Package ‘ssmodels’

May 9, 2026

Title Sample Selection Models

Version 2.0.1

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 sn (>= 2.1.0), numDeriv (>= 2016.8-1.1), pracma (>= 2.3.8), miscTools (>= 0.6-26), Rdpack (>= 2.4)

Suggests knitr (>= 1.24), testthat (>= 3.0.0), 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>), 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> and the Heckman-Skew model introduced in the literature by Ogundimu and Hutton (2016) <doi: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>, 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> 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

VignetteBuilder knitr

RoxygenNote 7.3.2

RdMacros Rdpack

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

Config/testthat/edition 3

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

NeedsCompilation no

Repository CRAN

Date/Publication 2025-06-02 12:50:01 UTC

Contents

HCinitial 2
HeckmanBS 4
HeckmanCL 5
HeckmanGe 7
HeckmanSK 9
HeckmantS 11
IMR 12
MEPS2001 13
Mroz87 14
nhanes 16
PSID2 17
RandHIE 19
ssmodels 21
step2 23
summary.HeckmanBS 24
summary.HeckmanCL 25
summary.HeckmanGe 26
summary.HeckmanSK 28
summary.HeckmantS 29
twostep 30

Index **32**

`HCinitial` *Two-Step Method for Parameter Estimation of the Classical Heckman Model*

Description

Estimates the parameters of the classical Heckman sample selection model using the two-step estimation method.

Usage

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

Arguments

`selection`	A formula specifying the selection equation.
`outcome`	A formula specifying the outcome equation.
`data`	A data frame containing the variables in the model.

Details

This function implements the two-step approach proposed by Heckman (1979) to estimate the parameters of the classic sample selection model. It is particularly useful for obtaining initial values for maximum likelihood estimation (MLE).

In the first step, a probit model is fitted to the selection equation to estimate the probability of selection. The second step involves estimating a linear regression of the outcome equation for the observed (selected) data, incorporating the inverse Mills ratio (IMR) as an additional regressor to correct for sample selection bias.

The function also estimates:

- `sigma`: The standard deviation of the outcome equation's error term.
- `rho`: The correlation coefficient between the errors of the selection and outcome equations.

Value

A named numeric vector containing:

- Coefficients from the selection equation (probit model),
- Coefficients from the outcome equation (excluding the IMR),
- Estimated `sigma`,
- Estimated `rho`.

References

James J Heckman (1979). "Sample selection bias as a specification error." *Econometrica: Journal of the econometric society*, 153–161.

Examples

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

HeckmanBS *Heckman-BS Model Fit Function*

Description

Fits the Heckman Sample Selection Model based on the Birnbaum-Saunders (BS) bivariate distribution. This function implements the maximum likelihood estimation of the model parameters.

Usage

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

Arguments

`selection`	A formula object specifying the selection equation.
`outcome`	A formula object specifying the primary outcome equation.
`data`	A data frame containing the variables in the model.
`start`	An optional numeric vector of initial parameter values. If not provided, default values are used.

Details

The function estimates the parameters of the Heckman-BS model, which extends the classical Heckman model by assuming that the error terms follow a bivariate Birnbaum-Saunders distribution. The model has the same number of parameters as the classical Heckman model, including the correlation coefficient between the error terms. The optimization is performed using the `optim` function with the BFGS method.

The estimated quantities include:

- Coefficients of the selection equation.
- Coefficients of the outcome equation.
- Estimated `sigma` (scale parameter of the outcome equation's error term).
- Estimated `rho` (correlation coefficient between the error terms).

Additional outputs include measures of model fit, standard errors (approximated by the square root of the diagonal of the inverse Fisher information matrix), and diagnostic information.

Value

A list containing:

- `coefficients`: A named numeric vector of estimated model parameters.
- `value`: The value of the likelihood function at the optimum.
- `loglik`: The (negative) maximum log-likelihood.
- `counts`: Number of gradient evaluations performed.

