Package 'icenReg'

October 2, 2015

Type Package

Title Regression Models for Interval Censored Data

Version 1.2.8

Date 2015-09-25

Author Clifford Anderson-Bergman; the Eigen team for Eigen library included; uses Maarloes Maathius's HeightMap algorithm (MLEcens::reduc)

Depends survival, MLEcens

Imports foreach, methods

Maintainer Clifford Anderson-Bergman <pistacliffcho@gmail.com>

Description Regression models for interval censored data. Currently supports Cox-PH and proportional odds models. Allows for both semi and fully parametric models. Includes functions for easy visual diagnostics of model fits. Includes functions for fitting both the univariate and bivariate NPMLE.

License LGPL (>= 2.0, < 3)

R topics documented:

diag_baseline *Compare parametric baseline distributions with semi-parametric baseline*

Description

Creates plots to diagnosis fit of different choices of parametric baseline model. Plots the semi paramtric model (which only uses up to max_n_use subjects from the dataset for speed) against different choices of parametric models.

Usage

```
diag_baseline(object, data, model = "ph", weights = NULL,
              dists = c("exponential", "weibull", "gamma",
                        "lnorm", "loglogistic"),
              max_n_use = 10000, cols = NULL,
              lgdLocation = "bottomleft", useMidCovars = T)
```

Arguments

object	Either a formula or a model fit with ic_sp or ic_par
data	data. Unnecessary if object is a fit
model	type of model. Choices are 'ph' or 'po'
dists	parametric baseline fits
max_n_use	maximum number of observations used to fit semi-parametric model
cols	colors of baseline distributions
weights	case weights
lgdLocation	where legend will be placed. See ?legend for more details
useMidCovars	Should the distribution plotted be for covariates = mean values instead of 0

Details

If useMidCovars = T, then the survival curves plotted are for fits with the mean covariate value, rather than 0. This is because often the baseline distribution (i.e. with all covariates = 0) will be far away from the majority of the data.

For larger datasets (n > 10,000), the semi-parametric model can be become slow. Because of this, diag_covar allows the user to select max_n_use samples from their data to assess the fit of their models.

Author(s)

Clifford Anderson-Bergman

Examples

```
data(essIncData_small)
useData <- essIncData_small

#Note to user: suggest replacing useData with essIncData
#instead of essIncData_small. Using small dataset to quickly
#pass CRAN tests

fit_po <- ic_sp(Surv(inc_l, inc_u, type = 'interval2') ~ eduLevel * cntry,
                data = useData, bs_samples = 0, model = 'po')

diag_baseline(fit_po)

#Weibull model appears best fit
```

`diag_covar` *Evaluate covariate effect for regression model*

Description

Creates plots to diagnosis fit of covariate effect in a regression model. For a given variable, stratifies the data across different levels of the variable and adjusts for all the other covariates included in `fit` and then plots a given function to help diagnosis where covariate effect follows model assumption (i.e. either proportional hazards or proportional odds). See `details` for descriptions of the plots.

If `varName` is not provided, will attempt to figure out how to divide up each covariate and plot all of them, although this may fail.

Usage

```
diag_covar(object, varName,
           data, model, weights = NULL,
           yType = 'meanRemovedTransform',
           factorSplit = TRUE,
           numericCuts, col,
           xlab, ylab, main,
           max_n_use = 10000,
           lgdLocation = NULL)
```

Arguments

`object`	Either a formula or a model fit with `ic_sp` or `ic_par`
`varName`	covariate to split data on. If left blank, will split on each covariate
`data`	data. Unnecessary if `object` is a fit
`model`	type of model. Choices are `'ph'` or `'po'`
`weights`	case weights
`yType`	type of plot created. See details
`factorSplit`	Should covariate be split as a factor (i.e. by levels)
`numericCuts`	If `fractorSplit == FALSE`, cut points of covariate to stratify data on
`col`	colors of each subgroup plot. If left blank, will auto pick colors

xlab	label of x axis
ylab	label of y axis
main	title of plot
max_n_use	maximum number of observations used to fit semi-parametric model
lgdLocation	where legend should be placed. See details

Details

For the Cox-PH and proportional odds models, there exists a transformation of survival curves such that the difference should be constant for subjects with different covariates. In the case of the Cox-PH, this is the log(-log(S(t|X))) transformation, for the proporitonal odds, this is the log(S(t|X) / (1 - S(t|X))) transformation.

