

Package ‘ssmodels’

May 30, 2026

Title Sample Selection Models

Version 2.0.2

Language en-US

Author Fernando de Souza Bastos [aut, cre],
Wagner Barreto de Souza [aut]

Maintainer Fernando de Souza Bastos <fernando.bastos@ufv.br>

Depends R (>= 3.6.0)

Imports MASS, sn (>= 2.1.0), miscTools (>= 0.6-26), Rdpack (>= 2.4)

Suggests knitr (>= 1.24), testthat (>= 3.0.0), numDeriv (>= 2016.8-1.1), maxLik (>= 1.3-6), mvtnorm (>= 1.0-11), sampleSelection (>= 1.2-6), kableExtra (>= 1.1.0), kfigr (>= 1.2), ggplot2 (>= 3.2.1), gridExtra (>= 2.3)

Description In order to facilitate the adjustment of the sample selection models existing in the literature, we created the 'ssmodels' package. Our package allows the adjustment of the classic Heckman model (Heckman (1976), Heckman (1979) <[doi:10.2307/1912352](https://doi.org/10.2307/1912352)>), and the estimation of the parameters of this model via the maximum likelihood method and two-step method, in addition to the adjustment of the Heckman-t models introduced in the literature by Marchenko and Genton (2012) <[doi:10.1080/01621459.2012.656011](https://doi.org/10.1080/01621459.2012.656011)> and the Heckman-Skew model introduced in the literature by Ogundimu and Hutton (2016) <[doi:10.1111/sjos.12171](https://doi.org/10.1111/sjos.12171)>. We also implemented functions to adjust the generalized version of the Heckman model, introduced by Bastos, Barreto-Souza, and Genton (2021) <[doi:10.5705/ss.202021.0068](https://doi.org/10.5705/ss.202021.0068)>, that allows the inclusion of covariables to the dispersion and correlation parameters, and a function to adjust the Heckman-BS model introduced by Bastos and Barreto-Souza (2020) <[doi:10.1080/02664763.2020.1780570](https://doi.org/10.1080/02664763.2020.1780570)> that uses the Birnbaum-Saunders distribution as a joint distribution of the selection and primary regression variables. This package extends and complements existing R packages such as 'sampleSelection' (Toomet and Henningsen, 2008) and 'ssmrob' (Zhelonkin et al., 2016), providing additional robust and flexible sample selection models.

License GPL (>= 2)

Encoding UTF-8

LazyData true

RdMacros Rdpack

BugReports <https://github.com/fsbmat-ufv/ssmodels/issues>

Config/testthat/edition 3

URL <https://fsbmat-ufv.github.io/ssmodels/>

NeedsCompilation no

Config/roxygen2/version 8.0.0

Repository CRAN

Date/Publication 2026-05-30 04:20:02 UTC

Contents

extract_model_components	2
HCinitial	3
HeckmanBS	4
HeckmanCL	5
HeckmanGe	7
HeckmanSK	7
HeckmantS	9
IMR	10
MEPS2001	11
Mroz87	12
nhanes	14
PSID2	15
RandHIE	17
step2	19
summary.HeckmanBS	20
summary.HeckmanCL	20
summary.HeckmanGe	21
summary.HeckmanSK	21
summary.HeckmantS	22
twostep	22
Index	24

extract_model_components

Extract model components from formulas

Description

This function extracts model matrices and response vectors used by the sample-selection model fitting functions. It standardizes the selection response to a numeric indicator, where 1 denotes selected/observed and 0 denotes not selected/not observed.

Usage

```
extract_model_components(
  selection,
  outcome,
  data,
  outcomeS = NULL,
  outcomeC = NULL,
  drop.levels = TRUE
)
```

Arguments

selection	A formula representing the selection equation.
outcome	A formula representing the outcome equation.
data	A data frame containing all variables used in the formulas.
outcomeS	Optional matrix, vector, one-sided formula, or 1 for the log-scale model in the generalized Heckman model.
outcomeC	Optional matrix, vector, one-sided formula, or 1 for the correlation model in the generalized Heckman model.
drop.levels	Logical; if TRUE, unused factor levels are dropped.

Value

A list with selection, outcome, dispersion and correlation matrices.

 HCinitial

Two-Step Method for Parameter Estimation of the Heckman Model

Description

Estimates classical Heckman starting values using Heckman's two-step method.