- `hessian`: The Hessian matrix at the optimum.
- `fisher_infoBS`: The (approximate) Fisher information matrix.
- `prop_sigmaBS`: Approximate standard errors (square root of the Fisher information diagonal).
- `level`: Levels of the selection variable.
- `nObs`: Number of observations in the dataset.
- `nParam`: Number of parameters estimated.
- `N0`: Number of observations where the selection variable is zero.
- `N1`: Number of observations where the selection variable is one.
- `NXS`: Number of parameters in the selection equation.
- `NXO`: Number of parameters in the outcome equation.
- `df`: Degrees of freedom (observations minus number of parameters).
- `aic`: Akaike Information Criterion.
- `bic`: Bayesian Information Criterion.
- `initial.value`: Initial values used in the optimization.

References

There are no references for Rd macro `\insertAllCites` on this help page.

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

Fits the classical Heckman sample selection model using Maximum Likelihood Estimation (MLE). Initial parameter estimates are obtained via the two-step method.

Usage

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

Arguments

`selection`	A formula specifying the selection equation.
`outcome`	A formula specifying the primary outcome equation.
`data`	A data frame containing the variables in the model.
`start`	An optional numeric vector of initial parameter values. If not provided, default values are used.

Details

This function estimates the parameters of the classical Heckman sample selection model via MLE, accounting for potential sample selection bias. It uses the `optim` function with the BFGS method to find the parameter estimates that maximize the log-likelihood function. The initial values for optimization are obtained using the two-step Heckman method.

The function returns a rich set of results, including:

- Estimated coefficients for the selection and outcome equations.
- Standard deviation of the outcome error term (`sigma`).
- Correlation between the errors of the selection and outcome equations (`rho`).
- Measures of model fit (AIC, BIC).
- Standard errors (approximated by the square root of the Fisher information diagonal).

Value

A list containing:

- `coefficients`: A named numeric vector of estimated parameters.
- `value`: The value of the (negative) log-likelihood at convergence.
- `loglik`: The maximized log-likelihood.
- `counts`: Number of gradient evaluations performed.
- `hessian`: Hessian matrix at the optimum.
- `fisher_infoHC`: The (approximate) Fisher information matrix.
- `prop_sigmaHC`: Approximate standard errors.
- `level`: Levels of the selection variable.
- `nObs`: Number of observations.
- `nParam`: Number of estimated parameters.
- `N0`: Number of unobserved (censored) observations.
- `N1`: Number of observed (uncensored) observations.
- `NXS`: Number of parameters in the selection equation.
- `NXO`: Number of parameters in the outcome equation.
- `df`: Degrees of freedom (observations minus parameters).
- `aic`: Akaike Information Criterion.
- `bic`: Bayesian Information Criterion.
- `initial.value`: Initial parameter values used in the optimization.

References

James J Heckman (1979). "Sample selection bias as a specification error." *Econometrica: Journal of the econometric society*, 153–161.

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

Fits a generalized Heckman sample selection model that allows for heteroskedasticity in the outcome equation and correlation of the error terms depending on covariates. The estimation is performed via Maximum Likelihood using the BFGS algorithm.

Usage

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

Arguments

`selection`	A formula specifying the selection equation.
`outcome`	A formula specifying the outcome equation.
`outcomeS`	A formula or matrix specifying covariates for the scale (variance) model.
`outcomeC`	A formula or matrix specifying covariates for the correlation model.
`data`	A data frame containing the variables in the model.
`start`	An optional numeric vector with starting values for the optimization.

Details

This function extends the classical Heckman selection model by incorporating models for the error term's variance (scale) and the correlation between the selection and outcome equations. The scale model (`outcomeS`) allows the error variance of the outcome equation to depend on covariates, while the correlation model (`outcomeC`) allows the error correlation to vary with covariates.

The optimization is initialized with default or user-supplied starting values, and the results include robust standard errors derived from the inverse of the observed Fisher information matrix.

