Package ‘SortedEffects’

September 7, 2021

Type Package

Title Estimation and Inference Methods for Sorted Causal Effects and Classification Analysis

Version 1.6.0

Author Shuowen Chen [aut, cre],
       Victor Chernozhukov [aut],
       Ivan Fernandez-Val [aut],
       Ye Luo [aut]

Maintainer Shuowen Chen <swchen@bu.edu>


License MIT + file LICENSE

Depends R (>= 2.10)

URL https://github.com/shuowencs/SortedEffects

Encoding UTF-8

LazyData true

Imports boot, graphics, Hmisc, pbapply, parallel, quantreg, SparseM, stats, dummies

RoxygenNote 7.1.1

Suggests knitr, rmarkdown

VignetteBuilder knitr

NeedsCompilation no

Repository CRAN

Date/Publication 2021-09-07 07:00:13 UTC

R topics documented:

  ca .................................................. 2
  mortgage ........................................ 5


1
**ca**

*Empirical Classification Analysis (CA) and Inference*

**Description**

ca conducts CA estimation and inference on user-specified objects of interest: first (weighted) moment or (weighted) distribution. Users can use t to specify variables in interest. When object of interest is moment, use cl to specify whether want to see averages or difference of the two groups.

**Usage**

```r
ca(
  fm,
  data,
  method = c("ols", "logit", "probit", "QR"),
  var_type = c("binary", "continuous", "categorical"),
  var,
  compare,
  subgroup = NULL,
  samp_weight = NULL,
  taus = c(5:95)/100,
  u = 0.1,
  interest = c("moment", "dist"),
  t = c(1, 1, rep(0, dim(data)[2] - 2)),
  cl = c("both", "diff"),
  cat = NULL,
  alpha = 0.1,
  b = 500,
  parallel = FALSE,
  ncores = detectCores(),
  seed = 1,
  bc = TRUE,
  range_cb = c(1:99)/100,
  boot_type = c("nonpar", "weighted")
)
```
Arguments

fm
Regression formula

data
The data in use: full sample or subpopulation in interset

method
Models to be used for estimating partial effects. Four options: "logit" (binary response), "probit" (binary response), "ols" (interactive linear with additive errors), "QR" (linear model with non-additive errors). Default is "ols".

var_type
The type of parameter in interest. Three options: "binary", "categorical", "continuous". Default is "binary".

var
Variable T in interset. Should be a character.

compare
If parameter in interest is categorical, then user needs to specify which two category to compare with. Should be a 1 by 2 character vector. For example, if the two levels to compare with is 1 and 3, then c="1", "3"), which will calculate partial effect from 1 to 3. To use this option, users first need to specify var as a factor variable.

subgroup
Subgroup in interest. Default is NULL. Specification should be a logical variable. For example, suppose data contain indicator variable for women (female if 1, male if 0). If users are interested in women SPE, then users should specify subgroup = data[, "female"] == 1.

samp_weight
Sampling weight of data. Input should be a n by 1 vector, where n denotes sample size. Default is NULL.

taus
Indexes for quantile regression. Default is c(5:95)/100.

u
Percentile of most and least affected. Default is set to be 0.1.

interest
Generic objects in the least and most affected subpopulations. Two options: (1) "moment": weighted mean of Z in the u-least/most affected subpopulation. (2) "dist": distribution of Z in the u-least/most affected subpopulation. Default is interest = "moment".

t
An index for ca object. Should be a 1 by ncol(data) indicator vector. Users can either (1) specify names of variables of interest directly, or (2) use 1 to indicate the variable of interest. For example, total number of variables is 5 and interested in the 1st and 3rd vars, then specify t = c(1,0,1,0,0).

cl
If moment = "interest", cl allows the user to get the variables of interest (specified in t option) of the most and least affected groups. The default is "both", which shows the variables of the two groups; the alternative is "diff", which shows the difference of the two groups. The user can use the summary.ca to tabulate the results, which also contain the standard errors and p-values. If interest = "dist", this option doesn’t have any bearing and user can leave it to be the default value.

cat
P-values in classification analysis are adjusted for multiplicity to account for joint testing of zero coefficients on for all variables within a category. Suppose we have selected specified 3 variables in interest: t = c("a", "b", "c"). Without loss of generality, assume "a" is not a factor, while "b" and "c" are two factors. Then users need to specify as cat = c("b", "c"). Default is NULL.

alpha
Size for confidence interval. Shoule be between 0 and 1. Default is 0.1
Number of bootstrap draws. Default is 500.