Usage

```
HCinitial(selection, outcome, data = sys.frame(sys.parent()))
```

Arguments

selection	A formula for the selection equation.
outcome	A formula for the outcome equation.
data	A data frame containing the variables.

Value

A numeric vector containing selection coefficients, outcome coefficients, sigma and rho.

 HeckmanBS

Heckman BS Model fit Function

Description

Estimates the parameters of the Heckman-BS model

Usage

```
HeckmanBS(selection, outcome, data = sys.frame(sys.parent()), start = NULL)
```

Arguments

selection	Selection equation.
outcome	Primary Regression Equation.
data	Database.
start	initial values.

Details

The HeckmanBS() function fits the Sample Selection Model based on the Birnbaum Saunders bivariate distribution, it has the same number of parameters as the classical Heckman model. For more information see Bastos and Barreto-Souza (2020) [doi:10.1080/02664763.2020.1780570](https://doi.org/10.1080/02664763.2020.1780570).

Value

Returns a list with the following components.

Coefficients: Returns a numerical vector with the best estimated values of the model parameters;

Value: The value of function to be minimized (or maximized) corresponding to par.

loglik: Maximized value of the log-likelihood function calculated from the estimated coefficients.

counts: Component of the Optim function. A two-element integer vector giving the number of calls to fn and gr respectively. This excludes those calls needed to compute the Hessian, if requested, and any calls to fn to compute a finite-difference approximation to the gradient.

hessian: Component of the Optim function, with pre-defined option hessian=TRUE. A symmetric matrix giving an estimate of the Hessian at the solution found. Note that this is the Hessian of the unconstrained problem even if the box constraints are active.

fisher_infoBS: Fisher information matrix

prop_sigmaBS: Square root of the Fisher information matrix diagonal

coefficients_link: Estimates on the optimization scale. The last component is rho_star, where $\rho = 2 / (1 + \exp(-\rho_star)) - 1$.

gradient_link: Analytical score vector evaluated at coefficients_link.

level: Selection variable levels

nObs: Numeric value representing the size of the database

nParam: Numerical value representing the number of model parameters
 N0: Numerical value representing the number of unobserved entries
 N1: Numerical value representing the number of complete entries
 NXS: Numerical value representing the number of parameters of the selection model
 NXO: Numerical value representing the number of parameters of the regression model
 df: Numerical value that represents the difference between the size of the response vector of the selection equation and the number of model parameters
 aic: Numerical value representing Akaike's information criterion.
 bic: Numerical value representing Schwarz's Bayesian Criterion
 initial.value: Numerical vector that represents the input values (Initial Values) used in the parameter estimation.

References

Bastos, F. S. and Barreto-Souza, W. (2020). Birnbaum-Saunders sample selection model. *Journal of Applied Statistics*. doi:[10.1080/02664763.2020.1780570](https://doi.org/10.1080/02664763.2020.1780570).

Examples

```
data(MEPS2001)
attach(MEPS2001)
selectEq <- dambexp ~ age + female + educ + blhisp + totchr + ins + income
outcomeBS <- ambexp ~ age + female + educ + blhisp + totchr + ins
HeckmanBS(selectEq, outcomeBS, data = MEPS2001)
```

HeckmanCL

Classic Heckman Model fit Function

Description

Estimates the parameters of the classic Heckman model via Maximum Likelihood method. The initial start is obtained via the two-step method.

Usage

```
HeckmanCL(selection, outcome, data = sys.frame(sys.parent()), start = NULL)
```

Arguments

selection	Selection equation.
outcome	Primary Regression Equation.
data	Database.
start	initial values.

Value

Returns a list with the following components.

Coefficients: Returns a numerical vector with the best estimated values of the model parameters;

Value: The value of function to be minimized (or maximized) corresponding to par.

loglik: Maximized value of the log-likelihood function calculated from the estimated coefficients.

counts: Component of the Optim function. A two-element integer vector giving the number of calls to fn and gr respectively. This excludes those calls needed to compute the Hessian, if requested, and any calls to fn to compute a finite-difference approximation to the gradient.

hessian: Component of the Optim function, with pre-defined option hessian=TRUE. A symmetric matrix giving an estimate of the Hessian at the solution found. Note that this is the Hessian of the unconstrained problem even if the box constraints are active.

