Package ‘sketching’

October 14, 2022

Type Package

Title Sketching of Data via Random Subspace Embeddings

Version 0.1.2

Description Construct sketches of data via random subspace embeddings.

   For more details, see the following papers.

License GPL-3

Encoding UTF-8

LazyData true

RoxygenNote 7.2.0

Imports stats, MASS, Rcpp (>= 1.0.7), phangorn (>= 2.8.1)

LinkingTo Rcpp

Suggests knitr, rmarkdown, testthat (>= 3.0.0), lmtest (>= 0.9), ivreg (>= 0.6), sandwich (>= 3.0)

VignetteBuilder knitr

Depends R (>= 4.1.0)

URL https://github.com/sokbae/sketching/

BugReports https://github.com/sokbae/sketching/issues

Config/testthat/edition 3

NeedsCompilation yes

Author Sokbae Lee [aut, cre],
    Serena Ng [aut]

Maintainer Sokbae Lee <sl3841@columbia.edu>

Repository CRAN

Date/Publication 2022-09-07 07:50:02 UTC
Description

Angrist-Krueger (AK) dataset is a data extract from US Censuses that was analyzed in Angrist and Krueger (1991). In particular, the current dataset is from the 1970 Census, consisting of men born 1920-1929 (Year 1929 is the omitted cohort group).

Usage

AK

Format

A data frame with 247,199 rows and 42 variables:

LWKLYWGE  Outcome: log weekly wages
EDUC  Covariate of interest: years of education
YR20  Indicator variable for the year of birth: equals 1 if yob = 1920
YR21  Indicator variable for the year of birth: equals 1 if yob = 1921
YR22  Indicator variable for the year of birth: equals 1 if yob = 1922
YR23  Indicator variable for the year of birth: equals 1 if yob = 1923
YR24  Indicator variable for the year of birth: equals 1 if yob = 1924
YR25  Indicator variable for the year of birth: equals 1 if yob = 1925
YR26  Indicator variable for the year of birth: equals 1 if yob = 1926
YR27  Indicator variable for the year of birth: equals 1 if yob = 1927
YR28  Indicator variable for the year of birth: equals 1 if yob = 1928
QTR120  Quarter-of-birth indicator interacted with year-of-birth indicator
QTR121  Quarter-of-birth indicator interacted with year-of-birth indicator
QTR122  Quarter-of-birth indicator interacted with year-of-birth indicator
QTR123  Quarter-of-birth indicator interacted with year-of-birth indicator
QTR124  Quarter-of-birth indicator interacted with year-of-birth indicator
QTR125  Quarter-of-birth indicator interacted with year-of-birth indicator
QTR126 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR127 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR128 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR129 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR220 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR221 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR222 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR223 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR224 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR225 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR226 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR227 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR228 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR229 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR320 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR321 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR322 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR323 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR324 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR325 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR326 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR327 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR328 Quarter-of-birth indicator interacted with year-of-birth indicator
QTR329 Quarter-of-birth indicator interacted with year-of-birth indicator

CNST Constant

Source

The dataset is publicly available on Joshua Angrist’s website at https://economics.mit.edu/ faculty/angrist/data1/data/angkru1991/.

References

simulation_dgp

Simulating observations from the data-generating process considered in Lee and Ng (2022)

Description

Simulates observations from the data-generating process considered in Lee and Ng (2022)

Usage

simulation_dgp(n, d, hetero = FALSE)

Arguments

n sample size
d dimension of regressors from a multivariate normal distribution
hetero TRUE if the conditional variance of the error term is heteroskedastic and FALSE if it is homoskedastic (default: FALSE)

Value

An S3 object has the following elements.

Y n observations of outcomes
X n times d matrix of regressors
beta d dimensional vector of coefficients

References


Examples

data <- simulation_dgp(100, 5, hetero = TRUE)
y <- data$Y
x <- data$X
model <- lm(y ~ x)
Description

Provides a subsample of data using sketches

Usage

```
sketch(data, m, method = "unif")
```

Arguments

- `data`: (n times d)-dimensional matrix of data.
- `m`: (expected) subsample size that is less than n
- `method`: method for sketching: "unif" uniform sampling with replacement (default); "unif_without_replacement" uniform sampling without replacement; "bernoulli" Bernoulli sampling; "gaussian" Gaussian projection; "countskeck" CountSketch; "srht" subsampled randomized Hadamard transform; "fft" subsampled randomized trigonometric transforms using the real part of fast discrete Fourier transform (stats::fft).

Value

(m times d)-dimensional matrix of data For Bernoulli sampling, the number of rows is not necessarily m.

Examples

```
## Least squares: sketch and solve
# setup
n <- 1e+6 # full sample size
d <- 5    # dimension of covariates
m <- 1e+3 # sketch size
# generate psuedo-data
X <- matrix(stats::rnorm(n*d), nrow = n, ncol = d)
beta <- matrix(rep(1,d), nrow = d, ncol = 1)
eps <- matrix(stats::rnorm(n), nrow = n, ncol = 1)
Y <- X %*% beta + eps
intercept <- matrix(rep(1,n), nrow = n, ncol = 1)
# full sample including the intercept term
fullsample <- cbind(Y,intercept,X)
# generate a sketch using CountSketch
s_cs <- sketch(fullsample, m, "countsketch")
# solve without the intercept
ls_cs <- lm(s_cs[,1] ~ s_cs[,2] - 1)
# generate a sketch using SRHT
s_srht <- sketch(fullsample, m, "srht")
# solve without the intercept
ls_srht <- lm(s_srht[,1] ~ s_srht[,2] - 1)
```
Synopsis

Sketch using leverage score type sampling

Description

Provides a subsample of data using sketches

Usage

```
sketch_leverage(data, m, method = "leverage")
```

Arguments

- `data` (n times d)-dimensional matrix of data. The first column needs to be a vector of the dependent variable (Y)
- `m` subsample size that is less than n
- `method` method for sketching: "leverage" leverage score sampling using X (default); "root_leverage" square-root leverage score sampling using X.

Value

An S3 object has the following elements.

- `subsample` (m times d)-dimensional matrix of data
- `prob` m-dimensional vector of probabilities

References


Examples

```r
## Least squares: sketch and solve
# setup
n <- 1e+6 # full sample size
d <- 5   # dimension of covariates
m <- 1e+3 # sketch size
# generate pseudo-data
X <- matrix(stats::rnorm(n*d), nrow = n, ncol = d)
beta <- matrix(rep(1,d), nrow = d, ncol = 1)
eps <- matrix(stats::rnorm(n), nrow = n, ncol = 1)
Y <- X %*% beta + eps
intercept <- matrix(rep(1,n), nrow = n, ncol = 1)
# full sample including the intercept term
```
fullsample <- cbind(Y, intercept, X)
# generate a sketch using leverage score sampling
s.lev <- sketch_leverage(fullsample, m, "leverage")
# solve without the intercept with weighting
ls.lev <- lm(s.lev$subsample[,1] - s.lev$subsample[,2] - 1, weights = s.lev$prob)
Index

* datasets
  AK, 2

AK, 2

simulation_dgp, 4
sketch, 5
sketch_leverage, 6