Value

A list containing:

- `coefficients`: Named vector of estimated model parameters.
- `value`: Negative of the maximum log-likelihood.
- `loglik`: Maximum log-likelihood.
- `counts`: Number of gradient evaluations performed.
- `hessian`: Hessian matrix at the optimum.
- `fisher_infoHG`: Approximate Fisher information matrix.
- `prop_sigmaHG`: Standard errors for the parameter estimates.
- `level`: Levels of the selection variable.
- `nObs`: Number of observations in the dataset.
- `nParam`: Number of estimated parameters.
- `N0`: Number of censored (unobserved) observations.
- `N1`: Number of uncensored (observed) observations.
- `NXS`: Number of covariates in the selection equation.
- `NXO`: Number of covariates in the outcome equation.
- `df`: Degrees of freedom (observations minus parameters).
- `aic`: Akaike Information Criterion.
- `bic`: Bayesian Information Criterion.
- `initial.value`: Starting values used for optimization.
- `NE`: Number of parameters in the scale model.
- `NV`: Number of parameters in the correlation model.

References

Fernando de Souza Bastos, Wagner Barreto-Souza, Marc G Genton (2022). "A Generalized Heckman Model With Varying Sample Selection Bias and Dispersion Parameters." *Statistica Sinica*.

Examples

```
## Not run:
data(MEPS2001)
attach(MEPS2001)
selectEq <- dambexp ~ age + female + educ + blhisp + totchr + ins + income
outcomeEq <- lnambx ~ age + female + educ + blhisp + totchr + ins
outcomeS <- ~ educ + income
outcomeC <- ~ blhisp + female
HeckmanGe(selectEq, outcomeEq, outcomeS = outcomeS, outcomeC = outcomeC, data = MEPS2001)

## End(Not run)
```

HeckmanSK *Skew-Normal Sample Selection Model Fit Function*

Description

Fits a sample selection model based on the Skew-Normal distribution using Maximum Likelihood Estimation (MLE). This model allows for asymmetry in the distribution of the outcome variable's error term, addressing potential skewness.

Usage

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

Arguments

selection	A formula specifying the selection equation.
outcome	A formula specifying the outcome equation.
data	A data frame containing the variables.
lambda	Initial start value for the skewness parameter (lambda).
start	Optional numeric vector of initial parameter values.

Details

The function implements MLE for a sample selection model where the outcome equation's errors follow a Skew-Normal distribution, as proposed in Ogundimu and Hutton (2016). The optimization is performed via the BFGS algorithm.

The results include estimates for:

- Selection equation coefficients.
- Outcome equation coefficients.
- Standard deviation of the error term (`sigma`).
- Correlation between the selection and outcome errors (`rho`).
- Skewness parameter (`lambda`).
- Robust standard errors from the Fisher information matrix.

Value

A list containing:

- `coefficients`: Named vector of estimated model parameters.
- `value`: The (negative) log-likelihood at convergence.
- `loglik`: The maximum log-likelihood.
- `counts`: Number of gradient evaluations.
- `hessian`: Hessian matrix at the optimum.
- `fisher_infoSK`: Approximate Fisher information matrix.
- `prop_sigmaSK`: Standard errors for the estimates.
- `level`: Levels of the selection variable.
- `nObs`: Number of observations.
- `nParam`: Number of model parameters.
- `N0`: Number of censored (unobserved) observations.
- `N1`: Number of observed (uncensored) observations.
- `NXS`: Number of covariates in the selection equation.
- `NXO`: Number of covariates in the outcome equation.
- `df`: Degrees of freedom (observations minus parameters).
- `aic`: Akaike Information Criterion.
- `bic`: Bayesian Information Criterion.
- `initial.value`: Initial parameter values used.

References

Emmanuel O Ogundimu, Jane L Hutton (2016). "A Sample Selection Model with Skew-normal Distribution." *Scandinavian Journal of Statistics*, **43**(1), 172–190.

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

Fits a sample selection model based on the Student's t-distribution, extending the classical Heckman model to account for heavy-tailed error terms. The estimation is performed via Maximum Likelihood using the BFGS algorithm.

Usage

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

Arguments

`selection`	A formula specifying the selection equation.
`outcome`	A formula specifying the outcome equation.
`data`	A data frame containing the variables in the model.
`df`	Initial value for the degrees of freedom parameter of the t-distribution.
`start`	Optional numeric vector of initial parameter values.

