Package 'Counterfactual'

July 21, 2025

Type Package

Title Estimation and Inference Methods for Counterfactual Analysis

Version 1.2

Author Mingli Chen, Victor Chernozhukov, Ivan Fernandez-Val, Blaise Melly

Maintainer Ivan Fernandez-Val <ivanf@bu.edu>

Description

Implements the estimation and inference methods for counterfactual analysis described in Chernozhukov, Fernandez-Val and Melly (2013) <DOI:10.3982/ECTA10582> ``Inference on Counterfactual Distributions,'' Econometrica, 81(6). The counterfactual distributions considered are the result of changing either the marginal distribution of covariates related to the outcome variable of interest, or the conditional distribution of the outcome given the covariates. They can be applied to estimate quantile treatment effects and wage decompositions.

License GPL (>= 2)

LazyLoad yes

Imports quantreg, survival, Hmisc, foreach, doRNG, doParallel, parallel

NeedsCompilation no

Repository CRAN

Date/Publication 2020-01-31 16:30:08 UTC

Contents

counterfactual *Estimation and Inference Methods for Counterfactual Analysis*

Description

Implements the estimation and inference methods for counterfactual analysis described in Chernozhukov, Fernandez-Val and Melly (2013). `Counterfactual` reports point estimates, pointwise confidence bands, and simultaneous confidence bands for function-valued quantile effects (QE). It also reports p-values for functional hypotheses such as no effect, constant effect and stochastic dominance. The uniform confidence bands and p-values are obtained by inverting Kolmogorov-Smirnov (KS) and Cramer-von-Misses-Smirnov (CMS) statistics. The distribution of these statistics is approximated by empirical or weighted bootstrap. We recommend the use of weighted bootstrap when the covariates X include discrete components with small cell sizes.

Usage

```
counterfactual(formula, data, weights, na.action = na.exclude,
  group, treatment =FALSE, decomposition = FALSE, counterfactual_var,
  transformation = FALSE, quantiles = c(1:9)/10,
  method = "qr", trimming = 0.005, nreg = 100,
  scale_variable, counterfactual_scale_variable, censoring = 0,
  right = FALSE, nsteps = 3, firstc = 0.1, secondc = 0.05,
  noboot = FALSE, weightedboot = FALSE, seed = 8, robust = FALSE,
  reps = 100, alpha = 0.05, first = 0.1, last = 0.9, cons_test = 0,
  printdeco = TRUE, sepcore = FALSE, ncore = 1)
```

Arguments

`formula`	a formula object, with the response Y on the left of a ~ operator, and the covariate terms X, separated by + operators, on the right.
`data`	a data.frame in which to interpret the variables named in the formula, or in the weights argument. If this is missing, then the variables in the formula should be on the search list.
`weights`	vector of observation weights.
`na.action`	a function to filter missing data. The default (with `na.fail`) is to create an error if any missing values are found. A possible alternative is `na.omit`, which deletes observations that contain one or more missing values.
`quantiles`	quantile indexes of interest for the QE. It should be a vector of values between 0 and 1 with default `c(1:9)/10`.
`group`	name of a binary variable defining the reference population (value 0) and counterfactual population (value 1).
`treatment`	logical: if `TRUE`, then computes the structure or treatment effect (only useful when `group` is specified); if `FALSE`, then computes the composition effect.
`decomposition`	logical: if `TRUE`, then computes the structure effect, composition effect and total effect; if `FALSE`, then computes the structure effect (only useful when `group` is specified, and `treatment=TRUE`).

`transformation` logical: if `TRUE`, then the counterfactual distribution of X is generated by transformation of the distribution of X in the reference population.

`counterfactual_var`
selects the values of X in the counterfactual population (only useful when `group` is not specified).

`method` selects the model to be used to estimate the conditional distribution. The following methods have been implemented: `qr` (quantile regression, the default), `loc` (location shift), `locsca` (location scale shift), `cqr` (censored quantile regression), `cox` (duration regression), `logit` (logit distribution regression), `probit` (probit distribution regression), and `lpm` (linear probability model).

`trimming` value between 0 and 0.5 specifying the amount of trimming to avoid tail estimation in `qr` method; default is 0.005.