fisher_infoHC: Fisher information matrix

prop_sigmaHC: Square root of the Fisher information matrix diagonal

level: Selection variable levels

nObs: Numeric value representing the size of the database

nParam: Numerical value representing the number of model parameters

N0: Numerical value representing the number of unobserved entries

N1: Numerical value representing the number of complete entries

NXS: Numerical value representing the number of parameters of the selection model

NXO: Numerical value representing the number of parameters of the regression model

df: Numerical value that represents the difference between the size of the response vector of the selection equation and the number of model parameters

aic: Numerical value representing Akaike's information criterion.

bic: Numerical value representing Schwarz's Bayesian Criterion

initial.value: Numerical vector that represents the input values (Initial Values) used in the parameter estimation.

Examples

```
data(MEPS2001)
attach(MEPS2001)
selectEq <- dambexp ~ age + female + educ + blhisp + totchr + ins + income
outcomeEq <- lnambx ~ age + female + educ + blhisp + totchr + ins
HeckmanCL(selectEq, outcomeEq, data = MEPS2001)
```

HeckmanGe

Generalized Heckman Model Estimation

Description

Estimates the parameters of a generalized Heckman selection model, allowing covariates in the scale and correlation components.

Usage

```
HeckmanGe(
  selection,
  outcome,
  outcomeS = 1,
  outcomeC = 1,
  data = sys.frame(sys.parent()),
  start = NULL
)
```

Arguments

selection	Formula for the selection equation.
outcome	Formula for the outcome equation.
outcomeS	Covariates for the log-scale model. It may be a matrix, vector, one-sided formula, or 1 for an intercept-only scale model.
outcomeC	Covariates for the correlation model. It may be a matrix, vector, one-sided formula, or 1 for an intercept-only correlation model.
data	Data frame containing all variables.
start	Optional starting values. If supplied on the natural classical scale, sigma and rho are internally mapped to $\log(\sigma)$ and $\operatorname{atanh}(\rho)$.

Value

A list with the fitted model parameters and diagnostics.

HeckmanSK

Normal Skew Model fit Function

Description

Estimates the parameters of the Sample Selection Model with Skew-Normal Distribution

Usage

```
HeckmanSK(
  selection,
  outcome,
  data = sys.frame(sys.parent()),
  lambda = c(-5, -3, -2, -1, -0.5, 0, 0.5, 1, 2, 3, 5),
  start = NULL
)
```

Arguments

selection	Selection equation.
outcome	Primary Regression Equation.
data	Database.
lambda	Numeric scalar or vector with initial values for the skewness parameter. If more than one value is supplied and start = NULL, the function uses a multi-start strategy and keeps the fit with the largest log-likelihood.
start	initial values.

Details

The HeckmanSK() function fits the Sample Selection Model based on the Skew-normal distribution. For more information see Ogundimu and Hutton (2016) [doi:10.1111/sjos.12171](https://doi.org/10.1111/sjos.12171).

Value

Returns a list with the following components.

Coefficients: Returns a numerical vector with the best estimated values of the model parameters;

Value: The value of function to be minimized (or maximized) corresponding to par.

loglik: Maximized value of the log-likelihood function calculated from the estimated coefficients.

gradient: Analytical score vector evaluated at the estimated coefficients.

counts: Component of the Optim function. A two-element integer vector giving the number of calls to fn and gr respectively. This excludes those calls needed to compute the Hessian, if requested, and any calls to fn to compute a finite-difference approximation to the gradient.

hessian: Component of the Optim function, with pre-defined option hessian=TRUE. A symmetric matrix giving an estimate of the Hessian at the solution found. Note that this is the Hessian of the unconstrained problem even if the box constraints are active.

fisher_infoSK: Fisher information matrix

prop_sigmaSK: Square root of the Fisher information matrix diagonal

level: Selection variable levels

nObs: Numeric value representing the size of the database

nParam: Numerical value representing the number of model parameters

N0: Numerical value representing the number of unobserved entries

NI: Numerical value representing the number of complete entries
 NXS: Numerical value representing the number of parameters of the selection model
 NXO: Numerical value representing the number of parameters of the regression model
 df: Numerical value that represents the difference between the size of the response vector of the selection equation and the number of model parameters
 aic: Numerical value representing Akaike's information criterion.
 bic: Numerical value representing Schwarz's Bayesian Criterion
 initial.value: Numerical vector that represents the input values (Initial Values) used in the parameter estimation.