Whether the user wants to use parallel computation. The default is FALSE and only 1 CPU will be used. The other option is TRUE, and user can specify the number of CPUs in the ncores option.

Number of cores for computation. Default is set to be detectCores(), which is a function from package parallel that detects the number of CPUs on the current host. For large dataset, parallel computing is highly recommended since bootstrap is time-consuming.

Pseudo-number generation for reproduction. Default is 1.

Whether want the estimate to be bias-corrected. Default is TRUE. If FALSE uncorrected estimate and corresponding confidence bands will be reported.

When interest = "dist", we sort and unique variables in interest to estimate weighted CDF. For large dataset there can be memory problem storing very many of observations, and thus users can provide a Sort value and the package will sort and unique based on the weighted quantile of Sort. If users don’t want this feature, set range_cb = NULL. Default is c(1:99)/100.

Type of bootstrap. Default is "nonpar", and the package implements nonparametric bootstrap. The alternative is "weighted", and the package implements weighted bootstrap.

Details

All estimates are bias-corrected and all confidence bands are monotonized. The bootstrap procedures follow algorithm 2.2 as in Chernozhukov, Fernandez-Val and Luo (2018).

Value

If subgroup = NULL, all outputs are whole sample. Otherwise output are subgroup results. When interest = "moment", the output is a list showing

- est Estimates of variables in interest.
- bse Bootstrap standard errors.
- joint_p P-values that are adjusted for multiplicity to account for joint testing for all variables.
- pointwise_p P-values that doesn’t adjust for join testing

If users have further specified cat (e.g., !is.null(cat)), the fourth component will be replaced with p_cat: P-values that are a djusted for multiplicity to account for joint testing for all variables within a category. Users can use summary.ca to tabulate the results.

When interest = "dist", the output is a list of two components:

- infresults A list that stores estimates, upper and lower confidence bounds for all variables in interest for least and most affected groups.
- sortvar A list that stores sorted and unique variables in interest.

We recommend using plot.ca command for result visualization.
Examples

data("mortgage")
### Regression Specification
fm <- deny ~ black + p_irat + hse_inc + cc_red + mcred + pubrec + ltv_med + ltv_high + denpmi + selfemp + single + hischl
### Specify characteristics of interest
t <- c("deny", "p_irat", "black", "hse_inc", "cc_red", "mcred", "pubrec", "denpmi", "selfemp", "single", "hischl", "ltv_med", "ltv_high")
### issue ca command
CA <- ca(fm = fm, data = mortgage, var = "black", method = "logit", cl = "diff", t = t, b = 50, bc = TRUE)

Description
Mortgage Denial

Usage
mortgage

Format
Contains the data on mortgage application in Boston from 1990, (Munnell et al., 1996.) We obtain
the data from the companion website of Stock and Watson (2011). The file contains the following
variables:
deny  indicator for mortgage application denied
p_irat monthly debt to income ratio
black  indicator for black applicant
hse_inc monthly housing expenses to income ratio
loan_val loan to assessed property value ratio
cc_red  consumer credit score with 6 categories. 1 if no "slow" payments or delinquencies, 2 if one or
two "slow" payments or delinquencies, 3 if more than two "slow" payments or delinquencies,
4 if insufficient credit history for determination, 5 if delinquent credit history with payment
60 days overdue, and 6 if delinquent credit history with payments 90 days overdue.
mcred  mortgage credit score with 4 categories. 1 if no late mortgage payments, 2 if no mortgage
payment history, 3 if one or two late mortgage payments, and 4 if more than two late mortgages
payments
pubrec  indicator for any public record of credit problems: bankruptcy, charge-offs, collection
actions
denpmi  indicator for applicant applied for mortgage insurance and was denied
selfemp  indicator for self-employed applicant
single indicator for single applicant
hischl indicator for high school graduated applicant
probunmp 1989 Massachusetts unemployment rate in the applicant’s history
condo indicator for unit is a condominium
ltv_med indicator for medium loan to property value ratio [.80, .95]
ltv_high indicator for high loan to property value ratio >.95

Source

---

plot.ca  

Distribution plotting

Description
Plots distributions and joint uniform confidence bands of variables in interest from ca command.