`nreg` sets the number of regressions estimated to approximate the conditional distribution; default is 100.

`scale_variable` selects the components of X that affect the scale in the `locsca` method.

`counterfactual_scale_variable`
selects the counterfactual values of the components of X that affect the scale in the `locsca` method (only useful when `counterfactual_var` is specified).

`censoring` variable specifying the censoring point for each observations (only useful when `method=cqr`).

`right` logical: if `TRUE`, then indicates that the variable is right-censored; if `TRUE`, then indicates that the variable is left-censored (only useful when `method=cqr`).

`nsteps` selects the number of steps performed in the `cqr` method; default and minimum is 3 (only useful when `method=cqr`).

`firstc` selects the percentage of observations thrown out during the second step in the `cqr` method; default is 0.1 (only useful when `method=cqr`).

`secondc` selects the percentage of observations thrown out during the third and further steps of the `cqr` method; default is 0.05 (only useful when `method=cqr`).

`noboot` logical: if `TRUE`, then suppresses the bootstrap; if `FALSE`, the default, then runs the bootstrap.

`weightedboot` logical: if `TRUE`, then implements weighted bootstrap with standard exponential weights; if `FALSE`, the default, then implements empirical bootstrap (only useful when `noboot=FALSE`).

`seed` sets the seed for the random number generation (only useful when `noboot=FALSE`).

`robust` logical: if `TRUE`, then uses the bootstrap interquartile range to estimate standard errors in the KS and CMS statistics; if `FALSE`, the default, then uses the bootstrap standard deviation to estimate standard errors in the KS and CMS statistics (only useful when `noboot=FALSE`).

`reps` number of bootstrap replications; default is 100 (only useful when `noboot=FALSE`).

`alpha` a real number between 0 and 1 reflecting the desired significance level for the confidence bands and hypotheses tests (only useful when `noboot=FALSE`).

`first` sets the lowest quantile that is used for functional inference; default is 0.1 (only useful when `noboot=FALSE`).

last	sets the highes quantile that is used for functional inference; default is 0.9 (only useful when noboot=FALSE).
cons_test	adds tests of the null hypothesis that the QEs = cons_test at all the specified quantiles (only useful when noboot=FALSE).
printdeco	logical: if FALSE, then suppresses table of results.
sepcore	logical: if TRUE, then multiple cores are used for parallel computing.
ncore	number of cores used for parallel computing (only useful when sepcore=TRUE).

Details

The populations to construct the observed and counterfactual distributions can be specified in two alternative ways. If the option group is specified and treatment=FALSE, then the observed distribution is estimated from the conditional and covariate distributions of group=0, and the counterfactual distribution is estimated from the conditional distribution of group=0 and the covariate distribution of group=1. If group is specified and treatment=TRUE, then the observed distribution is estimated from the conditional and covariate distributions of group=1, and the counterfactual distribution is estimated from the conditional distribution of group=0 and the covariate distribution of group=1. If group is specified, treatment=TRUE and decomposition=TRUE, then all the previous observed and counterfactual distributions are estimated. Alternatively, the option counterfactual_var can be specified. In this case, the variables specified in the right hand side of formula contain the covariate values used to estimate the observed distribution and the variables specified in counterfactual_var contain the covariate values to estimate the counterfactual distribution. Note that counterfactual_var must contain exactly the same number of variables as in the right hand side of formula and that the order matters. In addition, if counterfactual_var is a deterministic transformation of the covariates in the reference population, then transformation should be set to TRUE.

method:

qr is the default, selects the method based on the linear quantile regression estimator of Koenker and Bassett (1978).

loc selects the linear location shift method.

locsca selects the linear location-scale shift method. The logarithm of the variance of the residuals is assumed to be a linear function of the variables given in scale_variable.

cqr selects the method based on the censored linear quantile regression estimator of Chernozhukov and Hong (2002). The variable with the censoring values for each observation must be specified in censoring. By default, this estimator is a three-steps estimator. The number of steps can be increased by the option nsteps.

cox selects the methob based on the proportional hazard or duration regression estimator of Cox (1972).

logit selects the method based on the distribution regression estimator of Chernozhukov, Fernandez-Val and Melly (2013) with logit link function.

probit selects the method based on the distribution regression estimator of Chernozhukov, Fernandez-Val and Melly (2013) with probit link function.

lpm selects the method based on the distribution regression estimator of Chernozhukov, Fernandez-Val and Melly (2013) with linear link function.