References

Ogundimu, E. O. and Hutton, J. L. (2016). A sample selection model with skew-normal distribution. *Scandinavian Journal of Statistics*. doi:10.1111/sjos.12171.

Examples

```
data("Mroz87")
attach(Mroz87)
selectEq <- lfp ~ huswage + kids5 + mtr + fatheduc + educ + city
outcomeEq <- log(wage) ~ educ+city
HeckmanSK(selectEq, outcomeEq, data = Mroz87, lambda = -1.5)
```

 HeckmantS

Heckman-t Model fit Function

Description

Estimates the parameters of the Heckman-t model. The scale, correlation and degrees of freedom are optimized on unconstrained scales to improve numerical stability: $\sigma = \exp(\eta_{\sigma})$, $\rho = \tanh(\eta_{\rho})$, and $df = 2 + \exp(\eta_{df})$.

Usage

```
HeckmantS(selection, outcome, data = sys.frame(sys.parent()), df, start = NULL)
```

Arguments

selection	Selection equation.
outcome	Primary regression equation.
data	Database.
df	Initial value for the degrees of freedom. Values greater than 2 are recommended because the model uses a finite-variance Student-t error.
start	Optional starting values on the natural scale (selection, outcome, sigma, rho, df).

Value

A list with estimated coefficients, log-likelihood and diagnostics.

References

Marchenko, Y. V. and Genton, M. G. (2012). A Heckman selection-t model. *Journal of the American Statistical Association*, 107(497), 304-317.

Examples

```
data(MEPS2001)
dados <- MEPS2001[seq_len(500), ]
selectEq <- dambexp ~ age + female + educ + blhisp + totchr + ins + income
outcomeEq <- lnambx ~ age + female + educ + blhisp + totchr + ins
HeckmantS(selectEq, outcomeEq, data = dados, df = 12)
```

IMR

Inverse Mills Ratio

Description

Computes the column vector of the Inverse Mills Ratio (IMR) from a Probit selection equation.

Usage

```
IMR(selection, data = sys.frame(sys.parent()))
```

Arguments

selection	A formula for the selection equation.
data	A data frame containing the variables.

Value

A column vector (matrix with one column) containing the Inverse Mills Ratio computed from the Probit model fitted to the selection equation.

Examples

```
data(MEPS2001)
attach(MEPS2001)
selectEq <- dambexp ~ age + female + educ + blhisp + totchr + ins + income
IMR(selectEq, data = MEPS2001)
```

Description

The MEPS is a set of large-scale surveys of families, individuals and their medical providers (doctors, hospitals, pharmacies, etc.) in the United States. It has data on the health services Americans use, how often they use them, the cost of these services and how they are paid, as well as data on the cost and reach of health insurance available to American workers. The sample is restricted to persons aged between 21 and 64 years and contains a variable response with 3328 observations of outpatient costs, of which 526 (15.8%) correspond to unobserved expenditure values and identified as zero expenditure for adjustment of the models. It also includes the following explanatory variables:

- educ: education status
- age: Age
- income: income
- female: gender
- vgood: a numeric vector
- good: a numeric vector
- hospexp: a numeric vector
- totchr: number of chronic diseases
- ffs: a numeric vector
- dhospexp: a numeric vector
- age2: a numeric vector
- agefem: a numeric vector
- fairpoor: a numeric vector
- year01: a numeric vector
- instype: a numeric vector
- ambexp: a numeric vector
- lambexp: log ambulatory expenditures
- blhisp: ethnicity
- instype_s1: a numeric vector
- dambexp: dummy variable, ambulatory expenditures
- lnambx: a numeric vector
- ins: insurance status

Usage

MEPS2001

Format

An object of class `data.frame` with 3328 rows and 22 columns.

