Package 'SGDinference'

January 20, 2025

Type Package

Title Inference with Stochastic Gradient Descent

Version 0.1.0

Description Estimation and inference methods for large-scale mean and quantile regression models via stochastic (sub-)gradient descent (S-subGD) algorithms.
The inference procedure handles cross-sectional data sequentially:
(i) updating the parameter estimate with each incoming ``new observation",
(ii) aggregating it as a Polyak-Ruppert average, and
(iii) computing an asymptotically pivotal statistic for inference through random scaling.
The methodology used in the 'SGDinference' package is described in detail in the following papers:
(i) Lee, S., Liao, Y., Seo, M.H. and Shin, Y. (2022) <doi:10.1609/aaai.v36i7.20701> ``Fast and robust online inference with stochastic gradient descent via random scaling".
(ii) Lee, S., Liao, Y., Seo, M.H. and Shin, Y. (2023) <arXiv:2209.14502> ``Fast Inference for Quantile Regression with Tens of Millions of Observations".

License GPL-3

Imports stats, Rcpp (>= 1.0.5)

LinkingTo Rcpp, RcppArmadillo

RoxygenNote 7.2.3

Encoding UTF-8

Suggests knitr, rmarkdown, testthat (>= 3.0.0), lmtest (>= 0.9), sandwich (>= 3.0), microbenchmark (>= 1.4), conquer (>= 1.3.3)

VignetteBuilder knitr

Config/testthat/edition 3

Depends R (>= 3.5.0)

LazyData true

URL https://github.com/SGDinference-Lab/SGDinference/

BugReports https://github.com/SGDinference-Lab/SGDinference/issues

NeedsCompilation yes

Author Sokbae Lee [aut],
Yuan Liao [aut],
Myung Hwan Seo [aut],
Youngki Shin [aut, cre]

Maintainer Youngki Shin <shiny11@mcmaster.ca>

Repository CRAN

Date/Publication 2023-11-16 20:43:54 UTC

Contents

`Census2000` *Census2000*

Description

The Census2000 dataset from Acemoglu and Autor (2011) consists of observations on 26,120 nonwhite, female workers. This small dataset is constructed from "microwage2000_ext.dta" at https://economics.mit.edu/people/faculty/david-h-autor/data-archive. Specifically, observations are dropped if hourly wages are missing or years of education are smaller than 6. Then, a 5 percent random sample is drawn to make the dataset small.

Usage

```
Census2000
```

Format

A data frame with 26,120 rows and 3 variables:

ln_hrwage log hourly wages

edyrs years of education

exp years of potential experience

Source

The original dataset from Acemoglu and Autor (2011) is available at https://economics.mit.edu/people/faculty/david-h-autor/data-archive.

References

Acemoglu, D. and Autor, D., 2011. Skills, tasks and technologies: Implications for employment and earnings. In Handbook of labor economics (Vol. 4, pp. 1043-1171). Elsevier.

`SGDinference` *SGDinference*

Description

The 'SGDinference' package provides estimation and inference methods for large-scale mean and quantile regression models via stochastic (sub-)gradient descent (S-subGD) algorithms. The inference procedure handles cross-sectional data sequentially: (i) updating the parameter estimate with each incoming "new observation", (ii) aggregating it as a Polyak-Ruppert average, and (iii) computing an asymptotically pivotal statistic for inference through random scaling.

Author(s)

Sokbae Lee, Yuan Liao, Myung Hwan Seo, Youngki Shin

`sgdi_lm` *Averaged SGD and its Inference via Random Scaling*

Description

Compute the averaged SGD estimator and conduct inference via random scaling method.

Usage

```
sgdi_lm(
  formula,
  data,
  gamma_0 = NULL,
  alpha = 0.501,
  burn = 1,
  inference = "rs",
  bt_start = NULL,
  studentize = TRUE,
  no_studentize = 100L,
  intercept = TRUE,
  rss_idx = c(1),
  level = 0.95,
  path = FALSE,
  path_index = c(1)
)
```

