Package ‘PIMixture’

January 9, 2023

Title Prevalence Incidence Mixture Models

Version 0.4.4

Date 2023-01-06

Authors c(person(c(``Li",``C."), ``Cheung", email = ``li.cheung@nih.gov", role = c(``aut", ``cre")), person(c(``Noorie"), ``Hyun", email = ``nhyun@mcw.edu", role = c(``aut", ``cre")), person(c(``Xiaoqin"), ``Xiong", role = c(``aut")), person(c(``Qing"), ``Pan", role = ``aut", person(c(``Hormuzd",``A."), ``Katki", role = ``aut")

Author Li C. Cheung [aut, cre],
Noorie Hyun [aut, cre],
Xiaoqin Xiong [aut],
Qing Pan [aut],
Hormuzd A. Katki [aut]

Maintainer Bill Wheeler `<wheelerb@imsweb.com>`

Description This package fits Prevalence Incidence Mixture models to data for which the time to event is interval censored or
for which the event is prevalent at the time zero but is only partially observed. Such data often arises in medical screening
for asymptomatic disease or disease precursors, such as precancerous lesions. In such data, 1) onset of incident disease occurs
between screening visits (interval-censoring), 2) the disease may have already occurred before the initial screen (prevalent at time zero)
but may be initially missed and found some time after the initial screen. These models estimates absolute and relative risks.
Semi-parametric, weakly-parametric (integrated B-splines), and fully parametric members of the Prevalence-Incidence Mixture model
family are supported. A non-parametric estimator is provided and is useful for checking parametric assumptions of fully parametric
Prevalence Incidence Mixture models. Only weakly-parametric and semi-parametric models (no variance calculation) currently support
stratified random samples in two frames: a superpopulation and a finite population (where cases and controls at time zero can be a stratum factor).
A later version will add this functionality for the logistic-Weibull and logistic-exponential models.
Semi-parametric, weakly-parametric models, logistic-Weibull, and logistic-exponential uses a logistic regression model as the prevalence model
and a proportion hazard survival model as the incidence model. The semi-parametric model makes no assumptions regarding the baseline hazard function.
However, it can be computationally expensive. The weakly-parametric model approximates the bazeline hazard function using

integrated B-splines and is faster. When parametric assumptions can be made, the fully parametric models are fastest.
The following parametric assumptions are supported: logistic-Weibull, logistic-exponential, logistic-lognormal,
logistic-loglogistic, logistic-gengamma, and logistic-gamma. Variance estimates are available only for the weakly-parametric,
logistic-Weibull, logistic-exponential, and non-parametric. For the non-parametric, this is achieved through boot-strapping by setting
the conf.int parameter to ``TRUE" and can be very computationally expensive.
For identifiability of the mixture model, the data must contained observed prevalent disease and interval-censored incident disease.

Depends R (>= 3.4.4)

Imports survival, optimx, fdrtool, interval, Icens, plyr, flexsurv

License GPL-2

LazyData true

NeedsCompilation no

R topics documented:

`PIdata1` *PIdata includes two datasets, PIdata1 and PIdata2, which are simulation data.*

Description

PIdata1 and PIdata2 are from simple random samples and stratified simple random samples, respectively. The two variables of "samp.weight" and "strata" should be included on the dataset of stratified random samples. The variable of "strata.frac" is needed for the design-based variance calculation with the option "sample.design=2".

Usage

```
data(PIdata)
```

Format

Objects of class data.frame, and missing values are coded as -999.