Usage
```r
## S3 method for class 'ca'
plot(x, var, main = NULL, sub = NULL, xlab = NULL, ylab = NULL, ...)
```

Arguments
- `x` Output of `ca` command with `interest = "dist"`.
- `var` Name of variable for plotting
- `main` Main title of the plot. Default is NULL.
- `sub` Sub title of the plot. Default is NULL.
- `xlab` x-axis label. Default is NULL.
- `ylab` y-axis label. Default is NULL.
- `...` graphics parameters to be passed to the plotting routines.

Examples
```r
data("mortgage")
### Regression Specification
fm <- deny ~ black + p_irat + hse_inc + ccred + mcred + pubrec +
ltv_med + ltv_high + denpmi + selfemp + single + hischl
### Specify characteristics of interest for plotting
t2 <- "p_irat"
### issue ca command
CAdist <- ca(fm = fm, data = mortgage, var = "black", method = "logit",
t = "p_irat", b = 50, interest = "dist")
```
### plot.spe

```r
### plotting
plot(CAdist, var = "p_irat", ylab = "Prob",
     xlab = "Monthly Debt-to-Income Ratio", sub = "logit model")
```

---

#### Description

Plot output of `spe` command. The x-axis limits are set to the specified range of percentile index.

#### Usage

```r
## S3 method for class 'Var.spe'
plot(
  x,       # Output of `spe` command.
  ylim = NULL,       # y-axis limits. Default is NULL.
  main = NULL,       # Main title of the plot. Default is NULL.
  sub = NULL,        # Sub title of the plot. Default is NULL.
  xlab = "Percentile Index",   # x-axis label. Default is "Percentile Index".
  ylab = "Sorted Effects",   # y-axis label. Default is "Sorted Effects".
  ...                 # graphics parameters to be passed to the plotting routines.
)
```

#### Arguments

- `x` Output of `spe` command.
- `ylim` y-axis limits. Default is NULL.
- `main` Main title of the plot. Default is NULL.
- `sub` Sub title of the plot. Default is NULL.
- `xlab` x-axis label. Default is "Percentile Index".
- `ylab` y-axis label. Default is "Sorted Effects".
- `...` graphics parameters to be passed to the plotting routines.

#### Examples

```r
data("mortgage")
fm <- deny ~ black + p_irat + hse_inc + ccred + mcred + pubrec + ltv_med + ltv_high + denpmi + selfemp + single + hischl
test <- spe(fm = fm, data = mortgage, var = "black", method = "logit",
            us = c(2:98)/100, b = 50)

plot(x = test, main="APE and SPE of Being Black on the prob of Mortgage Denial", sub="Logit Model", ylab="Change in Probability")
```
plot.subpop  
*Plot 2-dimensional projections of variables in interest.*

Description

Takes output from `subpop` command as inputs and plots 2-dimensional projection plots of two specified variables. If a variable in interest is of type factor, then the user must put it on the y-axis. If the variable on the y-coordinate is a factor, range of y-axis is set to be the factor level. Otherwise, users can use `summary.subpop` to know the ranges of variables in the two groups.

Usage

```r
## S3 method for class 'subpop'
plot(
  x, varx, vary, xlim = NULL, ylim = NULL, main = NULL, sub = NULL, xlab = NULL, ylab = NULL, overlap = FALSE, ...
)
```