Source

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

References

- Cameron, A. C. and Trivedi, P. K. (2009). *Microeconometrics using Stata*. Stata Press.
- Zhelonkin, M., Genton, M. G. and Ronchetti, E. (2019). *ssmrob: Robust Estimation and Inference in Sample Selection Models*. R package version 0.7. <https://CRAN.R-project.org/package=ssmrob>
- Toomet, O. and Henningsen, A. (2008). Sample selection models in R: package `sampleSelection`. *Journal of Statistical Software*, 27(7). <https://www.jstatsoft.org/article/view/v027i07>

Examples

```
data(MEPS2001)
attach(MEPS2001)
hist(lnambx)
selectEq <- dambexp ~ age + female + educ + blhisp + totchr + ins + income
outcomeEq <- lnambx ~ age + female + educ + blhisp + totchr + ins
HeckmanCL(selectEq, outcomeEq, data = MEPS2001)
```

Mroz87

U.S. Women's Labor Force Participation

Description

The Mroz87 data frame contains data about 753 married women. These data are collected within the "Panel Study of Income Dynamics" (PSID). Of the 753 observations, the first 428 are for women with positive hours worked in 1975, while the remaining 325 observations are for women who did not work for pay in 1975. A more complete discussion of the data is given by Mroz (1987). It also includes the following explanatory variables:

- `lfp`: Dummy variable for labor-force participation.
- `hours`: Wife's hours of work in 1975.
- `kids5`: Number of children 5 years old or younger.
- `kids618`: Number of children 6 to 18 years old.
- `Age`: Wife's age.
- `Educ`: Wife's educational attainment, in years.
- `wage`: Wife's average hourly earnings, in 1975 dollars.
- `repwage`: Wife's wage reported at the time of the 1976 interview.
- `hushrs`: Husband's hours worked in 1975.

- `husage`: Husband's age.
- `huseduc`: Husband's educational attainment, in years.
- `huswage`: Husband's wage, in 1975 dollars.
- `faminc`: Family income, in 1975 dollars.
- `mtr`: Marginal tax rate facing the wife.
- `motheduc`: Wife's mother's educational attainment, in years.
- `fatheduc`: Wife's father's educational attainment, in years.
- `unem`: Unemployment rate in county of residence, in percentage points.
- `city`: Dummy variable = 1 if live in large city, else 0.
- `exper`: Actual years of wife's previous labor market experience.
- `nwifeinc`: Non-wife income.
- `wifecoll`: Dummy variable for wife's college attendance.
- `huscoll`: Dummy variable for husband's college attendance.

Usage

`Mroz87`

Format

An object of class `data.frame` with 753 rows and 22 columns.

Source

PSID Staff, The Panel Study of Income Dynamics, Institute for Social Research Panel Study of Income Dynamics, University of Michigan, <https://www.icpsr.umich.edu/web/ICPSR/series/131>

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, 765-799.

Zhelonkin, M., Genton, M. G. and Ronchetti, E. (2019). *ssmrob: Robust Estimation and Inference in Sample Selection Models*. R package version 0.7. <https://CRAN.R-project.org/package=ssmrob>

Toomet, O. and Henningsen, A. (2008). Sample selection models in R: package `sampleSelection`. *Journal of Statistical Software*, 27(7). <https://www.jstatsoft.org/article/view/v027i07>

Wooldridge, J. M. (2016). *Introductory Econometrics: A Modern Approach*. Nelson Education.

Examples

```
# Wooldridge(2016): page 247
data(Mroz87)
attach(Mroz87)
selectEq <- lfp ~ nwifeinc + educ + exper + I(exper^2) + age + kids5 + kids618
outcomeEq <- log(wage) ~ educ + exper + I(exper^2)
outcomeS <- cbind(educ, exper)
outcomeC <- cbind(educ, exper)
outcomeBS <- wage ~ educ + exper + I(exper^2)
outcomeBS <- wage ~ educ + exper + I(exper^2)
HeckmanCL(selectEq, outcomeEq, data = Mroz87)
HeckmanBS(selectEq, outcomeBS, data = Mroz87)
HeckmanSK(selectEq, outcomeEq, data = Mroz87, lambda = 1)
HeckmanTS(selectEq, outcomeEq, data = Mroz87, df=5)
HeckmanGe(selectEq, outcomeEq, outcomeS, outcomeC, data = Mroz87)
```

nhanes

US National Health and Nutrition Examination Study

Description

The US National Health and Nutrition Examination Study (NHANES) is a survey data collected by the US National Center for Health Statistics. The survey data dates back to 1999, where individuals of all ages are interviewed in their home annually and complete the health examination component of the survey. The study variables include demographic variables (e.g. age and annual household income), physical measurements (e.g. BMI – body mass index), health variables (e.g. diabetes status), and lifestyle variables (e.g. smoking status). This data frame contains the following columns:

- id: Individual identifier
- age: Age
- gender: Sex 1=male, 0=female
- educ: Education is dichotomized into high school and above versus less than high school
- race: categorical variable with five levels
- income: Household income (\$1000 per year) was reported as a range of values in dollar (e.g. 0–4999, 5000–9999, etc.) and had 10 interval categories.
- Income: Household income (\$1000 per year) was reported as a range of values in dollar (e.g. 0–4999, 5000–9999, etc.) and had 10 interval categories.
- bmi: body mass index
- sbp: systolic blood pressure

Usage

nhanes

Format

An object of class `data.frame` with 9643 rows and 9 columns.

Source

<https://www.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2003>

References

Ogundimu, E. O. and Collins, G. S. (2019). A robust imputation method for missing responses and covariates in sample selection models. *Statistical Methods in Medical Research*, 28(1), 102-116.

Little, R. J. and Zhang, N. (2011). Subsample ignorable likelihood for regression analysis with missing data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(4), 591-605.

Zhelonkin, M., Genton, M. G. and Ronchetti, E. (2019). *ssmrob: Robust Estimation and Inference in Sample Selection Models*. R package version 0.7. <https://CRAN.R-project.org/package=ssmrob>

Toomet, O. and Henningsen, A. (2008). Sample selection models in R: package `sampleSelection`. *Journal of Statistical Software*, 27(7). <https://www.jstatsoft.org/article/view/v027i07>

Examples

```
data("nhanes")
attach(nhanes)
hist(Income, prob= TRUE, breaks = seq(1, 99, 0.5), xlim = c(1,10),
ylim = c(0,0.35), main = "Histogram of Income", xlab = "Category")
data2 <- subset(nhanes, !is.na(sbp))
data3 <- subset(data2, !is.na(bmi))
attach(data3)
data <- data3
data$YS <- ifelse(is.na(data$Income),0,1)
data$educ <- ifelse(data$educ<=2,0,1)
attach(data)
selectionEq <- YS~age+gender+educ+race
outcomeEq <- sbp~age+gender+educ+bmi
```

 PSID2

Panel Study of Income Dynamics

Description

The data come from the Panel Study of Income Dynamics, years 1981 to 1992 (also contains earnings data from 1980). The sample consists of 579 white females, who were followed over the considered period. In total, there are 6,948 observations over the 12-year period (1981-1992). This data frame contains the following columns:

- `id`: Individual identifier

- year: Survey year
- age: Calculated age in years (based on year and month of birth)
- educ: Years of schooling
- children: Total number of children in family unit, ages 0-17
- s: Participation dummy, =1 if worked (hours>0)
- lnw: Log of real average hourly earnings
- lnw80: Log earnings in 1980
- agesq: Age squared
- children_lag1: Number of children in t-1
- children_lag2: Number of children in t-2
- lnw2: Log of real average hourly earnings
- Lnw: Log of real average hourly earnings

Usage

PSID2

Format

An object of class `data.frame` with 6948 rows and 13 columns.

Source

<https://www.icpsr.umich.edu/web/ICPSR/series/131>

References

Semykina, A. and Wooldridge, J. M. (2013). Estimation of dynamic panel data models with sample selection. *Journal of Applied Econometrics*, 28(1), 47-61.

Zhelonkin, M., Genton, M. G. and Ronchetti, E. (2019). *ssmrob: Robust Estimation and Inference in Sample Selection Models*. R package version 0.7. <https://CRAN.R-project.org/package=ssmrob>

Toomet, O. and Henningsen, A. (2008). Sample selection models in R: package `sampleSelection`. *Journal of Statistical Software*, 27(7). <https://www.jstatsoft.org/article/view/v027i07>

Examples

```
data(PSID2)
attach(PSID2)
hist(Lnw)
selectEq <- s ~ educ+ age+ children+ year
outcomeEq <- Lnw ~ educ+ age+ children
HCinitial(selectEq,outcomeEq, data = PSID2)
#Note that the estimated value of rho by the two-step
#method is greater than 1
summary(HeckmanGe(selectEq,outcomeEq, 1, 1, data = PSID2))
```