Examples

```
data(PIdata)
head(PIdata1)
head(PIdata2)
```

`PIMixture` *Prevalence-Incidence Mixture Models*

Description

This package fits Prevalence Incidence Mixture models to data for which the time to event is interval censored or for which the event is prevalent at the time zero but is only partially observed. Such data often arises in medical screening for asymptomatic disease or disease precursors, such as precancerous lesions. In such data, 1) onset of incident disease occurs between screening visits (interval-censoring), 2) the disease may have already occurred before the initial screen (prevalent at time zero) but may be initially missed and found some time after the initial screen. These models estimates absolute and relative risks. Semi-parametric (iterative convex minorant algorithm by Robertson,Wright and Dykstra, 1988), weakly-parametric (integrated B-splines), and fully parametric members of the Prevalence-Incidence Mixture model family are supported. A non-parametric estimator (Turnbull 1976) is provided and is useful for checking parametric assumptions of fully parametric Prevalence Incidence Mixture models. Only weakly-parametric and semi-parametric models (no variance calculation) currently support stratified random samples in the two viewpoints, a superpopulation and a finite population. The superpopulation is to view the finite population of interest as an independent sample of size N from an infinite superpopulation. A later version will add this functionality for the logistic-Weibull and logistic-exponential models. Semi-parametric, weakly-parametric models, logistic-Weibull, and logistic-exponential uses a logistic regression model as the prevalence model and a proportion hazard survival model as the incidence model. The semi-parametric model makes no assumptions regarding the baseline hazard function. However, it can be computationally expensive when the unique number of visit times are over hundreds. The weakly-parametric model approximates the bazeline hazard function using integrated B-splines and is faster. When parametric assumptions can be made, the fully parametric models are fastest. The following parametric assumptions are supported: logistic-Weibull, logistic-exponential, logistic-lognormal, logistic-loglogistic, logistic-gengamma, and logistic-gamma. Variance estimates are available only for the weakly-parametric, logistic-Weibull, logistic-exponential, and non-parametric. For the non-parametric, this is achieved through boot-strapping by setting the conf.int parameter to "TRUE" and can be very computationally expensive. For identifiability of the mixture model, the data must contained observed prevalent disease and interval-censored incident disease.

Usage

```
PIMixture(p.model, i.model, data, model = "semi-parametric",
  reg.initials = NULL, conf.int = FALSE,
  convergence.criteria = 0.001, iteration.limit = 250,
  time.interval = 0.01, design.out = TRUE, sample.design = NULL,
  N = NULL, n.knots = 5, order = 4, max.time, ...)
```

Arguments

`p.model`	The prevalence model to be fitted, specified using an expression of the form *c+l+r~model*. Elements in the expression are as followed: • c - Numeric variable indicating whether the event was prevalent at time zero, taking values of 1=="Yes", 0="No", -999="Latent"; • l - Numeric starting time of the interval in which event occurred, with -999 denoting known prevalent events; • r - Ending time of the interval in which event occurred, with -999 and Inf denoting known prevalent events and right-censoring, respectively;

	• model - Linear predictor consisting of a series of terms separated by + operators.
`i.model`	The incidence model to be fitted, specified using an expression of the form *c+l+r~model* (see p.model). Defaults to p.model.
`data`	Data used to fit the model containing columns for each term in p.model and i.model expressions. For stratified random sampling designs, columns denoted samp.weight and strata are expected indicating the sampling weights and sampling strata. For sample.design="superpopulation" option, an additional column denoted strata.frac is expected indicating the fraction of the population that consists of each strata. For example, if in the target population there are three strata that occurs with proportions 0.2, 0.4, and 0.6, then strata.frac will take values of 0.2, 0.4 or 0.6.
`model`	Character string indicating the specific member of the Prevalence-Incidence Mixture Model family to be fitted. Options are: • "semi-parametric" Fits logistic regression and proportional hazards model as the prevalence and incidence models, respectively. The baseline hazard funcion is non-parametrically estimated using the iterative convex minorant algorithm. • "weakly-parametric" Fits logistic regression and proportional hazards model as the prevalence and incidence models, respectively. The baseline hazard function is approximated using integrated B-splines. • "logistic-Weibull" Fits logistic regression and proportional hazards model as the prevalence and incidence models, respectively. The baseline hazard function is approximated using a Weibull distribution. • "logistic-exponential" Fits logistic regression and proportional hazards model as the prevalence and incidence models, respectively. The baseline hazard function is approximated using a exponential distribution. • "logistic-lognormal" Fits logistic regression and lognormal survival as the prevalence and incidence models, respectively. • "logistic-loglogistic" Fits logistic regression and loglogistic survival as the prevalence and incidence models, respectively. • "logistic-gengamma" Fits logistic regression and generalized-gamma survival as the prevalence and incidence models, respectively. • "logistic-gamma" Fits logistic regression and gamma survival as the prevalence and incidence models, respectively. • "non-parametric" Provides the non-parametric cumulative risk estimator. This is akin to the non-parametric estimates provided using the Turnbull methods. Covariates are not supported. Confidence intervals are obtained through bootstrapping but is computationally expensive. Variance estimates are not available for "semi-parametric", logistic-lognormal", "logistic-loglogistic", "logistic-gengamma", or "logistic-gamma", but they can be obtained using bootstrap methods. Defaults to "semi-parametric".
`reg.initials`	Initial parameter estimates. Defaults to NULL.
`conf.int`	For non-parametric model option, FALSE="Do not obtain bootstrap confidence intervals", TRUE="Obtain bootstrap confidence intervals. Defaults to FALSE.
`convergence.criteria`	Convergence of models occurs when reduction in the objective is within this convergence tolerance. Defaults to 0.001.

