

Predictiveness Curve Estimation

Description

Estimate a Predictiveness curve for a continuous disease marker(s) or diagnostic test variables. Bootstrap confidence intervals for estimates are optionally included. Baseline covariates can be accommodated in risk modelling.

Usage

```
predcurve(dataset=NULL, d, markers, covar=NULL, nested=FALSE, link="logit",
  riskl=NULL, riskh=NULL, ci=FALSE, rho=NULL,
  ylim=NULL, class=FALSE, offset=.005, class_offset=.005,
  plot_type = "predictiveness", nsamp=1000, cluster=NULL, level=95,
  boot_seed=51, densBW=2, pdf=NULL)
```

Arguments

<code>dataset</code>	optional character string specifying the name of the dataset to be used for analysis.
<code>d</code>	character string specifying the name of the 0/1 (disease) outcome vector.
<code>markers</code>	(vector of) character string(s) specifying the names of the test markers/variables.
<code>covar</code>	optional (vector of) character string(s) specifying variables other than the marker to be included in the risk model. default is NULL
<code>nested</code>	logical specifying whether to treat markers as nested models. When <code>nested==TRUE</code> , a covariate-only model will be done in addition to the model with the marker(s). default is FALSE
<code>link</code>	character string specifying the risk model link function as "logit" (default) or "probit". "logit" link is forced when <code>rho</code> is specified (see below).
<code>riskl</code>	lower risk threshold at which to estimate risk percentiles (and TPR/FPR when <code>class==TRUE</code>). default is NULL
<code>riskh</code>	upper risk threshold at which to estimate risk percentiles (and TPR/FPR when <code>class==TRUE</code>). default is NULL
<code>ci</code>	logical specifying whether to include bootstrap confidence intervals for risk, TPR, and FPR estimates. default is FALSE
<code>rho</code>	optional numeric specifying the population disease prevalence which should be used with case-control data. A cohort design is assumed by default. When <code>rho</code> is specified: (1) overall risk distribution is estimated as a weighted average of the case risk distribution and the control risk

	distribution (2) "logit" link is forced
<code>ylim</code>	optional upper limit for the y-axis scale. default is 1
<code>class</code>	logical specifying whether to plot TPR & FPR curves in a second panel. default is FALSE
<code>offset</code>	numeric specifying the vertical CI offset from risk thresholds used to avoid overlaying CIs from multiple markers. default is .005 (0.5 percent of the the full scale).
<code>class_offset</code>	numeric specifying the horizontal CI offset for FPR and TPR CI's. default is .005 (0.5 percent of the the full scale)
<code>plot_type</code>	numeric specifying the horizontal CI offset for FPR and TPR CI's. default is .005 (0.5 percent of the the full scale)
<code>nsamp</code>	numeric specifying the number of bootstrap samples to be drawn for estimating sampling variability of parameter estimates. default is 1000.
<code>cluster</code>	character string specifying variables that identify bootstrap resampling clusters.
<code>level</code>	numeric specifying the confidence level for CIs as a percentage. default is 95
<code>boot_seed</code>	random seed to set at the beginning of bootstrap sampling. default is 51
<code>densBW</code>	multiplier for the bandwidth used for smoothing when density plots are requested. default is 2
<code>pdf</code>	the name of the desired PDF when PDF output of the graphs is desired. default = NULL (output to display)

Details

`predcurve` plots predictiveness curves, i.e. plots of estimated disease risk vs. the risk distribution (empirical cdf). Risk estimates are based on a generalized linear binary model of disease risk as a function of the specified marker variable(s). `d` is the disease indicator used for the model dependent variable. Additional covariables (specified with `covar`) can be included with the marker variable in the risk model.

Risk percentiles and reference lines for specified high and/or low risk thresholds, `riskh` and `riskl`, are optionally included on the plot.

An additional plot of the true and false positive fractions, TPR and FPR, as functions of the risk distribution is optionally included as are estimates and reference lines corresponding to specified risk thresholds.

Risk calculations assume a cohort sampling design by default. A correction for case-control data will be employed if the population prevalence of disease is specified with `rho`, and bootstrap samples for optional CI calculation will be drawn separately from cases and controls.

Alternate graphical outputs (instead of the predictiveness curves) are density plots and CDF plots. These are specified with the `plottype` option.

