Package 'RenewGLM'

January 21, 2018

Type Package

Title Renewable Estimation and Incremental Inference in Generalized Linear Models with Streaming Datasets

Version 0.1.0

Author Lan Luo and Peter X.-K. Song

Maintainer Lan Luo <luolsph@umich.edu>

Description This package updates the regression coefficients and their standard errors in generalized linear models as data batches arrive sequentially.

License GPL-2

Depends MASS, stats

Encoding UTF-8

RoxygenNote 6.0.1

NeedsCompilation no

R topics documented:

`RenewGLM-package` *Renewable Estimation and Incremental Inference in Generalized Linear Models with Streaming Datasets*

Description

This package updates the regression coefficients and their standard errors in generalized linear models as data batches arrive sequentially.

Details

This package aims to update the regression coefficients as data batches arrive sequentially. There are two main functions in package. `RenewGLM_in` is used to processing a sequence of datasets that are stored in a given directory, and the major input is the name of the directory. `RenewGLM_out` is applied in the case where data batches are imported externally, and the form of the input is `X` and `y` of the current data batch.

Author(s)

Lan Luo and Peter X.-K. Song
Maintainer: Lan Luo <luolsph@umich.edu>

Examples

```
#Processing data batches internally
N=1000
B=10
p=5
n=N/B
beta<-c(0.2,-0.2,0.2,-0.2,0.2)

tempdatadir<-"~/Desktop/tempdata"
datagenerator_in(beta=beta,n=n, p=p, B=B, family="binomial", construct="cs",
    rho=0.5, tempdatadir=tempdatadir)
RenewGLM_in(B, tempdatadir=tempdatadir, "binomial", p=p, intercept=TRUE)
unlink(tempdatadir)

#Processing data batches externally
N=1000
B=10
p=5
n=N/B
beta<-c(0.2,-0.2,0.2,-0.2,0.2)
infomats<-diag(0,p,p);
betahat<-rep(0,p)
for(b in 1:10){
 data<-datagenerator_out(beta,b,n,"binomial","cs",0.5)
 y<-data[,1]
 X<-data[,-1]
 summary<-RenewGLM_out(X,y,"binomial",betahat,infomats,intercept=TRUE,s,phi)
 betahat<-summary[[1]]
 infomats<-summary[[2]]
 rm(data)
}
sd<-sqrt(diag(solve(infomats)));
pvalue<-2*pnorm(-abs(betahat)/sd)
result<-cbind(betahat=betahat,sd=sd,pvalue=pvalue)
colnames(result)<-c("Estimates","Std.Errors","p-values")
```

`datagenerator_in` *This function is used to generate B data batches each time to illustrate the usage of RenewGLM function*

Description

This function is used to generate B data batches each time to illustrate the usage of RenewGLM_in function

Usage

```
datagenerator_in(beta, n, p, B, family, construct, rho, tempdatadir,
  categorical = FALSE, seed = NA)
```

Arguments

beta	designed coefficients for simulated data batches, including the intercept
n	sample size of the bth data batch
p	number of coefficients to estimate including intercept
B	the terminal time to get the result
family	exponential family of the responses, choices include c("gaussian","binomial","poisson")
construct	structure of covariance matrix for generating the covariate matrix X
rho	the correlation coefficient in the covariance matrix, choices inlcude c("ind", "cs", "ar1")
tempdatadir	the directory for saving a total of B data batches
categorical	set to TRUE to let some X be dichotomized, the default is FALSE
seed	set random seed and the default is NA

Value

a data matrix containing response vector y and covariate matrix X

---

datagenerator_out *This function is used to generate one data batch each time to illustrate the usage of RenewGLM function*

Description

This function is used to generate one data batch each time to illustrate the usage of RenewGLM_out function

Usage

```
datagenerator_out(beta, b, n, family, construct, rho, categorical = FALSE,
  seed = NA)
```

Arguments

beta	designed coefficients for simulated data batches, including the intercept
b	index for the current data batch
n	sample size of the bth data batch
family	exponential family of the reponses, including c("gaussian","binomial","poisson")
construct	structure of covariance matrix for generating the covariate matrix X
rho	the correlation coefficient in the covariance matrix, choices inlcude c("ind", "cs", "ar1")
categorical	set to TRUE to let some X be dichotomized, the default is FALSE
seed	set random seed and the default is NA

Value

a data matrix containing response vector y and covariate matrix X

RenewGLM_in *Renewable GLM function processing data batches internally*

Description

Looping over data batches inside the function and output the final regression coefficients and their standard errors

Usage

```
RenewGLM_in(B, tempdatadir, type, init = NA, p, intercept = TRUE)
```

Arguments

B	index for the terminal data batch
tempdatadir	the directory of the streaming datasets, each data batch includes a covariate matrix X and a response vector y
type	the GLM family you want to fit your data to c("gaussian","binomial","poisson")
init	the initial value for regression coefficients (default is a vector of 0)
p	number of coefficients to be estimated, including intercept
intercept	if intercept is included in the model, default is TRUE

Value

coefficient estimates, standard errors and p-values at data batch B

Examples

```
N=1000
B=10
p=5
n=N/B
beta<-c(0.2,-0.2,0.2,-0.2,0.2)

tempdatadir<-"~/Desktop/tempdata"
datagenerator_in(beta=beta, n=n, p=p, B=B, family="binomial", construct="cs",
    rho=0.5, tempdatadir=tempdatadir)
RenewGLM_in(B, tempdatadir=tempdatadir, "binomial", p=p, intercept=TRUE)
unlink(tempdatadir)
```

RenewGLM_out *Renewable GLM function processing data batches externally*

Description

Take in data batches sequentially and update the regression coefficients and their standard errors

Usage

```
RenewGLM_out(X, y, type, betahat, infomats, intercept, s, phi)
```

Arguments

X	covariate matrix for the current data batch
y	response vector for the current data batch
type	the GLM family you want to fit your data to c("gaussian","binomial","poisson")
betahat	the old estimates that need to be updated
infomats	the old cumulative information matrix that need to be updated
intercept	if an intercept is included in the model
s	the cumulative sample size (only needs to be specified in Gaussian model, does not include the samples in the current data batch)
phi	the old estimate of the dispersion parameter in Gaussian model

Value

updated coefficient estimates and the cumulative information matrix

Examples

```
N=1000
B=10
p=5
n=N/B
beta<-c(0.2,-0.2,0.2,-0.2,0.2)
infomats<-diag(0,p,p);
betahat<-rep(0,p)
for(b in 1:10){
 data<-datagenerator_out(beta,b,n,"binomial","cs",0.5)
 y<-data[,1]
 X<-data[,-1]
 summary<-RenewGLM_out(X,y,"binomial",betahat,infomats,intercept=TRUE,s,phi)
 betahat<-summary[[1]]
 infomats<-summary[[2]]
 rm(data)
}
sd<-sqrt(diag(solve(infomats)))
pvalue<-2*pnorm(-abs(betahat)/sd)
result<-cbind(betahat=betahat,sd=sd,pvalue=pvalue)
colnames(result)<-c("Estimates","Std.Errors","p-values")
```

Index