Details

The function implements the Heckman sample selection model using the Student's t-distribution for the error terms, as proposed by Marchenko and Genton (2012). This extension allows for robustness against outliers and heavy-tailed distributions. Initial parameter values can be specified by the user or default to standard starting values.

Value

A list containing:

- `coefficients`: Named vector of estimated model parameters.
- `value`: Negative of the maximum log-likelihood.
- `loglik`: Maximum log-likelihood.
- `counts`: Number of gradient evaluations performed.
- `hessian`: Hessian matrix at the optimum.
- `fisher_infotS`: Approximate Fisher information matrix.
- `prop_sigmatS`: Standard errors for the parameter estimates.
- `level`: Levels of the selection variable.
- `nObs`: Number of observations.
- `nParam`: Number of model parameters.
- `N0`: Number of censored (unobserved) observations.

- N1: Number of uncensored (observed) observations.
- NXS: Number of parameters in the selection equation.
- NXO: Number of parameters in the outcome equation.
- df: Degrees of freedom (observations minus parameters).
- aic: Akaike Information Criterion.
- bic: Bayesian Information Criterion.
- initial.value: Initial parameter values used in the optimization.

References

Yulia V Marchenko, Marc G Genton (2012). "A Heckman selection-t model." *Journal of the American Statistical Association*, **107**(497), 304–317.

Examples

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

IMR — *Inverse Mills Ratio (IMR) Calculation*

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 specifying the selection equation.
data	A data frame containing the variables in the model.

Details

This function fits a Probit model to the provided selection equation and returns the Inverse Mills Ratio (IMR) for each observation. The IMR is useful for correcting sample selection bias in regression models, following the classical Heckman approach.

Value

A numeric matrix with one column containing the IMR values for each observation.

Examples

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

MEPS2001 *Medical Expenditure Panel Survey*

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 Colin, Pravin K Trivedi (2009). "Microeconometrics using STATA." *Lakeway Drive, TX: Stata Press Books*.

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

Ott Toomet, Arne Henningsen (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 found in 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 ResearchPanel Study of Income Dynamics, University of Michigan, https://psidonline.isr.umich.edu/

References

Thomas A Mroz (1987). "The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions." *Econometrica: Journal of the Econometric Society*, 765–799.

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

Ott Toomet, Arne Henningsen (2008). "Sample Selection Models in R: Package sampleSelection." *Journal of Statistical Software*, **27**(7). https://www.jstatsoft.org/article/view/v027i07.

Jeffrey M Wooldridge (2016). *Introductory econometrics: A modern approach*. Nelson Education.

Examples

```
# Wooldridge(2016): page 247
data(Mroz87)
attach(Mroz87)
Mroz87$lwage <- ifelse(Mroz87$wage>0,log(Mroz87$wage), NA)
selectEq <- lfp ~ nwifeinc + educ + exper + I(exper^2) + age + kids5 + kids618
outcomeEq <- lwage ~ educ + exper + I(exper^2)
outcomeS <- cbind(educ, exper)
outcomeC <- 1
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://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2003

References

Emmanuel O Ogundimu, Gary S Collins (2019). "A robust imputation method for missing responses and covariates in sample selection models." *Statistical methods in medical research*, **28**(1), 102–116.

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

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

Ott Toomet, Arne Henningsen (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://simba.isr.umich.edu/

References

Anastasia Semykina, Jeffrey M Wooldridge (2013). "Estimation of dynamic panel data models with sample selection." *Journal of Applied Econometrics*, **28**(1), 47–61.

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

Ott Toomet, Arne Henningsen (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))
```

RandHIE *RAND Health Insurance Experiment*

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 nonphysician 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(coins+1)`.
- fmde: 0 if `idp=1`, `ln(max(1,mdeoff/(0.01*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 healt 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

A Colin Cameron, Pravin K Trivedi (2005). *Microeconometrics: methods and applications*. Cambridge university press.

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

Ott Toomet, Arne Henningsen (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 — Wikipedia, The Free Encyclopedia." https://en.wikipedia.org/w/index.php?title=RAND_Health_Insurance_Experiment&oldid=909771077. [Online; accessed 9-September-2019].