Value

List containing function call information, estimates and confidence intervals. Returned list items include the following:

<code>coefs</code>	a dataframe of coefficients for the risk model(s) with bootstrap CI.
<code>riskPerc</code>	a dataframe of risk percentage estimates with bootstrap CI. (raw form of the matrix written to the display) NULL if neither <code>riskh</code> or <code>riskl</code> is specified.
<code>class</code>	a dataframe of classification estimates with bootstrap CI. (raw form of the matrix written to the display) NULL if <code>class=FALSE</code> or if neither <code>riskh</code> or <code>riskl</code> is specified.
<code>call</code>	a string showing how the function was called.
<code>"markerName"_ptEst</code>	the point estimates for each marker. (one list for each marker) The list contains: <code>row</code> (a vector showing the ordering of the other output vectors, eg <code>fitted[1]</code> corresponds to the <code>row[1]</code> 'th row in the dataset), <code>fitted</code> (sorted estimated risks from the model) <code>perc</code> (sorted risk percentiles of the risk estimate), <code>fpr</code> (sorted false positive rates using that risk estimate as a cutoff), <code>tpr</code> (sorted true positive rates using that risk estimate as a cutoff), <code>riskHperc</code> , <code>riskLperc</code> , <code>tprH</code> , <code>tprL</code> , <code>fprH</code> , <code>fprL</code> (estimates at the specified <code>riskh</code> and <code>riskl</code>), <code>coefs</code> (coefficients of the risk model)

Author(s)

Daryl Morris, Fred Hutchinson Cancer Research Center and University of Washington, Seattle, WA. darylm@u.washington.edu

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

References

Pepe M.S., Feng Z., Huang Y., Longton G., Prentice R., Thompson I.M., Zheng Y. 2008. Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology* **167**(3),362–368.

Pepe M.S., Gu, W., Morris, D.E. 2010. The Potential of Genes and other Markers to Inform about Risk. *Cancer Epidemiology, Biomarkers and Prevention (in press)*

See Also

[comproc](#), [roccurve](#).

Examples

```
janssens <- read.table("http://labs.fhcrc.org/pepe/data/janssens_c.csv",
                      sep=" ", header=T)

predcurve(dataset="janssens", d="disease", markers="logscr")

predcurve(dataset="janssens", d="disease", markers=c("logscr", "bmi"))

predcurve(dataset="janssens", d="disease", markers=c("logscr", "bmi"), riskh=.4)

predcurve(dataset="janssens", d="disease", markers=c("logscr", "bmi"),
          riskh=.4, riskl = .1, class=TRUE)

predcurve(dataset="janssens", d="disease", markers="logscr",
          covar=c("age", "hypertension", "bmi", "bruit", "vascular", "gender"),
          link="logit", riskl=.1, riskh=.4, ci=TRUE, class=TRUE)

#####
# example of nested marker models
m1 <- glm(disease ~ bmi +age +hypertension +bruit +vascular +gender,
          data=janssens, family=binomial)
janssens$m1Fit <- m1$linear.predictors
m2 <- glm(disease ~ bmi +age +hypertension +bruit +vascular +gender +logscr,
          data=janssens, family=binomial)
janssens$m2Fit <- m2$linear.predictors

predcurve(dataset="janssens", d="disease", markers=c("m1Fit", "m2Fit"),
          riskh=.4, riskl = .1, class=TRUE, ci=TRUE)

#####
# examples of alternate graphical outputs (cdf and density)
predcurve(dataset="janssens", d="disease", markers=c("m1Fit", "m2Fit"),
          riskh=.4, riskl = .1, class=TRUE, plot_type="cdf", ci=TRUE)

predcurve(dataset="janssens", d="disease", markers=c("m1Fit", "m2Fit"),
          riskh=.4, riskl = .1, class=TRUE, plot_type="density")

predcurve(dataset="janssens", d="disease", markers=c("m1Fit", "m2Fit"),
          riskh=.4, riskl = .1, class=TRUE, plot_type="density", densBW=1)

#####
# example using nested parameter (same as above except (1) no need for
# fitting models outside of predcurve and (2) CI may differ
# since covariates are being included in the bootstrap)
```

```
predcurve(dataset="janssens",d="disease", markers="logscr",  
  covar=c("bmi","age","hypertension","bruit","vascular","gender"),  
  link="logit", riskl=.1, riskh=.4, class=TRUE, ci=TRUE, nested=TRUE)
```

[Package *pcvsuite* version 1.0 [Index](#)]