Arguments

formula	formula. The response is on the left of a ~ operator. The terms are on the right of a ~ operator, separated by a + operator.
data	an optional data frame containing variables in the model.
gamma_0	numeric. A tuning parameter for the learning rate (gamma_0 x t ^ alpha). Default is NULL and it is determined by the adaptive method: 1/sd(y).
alpha	numeric. A tuning parameter for the learning rate (gamma_0 x t ^ alpha). Default is 0.501.
burn	numeric. A tuning parameter for "burn-in" observations. We burn-in up to (burn-1) observations and use observations from (burn) for estimation. Default is 1, i.e. no burn-in.
inference	character. Specifying the inference method. Default is "rs" (random scaling matrix for joint inference using all the parameters). "rss" is for ransom scaling subset inference. This option requires that "rss_indx" should be provided. "rsd" is for the diagonal elements of the random scaling matrix, excluding one for the intercept term.
bt_start	numeric. (p x 1) vector. User-provided starting value Default is NULL.
studentize	logical. Studentize regressors. Default is TRUE
no_studentize	numeric. The number of observations to compute the mean and std error for studentization. Default is 100.
intercept	logical. Use the intercept term for regressors. Default is TRUE. If this option is TRUE, the first element of the parameter vector is the intercept term.
rss_idx	numeric. Index of x for random scaling subset inference. Default is 1, the first regressor of x. For example, if we want to focus on the 1st and 3rd covariates of x, then set it to be c(1,3).
level	numeric. The confidence level required. Default is 0.95. Can choose 0.90 and 0.80.
path	logical. The whole path of estimation results is out. Default is FALSE.
path_index	numeric. A vector of indices to print out the path. Default is 1.

Value

An object of class "sgdi", which is a list containing the following

coefficient A (p + 1)-vector of estimated parameter values including the intercept.

var A (p+1)x (p+1) variance-covariance matrix of coefficient

ci.lower The lower part of the 95% confidence interval

ci.upper The upper part of the 95% confidence interval

level The confidence level required. Default is 0.95.

path_coefficients The path of coefficients.

Examples

```
n = 1e05
p = 5
bt0 = rep(5,p)
x = matrix(rnorm(n*(p-1)), n, (p-1))
y = cbind(1,x) %*% bt0 + rnorm(n)
my.dat = data.frame(y=y, x=x)
sgdi.out = sgdi_lm(y~., data=my.dat)
```

`sgdi_qr` *Averaged S-subGD and its Inference via Random Scaling in Linear Quantile Regression*

Description

Compute the averaged S-subGD (stochastic subgradient) estimator for the coefficients in linear quantile regression and conduct inference via random scaling method.

Usage

```
sgdi_qr(
  formula,
  data,
  gamma_0 = NULL,
  alpha = 0.501,
  burn = 1,
  inference = "rs",
  bt_start = NULL,
  qt = 0.5,
  studentize = TRUE,
  no_studentize = 100L,
  intercept = TRUE,
  rss_idx = c(1),
  level = 0.95,
  path = FALSE,
  path_index = c(1)
)
```

Arguments

`formula`	formula. The response is on the left of a ~ operator. The terms are on the right of a ~ operator, separated by a + operator.
`data`	an optional data frame containing variables in the model.
`gamma_0`	numeric. A tuning parameter for the learning rate (gamma_0 x t ^ alpha). Default is NULL and it is determined by the adaptive method in Lee et al. (2023).
`alpha`	numeric. A tuning parameter for the learning rate (gamma_0 x t ^ alpha). Default is 0.501.

burn	numeric. A tuning parameter for "burn-in" observations. We burn-in up to (burn-1) observations and use observations from (burn) for estimation. Default is 1, i.e. no burn-in.
inference	character. Specifying the inference method. Default is "rs" (random scaling matrix for joint inference using all the parameters). "rss" is for ransom scaling subset inference. This option requires that "rss_indx" should be provided. "rsd" is for the diagonal elements of the random scaling matrix, excluding one for the intercept term.
bt_start	numeric. (p x 1) vector, excluding the intercept term. User-provided starting value. Default is NULL. Then, it is estimated by conquer.
qt	numeric. Quantile. Default is 0.5.
studentize	logical. Studentize regressors. Default is TRUE.
no_studentize	numeric. The number of observations to compute the mean and std error for studentization. Default is 100.
intercept	logical. Use the intercept term for regressors. Default is TRUE. If this option is TRUE, the first element of the parameter vector is the intercept term.
rss_idx	numeric. Index of x for random scaling subset inference. Default is 1, the first regressor of x. For example, if we want to focus on the 1st and 3rd covariates of x, then set it to be c(1,3).
level	numeric. The confidence level required. Default is 0.95. Can choose 0.90 and 0.80.
path	logical. The whole path of estimation results is out. Default is FALSE.
path_index	numeric. A vector of indices to print out the path. Default is 1.

Value

An object of class "sgdi", which is a list containing the following

coefficients a vector of estimated parameter values

V a random scaling matrix depending on the inference method

ci.lower a vector of lower confidence limits

ci.upper a vector of upper confidence limits

inference character that specifies the inference method

level The confidence level required. Default is 0.95.

path_coefficients The path of coefficients.

Note

The dimension of coefficients is (p+1) if intercept=TRUE or p otherwise. The random scaling matrix V is a full matrix if "rs" is chosen; it is a scalar or smaller matrix, depending on the specification of "rss_indx" if "rss" is selected; it is a vector of diagonal elements of the full matrix if "rsd" is selected. In this case, the first element is missing if the intercept is included. The confidence intervals may contain NA under "rss" and "rsd".