We refer the user to Chen, Chernozhukov, Fernandez-Val and Melly (2016) for a more detailed description of the methods.

Value

Return a list of results

`quantiles` quantile indexes of interest for the QE.

`structure_effect`
: a vector with the estimated structure effects at the quantile indexes specified with `quantiles`. This vector is reported when `group` is specified and `treatment=TRUE`.

`composition_effect`
: a vector with the estimated composition effects at the quantile indexes specified with `quantiles`. If `group` is specified, then this vector is reported when `treatment=FALSE`, or `treatment=TRUE` and `decomposition=TRUE`.

`total_effect` a vector with the estimated total effects at the quantile indexes specified with `quantiles`. This vector is reported when `group` is specified, `treatment=TRUE` and `decomposition=TRUE`.

`sample_quantile_ref0`
: a matrix with 4 columns. The columns contain the point estimates, standard errors, uniform lower end of confidence band, and uniform upper end of confidence band for the quantiles of Y in the observed distribution estimated using sample quantiles at the quantile indexes specified with `quantiles`. If `group` is specified, then this matrix is reported when `treatment=FALSE`, or `treatment=TRUE` and `decomposition=TRUE`.

`model_quantile_ref0`
: a matrix with 4 columns. The columns contain the point estimates, standard errors, uniform lower end of confidence band, and uniform upper end of confidence band for the quantiles of Y in the observed distribution estimated using the conditional model at the quantile indexes specified with `quantiles`. If `group` is specified, then this matrix is reported when `treatment=FALSE`, or `treatment=TRUE` and `decomposition=TRUE`.

`model_quantile_counter`
: a matrix with 4 columns. The columns contain the point estimates, standard errors, uniform lower end of confidence band, and uniform upper end of confidence band for the quantiles of Y in the counterfactual distribution estimated using the conditional model at the quantile indexes specified with `quantiles`.

`sample_quantile_ref1`
: a matrix with 4 columns. The columns contain the point estimates, standard errors, uniform lower end of confidence band, and uniform upper end of confidence band for the quantiles of Y in the observed distribution of the population defined by $group=1$ estimated using sample quantiles at the quantile indexes specified with `quantiles`. This matrix is reported when `group` is specified and `treatment=TRUE`.

`model_quantile_ref1`
: a matrix with 4 columns. The columns contain the point estimates, standard errors, uniform lower end of confidence band, and uniform upper end of confidence band for the quantiles of Y in the observed distribution of the population

defined by $group=1$ estimated using the conditional model at the quantile indexes specified with quantiles. This matrix is reported when group is specified and treatment=TRUE.

nreg — number of regressions estimated to approximate the conditional distribution.

resSE — a matrix with 6 columns. The columns contain the point estimates, standard errors, pointwise lower end of confidence band, pointwise upper end of confidence band, uniform lower end of confidence band, and uniform upper end of confidence band for the structure or treatment quantile effect at the quantile indexes specified with quantiles. This matrix is reported when group is specified and treatment=TRUE.

testSE — a matrix with 2 columns including the p-values based on the KS and CMS statistics for several functional hypotheses on the structure or treatment effect. The first row tests the null-hypothesis of correct specification of the conditional model. The second row tests the null hypothesis that the change in the distribution of the covariates has no effect. The following rows tests the null hypotheses of constant QE, positive QE, and negative QE. An additional row testing the null hypotheses of constant QE (but at a different level than 0) is added if the option cons_test is specified. This matrix is reported when group is specified and treatment=TRUE.

resCE — a matrix with 6 columns. The columns contain the point estimates, standard errors, pointwise lower end of confidence band, pointwise upper end of confidence band, uniform lower end of confidence band, and uniform upper end of confidence band for the composition quantile effect at the quantile indexes specified with quantiles. If group is specified, then this matrix is reported when treatment=FALSE, or treatment=TRUE and decomposition=TRUE.

