Statistics/Data Analysis

help for **predcurve** (version 1.0.9 Beta)

## Title

**predcurve** —— Predictiveness curve

## Syntax

**predcurve** disease_var test_var1 [test_var2 ] [*if*] [*in*] [**,** *options* ]

| *options* | Description |
|---|---|
| **Risk estimation** | |
| **link(***function* **)** | **logit** (default) or **probit** link function. |
| **riskl(***p***)** | lower risk threshold. |
| **riskh(***p***)** | upper risk threshold. |
| **covar(***varlist* **)** | covariables to include in the risk model. |
| **ci** | include bootstrap confidence intervals for risk, TPR, and FPR estimates. |
| **Case-control adjustment** | |
| **rho(***#***)** | the population disease prevalence should be specified when case-control data are used. A cohort design is assumed by default. |
| **Graph** | |
| **ylim(***#***)** | upper limit for the y-axis scale. |
| **class** | plot TPR & FPR curves in a second panel. |
| **offset(***#***)** | specify CI offset from risk thresholds. default is .004 |
| **class_offset(***#***)** | specify x axis offset for FPR and TPR CI's. default is .25. |
| **Bootstrap options** | |
| **nsamp(***#***)** | number of bootstrap samples; default is 1000. |
| **cluster(***varlist* **)** | variables identifying bootstrap resampling clusters |
| **level(***#***)** | set confidence level; The default is **level(95)** or as set by set level |
| **New variable options (pending)** | |
| **genprvvars** | generate new variables to hold model p_hat. |

## Description

**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 continuous marker variable(s), *test_var1* [and *test_var2* ]. *disease_var* is the disease indicator used for the model dependent variable. Additional covariables specified with **covar(***varlist* **)** 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(***p***)** and **riskl(***p***)**, 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.

## Options

### Risk estimation

**link(***function* **)** Specifies the binary GLM link function. *function* options include **logit** (the default) and **probit**.

**riskl(***p***)** lower risk threshold. Calculate the percentage of subjects with risk ≤ p; also calculate the percentages of cases (TPR) and controls (FPR) with risk ≥ p if **class** is specified.

**riskh(***p***)** upper risk threshold. Calculate the percentage of subjects with risk > p; also calculate the percentages of cases (TPR) and controls (FPR) with risk ≥ p if **class** is specified.

**covar(***varlist* **)** covariables to include in the risk model.

### Case-control adjustment

**rho(***#***)** The population disease prevalence should be specified when case-control data are used. rho will be used for risk calculation adjustment, and bootstrap sampling for optional CI's will done separately from case and control samples. Only the **logit** GLM link may be used with case-control data; including the **link(probit)** with the **rho(***#***)** option will return an error. Risk calculations assume a cohort sampling design by default.

### Graph options

**class** specifies that TPR and FPR curves should be plotted.

**ylim(***#***)** set the upper limit for y-axis scale. Must be between 0 and 1. If not specified, this is 10% larger than the largest observed risk estimate.

**ci** indicates that bootstrap percentile-based confidence intervals for risk percentiles at specified thresholds **riskh(***p***)** or **riskl(***p***)** be calculated. CI's for TPR and FPR will be calculated for the specified thresholds if the **class** option is included.

**offset(***#***)** specifies the offset from specified risk thresholds for placement of 2nd CI if 2 markers are specified, for avoidance of superimposed interval bars. The argument must be between 0 and .02; default is **offset(.004)**. # is a proportion of the yaxis range if **ylim(***#***)** is included.

**class_offset(***#***)** specifies the x-axis (risk percentile) offset for placement of TPR and FPR CI's in order to avoid superimposed interval bars. The argument must be between 0 and 2; default is **offset(.25)**. relevant only if both the **class** and **ci** options are specified.

### Bootstrap options

These options are relevant only if the **ci** option is specified.

**nsamp(***#***)** specifies the number of bootstrap samples for risk model estimation be performed to obtain confidence intervals. The default is 1000 replications.

**cluster(***varlist* **)** specifies variables identifying bootstrap resampling clusters. See the cluster option of the bootstrap command.

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

New variable options

**genprvvars (***pending***)** generate new variables, ph#, for each marker in the *test_varlist* to hold predicted probabilities for each subject based on the binary risk model fit. New variable names are numbered (#) according to marker variable order in the *test_varlist*

**replace** requests that existing variables *ph#* be overwritten by **genpcv** or **genrocvars**.

## Saved results

If risk threshold options, **riskl(***p***)** or **riskh(***p***)** are specified, **predcurve** saves the following in **r()**:

Scalars

| | |
|---|---|
| **r(*rpc_l_#*)** | lower threshold risk percentile estimate, marker number #. |
| **r(*rpc_u_#*)** | upper threshold risk percentile estimate, marker number #. |

If the class option, **class**, is additionally specifed, TPR and FPR estimates for the specified thresholds are returned in:

| | |
|---|---|
| **r(*tpr_l_#*)** | lower threshold TPR estimate, marker number #. |
| **r(*tpr_u_#*)** | upper threshold TPR estimate, marker number #. |
| **r(*fpr_l_#*)** | lower threshold FPR estimate, marker number #. |
| **r(*fpr_u_#*)** | upper threshold FPR estimate, marker number #. |

If the **ci** option is included, bootstrap postestimation results left behind by bstat are available.

Returned matrices include:

**e(ci_percentile)** 2 x k matrix of bootstrap percentile CI's, where k = (# markers) * (# threholds) * (# estimators), and rows correspond to upper and lower bounds. Matrix column names indicate estimator, marker #, and threhold.

## Examples

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

. predcurve d logscr

. predcurve d logscr bmi

. predcurve d logscr bmi, riskh(.40)

. predcurve d logscr bmi, riskh(.40) riskl(.10) class

. predcurve d logscr, cov(age hyper bmi bruit vas gender) riskh(.40) riskl(.10) class
    ci

. logistic d age hyper bmi bruit vasc gender, coef
. predict mod1, xb
. la var mod1 "risk(X)"

. logistic d age hyper bmi bruit vasc gender logscr, coef
. predict mod2, xb
. la var mod2 "risk(X,Y)"

. predcurve d mod1 mod2, riskh(.40) riskl(.10) class
```

```
. predcurve d mod1 mod2, riskh(.40) riskl(.10) class ylim(.5)
```

## References

## 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