Arguments

- `x`  
  Output of `subpop` command.
- `varx`  
  The name of the variable to be plotted on the x-axis.
- `vary`  
  The name of the variable name to be plotted on the y-axis.
- `xlim`  
  The range of x-axis. Default is `NULL`.
- `ylim`  
  The range of y-axis. Default is `NULL`. If the variable on the y-coordinate is a factor, the default will set it to be the factor level, and users don’t need to specify `ylim`.
- `main`  
  Main title of the plot. Default is `NULL`.
- `sub`  
  Sub title of the plot. Default is `NULL`.
- `xlab`  
  x-axis label. Default is `NULL`.
- `ylab`  
  y-axis label. Default is `NULL`.
- `overlap`  
  Whether user wants to allow observations included in both confidence sets. Default is `FALSE`, and the plot drops the overlapped observations.
- `...`  
  Graphics parameters to be passed to the plotting routines.
**Examples**

```r
data("mortgage")

### Regression Specification
fm <- deny ~ black + p_irat + hse_inc + ccred + mcred + pubrec +
  ltv_med + ltv_high + denpmi + selfemp + single + hischl

### Issue the subpop command
set_b <- subpop(fm, data = mortgage, method = "logit", var = "black",
  u = 0.1, alpha = 0.1, b = 50)

### Plotting
plot(set_b, varx = mortgage$p_irat, vary = mortgage$hse_inc,
  xlim = c(0, 1.5), ylim = c(0, 1.5), xlab = "Debt/Income",
  ylab = "Housing expenses/Income", overlap = TRUE)
```

---

**Description**

`spe` conducts SPE estimation and inference at user-specified quantile index. The bootstrap procedures follows algorithm 2.1 as in Chernozhukov, Fernandez-Val and Luo (2018). All estimates are bias-corrected and all confidence bands are monotonized. For graphical results, please use `plot.spe`.

**Usage**

```r
spe(
  fm,
  data,
  method = c("ols", "logit", "probit", "QR"),
  var_type = c("binary", "continuous", "categorical"),
  var,
  compare,
  subgroup = NULL,
  samp_weight = NULL,
  us = c(1:9)/10,
  alpha = 0.1,
  taus = c(5:95)/100,
  b = 500,
  parallel = FALSE,
  ncores = detectCores(),
  seed = 1,
  bc = TRUE,
  boot_type = c("nonpar", "weighted")
)
```
**Arguments**

- **fm**: Regression formula.
- **data**: Data in use.
- **method**: Models to be used for estimating partial effects. Four options: "logit" (binary response), "probit" (binary response), "ols" (interactive linear with additive errors), "QR" (linear model with non-additive errors). Default is "ols".
- **var_type**: The type of parameter in interest. Three options: "binary", "categorical", "continuous". Default is "binary".
- **var**: Variable T in interest. Should be a character type.
- **compare**: If parameter in interest is categorical, then user needs to specify which two category to compare with. Should be a 1 by 2 character vector. For example, if the two levels to compare with is 1 and 3, then c="1","3"), which will calculate partial effect from 1 to 3. To use this option, users first need to specify var as a factor variable.
- **subgroup**: Subgroup in interest. Default is NULL. Specification should be a logical variable. For example, suppose data contains indicator variable for women (female if 1, male if 0). If users are interested in women SPE, then users should specify subgroup = data[,"female"] == 1.
- **samp_weight**: Sampling weight of data. Input should be a n by 1 vector, where n denotes sample size. Default is NULL.
- **us**: Percentile of interest for SPE. Should be a vector of values between 0 and 1. Default is c(1:9)/10.
- **alpha**: Size for confidence interval. Should be between 0 and 1. Default is 0.1
- **taus**: Indexes for quantile regression. Default is c(5:95)/100.
- **b**: Number of bootstrap draws. Default is set to be 500.
- **parallel**: Whether the user wants to use parallel computation. The default is FALSE and only 1 CPU will be used. The other option is TRUE, and user can specify the number of CPUs in the ncores option.
- **ncores**: Number of cores for computation. Default is set to be detectCores(), which is a function from package parallel that detects the number of CPUs on the current host. For large dataset, parallel computing is highly recommended since bootstrap is time-consuming.
- **seed**: Pseudo-number generation for reproduction. Default is 1.
- **bc**: Whether want the estimate to be bias-corrected. Default is TRUE. If FALSE uncorrected estimate and corresponding confidence bands will be reported.
- **boot_type**: Type of bootstrap. Default is "nonpar", and the package implements nonparametric bootstrap. The other alternative is "weighted", and the package implements weighted bootstrap.