The function diag_covar allows the user to easily use these transformations to diagnosis whether such a model is appropriate. In particular, it takes a single covariate and stratifies the data on different levels of that covariate. Then, it fits the semi-parametric regression model (adjusting for all other covariates in the data set) on each of these stratas and extracts the baseline survival function. If the stratified covariate does follow the regression assumption, the difference between these transformed baseline survival functions should be approximately constant.

To help diagnosis, the default function plotted is the transformed survival functions, with the overall means subtracted off. If the assumption holds true, then the difference between the plotted lines should be approximately constant. Other choices of `yType`, the function to plot, are `"transform"`, which is the transformed functions without the means subtracted and `"survival"`, which is the baseline survival distribution is plotted for each strata.

Currently does not support stratifying covariates that are involved in an interaction term.

For variables that are factors, it will create a strata for each level of the covariate, up to 20 levels. If `factorSplit == FALSE`, will divide up numeric covariates according to the cuts provided to numericCuts.

For larger datasets (n > 10,000), the semi-parametric model can become slow. Because of this, `diag_covar` allows the user to select `max_n_use` samples from their data to assess the fit of their models.

`lgdLocation` is an argument placed to `legend` dictating where the legend will be placed. If `lgdLocation = NULL`, will use standard placement given `yType`. See `?legend` for more details.

Author(s)

Clifford Anderson-Bergman

Examples

```
data(essIncData_small)
useData <- essIncData_small

#Note to user: suggest replacing useData with essIncData
#instead of essIncData_small. Using small dataset to quickly
#pass CRAN tests
par(mfrow = c(1,2))

diag_covar(Surv(inc_l, inc_u, type = 'interval2') ~ eduLevel * cntry,
        data = useData, model = 'po')

diag_covar(Surv(inc_l, inc_u, type = 'interval2') ~ eduLevel * cntry,
        data = useData, model = 'ph')
```

essIncData *Interval Censored Income Data from European Social Survey*

Description

Dataset containing a sample from the European Social Survey.

In the European Social Survey, income is reported up a to discrete intervals. Within each country, these intervals are disjoint (i.e. [0, 1k), [1k,2k)...). However, across countries, the intervals are not disjoint and so interval censored methods should be used to compare subjects across different countries.

Usage

```
data(essIncData)
```

Format

A data frame with 6712 rows and 4 variables

- `cntry` Country
- `eduLevel` Categorical variable for number of years of reported education
- `inc_l` Lower limit of reported income level (in Euros)
- `inc_u` Upper limit of reported income level (in Euros)

Source

ESS Round 5: European Social Survey Round 5 Data (2010). Data file edition 3.2. Norwegian Social Science Data Services, Norway\- Data Archive and distributor of ESS data.

Examples

```
# Not run

# data(essIncData)
#
# lnormFit <- ic_par(Surv(inc_l, inc_u, type = 'interval2') ~ eduLevel * cntry,
#                    data = essIncData,
#                    model = 'po',
#                    dist = 'loglogistic')
#
# summary(lnormFit)
```

essIncData_small *Interval Censored Income Data from European Social Survey*

Description

Dataset containing a sample from the European Social Survey. This is a small subsample from the dataset essIncData, used only to run the example very quickly for CRAN. Using the full dataset, running the examples should still be fairly fast; at most a few seconds.

In the European Social Survey, income is reported up a to discrete intervals. Within each country, these intervals are disjoint (i.e. [0, 1k), [1k,2k)...). However, across countries, the intervals are not disjoint and so interval censored methods should be used to compare subjects across different countries.

