Package 'cchs'

October 12, 2022

Type Package

Title Cox Model for Case-Cohort Data with Stratified Subcohort-Selection

Version 0.4.2

Date 2020-09-10

Author E. Jones

Maintainer E. Jones <edmundjones79@gmail.com>

Description Contains a function, also called 'cchs', that calculates Estimator III of Borgan et al (2000), <DOI:10.1023/A:1009661900674>. This estimator is for fitting a Cox proportional hazards model to data from a case-cohort study where the subcohort was selected by stratified simple random sampling.

License GPL-3

Depends R (>= 2.15.0), survival (>= 2.36-12)

LazyData yes

NeedsCompilation no

Repository CRAN

Date/Publication 2020-09-10 20:50:02 UTC

R topics documented:

cchs *Cox model for case–cohort data with stratified subcohort-selection*

Description

cchs fits a Cox proportional-hazards regression model to case-cohort data where the subcohort was selected by stratified simple random sampling. It uses Estimator III of Borgan et al (2000).

Usage

```
cchs(formula, data=parent.frame(), inSubcohort, stratum,
samplingFractions, cohortStratumSizes, precision=NULL,
returnAdjustedTimes=FALSE, swap=TRUE, dropNeverAtRiskRows=TRUE,
dropSubcohEventsDfbeta=FALSE, adjustSampFracIfAnyNAs=FALSE,
keepAllCoxphElements=FALSE, confidenceLevel=0.95, verbose=FALSE,
annotateErrors=TRUE, coxphControl, ...)
```

Arguments

formula	An object of class formula that specifies the terms in the model. The left-hand side must be a Surv object. The special terms cluster and strata are not allowed.
data	A data-frame or environment that contains the variables used in the formula. The variables named in inSubcohort, stratum, samplingFractions, and cohortStratumSizes will be looked for first in data, if that is a data-frame, and then in the environment that cchs was called from.
inSubcohort	A vector of logical variables that shows whether each observation/row is in the subcohort (TRUE) or not (FALSE).
stratum	A vector that defines the strata within which the subcohort was selected. Each element of stratum corresponds to one observation/row in the data. The elements can be character strings, integers, or any other type of variable that can be converted to a factor.
samplingFractions, cohortStratumSizes	samplingFractions is a vector of the sampling fractions in the different strata, and cohortStratumSizes is a vector of the sizes of the strata in the full cohort. Exactly one of these must be given. There are two possible forms for the vector: if it has names, then these must all be distinct and include the names of the strata (and if one value of stratum is "France", then samplingFraction["France"] should be the sampling fraction for that stratum); if it does not have names, then it must have one element for each observation/row in the data.
precision	For example, if the times were recorded to the nearest day but are stored as numbers of years, then precision should be 1/365.25. If there are no tied event-times, then it makes no difference what precision is. If there are tied event-times and precision is a number, then the tied event-times will be slightly changed before the estimator is calculated. If there are tied event-times and precision is NULL (meaning unspecified), then the estimator cannot be calculated and an error will be thrown.

`returnAdjustedTimes`
: If this is `TRUE`, the object returned by `cchs` will contain the exit-times after they have been adjusted to deal with any tied event-times. If a row is dropped because of missing data (`NA`s) then its exit-time is not adjusted.

`swap`
: If this is `FALSE` then the swapping will be omitted (in the formula for Estimator III in Borgan et al 2000, the randomly selected observation/row will not be removed). This is only intended to be used for testing or development.

`dropNeverAtRiskRows`
: If this is `TRUE`, observations/rows whose at-risk periods do not include any of the event-times will be dropped just before `cchs` internally calls `coxph`. These observations/rows make no difference to the regression coefficients produced by `coxph`, but they do affect the dfbeta residuals (see Langholz & Jiao 2007) and therefore the variance-estimates, because `coxph` calculates the dfbeta residuals using an approximation.

`dropSubcohEventsDfbeta`
: If this is `FALSE`, which is the default, the dfbeta residuals and therefore the variance-estimates will be calculated exactly as described by Borgan et al (2000). If it is `TRUE`, they will be calculated as described by Langholz & Jiao (2007) (see "There is a slight approximation ..." in section 2.4).