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)
```

`ssmodels` *ssmodels: Sample Selection Models in R*

Description

The `ssmodels` package provides functions to fit data affected by sample selection bias. It includes several extensions of the classical Heckman selection model, allowing for different assumptions about the joint distribution of the selection and outcome equations.

Details

The following models are implemented:

`HeckmanCL` Classic Heckman model (Tobit-2).

`HeckmantS` Heckman model with Student's t-distribution.

`HeckmanSK` Heckman model with Skew-Normal distribution.

`HeckmanBS` Heckman model with Birnbaum-Saunders distribution.

`HeckmanGe` Generalized Heckman model with covariates in the dispersion and correlation structures.

The package also includes helper functions for computing Inverse Mills Ratios (IMR), post-processing parameter vectors, and two-step initial value estimation.

Author(s)

Fernando de Souza Bastos, Wagner Barreto de Souza

References

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

James J Heckman (1979). "Sample selection bias as a specification error." *Econometrica: Journal of the econometric society*, 153–161.

Thomas A Mroz (1987). "The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions." *Econometrica: Journal of the Econometric Society*, 765–799.

Ott Toomet, Arne Henningsen (2008). "Sample Selection Models in R: Package sampleSelection." *Journal of Statistical Software*, **27**(7). https://www.jstatsoft.org/article/view/v027i07.

Yulia V Marchenko, Marc G Genton (2012). "A Heckman selection-t model." *Journal of the American Statistical Association*, **107**(497), 304–317.

Emmanuel O Ogundimu, Jane L Hutton (2016). "A Sample Selection Model with Skew-normal Distribution." *Scandinavian Journal of Statistics*, **43**(1), 172–190.

Mikhail Zhelonkin, Marc G Genton, Elvezio Ronchetti (2016). "Robust inference in sample selection models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **78**(4), 805–827.

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

Emmanuel O Ogundimu, Gary S Collins (2019). "A robust imputation method for missing responses and covariates in sample selection models." *Statistical methods in medical research*, **28**(1), 102–116.

Fernando de Souza Bastos, Wagner Barreto-Souza (2020). "Birnbaum–Saunders sample selection model." *Journal of Applied Statistics*.

Fernando de Souza Bastos, Wagner Barreto-Souza, Marc G Genton (2022). "A Generalized Heckman Model With Varying Sample Selection Bias and Dispersion Parameters." *Statistica Sinica*.

See Also

HeckmanCL, HeckmantS, HeckmanSK, HeckmanBS, HeckmanGe

step2 *Heckman's Two-Step Method*

Description

Estimates the parameters of the classical Heckman selection model using the two-step method. The first step fits a probit model for the selection equation. In the second step, the inverse Mills ratio (IMR) is included as an additional regressor in the outcome equation.

Usage

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

Arguments

YS	A binary vector indicating selection (1 if observed, 0 otherwise).
XS	A matrix of covariates for the selection equation.
YO	A numeric vector representing the outcome variable of interest.
XO	A matrix of covariates for the outcome equation.

Details

This function implements the two-step estimation procedure of the classical Heckman model. In the first step, a probit model is estimated to predict the selection indicator YS using the selection covariates XS. The IMR is calculated from this model. In the second step, an ordinary least squares (OLS) regression of the observed outcome YO on XO and the IMR is performed for the uncensored observations (YS == 1).

The function also calculates:

- sigma: The estimated standard deviation of the outcome equation's error term.
- rho: The estimated correlation between the error terms of the selection and outcome equations.

Value

A numeric vector containing the parameter estimates from the two-step Heckman model:

- Coefficients of the selection equation (probit model).
- Coefficients of the outcome equation (excluding the IMR term).
- Estimated sigma.
- Estimated rho.

References

There are no references for Rd macro \insertAllCites on this help page.

Examples