**Value**

The output is a list with 4 components: (1) spe stores spe estimates, the upper and lower confidence bounds, and standard errors; (2) ape stores ape estimates, the upper and lower confidence bounds, and the standard error; (3) us stores percentile index as in \codespe command; (4) alpha stores significance level as in spe command.
Examples

```r
data("mortgage")
fm <- deny ~ black + p_irat + hse_inc + ccred + mcred + pubrec + ltv_med +
ltv_high + denpmi + selfemp + single + hischl

test <- spe(fm = fm, data = mortgage, var = "black", method = "logit",
us = c(2:98)/100, b = 50)
```

Description

`subpop` conducts set inference on the groups of most and least affected. When `subgroup = NULL`, output is for whole sample. Otherwise the results are subgroup. The output of `subpop` is a list containing six components: `cs_most`, `cs_least`, `u`, `subgroup`, `most` and `least`. As the names indicate, `cs_most` and `cs_least` denote the confidence sets for the most and least affected units. `u` stores the u-th most and least affected index. `subgroup` stores the indicators for subpopulations. `most` and `least` store the data of the most and least affected groups. The confidence sets can be visualized using the `plot.subpop` command while the two groups can be tabulated via the `summary.subpop` command.

Usage

```r
subpop(
  fm,
  data,
  method = c("ols", "logit", "probit", "QR"),
  var_type = c("binary", "continuous", "categorical"),
  var,
  compare,
  subgroup = NULL,
  samp_weight = NULL,
  taus = c(5:95)/100,
  u = 0.1,
  alpha = 0.1,
  b = 500,
  seed = 1,
  parallel = FALSE,
  ncores = detectCores(),
  boot_type = c("nonpar", "weighted")
)
```
Arguments

fm Regression formula
data The data in use
method Models to be used for estimating partial effects. Four options: "logit" (binary response), "probit" (binary response), "ols" (interactive linear with additive errors), "QR" (linear model with non-additive errors). Default is "ols".
var_type The type of parameter in interest. Three options: "binary", "categorical", "continuous". Default is "binary".
var Variable T in interset. Should be a character.
compare If parameter in interest is categorical, then user needs to specify which two category to compare with. Should be a 1 by 2 character vector. For example, if the two levels to compare with is 1 and 3, then c=("1", "3"), which will calculate partial effect from 1 to 3. To use this option, users first need to specify var as a factor variable.
subgroup Subgroup in interest. Default is NULL. Specification should be a logical variable. For example, suppose data contains indicator variable for women (female if 1, male if 0). If users are interested in women SPE, then users should specify subgroup = data[, "female"] == 1.
samp_weight Sampling weight of data. Input should be a n by 1 vector, where n denotes sample size. Default is NULL.
taus Indexes for quantile regression. Default is c(5:95)/100.
u Percentile of most and least affected. Default is set to be 0.1.
alpha Size for confidence interval. Should be between 0 and 1. Default is 0.1
b Number of bootstrap draws. Default is set to be 500.
seed Pseudo-number generation for reproduction. Default is 1.
parallel Whether the user wants to use parallel computation. The default is FALSE and only 1 CPU will be used. The other option is TRUE, and user can specify the number of CPUs in the ncores option.
ncores Number of cores for computation. Default is set to be detectCores(), which is a function from package parallel that detects the number of CPUs on the current host. For large dataset, parallel computing is highly recommended since bootstrap is time-consuming.
boot_type Type of bootstrap. Default is "nonpar", and the package implements nonparametric bootstrap. The alternative is "weighted", and the package implements weighted bootstrap.

Examples

data("mortgage")
### Regression Specification
fm <- deny ~ black + p_irat + hse_inc + ccred + mcred + pubrec +
     ltv_med + ltv_high + denpmi + selfemp + single + hischl
### Issue the subpop command
set_b <- subpop(fm, data = mortgage, method = "logit", var = "black",
                 u = 0.1, alpha = 0.1, b = 50)
**summary.ca**

*Return the output of ca function.*

---

**Description**

Return the output of ca function.

**Usage**

```r
## S3 method for class 'ca'
summary(object, ...)
```

**Arguments**

- `object` Output of ca command with interest = "moments".
- `...` additional arguments affecting the summary produced.