`iteration.limit`	Maximum number of iterations allowed to achieve convergence. Defaults to 250.
`time.interval`	Define time intervals to output baseline hazards for. Defaults to .01.
`design.out`	Option to include the design matrix of data used for model fitting in the output. Defaults to TRUE.
`sample.design`	Sampling design of the NULL="simple random sampling", 1="finite population", 2="superpopulation". Defaults to NULL. For "superpopulation", N is required for variance calculation (only provided when model="weakly-parametric" is used)
`N`	Population size, required for superpopulation. Defaults to NULL.
`n.knots`	Number of knots for splines for "weakly-parametric" model. Defaults to 5.
`order`	Degree of splines for "weakly-parametric" model. Defaults to 4 (cubic splines).
`max.time`	Define maximum time to output baseline hazards for. Defaults to the largest finite start/end time given in the data.
`...`	Additional arguments.

Value

The output is a list of class PIMix which contains the following elements.

- data.summary A data frame containing the following: Num. of observation - total number of observations in data set; Included subjects - number of observations used in fitting model; Known prevalent cases - the number of events known to be prevalent at time zero; Interval censoring - the number of event times occuring in the interval (L>0,R<Inf]; Left censoring - the number of event times known to occur by R<Inf, but can also have been prevalent at time zero; Right censoring - the number of observations right-censored with event time occurring in the interval (L>0,Inf); Missing prevalent status - the number of observations where it is unknown whether the event was prevalent at time zero; Non-informative intervals - the number of observations with intervals (0,Inf) or [0,Inf) (denoting missing prevalent status).
- regression.coef A data frame summarizing parameter values, standard errors, and 95 percent confidence intervals.
- OR A data frame summary odds ratios, , standard errors, and 95 percent confidence intervals.
- HR A data frame summary hazard ratios, , standard errors, and 95 percent confidence intervals.
- knots If model="weakly-parametric" is specified, this is a numeric vector of starting time points for each exponential spline.
- exp.spline.coeff If model="weakly-parametric" is specified, this is a numeric vector of coefficients for each exponential spline.
- cum.hazard If model="semi-parametric" or model="weakly-parametric" is specified, this is a data frame containing the baseline cumulative hazard.
- covariance A matrix containing the covariance matrix for the parameters (not produced for model="semi-parametric").
- hessian A matrix containing the hessian matrix for the parameters (not produced for model="semi-parametric").
- model Character string indicating the specific member of the Prevalence-Incidence Mixture Model family fitted.
- p.model The prevalence model.

- prev.design The design matrix for the prevalence model.
- i.model The incidence model.
- incid.design The design matrix for the incidence model.
- loglikelihood For random samples, this is the log-likelihood of the fitted model. For stratified random samples, the weighted-likelihood approach is used and a log-pseudolikelihood (weighted log-likelihood) is reported.
- convergence Convergence statistics.

Warning

The model="semi-parametric" option is very computationally expensive when the unique visit times are over hundreds.

