Package 'HeckmanStan'

July 21, 2025

Type Package

Title Heckman Selection Models Based on Bayesian Analysis

Version 1.0.0

Maintainer Heeju Lim <heeju.lim@uconn.edu>

Description Implements Heckman selection models using a Bayesian approach via 'Stan' and compares the performance of normal, Student's t, and contaminated normal distributions in addressing complexities and selection bias (Heeju Lim, Victor E. Lachos, and Victor H. Lachos, Bayesian analysis of flexible Heckman selection models using Hamiltonian Monte Carlo, 2025, under submission).

Imports rstan (>= 2.26.23), mvtnorm (>= 1.2-3), loo, stats

License GPL-3

Encoding UTF-8

LazyData true

RoxygenNote 7.3.2

NeedsCompilation no

Author Heeju Lim [aut, cre],
Victor E. Lachos [aut],
Victor H. Lachos [aut]

Depends R (>= 3.5.0)

Repository CRAN

Date/Publication 2025-05-06 08:50:05 UTC

Contents

`geraHeckman` *Generating Heckman data : Normal, Student-t, Slash and Laplace*

Description

'geraHeckman()' generates a random sample from the Heckman selection model (Normal, Student-t or CN).

Usage

```
geraHeckman(x, w, beta, gamma, sigma2, rho, nu, family = "T")
```

Arguments

`x`	A covariate matrix for the response y.
`w`	A covariate matrix for the missing indicator cc.
`beta`	Values for the beta vector.
`gamma`	Values for the gamma vector.
`sigma2`	Value for the variance.
`rho`	Value for the dependence between the response and missing value.
`nu`	When using the t- distribution, the initial value for the degrees of freedom.
`family`	The distribution family to be used (Normal, T, or CN).

Value

Return an object with the response (y) and missing values (cc).

References

Lachos, V. H., Prates, M. O., & Dey, D. K. (2021). Heckman selection-t model: Parameter estimation via the EM-algorithm. Journal of Multivariate Analysis, 184, 104737.

Examples

```
n <- 100
rho <- .6
cens <- 0.25
nu <- 4
set.seed(20200527)
w <- cbind(1,runif(n,-1,1),rnorm(n))
x <- cbind(w[,1:2])

family <- "T"
c <- qt(cens, df=nu)

sigma2 <- 1
```

```
beta <- c(1,0.5)
gamma<- c(1,0.3,-.5)
gamma[1] <- -c*sqrt(sigma2)

data <- geraHeckman(x,w,beta,gamma,sigma2,rho,nu,family=family)
```

HeckmanStan *Fit the Heckman Selection Stan model using the Normal, Student-t or Contaminated Normal distributions.*

Description

'HeckmanStan()' fits the Heckman selection model using a Bayesian approach to address sample selection bias.

Usage

```
HeckmanStan(
  y,
  x,
  w,
  cc,
  family = "CN",
  init = "random",
  thin = 5,
  chains = 1,
  iter = 10,
  warmup = 5
)
```

Arguments

y	A response vector.
x	A covariate matrix for the response y.
w	A covariate matrix for the missing indicator cc.
cc	A missing indicator vector (1=observed, 0=missing) .
family	The distribution family to be used (Normal, T, or CN).
init	Parameters specifies the initial values for model parameters.
thin	An Interval at which samples are retained from the MCMC process to reduce autocorrelation.
chains	The number of chains to run during the MCMC sampling. Running multiple chains is useful for checking convergence.
iter	The total number of iterations for the MCMC sampling, determining how many samples will be drawn.
warmup	The number of initial iterations that will be discarded as the algorithm stabilizes before collecting samples.

Value

An object of class `HeckmanStan`, which is a list containing two elements:

- `list[[1]]`: Includes inference results from the Stan model, along with EAIC and EBIC.
- `list[[2]]`: Includes the HPC confidence intervals, along with LOOIC, WAIC, and CPO.

Examples

```
################################################################################
# Simulation
################################################################################
library(mvtnorm)
n<- 100
w<- cbind(1,rnorm(n),rnorm(n))
x<- cbind(w[,1:2])
family="CN"
sigma2<- 1
rho<-0.7
beta<- c(1,0.5)
gamma<- c(1,0.3,-.5)
nu=c(0.1,0.1)
data<-geraHeckman(x,w,beta,gamma,sigma2,rho,nu,family=family)
y<-data$y
cc<-data$cc
# Fit Heckman Normal Stan model
fit.n_stan <- HeckmanStan(y, x, w, cc, family="Normal"
                          , thin = 5, chains = 1, iter = 10000, warmup = 1000)
qoi=c("beta","gamma","sigma_e","sigma2", "rho","EAIC","EBIC")
print(fit.n_stan[[1]],par=qoi)
print(fit.n_stan[[2]])

require(rstan)
plot(fit.n_stan[[1]], pars=qoi)
plot(fit.n_stan[[1]], plotfun="hist", pars=qoi)
plot(fit.n_stan[[1]], plotfun="trace", pars=qoi)
plot(fit.n_stan[[1]], plotfun = "rhat")
```

`MEPS2001` *MEPS 2001: Ambulatory Expenditures Data*

Description

This dataset is an extract from the 2001 Medical Expenditure Panel Survey (MEPS), providing information on ambulatory expenditures and various demographic and health-related variables. It has been used for illustrative examples by Cameron and Trivedi (2009, Chapter 16).

Usage

```
data(MEPS2001)
```

Format

A data frame with 3,328 observations on the following 22 variables.

educ Education status

age Age

income Income

female Gender

vgood Self-reported health status, very good

good Self-reported health status, good

hospexp Hospital expenditures

totchr Total number of chronic diseases

ffs Family support

dhospexp Dummy variable for hospital expenditures

age2 Age squared

agefem Interaction between age and gender

fairpoor Self-reported health status, fair or poor

year01 Year of survey

instype Type of insurance

ambexp Ambulatory expenditures

lambexp Log of ambulatory expenditures

blhisp Ethnicity

instype_s1 Insurance type, version 1

dambexp Dummy variable for ambulatory expenditures

lnambx Log-transformed ambulatory expenditures

ins Insurance status

Source

2001 Medical Expenditure Panel Survey by the Agency for Healthcare Research and Quality.

References

Cameron, C.A. and Trivedi, P.K. (2009). *Microeconometrics Using Stata*. College Station, TX: Stata Press.

Examples

```
data(MEPS2001)
head(MEPS2001)
```

PSID1976 *Panel Study of Income Dynamics 1976 Extract*

Description

Cross-section data originating from the 1976 Panel Study of Income Dynamics (PSID). The dataset includes demographic and economic characteristics of married women and their husbands, and is commonly used for analyzing female labor force participation.

Usage

```
data(PSID1976)
```

Format

A data frame with 753 observations on the following 22 variables.

age age of the woman

city dummy for living in a city

college dummy for college education (woman)

education years of education (woman)

experience years of labor market experience

feducation father's years of education

fincome family income in 1,000s

hage husband's age

hcollege dummy for husband's college education

heducation husband's years of education

hhours husband's weekly working hours

hours woman's weekly working hours

hwage husband's log hourly wage

meducation mother's years of education

oldkids number of children older than 6

participation dummy for woman's labor force participation

repwage replacement wage (predicted wage if not employed)

tax marginal tax rate

unemp state unemployment rate

wage log hourly wage of the woman

youngkids number of children 6 or younger

References

Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica*, 55(4), 765–799.

Examples

```
data(PSID1976)
head(PSID1976)
```

Index