```
data(MEPS2001)
attach(MEPS2001)
YS <- dambexp
XS <- cbind(age, female, educ, blhisp, totchr, ins)
YO <- lnambx
XO <- cbind(age, female, educ, blhisp, totchr, ins, income)
step2(YS, XS, YO, XO)
```

---

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

Description

Prints a detailed summary of the parameter estimates and model fit statistics for an object of class `HeckmanBS`.

Usage

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

Arguments

`object`	An object of class `HeckmanBS`, containing the fitted model results.
`...`	Additional arguments (currently unused).

Details

This method provides a summary of the maximum likelihood estimation results for the Heckman sample selection model with Birnbaum-Saunders errors. It includes separate coefficient tables for:

- Selection equation (Probit model),
- Outcome equation,
- Error terms (`sigma` and `rho`).

Model fit criteria such as the log-likelihood, AIC, and BIC are also reported.

Value

Prints to the console:

- Model fit statistics (log-likelihood, AIC, BIC, number of observations).
- Coefficient tables with standard errors and significance stars.

Invisibly returns `NULL`.

See Also

HeckmanBS

Examples

```
## Not run:
data(MEPS2001)
attach(MEPS2001)
selectEq <- dambexp ~ age + female + educ + blhisp + totchr + ins + income
outcomeEq <- ambexp ~ age + female + educ + blhisp + totchr + ins
model <- HeckmanBS(selectEq, outcomeEq, data = MEPS2001)
summary(model)

## End(Not run)
```

---

summary.HeckmanCL *Summary of Classic Heckman Model*

Description

Prints a detailed summary of the parameter estimates and model fit statistics for an object of class HeckmanCL.

Usage

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

Arguments

object	An object of class HeckmanCL, containing the fitted model results.
...	Additional arguments (currently unused).

Details

This method displays the maximum likelihood estimation results for the classical Heckman sample selection model. It includes separate coefficient tables for:

- Selection equation (Probit model),
- Outcome equation,
- Error terms (sigma and rho).

Additionally, it reports the model fit statistics (log-likelihood, AIC, BIC, and number of observations).

Value

Prints to the console:

- Model fit statistics (log-likelihood, AIC, BIC, number of observations).
- Coefficient tables with standard errors and significance stars.

Invisibly returns `NULL`.

See Also

`HeckmanCL`

Examples

```
## Not run:
data(MEPS2001)
attach(MEPS2001)
selectEq <- dambexp ~ age + female + educ + blhisp + totchr + ins + income
outcomeEq <- lnambx ~ age + female + educ + blhisp + totchr + ins
model <- HeckmanCL(selectEq, outcomeEq, data = MEPS2001)
summary(model)

## End(Not run)
```

`summary.HeckmanGe` *Summary of Generalized Heckman Model*

Description

Prints a detailed summary of the parameter estimates and model fit statistics for an object of class `HeckmanGe`.

Usage

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

Arguments

`object`	An object of class `HeckmanGe`, containing the fitted model results.
`...`	Additional arguments (currently unused).

Details

This method displays the maximum likelihood estimation results for the generalized Heckman sample selection model. It includes separate coefficient tables for:

- Selection equation (Probit model),
- Outcome equation,
- Dispersion (scale) model parameters,
- Correlation model parameters.

Model fit statistics (log-likelihood, AIC, BIC, and number of observations) are also reported for interpretation and model assessment.

Value

Prints to the console:

- Model fit statistics (log-likelihood, AIC, BIC, number of observations).
- Coefficient tables with standard errors and significance stars.

Invisibly returns `NULL`.

See Also

`HeckmanGe`

Examples

```
## Not run:
data(MEPS2001)
attach(MEPS2001)
selectEq <- dambexp ~ age + female + educ + blhisp + totchr + ins + income
outcomeEq <- lnambx ~ age + female + educ + blhisp + totchr + ins
outcomeS <- ~ educ + income
outcomeC <- ~ blhisp + female
model <- HeckmanGe(selectEq, outcomeEq, outcomeS = outcomeS, outcomeC = outcomeC, data = MEPS2001)
summary(model)