**Examples**

```r
data("mortgage")
### Regression Specification
fm <- deny ~ black + p_irat + hse_inc + ccred + mcred + pubrec + 
ltv_med + ltv_high + denpmi + selfemp + single + hischl
### Specify characteristics of interest
t <- c("deny", "p_irat", "black", "hse_inc", "ccred", "mcred", "pubrec", 
"denpmi", "selfemp", "single", "hischl", "ltv_med", "ltv_high")
### Issue ca command
CA <- ca(fm = fm, data = mortgage, var = "black", method = "logit", 
c1 = "both", t = t, b = 50, bc = TRUE)
### Report summary table
summary(CA)
```

---

**summary.spe**

*Tabulate the output of spe function.*

---

**Description**

The option result allows user to tabulate either sorted estimates or average estimates. For sorted estimates, the table shows user-specified quantile indices, sorted estimates, standard errors, point-wise confidence intervals, and uniform confidence intervals. For average estimates, the table shows average estimates, standard errors, and confidence intervals.

**Usage**

```r
## S3 method for class 'spe'
summary(object, result = c("sorted", "average"), ...)
```
Arguments

object  The output of `spe` function.
result  Whether the user wants to see the sorted or the average estimates. Default is sorted, which shows the sorted estimates.
...
additional arguments affecting the summary produced.

Examples

data("mortgage")
fm <- deny ~ black + p_irat + hse_inc + ccredi + mcredd + pubrec + ltv_med +
ltv_high + denpmi + selfemp + single + hischl

## S3 method for class 'subpop'
summary(object, vars = NULL, ...)

Arguments

object  Output of `subpop` command.
vars  The variables that users want to see the summary statistics. The default is NULL and the command shows all variables. The summary statistics include min 1st quartile, median, mean, 3rd quartile and the max.
...
additional arguments affecting the summary produced.

Examples

data("mortgage")
### Regression Specification
fm <- deny ~ black + p_irat + hse_inc + ccredi + mcredd + pubrec +
ltv_med + ltv_high + denpmi + selfemp + single + hischl

### Issue the subpop command
set_b <- subpop(fm, data = mortgage, method = "logit", var = "black", u = 0.1, alpha = 0.1, b = 50)

### Produce summary of two variables
summary(set_b, vars = c("p_irat", "hse_inc"))
Description

Wage Data

Usage

wage2015

Format

Consists of white, non-hispanic individuals aging from 25 to 64 and working more than 35 hours per week during at least 50 weeks of the year. Excludes self-employed, individuals living in group quarters; individuals in the military, agricultural or private household sectors; individuals with inconsistent reports on earnings and employment status; individuals with allocated or missing information in any of the variables used in the analysis; individuals with hourly wage rate below $3. Contains 32,523 workers including 18,137 men and 14,382 women. The file contains the following variables:

- lnw  log of hourly wages
- weight  CPS sampling weight
- female  gender indicator: 1 if female
- exp1  max(age-years of educ-7, 0)
- exp2  exp1^2/100
- exp3  exp1^3/100
- exp4  exp1^4/100
- occ  Aggregated occupation with 5 categories: managers, service, sales, construction and production.
- ind  Aggregated industry with 12 categories: minery, construction, manufacture, retail, transport, information, finance, professional, education, leisure, services, public.
- educ  Education attainment with 5 categories: lhs (less than high school graduate, years of educ < 12), hsg (high school graduate: years of educ = 12), sc (some college: 13<=years of educ<=15), cg (college: 16<=years of educ<=17), ad (advanced degree: years of educ>=18).
- ms  Marital Status with 5 categories: married, widowed, separated, divorced, and nevermarried.
- region  Regions with 4 categories: mw (midwest), so (south), we (west), ne (northeast).

Source

Index

* datasets
  mortgage, 5
  wage2015, 15
ca, 2, 6, 13
mortgage, 5
plot.ca, 4, 6
plot.spe, 7, 9
plot.subpop, 8, 11, 14
spe, 7, 9, 13, 14
subpop, 8, 11, 14
summary.ca, 3, 4, 13
summary.spe, 13
summary.subpop, 8, 11, 14
wage2015, 15