Examples

```
n = 1e05
p = 5
bt0 = rep(5,p)
x = matrix(rnorm(n*(p-1)), n, (p-1))
y = cbind(1,x) %*% bt0 + rnorm(n)
my.dat = data.frame(y=y, x=x)
sgdi.out = sgdi_qr(y~., data=my.dat)
```

---

sgd_lm *Averaged SGD in Linear Mean Regression*

Description

Compute the averaged SGD estimator for the coefficients in linear mean regression.

Usage

```
sgd_lm(
  formula,
  data,
  gamma_0 = NULL,
  alpha = 0.501,
  burn = 1,
  bt_start = NULL,
  studentize = TRUE,
  no_studentize = 100L,
  intercept = TRUE,
  path = FALSE,
  path_index = c(1)
)
```

Arguments

formula	formula. The response is on the left of a ~ operator. The terms are on the right of a ~ operator, separated by a + operator.
data	an optional data frame containing variables in the model.
gamma_0	numeric. A tuning parameter for the learning rate (gamma_0 x t ^ alpha). Default is NULL and it is determined by the adaptive method: 1/sd(y).
alpha	numeric. A tuning parameter for the learning rate (gamma_0 x t ^ alpha). Default is 0.501.
burn	numeric. A tuning parameter for "burn-in" observations. We burn-in up to (burn-1) observations and use observations from (burn) for estimation. Default is 1, i.e. no burn-in.
bt_start	numeric. (p x 1) vector, excluding the intercept term. User-provided starting value. Default is NULL.

studentize	logical. Studentize regressors. Default is TRUE.
no_studentize	numeric. The number of observations to compute the mean and std error for studentization. Default is 100.
intercept	logical. Use the intercept term for regressors. Default is TRUE. If this option is TRUE, the first element of the parameter vector is the intercept term.
path	logical. The whole path of estimation results is out. Default is FALSE.
path_index	numeric. A vector of indices to print out the path. Default is 1.

Value

An object of class "sgdi", which is a list containing the following

coefficients a vector of estimated parameter values

path_coefficients The path of coefficients.

Note

The dimension of coefficients is (p+1) if intercept=TRUE or p otherwise.

Examples

```
n = 1e05
p = 5
bt0 = rep(5,p)
x = matrix(rnorm(n*(p-1)), n, (p-1))
y = cbind(1,x) %*% bt0 + rnorm(n)
my.dat = data.frame(y=y, x=x)
sgd.out = sgd_lm(y~., data=my.dat)
```

---

sgd_qr *Averaged S-subGD Estimator in Linear Quantile Regression*

Description

Compute the averaged S-subGD (stochastic subgradient) estimator for the coefficients in linear quantile regression.

Usage

```
sgd_qr(
  formula,
  data,
  gamma_0 = NULL,
  alpha = 0.501,
  burn = 1,
  bt_start = NULL,
  qt = 0.5,
```

```
  studentize = TRUE,
  no_studentize = 100L,
  intercept = TRUE,
  path = FALSE,
  path_index = c(1)
)
```

Arguments

formula	formula. The response is on the left of a ~ operator. The terms are on the right of a ~ operator, separated by a + operator.
data	an optional data frame containing variables in the model.
gamma_0	numeric. A tuning parameter for the learning rate (gamma_0 x t ^ alpha). Default is NULL and it is determined by the adaptive method in Lee et al. (2023).
alpha	numeric. A tuning parameter for the learning rate (gamma_0 x t ^ alpha). Default is 0.501.
burn	numeric. A tuning parameter for "burn-in" observations. We burn-in up to (burn-1) observations and use observations from (burn) for estimation. Default is 1, i.e. no burn-in.
bt_start	numeric. (p x 1) vector, excluding the intercept term. User-provided starting value. Default is NULL. Then, it is estimated by conquer.
qt	numeric. Quantile. Default is 0.5.
studentize	logical. Studentize regressors. Default is TRUE.
no_studentize	numeric. The number of observations to compute the mean and std error for studentization. Default is 100.
intercept	logical. Use the intercept term for regressors. Default is TRUE. If this option is TRUE, the first element of the parameter vector is the intercept term.
path	logical. The whole path of estimation results is out. Default is FALSE.
path_index	numeric. A vector of indices to print out the path. Default is 1.

Value

An object of class "sgdi", which is a list containing the following

coefficients a vector of estimated parameter values

path_coefficients The path of coefficients.

Note

The dimension of coefficients is (p+1) if intercept=TRUE or p otherwise.

Examples

```
n = 1e05
p = 5
bt0 = rep(5,p)
x = matrix(rnorm(n*(p-1)), n, (p-1))
y = cbind(1,x) %*% bt0 + rnorm(n)
my.dat = data.frame(y=y, x=x)
sgd.out = sgd_qr(y~., data=my.dat)
```

Index