`adjustSampFracIfAnyNAs`
: If this is `TRUE`, and if any observations are dropped because of missing data (`NA`s), then the sampling fractions will be recalculated using the numbers of observations after those observations are dropped.

`keepAllCoxphElements`
: If this is `TRUE`, then the object returned by `cchs` will contain elements such as `loglik` and `linear.predictors` from the object that was produced by `cchs`'s internal call to `coxph`. These are not likely to be relevant or correct, since `cchs` manipulates and changes the dataset in many ways before passing it to `coxph`. (For a list of the elements produced by `coxph`, see `coxph.object`.)

`confidenceLevel`
: The level for the hazard-ratio confidence intervals (a number in the interval [0,1]).

`verbose`
: If this is `TRUE`, detailed information about the internal manipulations and calculations will be displayed.

`annotateErrors`
: If this is `TRUE`, and if certain functions that are called internally by `cchs` produce errors or warnings, then extra messages will be added to make those easier to understand. The disadvantage of this is that the call stack produced by `traceback` is more complicated.

`coxphControl, ...`
: These are optional arguments to control the working of `coxph` when it is called internally by `cchs`. If `coxphControl` is supplied then it must be a list produced by `coxph.control`, and if "`...`" arguments are supplied then it must be possible to pass them to `coxph.control`.

Details

In a case–cohort study, the dataset consists only of the cases (the participants who have an event) and the participants who are in the subcohort, which is a randomly selected subset of the cohort. In

a stratified case–cohort study, the subcohort is selected by stratified simple random sampling. This means that the cohort is divided into strata, and from each stratum a proportion of the participants equal to that stratum's sampling fraction is selected to be in the subcohort (and within each stratum, each participant is selected with equal probability). For more on stratified case–cohort studies see any of the references listed below.

`cchs` fits a Cox proportional-hazards regression model to data from a stratified case–cohort study, using the time-fixed version of Estimator III from Borgan et al (2000). Estimators I and II from Borgan et al (2000) are available by using `cch` with the options `method="I.Borgan"` and `method="II.Borgan"`, but only Estimator III is score-unbiased, which is the main desirable criterion. The data must be in the usual form where each row corresponds to one observation (that is, one participant). `cchs` works by manipulating the data in various ways, then passing it to `coxph` (which is suitable for fitting a Cox model to data from a cohort study), and finally making corrections to the variance-estimates. It is planned that a vignette will be produced and this will contain more detail.

For normal use, the logical (boolean) arguments should have their default values. `cchs` performs a complete-case analysis, meaning that rows will be dropped if they contain `NA`s in any of the variables that appear in the model, including inside the `Surv()`, or in `inSubcohort` or `stratum`. `NA`s are not allowed in `samplingFractions` or `cohortStratumSizes`, unless that vector has names and any of those names are not equal to values of `stratum`, in which case the corresponding elements can be `NA`.

`cchs` does not normally give replicable results, because the swapping and the small changes to tied event-times are random (see `swap` and `precision` in the Arguments section). To get exactly the same results every time, use `set.seed` with a fixed seed just before calling `cchs`.

For more information about `cchs` see the article in *R Journal*, Jones (2018).

Value

An S3 object of class `cchs`. This is a list that contains the following elements:

`coefficients`	The vector of coefficients.
`var`	The variance matrix of the coefficients.
`loglik`	A vector of two elements: the first is the log-likelihood with the initial values of the coefficients that were used in the iteration to find the maximum likelihood, and the second is the maximized log-likelihood—that is, the log-likelihood with the final values of the coefficients. (Strictly speaking these should all say "pseudo-likelihood" instead of "likelihood".)
`iter`	The number of iterations used by `coxph`.
`n`	The number of observations (that is, rows), that were used in the call to `coxph`.
`nevent`	The number of events (also called failures).
`call`	The call that was used to create the `cchs` object (an object of mode `call`).
`coeffsTable`	A summary of the main output. This is a matrix that contains the hazard ratios, confidence intervals for them, *p*-values for the Wald tests, log hazard ratios (which are the coefficients in the Cox model), and standard errors of the log hazard ratios.
`confidenceLevel`	The level for the confidence intervals in `coeffsTable`. (This is a copy of the `confidenceLevel` argument.)