testCE — a matrix with 2 columns including the p-values based on the KS and CMS statistics for several functional hypotheses on the composition effect. The first row tests the null-hypothesis of correct specification of the conditional model. The second row tests the null hypothesis that the change in the distribution of the covariates has no effect. The following rows tests the null hypotheses of constant QE, positive QE, and negative QE. An additional row testing the null hypotheses of constant QE (but at a different level than 0) is added if the option cons_test is specified. If group is specified, then this matrix is reported when treatment=FALSE, or treatment=TRUE and decomposition=TRUE.

resTE — a matrix with 6 columns. The columns contain the point estimates, standard errors, pointwise lower end of confidence band, pointwise upper end of confidence band, uniform lower end of confidence band, and uniform upper end of confidence band for the total quantile effect at the quantile indexes specified with quantiles. This matrix is reported when group is specified, treatment=TRUE and decomposition=TRUE

testTE — a matrix with 2 columns including the p-values based on the KS and CMS statistics for several functional hypotheses on the total effect. The first row tests the null-hypothesis of correct specification of the conditional model. The second row tests the null hypothesis that the change in the distribution of the covariates has no effect. The following rows tests the null hypotheses of constant QE, positive QE, and negative QE. An additional row testing the null hypotheses of constant QE (but at a different level than 0) is added if the option cons_test

is specified. This matrix is reported when group is specified, treatment=TRUE and decomposition=TRUE.

Author(s)

Mingli Chen, Victor Chernozhukov, Ivan Fernandez-Val, Blaise Melly

References

Chen, M., Chernozhukov, V., I. Fernandez-Val, and B. Melly (2016). Counterfactual Analysis in R: A Vignette.

Chernozhukov, V., I. Fernandez-Val, and B. Melly (2013). Inference on Counterfactual Distributions. Econometrica 81(6), 2205-2268.

Chernozhukov, V., and H. Hong (2002). Three-step Censored Quantile Regression and Extramarital Affairs.Journal of the American Statistical Association, 97, 872-881.

Cox, D. R. (1972). Regression Models and Life Tables. Journal of the Royal Statistical Society, Ser. B, 34, 187-220.

Koenker, R., and G. Bassett (1978). Regression Quantiles. Econometrica, 46(1), 33-50.

Examples

```
#Counterfactual distribution of X constructed by transformation of reference distribution
## Not run:

data(engel)
attach(engel)
counter_income <- mean(income)+0.75*(income-mean(income))
rqres <- counterfactual(foodexp~income, counterfactual_var=counter_income,
nreg=100, transformation=TRUE, sepcore = TRUE, ncore=2)

## End(Not run)

# Wage decomposition: counterfactual and reference populations correspond to different groups
data(nlsw88)
attach(nlsw88)
lwage <- log(wage)

# method: logit
logitres<-counterfactual(lwage~tenure+ttl_exp+grade, group=union, treatment=TRUE,
decomposition=TRUE, method="logit", noboot=TRUE, sepcore = TRUE,ncore=2)
```

---

nlsw88 *NLSW, 1988 extract*

Description

National Longitudinal Surveys, Women sample

Usage

```
data(nlsw88)
```

Format

A data frame with 2246 observations on the following 17 variables:

`idcode` a numeric vector, NLS id

`age` a numeric vector, age at current year

`race` a numeric vector, race

`married` a numeric vector

`never_married` a numeric vector

`grade` a numeric vector, current grade completed

`collgrad` a numeric vector, college graduate

`south` a numeric vector, lives in south

`smsa` a numeric vector, lives in SMSA

`c_city` a numeric vector, lives in central city

`industry` a numeric vector

`occupation` a numeric vector

`union` a numeric vector, union worker

`wage` a numeric vector, hourly wage

`hours` a numeric vector, usual hours worked

`ttl_exp` a numeric vector, total work experience

`tenure` a numeric vector, job tenure(years)

Details

The NLSW88 data contains data of a group of women in their 30s and early 40s to study labor force patterns.

Source

Stata website

References

Stata website: http://www.stata-press.com/data/r10/g.html

Examples

```
data(nlsw88)
attach(nlsw88)
plot(wage,tenure)
```

Index