Usage

```
data(essIncData_small)
```

Format

A data frame with 500 rows and 4 variables

- cntry Country
- eduLevel Categorical variable for number of years of reported education
- inc_l Lower limit of reported income level (in Euros)
- inc_u Upper limit of reported income level (in Euros)

Source

ESS Round 5: European Social Survey Round 5 Data (2010). Data file edition 3.2. Norwegian Social Science Data Services, Norway\- Data Archive and distributor of ESS data.

Examples

```
data(essIncData_small)

lnormFit <- ic_par(Surv(inc_l, inc_u, type = 'interval2') ~ eduLevel * cntry,
                   data = essIncData_small,
                   model = 'po',
                   dist = 'loglogistic')

summary(lnormFit)
```

getFitEsts *Get Estimates from icenReg Regression Model*

Description

Gets estimates from a ic_par or ic_sp object. Provided estimates conditional on regression parameters found in newdata.

Usage

```
getFitEsts(fit, newdata, p, q)
```

Arguments

fit	model fit with ic_par or ic_sp
newdata	data.frame containing covariates
p	percentiles
q	quantiles

Details

If newdata is left blank, baseline estimates will be returned (i.e. all covariates = 0). If p is provided, will return the estimated F^-1(p | x). If q is provided, will return the estimated F(q | x). If neither p nor q are provided, the estimated conditional median is returned.

For ic_par fits, it is worth noting that F^-1(p | x) is approximated using R's optimize function, so there may be some numerical error.

In the case of ic_sp, the MLE of the baseline survival is not necessarily unique, as probability mass is assigned to disjoint Turnbull intervals, but the likelihood function is indifferent to how probability mass is assigned within these intervals. In order to have a well defined estimate returned, we assume probability is assigned uniformly in these intervals. In otherwords, we return *a* maximum likelihood estimate, but don't attempt to characterize *all* maximum likelihood estimates with this function. If that is desired, all the information needed can be extracted with getSCurves.

Author(s)

Clifford Anderson-Bergman

Examples

```
simdata <- simIC_weib(n = 500, b1 = .3, b2 = -.3,
                          inspections = 6, inspectLength = 1)
fit <- ic_par(Surv(l, u, type = 'interval2') ~ x1 + x2,
                  data = simdata)
new_data <- data.frame(x1 = c(1,2), x2 = c(-1,1))
rownames(new_data) <- c('grp1', 'grp2')

estQ <- getFitEsts(fit, new_data, p = c(.25, .5, .75))

estP <- getFitEsts(fit, q = 400)
```

`getSCurves` *Get Estimated Survival Curves from Semi-parametric Model for Interval Censored Data*

Description

Extracts the estimated survival curve(s) from a ic_sp model for interval censored data. Output will be a list with two elements: the first item will be `$Tbull_ints`, which is the Turnbull intervals. This is a k x 2 matrix, with the first column being the beginning of the Turnbull interval and the second being the end. This is necessary due to the *representational non-uniqueness*; any survival curve that lies between the survival curves created from the upper and lower limits of the Turnbull intervals will have equal likelihood. See example for proper display of this. The second item is `$S_curves`, or the estimated survival probability at each Turnbull interval for individuals with the covariates provided in `newdata`. Note that multiple rows may be provided to newdata, which will result in multiple S_curves.

Usage

```
getSCurves(fit, newdata)
```

Arguments

`fit`	model fit with `ic_sp`
`newdata`	data.frame containing covariates for which the survival curve will be fit to. Rownames from `newdata` will be used to name survival curve. If left blank, baseline covariates will be used

Author(s)