`nEachStatus` A vector with three elements: the numbers of subcohort non-cases, subcohort cases, and non-subcohort cases. The sum of these is `n`.

`nStrata` The number of strata that appear in the data.

`message` A message about observations that have been dropped because of NAs and event-times that have been changed to deal with ties, if either of these happened.

If `keepAllCoxphElements` is `TRUE`, then the `cchs` object will also contain the other elements listed under `coxph.object`. If `returnAdjustedTimes` is `TRUE`, then it will contain an `adjustedTimes` element, which is a vector of the adjusted exit-times (with elements in the same order as the observations/rows in the data).

References

Borgan, Ø., Langholz, B., Samuelsen S.O., Goldstein, L., Pogoda, J. (2000). Exposure stratified case–cohort designs. *Lifetime Data Analysis* **6** (1), 39–58. (link)

Cologne, J., Preston, D.L., Imai, K., Misumi, M., Yoshida, K., Hayashi, T., Nakachi, K. (2012). Conventional case–cohort design and analysis for studies of interaction. *International Journal of Epidemiology* **41** (4), 1174–1186. (link)

Jones, E. (2018). cchs: An R package for stratified case–cohort studies. *R Journal* **10** (1), 484–494. (link)

Langholz, B., Jiao, J. (2007). Computational methods for case–cohort studies. *Computational Statistics and Data Analysis* **51** (8), 3737–3748. (link)

See Also

`cch`, which can calculate Estimators I and II from Borgan et al (2000), `coxph`, which `cchs` uses internally, and `coxph.control`, a container for certain parameters that are passed to `coxph`. These are all in the **survival** package.

`cchsData`, an example dataset that `cchs` can be used on.

Examples

```
# Analyze the relation between survival and three covariates in cchsData.
# The times are stored as numbers of days, so precision has to be 1. The
# selection of the subcohort was stratified according to two strata, defined
# by cchsData$localHistol, and the sampling fractions are stored in
# cchsData$sampFrac.

cchs(Surv(time, isCase) ~ stage + centralLabHistol + ageAtDiagnosis,
      data=cchsData, inSubcohort=inSubcohort, stratum=localHistol,
      samplingFractions=sampFrac, precision=1)

# Do the same analysis using cohortStratumSizes instead of samplingFractions.
# For the value of cohortStratumSizes see the Details section of ?cchsData.
# These two calls to cchs will give slightly different results unless set.seed
# is used with the same seed just before both of them.

cchs(Surv(time, isCase) ~ stage + centralLabHistol + ageAtDiagnosis,
      data=cchsData, inSubcohort=inSubcohort, stratum=localHistol,
```

```
cohortStratumSizes=c(favorable=3622, unfavorable=406), precision=1)
```

`cchsData` *Data from a case–cohort study with stratified subcohort-selection*

Description

A case–cohort dataset where the subcohort was selected by stratified simple random sampling. This is an artificial dataset that was made from `nwtco`, a real dataset from the National Wilms Tumor Study (NWTS). It is designed for demonstrating the use of `cchs`.

Format

`id`	An ID number.
`localHistol`	Result of the histology from the local institution.
`centralLabHistol`	Result of the histology from the central laboratory.
`stage`	Stage of the cancer (I, II, III, or IV).
`study`	The study (NWTS-3 or NWTS-4). For details see this NWTS webpage.
`isCase`	Indicator for whether this participant had a relapse or not.
`time`	Number of days from diagnosis of Wilms tumor to relapse or censoring.
`ageAtDiagnosis`	Age in years at diagnosis of Wilms tumor.
`inSubcohort`	Indicator for whether this participant is in the subcohort or not.
`sampFrac`	The sampling fraction for the stratum that contains this participant.

Details

The `nwtco` data is from two clinical trials but can be regarded as cohort data. `cchsData` can be created from it by running the code in the Source section below, which is partly based on the Examples section of the `cch` documentation.