Description

The RAND Health Insurance Experiment (RAND HIE) was a comprehensive study of health care cost, utilization and outcome in the United States. It is the only randomized study of health insurance, and the only study which can give definitive evidence as to the causal effects of different health insurance plans. For more information about the database visit: https://en.wikipedia.org/w/index.php?title=RAND_Health_Insurance_Experiment&oldid=110166949 accessed september 09, 2019). This data frame contains the following columns:

- plan: HIE plan number.
- site: Participant's place of residence when the participant was initially enrolled.
- coins: Coinsurance rate.
- tookphys: Took baseline physical.
- year: Study year.
- zper: Person identifier.
- black: 1 if race of household head is black.
- income: Family income.
- xage: Age in years.
- female: 1 if person is female.
- educdec: Education of household head in years.
- time: Time eligible during the year.
- outpdol: Outpatient expenses: all covered outpatient medical services excluding dental care, outpatient psychotherapy, outpatient drugs or supplies.
- drugdol: Drug expenses: all covered outpatient and dental drugs.
- suppdol: Supply expenses: all covered outpatient supplies including dental.
- mentdol: Psychotherapy expenses: all covered outpatient psychotherapy services including injections excluding charges for visits in excess of 52 per year, prescription drugs, and inpatient care.
- inpdol: Inpatient expenses: all covered inpatient expenses in a hospital, mental hospital, or nursing home, excluding outpatient care and renal dialysis.
- meddol: Medical expenses: all covered inpatient and outpatient services, including drugs, supplies, and inpatient costs of newborns excluding dental care and outpatient psychotherapy.
- totadm: Hospital admissions: annual number of covered hospitalizations.
- inpmis: Incomplete Hospital Records: missing inpatient records.
- mentvis: Psychotherapy visits: indicates the annual number of outpatient visits for psychotherapy. It includes billed visits only. The limit was 52 covered visits per person per year. The count includes an initial visit to a psychiatrist or psychologist.

- mdvis: Face-to-Face visits to physicians: annual covered outpatient visits with physician providers (excludes dental, psychotherapy, and radiology/anesthesiology/pathology-only visits).
- notmdvis: Face-to-Face visits to nonphysicians: annual covered outpatient visits with non-physician providers such as speech and physical therapists, chiropractors, podiatrists, acupuncturists, Christian Science etc. (excludes dental, healers, psychotherapy, and radiology/anesthesiology/pathology-only visits).
- num: Family size.
- mhi: Mental health index.
- disea: Number of chronic diseases.
- physlm: Physical limitations.
- ghindx: General health index.
- mdeoff: Maximum expenditure offer.
- pioff: Participation incentive payment.
- child: 1 if age is less than 18 years.
- fchild: female * child.
- lfam: log of num (family size).
- lpi: log of pioff (participation incentive payment).
- idp: 1 if individual deductible plan.
- logc: $\log(\text{coins}+1)$.
- fmde: 0 if $\text{idp}=1$, $\ln(\max(1, \text{mdeoff}/(0.01*\text{coins})))$ otherwise.
- hlthg: 1 if self-rated health is good – baseline is excellent self-rated health.
- hlthf: 1 if self-rated health is fair – baseline is excellent self-rated health.
- hlthp: 1 if self-rated health is poor – baseline is excellent self-rated health.
- xghindx: ghindx (general health index) with imputations of missing values.
- linc: log of income (family income).
- lnum: log of num (family size).
- lnmeddol: log of meddol (medical expenses).
- binexp: 1 if meddol > 0.

Usage

RandHIE

Format

An object of class `data.frame` with 20190 rows and 45 columns.

Source

<https://cameron.econ.ucdavis.edu/mmabook/mmadata.html>

References

- Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.
- Zhelonkin, M., Genton, M. G. and Ronchetti, E. (2019). *ssmrob: Robust Estimation and Inference in Sample Selection Models*. R package version 0.7. <https://CRAN.R-project.org/package=ssmrob>
- Toomet, O. and Henningsen, A. (2008). Sample selection models in R: package sampleSelection. *Journal of Statistical Software*, 27(7). <https://www.jstatsoft.org/article/view/v027i07>
- Wikipedia contributors (2019). RAND Health Insurance Experiment. https://en.wikipedia.org/wiki/RAND_Health_Insurance_Experiment

Examples

```
##Cameron and Trivedi (2005): Section 16.6
data(RandHIE)
subsample <- RandHIE$year == 2 & !is.na( RandHIE$educdec )
selectEq <- binexp ~ logc + idp + lpi + fmde + physlm + disea +
  hlthg + hlthf + hlthp + linc + lfam + educdec + xage + female +
  child + fchild + black
outcomeEq <- lnmeddol ~ logc + idp + lpi + fmde + physlm + disea +
  hlthg + hlthf + hlthp + linc + lfam + educdec + xage + female +
  child + fchild + black
cameron <- HeckmanCL(selectEq, outcomeEq, data = RandHIE[subsample, ])
summary(cameron)
```

step2

Heckman's two-step method

Description

Estimate classical Heckman model starting values by Heckman's two-step method.

