Package 'SurveyCC'

January 9, 2024

Title Canonical Correlation for Survey Data

Version 0.1.1

Description Performs canonical correlation for survey data, including multiple tests of significance for secondary canonical correlations. A key feature of this package is that it incorporates survey data structure directly in a novel test of significance via a sequence of simple linear regression models on the canonical variates. See reference - Cruz-Cano, Cohen, and Mead-Morse (2024) ``Canonical Correlation Analysis of Survey data: the SurveyCC R package" The R Journal under review.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.2.3

Depends R (>= 2.10)

LazyData true

Imports candisc, graphics, stats, survey

NeedsCompilation no

Author Raul Cruz-Cano [aut, cre] (<https://orcid.org/0000-0001-7715-1198>)

Maintainer Raul Cruz-Cano <raulcruz@iu.edu>

Repository CRAN

Date/Publication 2024-01-09 08:20:29 UTC

R topics documented:

reducedNYTS2021data *Reduced NYTS data set*

Description

A subset of data from the National Youth Tobacco Survey (NYTS) Study

Usage

```
reducedNYTS2021data
```

Format

`reducedNYTS2021data`**:**
A data frame with 1150 rows and 24 columns:

psu2 Primary sampling unit
v_stratum2 Strata information ...

Source

https://www.cdc.gov/tobacco/data_statistics/surveys/nyts/data/index.html

reducedPATHdata *Reduced PATH tobacco use data set*

Description

A subset of data from the Population Assessment of Tobacco and Health (PATH) Study

Usage

```
reducedPATHdata
```

Format

`reducedPATHdata`**:**
A data frame with 132 rows and 107 columns:

PERSONID Participant ID
R01_AC1002 Ever smoked a cigarette ...

Source

https://www.icpsr.umich.edu/web/NAHDAP/studies/36498

surveycc *Canonical correlation analysis for complex survey data*

Description

This command extends the functionality of candisc::cancor by calculating the test statistics, degrees of freedom and p-values necessary to estimate and interpret the statistical significance of the secondary canonical corr according to the methods Wilks' lambda, Pillai's trace, and Hotelling-Lawley trace (Caliński et al., 2006) and Roy's largest root (Johnstone, 2009). The units and variables graphs (Gittins, 1986) can also be drawn by `surveycc` further complementing the information listed by the existing `cancor`.

Moreover, `csdcanon` implements an algorithm (Cruz-Cano, Cohen, and Mead-Morse, 2024) that allows the inclusion of complex survey design elements, e.g. strata, cluster and replicate weights, in the estimation of the statistical significance of the canonical correlations. The core idea of the algorithm is to reduce the problem of finding the correlations among the canonical variates and their corresponding statistical significance to calculating an equivalent sequence of univariate linear regression. This switch allows the user to take advantage of the existing theoretical and computational resources that integrate the complex survey design elements into these regression models (Valliant and Dever, 2018). Hence, this algorithm can include the same complex design elements as in `survey`.

Usage

```
surveycc(
  design_object,
  var.x,
  var.y,
  howmany = NA,
  dim1 = NA,
  dim2 = NA,
  selection = "FREQ"
)
```

Arguments

`design_object`	a survey design object generated from package `survey`, eg survey::svydesign
`var.x`	the first set of variables; a vector of names
`var.y`	the second set of variables; a vector of names
`howmany`	positive integer; allows the user to choose the number of canonical correlations for which the statistical significance statistics are displayed. Default is to choose the minimum of `length(var.x)` and `length(var.y)`. Cannot exceed this value.
`dim1, dim2`	determines which canonical variates serve as the horizontal and vertical axes in the optional plot. NOTE: if dim1 and dim2 not provided, no graph will be displayed.

`selection` allows the user to choose whether the type of sample size is equal to the number of rows in the data set or the sum of the survey weights.

Value

A list, containing the canonical correlation object, as well as tables of the various tests of significance. This includes the test statistics, degrees of freedom, and p-values for:

- Wilk's lambda
- Pillai's trace
- Hotelling-Lawley
- Roy's greatest root
- the Cruz-Cano algorithm using the survey design object

NOTE: For more information on the statistics presented, i.e. test statistic, df1, df2, Chi-Sq/F and p-val, please see the documentation in candisc::cancor for Wilk's Lambda, Pillai's Trace and Hotelling-Lawley Trace (although the present package uses a Chi-squared approximation to the F-distribution), and see the documentation in survey::svyglm for the Weighted/Complex Survey Design regression.

References

- Cruz-Cano, Cohen, and Mead-Morse. Canonical Correlation Analysis of Survey data: The SurveyCC R package. The R Journal under review; 2024.
- Gentzke AS, Wang TW, Cornelius M, Park-Lee E, Ren C, Sawdey MD, Cullen KA, Loretan C, Jamal A, Homa DM. Tobacco Product Use and Associated Factors among Middle and High School Students - National Youth Tobacco Survey, United States, 2021. rveill Summ. 2022;71(5):1-29. doi: 10.15585/mmwr.ss7105a1. PubMed PMID: 35271557.
- Gittins R. Canonical Analysis: A Review with Applications in Ecology: Springer Berlin Heidelberg; 1986.
- Caliński T., Krzyśko M. and WOłyński W. (2006) A Comparison of Some Tests for Determining the Number of Nonzero Canonical Correlations, Communications in Statistics -Simulation and Computation, 35:3, 727-749, DOI: 10.1080/03610 6290.
- Hyland A, Ambrose BK, Conway KP, et al. Design and methods of the Population Assessment of Tobacco and Health (PATH) StudyTobacco Control 2017;26:371-378.
- Johnstone IM. Approximate Null Distribution of the largest root in a Multivariate Analysis. Ann Appl Stat. 2009;3(4):1616-1633. doi: 10.1214/08-AOAS220. PMID: 20526465; PMCID: PMC2880335.
- Valliant R. and Dever JA. Survey Weights: A Step-by-Step Guide to Calculation: Stata Press; 2018. ISBN-13: 978-1-59718-260-7.

Examples

```
# PATH example
design_object <-
 survey::svrepdesign(
 id = ~PERSONID,
```

```
 weights = ~R01_A_PWGT,
 repweights = "R01_A_PWGT[1-9]+",
 type = "Fay",
 rho = 0.3,
 data=reducedPATHdata,
 mse = TRUE
 )
var.x <- c("R01_AC1022", "R01_AE1022", "R01_AG1022CG")
var.y <- c("R01_AX0075", "R01_AX0076")
howmany <- 2
dim1 <- 1
dim2 <- 2
surveycc(design_object, var.x, var.y, howmany = howmany,
  dim1 = dim1, dim2 = dim2, selection = "x")

# NYTS example
design_object <-
  survey::svydesign(
  ids = ~psu2,
  nest = TRUE,
  strata = ~v_stratum2,
  weights = ~finwgt,
  data = reducedNYTS2021data
)
var.x <- c("qn9", "qn38", "qn40", "qn53", "qn54", "qn64", "qn69", "qn74",
           "qn76", "qn78", "qn80", "qn82", "qn85", "qn88", "qn89")
var.y <- c("qn128", "qn129", "qn130", "qn131", "qn132", "qn134")
howmany <- 3
surveycc(design_object = design_object, var.x = var.x,
  var.y = var.y, howmany = howmany, selection = "x")
```

Index