## End(Not run)
```

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

Description

Prints a detailed summary of the parameter estimates and model fit statistics for an object of class HeckmanSK.

Usage

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

Arguments

object	An object of class HeckmanSK, containing the fitted model results.
...	Additional arguments (currently unused).

Details

This method displays the maximum likelihood estimation results for the Heckman sample selection model with Skew-Normal errors. It includes separate coefficient tables for:

- Selection equation (Probit model),
- Outcome equation,
- Error terms (sigma, rho, and lambda).

Additionally, it reports model fit statistics such as the log-likelihood, AIC, BIC, and the number of observations.

Value

Prints to the console:

- Model fit statistics (log-likelihood, AIC, BIC, number of observations).
- Coefficient tables with standard errors and significance stars.

Invisibly returns NULL.

See Also

HeckmanSK

Examples

```
## Not run:
data(Mroz87)
attach(Mroz87)
selectEq <- lfp ~ huswage + kids5 + mtr + fatheduc + educ + city
outcomeEq <- log(wage) ~ educ + city
model <- HeckmanSK(selectEq, outcomeEq, data = Mroz87, lambda = -1.5)
summary(model)

## End(Not run)
```

`summary.HeckmantS` *Summary of Heckman-t Model*

Description

Prints a detailed summary of the parameter estimates and model fit statistics for an object of class `HeckmantS`.

Usage

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

Arguments

`object`	An object of class `HeckmantS`, containing the fitted model results.
`...`	Additional arguments (currently unused).

Details

This method displays the maximum likelihood estimation results for the Heckman sample selection model with Student's t-distributed errors. It includes separate coefficient tables for:

- Selection equation (Probit model),
- Outcome equation,
- Error terms (including `sigma`, `rho`, and `df`).

Model fit statistics (log-likelihood, AIC, BIC, and number of observations) are also provided for model evaluation.

Value

Prints to the console:

- Model fit statistics (log-likelihood, AIC, BIC, number of observations).
- Coefficient tables with standard errors and significance stars.

Invisibly returns `NULL`.

See Also

HeckmantS

Examples

```
## Not run:
data(MEPS2001)
attach(MEPS2001)
selectEq <- dambexp ~ age + female + educ + blhisp + totchr + ins + income
outcomeEq <- lnambx ~ age + female + educ + blhisp + totchr + ins
model <- HeckmantS(selectEq, outcomeEq, data = MEPS2001, df = 12)
summary(model)

## End(Not run)
```

---

twostep — *Two-Step Estimation of the Classic Heckman Model*

Description

Estimates the parameters of the classical Heckman sample selection model using the two-step procedure.

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.

Details

The two-step method first estimates a Probit model for the selection equation, then fits an outcome equation that includes the Inverse Mills Ratio (IMR) as an additional regressor to correct for sample selection bias.

Value

A numeric vector containing:

- Estimated coefficients of the selection equation (Probit model),
- Estimated coefficients of the outcome equation (excluding IMR),
- Estimated standard deviation of the outcome errors (phi),
- Estimated correlation between the error terms (cor).

References

There are no references for Rd macro `\insertAllCites` on this help page. For details, see Heckman (1979).

Examples

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

Index

∗ **Heckman**
    ssmodels, 21
∗ **RandHIE**
    RandHIE, 19
∗ **datasets**
    MEPS2001, 13
    Mroz87, 14
    nhanes, 16
    PSID2, 17
∗ **econometrics**
    ssmodels, 21
∗ **likelihood**
    ssmodels, 21
∗ **maximum**
    ssmodels, 21
∗ **sample**
    ssmodels, 21
∗ **selection**
    ssmodels, 21

HCinitial, 2
HeckmanBS, 4, *22*, *25*
HeckmanCL, 5, *22*, *26*
HeckmanGe, 7, *22*, *27*
HeckmanSK, 9, *22*, *28*
HeckmantS, 11, *22*, *30*

IMR, 12

MEPS2001, 13
Mroz87, 14

nhanes, 16

PSID2, 17

RandHIE, 19

ssmodels, 21
ssmodels-package *(*ssmodels*)*, 21
step2, 23
summary.HeckmanBS, 24
summary.HeckmanCL, 25
summary.HeckmanGe, 26
summary.HeckmanSK, 28
summary.HeckmantS, 29

twostep, 30