Usage

```
step2(YS, XS, YO, XO)
```

Arguments

YS	Selection vector coded as 0/1.
XS	Selection matrix.
YO	Outcome vector.
XO	Covariate matrix for the outcome equation.

Value

A numeric vector with selection coefficients, outcome coefficients, sigma and rho.

summary.HeckmanBS *Summary of Birnbaum-Saunders Heckman Model*

Description

Print estimates of the parameters of the Heckman-BS model using Maximum Likelihood Estimation.

Usage

```
## S3 method for class 'HeckmanBS'
summary(object, ...)
```

Arguments

object An object of class HeckmanBS.
 ... Additional arguments (currently unused).

Value

Prints the summary output including coefficient tables and model fit statistics.

summary.HeckmanCL *Summary of Classic Heckman Model*

Description

Print estimates of the parameters of the Classic Heckman model using Maximum Likelihood Estimation.

Usage

```
## S3 method for class 'HeckmanCL'
summary(object, ...)
```

Arguments

object An object of class HeckmanCL.
 ... Additional arguments (currently unused).

Value

Prints the summary output including coefficient tables and model fit statistics.

summary.HeckmanGe *Summary of Generalized Heckman Model*

Description

Print estimates of the parameters of the Generalized Heckman model using Maximum Likelihood Estimation.

Usage

```
## S3 method for class 'HeckmanGe'  
summary(object, ...)
```

Arguments

object An object of class HeckmanGe.
... Additional arguments (currently unused).

Value

Prints the summary output including coefficient tables and model fit statistics.

summary.HeckmanSK *Summary of Skew-Normal Heckman Model*

Description

Print estimates of the parameters of the Heckman-SK model using Maximum Likelihood Estimation.

Usage

```
## S3 method for class 'HeckmanSK'  
summary(object, ...)
```

Arguments

object An object of class HeckmanSK.
... Additional arguments (currently unused).

Value

Prints the summary output including coefficient tables and model fit statistics.

summary.HeckmantS	<i>Summary of Heckman-t Model</i>
-------------------	-----------------------------------

Description

Print estimates of the parameters of the Heckman-t model using Maximum Likelihood Estimation.

Usage

```
## S3 method for class 'HeckmantS'
summary(object, ...)
```

Arguments

object	An object of class HeckmantS.
...	Additional arguments (currently unused).

Value

Prints the summary output including coefficient tables and model fit statistics.

twostep	<i>Heckman's two-step method</i>
---------	----------------------------------

Description

Estimate the classical Heckman model parameters by the two-step method.

Usage

```
twostep(selection, outcome, data = sys.frame(sys.parent()))
```

Arguments

selection	A formula for the selection equation.
outcome	A formula for the outcome equation.
data	A data frame containing the variables used in the model.

Value

A numeric vector with the two-step estimates.

References

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5(4), 475-492.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153-161.

Index

* **RandHIE**

RandHIE, [17](#)

* **datasets**

MEPS2001, [11](#)

Mroz87, [12](#)

nhanes, [14](#)

PSID2, [15](#)

extract_model_components, [2](#)

HCinitial, [3](#)

HeckmanBS, [4](#)

HeckmanCL, [5](#)

HeckmanGe, [7](#)

HeckmanSK, [7](#)

HeckmantS, [9](#)

IMR, [10](#)

MEPS2001, [11](#)

Mroz87, [12](#)

nhanes, [14](#)

PSID2, [15](#)

RandHIE, [17](#)

step2, [19](#)

summary.HeckmanBS, [20](#)

summary.HeckmanCL, [20](#)

summary.HeckmanGe, [21](#)

summary.HeckmanSK, [21](#)

summary.HeckmantS, [22](#)

twostep, [22](#)