Two strata are used for the subcohort-selection, corresponding to the two values of `localHistol`. The sampling fraction is 5% for the stratum defined by `localHistol="favorable"` and 20% for the stratum defined by `localHistol="unfavorable"`. After the subcohort is selected, the sampling fractions are recalculated using the exact integer numbers of participants in the subcohort and the full cohort, and then stored in the data-frame.

As an alternative to the sampling fractions, the stratum sizes in the full cohort could be used. A suitable value for the `cohortStratumSizes` argument to `cchs` would be `c(favorable=3622, unfavorable=406)`. This can be worked out by entering `table(nwtco$instit, useNA="always")` and noting that for `nwtco$instit` and `nwtco$histol`, a value of 1 means "favorable histology result" and 2 means "unfavorable"—this is not stated in the `nwtco` documentation but can be deduced from the line in the `cch` examples that contains `labels=c("FH","UH")`, or by comparing the output of the `table` command with the numbers in Table 1 of Breslow & Chatterjee (1999).

For information about the two clinical trials, NWTS-3 and NWTS-4, see D'Angio et al. (1989) and Green et al. (1998) respectively, or the National Wilms Tumor Study website.

Source

```
# Starting with nwtco, rename variables, convert some to factors, drop
# in.subcohort (which is used elsewhere for a different simulated dataset), etc.
library(survival, quietly=TRUE)
cchsData <- data.frame(
   id = nwtco$seqno,
   localHistol = factor(nwtco$instit, labels=c("favorable", "unfavorable")),
  centralLabHistol = factor(nwtco$histol, labels=c("favorable", "unfavorable")),
   stage = factor(nwtco$stage, labels=c("I", "II", "III", "IV")),
   study = factor(nwtco$study, labels=c("NWTS-3", "NWTS-4")),
   isCase = as.logical(nwtco$rel),
   time = nwtco$edrel,
   ageAtDiagnosis = nwtco$age / 12  # nwtco$age is in months
)

# Define the intended sampling fractions for the two strata.
samplingFractions <- c(favorable=0.05, unfavorable=0.2)

# Select participants/rows to be in the subcohort by stratified simple random
# sampling.
cchsData$inSubcohort <- rep(FALSE, nrow(cchsData))
set.seed(1)
for (stratumName in levels(cchsData$localHistol)) {
   inThisStratum <- cchsData$localHistol == stratumName
   stratumSubcohortSize <-
         round(samplingFractions[stratumName] * sum(inThisStratum))
   rowsToSetTrue <- sample(which(inThisStratum), size=stratumSubcohortSize)
   cchsData$inSubcohort[rowsToSetTrue] <- TRUE
}

# Change the sampling fractions to their exact values.
stratumSubcohortSizes <- table(cchsData$localHistol[cchsData$inSubcohort])
stratumCohortSizes <- table(cchsData$localHistol)
samplingFractions <- stratumSubcohortSizes / stratumCohortSizes
samplingFractions <- c(samplingFractions)  # make it a vector, not a table

# Keep only the cases and the subcohort.
cchsData <- cchsData[cchsData$isCase | cchsData$inSubcohort,]

# Put the sampling fraction in each row of the data-frame.
cchsData$sampFrac <-
      samplingFractions[match(cchsData$localHistol, names(samplingFractions))]
```

References

Breslow, N.E., Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **48** (4), 457–468. (link)

D'Angio, G.J., Breslow, N., Beckwith, J.B., Evans, A., Baum, E., Delorimier, A., Fernbach, D., Hrabovsky, E., Jones, B., Kelalis, P., Othersen, H.B., Tefft, M., Thomas, P.R.M. (1989). Treatment of Wilms' tumor: Results of the third national Wilms' tumor study. *Cancer* **64** (2), 349–360. (link)

Green, D.M., Breslow, N.E., Beckwith, J.B., Finklestein, J.Z., Grundy, P.E., Thomas, P.R., Kim, T., Shochat, S.J., Haase, G.M., Ritchey, M.L., Kelalis, P.P., D'Angio, G.J. (1998). Comparison between single-dose and divided-dose administration of dactinomycin and doxorubicin for patients with Wilms' tumor: a report from the National Wilms' Tumor Study Group. *Journal of Clinical Oncology* **16** (1), 237–245. (link)

Index