Author(s)

Li C. Cheung, <`li.cheung@nih.gov`>, Noorie Hyun <`nhyun@mcw.edu`> Xiaoqin Xiong, Qing Pan, Hormuzd A. Katki

References

- Cheung LC, Qing P, Hyun N, Schiffman M, Fetterman B, Castle P, Lorey T, Katki H. Mixture models for undiagnosed prevalent disease and interval-censored incident disease: Applications to a cohort assembled from electronic health records. Statistics in Medicine 2017; 36(22):3583-95.
- Hyun N, Cheung LC, Pan Q, Katki H. Flexible risk prediction models for left or interval-censored data from electronic health records. Annals of Applied Statistics 11(2), 1063-1084.
- Turnbull BW (1976). The empirical distribution with arbitrary grouped censored and truncated data. Journal of the Royal Statistical Society - Series B (Statistical Methodology) 38, 290-295.
- Robertson T, Wright FT, and Dykstra RL (1988). Order Restricted Statistical Inference. Wiley.

Examples

```
#PIMixture includes "PIdata" RData file, and PIdata includes the two datasets, PIdata1 and PIdata2
data(PIdata)
model<-"C_CIN3PLUS+L_CIN3PLUS+R_CIN3PLUS~RES_HPV16"
fit1<-PIMixture(p.model=model,data=PIdata1, model="logistic-Weibull")
fit2<-PIMixture(p.model=model,data=PIdata1, model="weakly-parametric",n.knots=5,order=4)
fit3<-PIMixture(p.model=model,data=PIdata1, model="semi-parametric")

model2<-"C_CIN3PLUS+L_CIN3PLUS+R_CIN3PLUS~1"
fit4<-PIMixture(p.model=model2,data=PIdata1, model="non-parametric", conf.int=TRUE)


#For stratified random samples
model3<-"C+L+R~X1+X2"
#sample.design=1 indicates the target population is a finite population, and the variance is design-based.
fit5<-PIMixture(p.model=model3,data=PIdata2, model="weakly-parametric",n.knots=7,order=4,sample.design=1)

#sample.design=2 indicates the target population is a superpopulation, and the variance consists of
#design-based and model-based variances. Generally, the Variance in the superpopulation frame is slightly larg
fit6<-PIMixture(p.model=model3,data=PIdata2, model="weakly-parametric",n.knots=7,order=4,sample.design=2,N

fit7<-PIMixture(p.model=model3,data=PIdata2, model="semi-parametric",sample.design=1)
```

`PIMixture.predict` *Prevalence-Incidence Mixture Models Predictions*

Description

This function produces cumulative risk predictions given an object of class PIMix a vector of times, and a data set of covariates from which to make predictions. In place of a PIMix object, all elements of the fitted model can be specified instead.

Usage

```
PIMixture.predict(x = NULL, data, time.points,
  model = "semi-parametric", prev.coef = NULL, incid.coef = NULL,
  Lambda.data = NULL, knots = NULL, order = NULL,
  spline.para.est = NULL, cov.mat = NULL, ...)
```

Arguments

`x`	object of class PIMix.
`data`	data set of covariates from which to make predictions.
`time.points`	numeric vector of times points to produce cumulative risk estimates for.
`model`	Character string indicating the specific member of the Prevalence-Incidence Mixture Model family to be fitted. Options are: • "semi-parametric" Fits logistic regression and proportional hazards model as the prevalence and incidence models, respectively. The baseline hazard function is estimated using the iterative convex minorant algorithm. Variance estimates are obtained using bootstrap methods. Can be computationally expensive; • "weakly-parametric" Fits logistic regression and proportional hazards model as the prevalence and incidence models, respectively. The baseline hazard function is approximated using integrated B-splines. • "logistic-Weibull" Fits logistic regression and proportional hazards model as the prevalence and incidence models, respectively. The baseline hazard function is approximated using a Weibull distribution. • "logistic-exponential" Fits logistic regression and proportional hazards model as the prevalence and incidence models, respectively. The baseline hazard function is approximated using a exponential distribution. • "logistic-lognormal" Fits logistic regression and lognormal survival as the prevalence and incidence models, respectively. • "logistic-loglogistic" Fits logistic regression and loglogistic survival as the prevalence and incidence models, respectively. • "logistic-gengamma" Fits logistic regression and generalized-gamma survival as the prevalence and incidence models, respectively. • "logistic-gamma" Fits logistic regression and gamma survival as the prevalence and incidence models, respectively.
`prev.coef`	A vector containing coefficients for the prevalence model.