Clifford Anderson-Bergman

Examples

```
set.seed(1)

sim_data <- simIC_weib(n = 500, b1 = .3, b2 = -.3,
                       shape = 2, scale = 2,
                       inspections = 6, inspectLength = 1)
fit <- ic_sp(Surv(l, u, type = 'interval2') ~ x1 + x2, data = sim_data, bs_samples = 0)

new_data <- data.frame(x1 = c(0,1), x2 = c(1, 1) )
#want to fit survival curves with above covariates
rownames(new_data) <- c('group 1', 'group 2')
#getSCurves will name the survival curves according to rownames

curveInfo <- getSCurves(fit, new_data)
xs <- curveInfo$Tbull_ints
#Extracting Turnbull intervals
sCurves <- curveInfo$S_curves
#Extracting estimated survival curves

plot(xs[,1], sCurves[[1]], xlab = 'time', ylab = 'S(t)',
     type = 's', ylim = c(0,1),
```

```
     xlim = range(as.numeric(xs), finite = TRUE))
#plotting upper survival curve estimate
lines(xs[,2], sCurves[[1]], type = 's')
#plotting lower survival curve estimate

lines(xs[,1], sCurves[[2]], col = 'blue', type = 's')
lines(xs[,2], sCurves[[2]], col = 'blue', type = 's')
#plotting upper and lower survival curves for group 2

# Actually, all this plotting is a unnecessary:
# plot(fit, new_data) will bascially do this all
# But this is more of a tutorial in case custom
# plots were desired
```

`ICNPMLE` *Computes the NPMLE for Univariate or Bivariate Interval Censored Data*

Description

Computes the MLE for a Interval Censored Data with a squeezing EM algorithm (*much* faster than the standard EM). Accepts either univariate interval censored data (where `times` is an n x 2 matrix with `times[,1]` being the left side of the interval and `times[,2]` is the right side), or bivariate interval censored data (where `times` is an n x 4 matrix with `times[,1:2]` being left and right side of the interval for event 1 and `times[,3:4]` being the left and right side of the interval for event 2).

Usage

```
ICNPMLE(times, B = c(1,1), max.inner = 100, max.outer = 100, tol = 1e-10)
```

Arguments

`times`	either an n x 2 or n x 4 `data.frame` or `matrix` of censoring intervals
`B`	A vector indicating whether each end of the intervals are open (0) or closed (1). Alternatively, this could be an n x 2 or n x 4 matrix of indicators for each individual interval
`max.inner`	number of inner loops used in optimization algorithm
`max.outer`	number of outer loops used in optimization algorithm
`tol`	numerical tolerance

Author(s)

Clifford Anderson-Bergman

Also uses Marloes Maathius's `MLEcens::reduc` function for calculation of the clique matrix.

References

Anderson-Bergman, C., (2014) Semi- and non-parametric methods for interval censored data with shape constraints, Ph.D. Thesis

Yu, Y., (2010), Improved EM for Mixture Proportions with Applications to Nonparametric ML Estimation for Censored Data, *preprint*

Maathuis, M., (2005). Reduction algorithm for the NPMLE for the distribution function of bivariate interval censored data. *Journal of Computational and Graphical Statistics* Vol 14 pp 252\- 262

Examples

```
simData <- simBVCen(500)

fit <- ICNPMLE(simData)

fit
```

---

ic_par *Parametric Regression Models for Interval Censored Data*

Description

Fits a parametric regression model for interval censored data. Can fit either a Cox-PH model or a proportional odds model.

Usage

```
ic_par(formula, data, model = 'ph', dist = 'weibull', weights = NULL)
```

Arguments

formula	regression formula. Response must be a Surv object of type 'interval2'
data	dataset
model	What type of model to fit. Current choices are "ph" (Cox PH) or "po" (proportional odds)
dist	What baseline parametric distribution to use. See details for current choices
weights	vector of case weights. Not standardized; see details

Details

Currently supported distributions choices are "exponential", "weibull", "gamma", "lnorm" and "loglogistic".

Does not allow uncensored data points at t = 0 (i.e. where left side of interval == right side == 0), as this will lead to a degenerate estimator for most parametric families. Unlike the current implementation of survreg, does allow left side of intervals of positive length to 0 and right side to be Inf.

