Statistics/Data Analysis

help incrisk (version 1.0.5)

Title

incrisk —— Incremental value of one or more markers or predictors relative to a list of existing predictors. Evaluation is with respect to various risk improvement measures.

Syntax

incrisk *disease_var* [*if*] [*in*], **x(***varlist* **)** **y(***varlist* **)** [*options*]

options	Description
Models	
* **x(***varlist* **)**	base model predictor variables
* **y(***varlist* **)**	additional marker or predictor variables
Optional Statistics	
cook	Cook Reclassification calibration statistic.
GLM alternatives	
glm(*type* **)**	specify GLM model type; **logit** (default), **poisson**, or **logbin**
Risk classification	
rcat(*numlist* **)**	specify up to 4 thresholds for risk category classification.
categorical	include categorical risk classification tables and statistics even if **rcat()** is not specified. Define two categories with single threshold *rho* = disease prevalence.
rho(*r***)**	population disease prevalence; sample prevalence is used by default. Required if case-control sampling is specified.
Cross-validation	
nsplit(*k***)**	specify number of random partitions for k-fold cross-validation; default is **nsplit(10)**
nocv	omit cross-validation
Summary statistics	
noauc	omit calculation of the area under ROC curve (AUC)
lz(*fp***)**	Lorenz measure at specified population proportion *fp*; default is **lz(.15)**
lzinv(*fc***)**	inverse Lorenz measure at specified case proportion *fc*; default is **lzinv(.70)**
Sampling variability	
nsamp(*#***)**	specify number of bootstrap samples; default is **nsamp(1000)**
nobstrap	omit bootstrap sampling
ccsamp	specify case-control rather than cohort bootstrap sampling
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_xb	overwrite existing **xb***#* variables

* **x(***varlist* **)** and **y(***varlist* **)** are required.

Description

incrisk compares predicted risks from base (X) and enhanced (XY) models of disease outcome with respect to various performance improvement measures including Net Reclassification Improvement (NRI), Integrated Discrimination Improvement (IDI), and changes in each of the Mean Risk Difference (MRD), Lorenz measure (L and L^(-1)), Above Average Risk Difference (AARD), standardized Total Gain (TG), and Area under the ROC curve (AUC).

If categorical risk classification is specified, additional measures include the categorical version of the Net Reclassification Improvement (NRI), Reclassification percent (RC), changes in percent of cases and controls classified as high risk (HRc and HRn), Net Benefit (NB), and optionally, the Cook Reclassification Calibration statistic (RCC).

Base model variables are specified with **x(***varlist***)**. The enhanced model additionally includes variables specified with **y(***varlist***)**. *disease_var* is the 0/1 event or disease indicator variable.

All estimators are calculated empirically from model based predicted values. The model based predicted values used for classification and for calculation of the improvement mesasures are obtained by k-fold cross-validation unless specified otherwise with the **nocv** option.

Bootstrap standard errors and confidence intervals (CIs) for the requested statistics are calculated. Percentile CI's are displayed. Normal-based, and bias-corrected CIs are available (see estat bootstrap and **[R] bootstrap**).

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

Options

Models

x(*varlist***)** specifies the variables to be included in the base model. **x()** is required.

y(*varlist***)** specifies additional variables to be included in the enhanced model. **y()** is required.

Optional Statistics

cook includes the Cook Reclassification calibration statistic in the output. Entries for the mean model-based risk from base and enhanced models are additionally included in the Reclassification table. Valid only when either of the **rcat()** or **categorical** options are also specified.

GLM alternatives

glm(*type***)** specifies alternate GLM model types for estimation of disease risk.

logit, the default, uses logistic regression models (GLM binomial family and logit link) and resulting linear predictor for calculation of improvement measures and risk classification.

poisson uses poisson regression models (GLM poisson family and log link) and the resulting linear predictor.

logbin uses a GLM model with binomial family and log link instead of a logit link. Convergence problems are often encountered with this option.

Risk Classification

rcat(*numlist***)** specifies between 1 and 4 thresholds for risk category classification into 2-5 categories. Values must be between 0 and 1. If **categorical** is specified and **rcat()** is not, two categories with single threshold at either the sample disease prevalence, *rho_hat*, or if specified, the population prevalence, **rho()**, are used.

categorical specifies that categorical risk classification tables and statistics are to be included when **rcat()** is not specified. Two categories are defined by threshold *rho* = disease prevalence.

rho(*r***)** specifies the population disease prevalence to be used for default categorical risk classfication, Total Gain statistic calculation, and model-based risk estimation. Value must be between 0 and 1. Required when case-control bootstrap sampling is specified with **ccsamp**. The sample disease prevalence is used by default.

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

noauc omits calculation and report of the ROC area under the curve (AUC).

lz(*fp***)** specifies *fp* for calculation of the Lorenz measure, the proportion of cases with risk above the threshold exceeded by the proportion *fp* of subjects in the population. The argument must be between 0 and 1. The default is **lz(.15)**.

lzinv(*fc***)** specifies *fc* for calculation of the inverse Lorenz measure, the proportion of the population with risk above the threshold exceeded by the proportion *fc* of cases. The argument must be between 0 and 1. The default is **lzinv(.70)**.

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.

ccsamp specifies that bootstrap samples be drawn separately from cases and controls rather than sampling from the combined sample; cohort sampling is the default.

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

Saved results

Examples

load simulated data

```
. use http://labs.fhcrc.org/pepe/dabs/risk_reclass_b
```

incrisk with continuous measures only, and with default settings otherwise.

```
. incrisk d, x(x) y(y)
```

include categorical measures with two categories defined by the sample disease prevalence

```
. incrisk d, x(x) y(y) categorical
```

include categorical measures with 4 categories defined by specified thresholds

```
. incrisk d, x(x) y(y) rcat(.15 .25 .5) cook nsamp(100)
```

list returned results

```
. return list
```

save bootstrap results to a file and display the saved results

```
. incrisk d, x(x) y(y) rcat(.15 .25.5) cook nocv resfile(result1) replace

. bstat using result1
```

References

Pepe MS. slides for a shortcourse, *Current Methods for Evaluating Prediction Performance of Biomarkers and Tests* , can be obtained at the *FHCRC DABS site*

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

Also See

Online: **predcurve**, **incroc** (if installed)