribution of the test measure among controls (D=0) as the reference distribution for the calculation of case PVs. The PV for the case measure y_i is the proportion of control measures Y_Db < y_i.

normal models the test measure among controls with a normal distribution. The PV for the case measure y_i is the standard normal cumulative distribution function of (y_i - mean)/sd, where the mean and the standard deviation are calculated by using the control sample.

tiecorr indicates that a correction for ties between case and control values is included in the empirical PV calculation. The correction is important only in calculating summary indices such as the AUC. The tie-corrected PV for a case with the model predicted value y_i is the proportion of control values Y_Db < y_i plus one half the proportion of control values Y_Db = y_i, where Y_Db denotes controls. By default, the PV calculation includes only the first term, i.e., the proportion of control values Y_Db < y_i. This option applies only to the empirical PV calculation method.

Covariate adjustment

adjcov(*varlist***)** specifies the variables to be included in the adjustment.

adjmodel(*model***)** specifies how the covariate adjustment is to be done. *model* can be one of the following:

stratified PVs are calculated separately for each stratum defined by *varlist* in **adjcov()**. This is the default if **adjmodel()** is not specified and **adjcov()** is. Each case-containing stratum must include at least two controls. Strata that do not include cases are excluded from calculations.

linear fits a linear regression of the predictor distribution on the adjustment covariates among controls. Standardized residuals based on this fitted linear model are used in place of the predicted values for cases and controls.

Sampling variability

nsamp(*#***)** specifies the number of bootstrap samples to be drawn for estimation of standard errors and CIs. The default is **nsamp(1000)**.

nobstrap omits bootstrap sampling and estimation of standard errors and CIs. If **nsamp()** is specified, **nobstrap** will override it.

noccsamp specifies that bootstrap samples be drawn from the combined sample rather than sampling separately from cases and controls; case-control sampling is the default.

nostsamp draws bootstrap samples without respect to covariate strata. By default, samples are drawn from within covariate strata when stratified covariate adjustment is requested via the **adjcov()** and **adjmodel()** options.

cluster(*varlist***)** specifies variables identifying bootstrap resampling clusters. See the **cluster()** option in **[R] bootstrap**.

resfile(*filename***)** creates a Stata file (a **.dta** file) with the bootstrap results for the included statistics. **bstat** can be run on this file to view the bootstrap results again.

replace specifies that if the specified file already exists, then the existing file should be overwritten.

level(*#***)** specifies the confidence level for CIs as a percentage. The default is **level(95)** or as set by **set level**.

New variable

genxb generates new variables **xb1** [and **xb2**], to hold the x-validation generated model predicted values from models specified in **mod1(***varlist***)** [and **mod2(***varlist***)**].

replace requests that existing variables **xb1** [and **xb2**] be overwritten by **genxb**.

Saved results

incroc saves the following in **r()**, where *stat* is one or more of **auc**, **pauc**, **roc**, or **rocinv**, corresponding to the requested summary statistics :

Scalars

r(*stat***1)**	statistic estimate for first model predictor
r(*stat***2)**	statistic estimate for second model predictor
r(*stat***delta)**	estimate difference, *stat***2** - *stat***1**
r(se_*stat***1)**	bootstrap standard-error estimate for first predictor statistic
r(se_*stat***2)**	bootstrap standard-error estimate for second predictor statistic
r(se_*stat***delta)**	bootstrap standard-error estimate for the difference, *stat***2** - *stat***1**

Examples

Use Norton neonatal audiology dataset

```
. use http://labs.fhcrc.org/pepe/book/data/nnhs2, clear
```

Single model

```
. incroc_v102 d if ear == 1, mod1(currage gender)
```

Clustered bootstrap sampling; observations for both ears

```
. incroc_v102 d, mod1(currage gender) cluster(id) auc pauc(.20) roc(.20)
```

Nested models

```
. incroc_v102 d if ear == 1 & site < 4, mod1(currage gender) mod2(currage gender
      y1) adjcov(site)adjmodel(strat) auc pauc(.20) roc(.20)
```

Predictors obtained without cross-validation; save and plot predicted values

```
. incroc_v102 d if ear == 1 & site < 4, mod1(currage gender) mod2(currage gender
      y1) nocv adjcov(site) adjmodel(strat) genxb
. roccurve d xb1 xb2
```

References

Dodd L and Pepe MS. 2003. Partial AUC estimation and regression. *Biometrics* 59:614-623.

Huang Y, Pepe MS. Biomarker evaluation using the controls as a reference population. *Biostatistics* (in press)

Janes H, Pepe MS. 2008. Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: an old concept in a new setting. *American Journal of Epidemiology* 168: 89-97.

Janes H, Pepe MS. Adjusting for covariate effects on classification accuracy using the covariate-adjusted roc curve. *Biometrika* (in press)

Pepe MS, Longton G. 2005. Standardizing markers to evaluate and compare their performances. *Epidemiology* . 16(5):598–603.

Pepe MS. 2003. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press.

Authors

Gary Longton
Fred Hutchinson Cancer Research Center
Seattle, WA
glongton@fhcrc.org

Margaret Pepe
Fred Hutchinson Cancer Research Center and University of Washington
Seattle, WA
mspepe@u.washington.edu

Holly Janes
Fred Hutchinson Cancer Research Center
Seattle, WA
hjanes@fhcrc.org

Also See

Online: **roccurve**, **comproc**, **rocreg** (if installed)