In regards to weights, they are not standardized. This means that if weight[i] = 2, this is the equivalent to having two observations with the same values as subject i.

For numeric stability, if abs(right - left) < 10^-6, observation is considered uncensored rather than interval censored with an extremely small interval.

Author(s)

Clifford Anderson-Bergman

Examples

```
data(miceData)

logist_ph_fit <- ic_par(Surv(l, u, type = 'interval2') ~ grp,
                        data = miceData, dist = 'loglogistic')

logist_po_fit <- ic_par(Surv(l, u, type = 'interval2') ~ grp,
                        data = miceData, dist = 'loglogistic', model = 'po')

summary(logist_ph_fit)
summary(logist_po_fit)
```

`ic_sp` *Semi-Parametric models for Interval Censored Data*

Description

Fits a semi-parametric model for interval censored data. Can fit either a Cox-PH model or a proportional odds model.

The covariance matrix for the regression coefficients is estimated via bootstrapping. For large datasets, this can become slow so parallel processing can be used to take advantage of multiple cores via the `foreach` package.

Usage

```
ic_sp(formula, data, model = 'ph', weights = NULL,
      bs_samples = 0, useMCores = F, seed = NULL)
```

Arguments

`formula`	regression formula. Response must be a `Surv` object of type `'interval2'`
`data`	dataset
`model`	What type of model to fit. Current choices are "ph" (Cox PH) or "po" (proportional odds)
`weights`	vector of case weights. Not standardized; see details
`bs_samples`	Number of bootstrap samples used for estimation of standard errors
`useMCores`	Should multiple cores be used for bootstrap sample? Does not register cluster (see example)
`seed`	seed for bootstrap. If `seed == NULL`, a random seed is still used. See details

Details

In regards to weights, they are not standardized. This means that if weight[i] = 2, this is the equivalent to having two observations with the same values as subject i.

It is very important to note that a random seed is *always* set if bs_samples > 0 (via `set.seed(seed)`), which can create problems in simulation studies if the same seed is set in every call to ic_sp during a simulation study. If `seed == NULL`, then the starting seed will be `round(runif(0, max = 10^8))`, which should be approximately equivalent to not setting a seed.

Likelihood function is maximized using an active set algorithm.

Author(s)

Clifford Anderson-Bergman

Examples

```
set.seed(1)

sim_data <- simIC_weib(n = 500, inspections = 5, inspectLength = 1)
ph_fit <- ic_sp(Surv(l, u, type = 'interval2') ~ x1 + x2, data = sim_data)
# Default fits a Cox-PH model

summary(ph_fit)
# Regression estimates close to true 0.5 and -0.5 values


new_data <- data.frame(x1 = c(0,1), x2 = c(1, 1) )
rownames(new_data) <- c('group 1', 'group 2')
plot(ph_fit, new_data)
# plotting the estimated survival curves

po_fit <- ic_sp(Surv(l, u, type = 'interval2') ~ x1 + x2, data = sim_data,
                model = 'po')
# fits a proportional odds model

summary(po_fit)

# Not run: how to set up multiple cores
# library(doParallel)
# myCluster <- makeCluster(2, type = 'FORK')
# registerDoParallel(myCluster)
# fit <- ic_sp(Surv(l, u, type = 'interval2') ~ x1 + x2,
#              data = sim_data, useMCores = TRUE)
```

---

miceData *Lung Tumor Interval Censored Data from Hoel and Walburg 1972*

Description

RFM mice were sacrificed and examined for lung tumors. This resulted in current status interval censored data: if the tumor was present, this implied left censoring and if no tumor was present this implied right censoring. Mice were placed in two different groups: conventional environment or germ free environment.