`incid.coef`	A vector containing coefficients for the incidence model.
`Lambda.data`	For semi-parametric or weakly-parametric models, this is a data frame containing the times and baseline cumulative hazard.
`knots`	For weakly-parametric models, this is a numeric vector of starting time points for each exponential spline.
`order`	Degree of splines for "weakly-parametric" model. Defaults to 4 (cubic splines).
`spline.para.est`	For weakly-parametric models, this is a numeric vector of coefficients for each exponential spline.
`cov.mat`	A matrix containing the covariance matrix for the parameters (not required for semi-parametric models).
`...`	Additional arguments.

Value

A data frame containing the following columns

- prev.predictor The design matrix for the prevalence model.
- incid.predictor The design matrix for the incidence model.
- time Time.
- CR Cumulative risk at the time specified.
- CR.se Standard error for the cumulative risk.
- LL95 Lower 95 percent confidence limit for the cumulative risk.
- LL95 Upper 95 percent confidence limit for the cumulative risk.

Author(s)

Li C. Cheung, <`li.cheung@nih.gov`>, Noorie Hyun <`nhyun@mcw.edu`> Xiaoqin Xiong, Qing Pan, Hormuzd A. Katki

References

- Cheung LC, Qing P, Hyun N, Schiffman M, Fetterman B, Castle P, Lorey T, Katki H. Mixture models for undiagnosed prevalent disease and interval-censored incident disease: Applications to a cohort assembled from electronic health records. Statistics in Medicine 2017; 36(22):3583-95.
- Hyun N, Cheung LC, Pan Q, Katki H. Flexible risk prediction models for left or interval-censored data from electronic health records. Annals of Applied Statistics 11(2), 1063-1084.

Examples

```
#PIMixture includes "PIdata" RData file, and PIdata includes the two datasets, PIdata1 and PIdata2
data(PIdata)
model<-"C_CIN3PLUS+L_CIN3PLUS+R_CIN3PLUS~RES_HPV16"
fit1<-PIMixture(p.model=model,data=PIdata1, model="logistic-Weibull")
fit2<-PIMixture(p.model=model,data=PIdata1, model="weakly-parametric",n.knots=5,order=4)
fit3<-PIMixture(p.model=model,data=PIdata1, model="semi-parametric")

model2<-"C_CIN3PLUS+L_CIN3PLUS+R_CIN3PLUS~1"
fit4<-PIMixture(p.model=model2,data=PIdata1, model="non-parametric", conf.int=TRUE)
fit5<-PIMixture(p.model=model2,data=PIdata1, model="semi-parametric", conf.int=TRUE)
```

```
test.data<- data.frame(rbind(1,0))
names(test.data)<- ”RES_HPV16”
time.points=c(0,12,36,60)
predict1<-PIMixture.predict(x=fit1, data=test.data, time.points=time.points)
predict2<-PIMixture.predict(x=fit2, data=test.data, time.points=time.points)
predict3<-PIMixture.predict(x=fit3, data=test.data, time.points=time.points)
predict4<-PIMixture.predict(x=fit4, data=test.data, time.points=time.points)
predict5<-PIMixture.predict(x=fit5, data=test.data, time.points=time.points)
predict1
predict2
predict3

predict4
predict5

#For stratified random samples
model3<-”C+L+R~X1+X2”
output1<-PIMixture(p.model=model3,data=PIdata2, model=”semi-parametric”,sample.design=1)
output2<-PIMixture(p.model=model3,data=PIdata2, model=”weakly-parametric”, n.knots=7,order=4,sample.design=
test.data<- data.frame(X1=1,X2=0.5)
time.points<-seq(0,10,by=2)
predict6<-PIMixture.predict(x=output1, data=test.data, time.points=time.points)
predict7<-PIMixture.predict(x=output2, data=test.data, time.points=time.points)
predict6
predict7
```

Index