Usage

```
data(miceData)
```

Format

A data frame with 144 rows and 3 variables

- `l` left side of observation interval
- `u` right side of observation interval
- `grp` Group for mouse. Either ce (conventional environment) or ge (grem-free environment)

Source

Hoel D. and Walburg, H.,(1972), Statistical analysis of survival experiments, *The Annals of Statistics*, 18, 1259-1294

Examples

```
data(miceData)

coxph_fit <- ic_sp(Surv(l, u, type = 'interval2') ~ grp,
                   bs_samples = 50,
                   data = miceData)

#In practice, more bootstrap samples should be used for inference
#Keeping it quick for CRAN testing purposes

summary(coxph_fit)
```

`optCliq` *Computes the MLE for a Binary Mixture Model.*

Description

Computes the MLE for a Binary Mixture Model. Used internally for ICNPMLE, but may be useful in other problems. In the abstraction, solves the problem

$$\arg\max_p \sum_{i=1}^{n} \log\left(\sum_{j=1} p_j C_{ij}\right)$$

where C_{ij} is an indicator of whether the i-th observation could have come from the j-th source. C is referred to as the *clique matrix*.

Usage

```
optCliq(cliqMat, tol = 10^-10,
        inner_loops = 100, outer_loops = 20)
```

Arguments

`cliqMat`	n x m clique matrix. n = number of observations, m = number of components
`tol`	numerical tolerance
`inner_loops`	number of inner loops used
`outer_loops`	number of outer loops used

Examples

```
testData <- simBVCen()
#simulate bivariate interval censored data

cliqMat <- MLEcens::reduc(testData, cm = TRUE)$cm
#computes the cliqMat associated with data

cliqFit <- optCliq(cliqMat)
#optimizes the component weights for clique matrix

cliqFit
```

`simBVCen` *Simulates Bivariate Interval Censored Data*

Description

Simulates Bivariate Interval Censored Data

Usage

```
simBVCen(n = 1000)
```

Arguments

`n`	number of observations simulated

Examples

```
testData <- simBVCen()
#simulate bivariate interval censored data

bvcenFit <- ICNPMLE(testData)

bvcenFit
```

`simIC_weib` *Simulates interval censored data from regression model with a Weibull baseline*

Description

Simulates interval censored data from a regression model with a weibull baseline distribution. Used for demonstration

Usage

```
simIC_weib(n = 100, b1 = 0.5, b2 = -0.5, model = "ph", shape = 2,
  scale = 2, inspections = 2, inspectLength = 2.5, rndDigits = NULL)
```

Arguments

n	Number of samples simulated
b1	Value of first regression coefficient
b2	Value of second regression coefficient
model	Type of regression model. Options are 'po' (prop. odds) and 'ph' (Cox PH)
shape	shape parameter of baseline distribution
scale	scale parameter of baseline distribution
inspections	number of inspections times of censoring process
inspectLength	max length of inspection interval
rndDigits	number of digits to which the inspection time is rounded to, creating a discrete inspection time. If `rndDigits = NULL`, the inspection time is not rounded, resulting in a continuous inspection time

Details

Exact event times are simulated according to regression model: covariate `x1` is distributed `rnorm(n)` and covariate `x2` is distributed `1 - 2 * rbinom(n, 1, 0.5)`. Event times are then censored with a case II interval censoring mechanism with `inspections` different inspection times. Time between inspections is distributed as `runif(min = 0, max = inspectLength)`. Note that the user should be careful in simulation studies not to simulate data where nearly all the data is right censored (or more over, all the data with x2 = 1 or -1) or this can result in degenerate solutions!

Author(s)

Clifford Anderson-Bergman

Examples

```
set.seed(1)
sim_data <- simIC_weib(n = 500, b1 = .3, b2 = -.3, model = 'ph',
                       shape = 2, scale = 2, inspections = 6, inspectLength = 1)
#simulates data from a cox-ph with beta weibull distribution.

diag_covar(Surv(l, u, type = 'interval2') ~ x1 + x2, data = sim_data, model = 'po')
diag_covar(Surv(l, u, type = 'interval2') ~ x1 + x2, data = sim_data, model = 'ph')
#'ph' fit looks better than 'po'; the difference between the transformed survival
#function looks more constant
```

Index