Package 'endogenous'

July 22, 2025

Type Package

Title Classical Simultaneous Equation Models

Version 1.0

Date 2016-10-25

Maintainer Andrew J. Spieker <aspieker@upenn.edu>

Description Likelihood-based approaches to estimate linear regression parameters and treatment effects in the presence of endogeneity. Specifically, this package includes James Heckman's classical simultaneous equation models-the sample selection model for outcome selection bias and hybrid model with structural shift for endogenous treatment. For more information, see the seminal paper of Heckman (1978) <DOI:10.3386/w0177> in which the details of these models are provided. This package accommodates repeated measures on subjects with a working independence approach. The hybrid model further accommodates treatment effect modification.

Imports mvtnorm

License GPL-2

NeedsCompilation no

Author Andrew J. Spieker [aut, cre]

Repository CRAN

Date/Publication 2016-10-29 10:48:20

Contents

---

hybrid	*Hybrid model with structural shift (permits covariate-specific treatment effects)*

Description

James Heckman's Hybrid Model with Structural Shift (also known as the Treatment Effects Model). Jointly models outcome regression model and endogenous variable probit model (e.g., outcome associations in the presence of endogenous treatment in observational data). Can handle clustered data. Accommodates treatment effect modification based on observable covariates.

Usage

```
## S3 method for class "hybrid"
hybrid(outcome, probit, modifiers = NULL, init = NULL, id = NULL, se = "R")
```

Arguments

outcome	an object of class "formula" with a numeric vector on the left hand side, and predictors of interest on the right hand side.
probit	an object of class "formula" with a binary (0/1) numeric vector on the left hand side (1 indicating medication use), and predictors of medication use on the right hand side (right hand side permitted to contain variables on the right hand side of the outcome equation).
modifiers	an object of class "formula" with a binary numeric vector indicating medication use on the left hand side, and treatment effect modifiers on the right hand side. If effect modifiers are treatment group specific (e.g., medication dose), set the effect modifier variables to zero for the untreated observations. If any other numeric values are used, they will ultimately be zet to zero. If NULL, the average treatment effect will be estimated under the assumption of no effect modification.
init	a vector of initial values. The ordering of subparameters is: alpha (probit model parameters), beta (outcome model parameters), eta (an intercept, with or without effect mofidier paramters), sigmay (outcome error standard deviation), rho (error correlation). If NULL, an initial value will be chosen through OLS linear regression and probit-link GLM without regard to endogeneity.
id	a numeric vector indicating subject IDs if data are clustered. In the absence of clustered data, this can be left blank (defaults to NULL).
se	a string, either "M" for model-based standard errors (based on inverse observed Fisher information), or "R" for robust standard errors (based on methods of Huber and White). Defaults to "R". If id is provided for clustered data, the cluster-robust variance estimator (with working independence) will be used even if the user specifies type "M".

Details

The model is evaluated with numerical minimization of the negative log-likelihood (the BFGS is used). The probit model and error correlation parameters are weakly identified and hence the error variance is set at unity. The data must be complete (no missing values) and numeric, with the exception of factors, which may be used on the right hand side of equations.

Value

`hybrid` prints a summary of the coefficient estimates, standard errors, Wald-based confidence intervals, and p-values for the outcome model, the treatment effects (and potentially effect modifiers), and the medication use probit model.

`alpha`	estimate of the medication use probit model parameters.
`beta`	estimate of the outcome model parameters.
`eta`	estimate of the treatment effect, with or without effect modifier parameters.
`sigma`	estimate of the standard deviation of the outcome error.
`rho`	estimate of the correlation between the errors.
`vcov`	entire estimated variance-covariance matrix, provided if the user wishes to perform any more specific hypothesis tests.
`init`	initial value ultimately used, whether specified by the user or generated through the default approach.
`fitted`	vector of fitted outcome values.
`call`	the matched call.
`out.form`	the formula used for the outcome model.
`prob.form`	the formula used for the medication use probit model.
`tx.form`	the formula used for the treatment effects model (potentially with effect modifiers).
`sterr`	the choice of the variance estimate procedure (either model-based or robust).
`labels`	labels for predictors to be passed into output.

Author(s)

Andrew J. Spieker, Ph.D.

References

Heckman JJ. Dummy endogenous variables in a simultaneous equation system. *Econometrica* 46(4), 931-959.

Maddala GS. *Limited-dependent and qualitative variables in econometrics.* Cambridgeshire: Cambridge University Press; 1983.

Spieker AJ, Delaney JAC, and McClelland RL. Evaluating the treatment effects model for estimation of cross-sectional associations between risk factors and cardiovascular biomarkers influenced by medication use. *Pharmacoepidemiology and Drug Safety* 24(12), 1286-1296.

Examples

```
#- Generate Data -#
require(mvtnorm)
set.seed(1)
N <- 2000
X1 <- rnorm(N, 0, 1);
X2 <- rnorm(N, 0, 1);
X3 <- rnorm(N, 0, 1);
errors <- rmvnorm(N, sigma = 50*matrix(c(1, 0.5, 0.5, 1), nrow = 2))
Y <- 50 + X1 + X2 + errors[,1]
Z <- rep(0, N)
Z[(-5 + X1 + X3 + errors[,2]) > 0] <- 1
Y[Z == 1] <- Y[Z == 1] - 0.5*X1[Z == 1]

#- Estimate Model with No Effect Modification -#
hybrid(Y ~ X1 + X2, probit = Z ~ X1 + X3)

#- Estimate Model with Effect Modification -#
hybrid(Y ~ X1 + X2, probit = Z ~ X1 + X3, modifiers = Z ~ X1)

#- Estimate Model with Effect Modification and Model-Based Variance -#
hybrid(Y ~ X1 + X2, probit = Z ~ X1 + X3, modifiers = Z ~ X1, se = "M")
```

---

sampselect   *Sample selection model (endogenous probit).*

Description

James Heckman's Classical Simultaneous Equation Model (also known as the Sample Selection Model). Used to account for endogenous sample selection. Jointly models outcome model with propensity of selection, in which some of the outcomes are unobserved. Can also handle clustered data.

Usage

```
## S3 method for class "sampselect"
sampselect(outcome, probit, init = NULL, id = NULL, se = "R")
```

Arguments

outcome   an object of class "formula" with a numeric vector on the left hand side, and predictors of interest on the right hand side. Values on the left hand side that correspond to unobserved outcomes should be set to numeric values (to zero, for example, although they can be set to any numeric values).

probit   an object of class "formula" with a binary (0/1) numeric vector on the left hand side (1 indicating unobserved outcome), and predictors of selection on the right hand side (right hand side permitted to contain variables on the right hand side of the outcome equation).

init
: a vector of initial values. The ordering of subparameters is: `alpha` (probit model parameters), `beta` (outcome model parameters), `sigmay` (outcome error standard deviation), `rho` (error correlation). If `NULL`, an initial value will be chosen through OLS linear regression and probit-link GLM without regard to endogeneity.

id
: a numeric vector indicating subject IDs if data are clustered. In the absence of clustered data, this can be left blank (defaults to `NULL`).

se
: a string, either `"M"` for model-based standard errors (based on inverse observed Fisher information), or `"R"` for robust standard errors (based on methods of Huber and White). Defaults to `"R"`. If `id` is provided for clustered data, the cluster-robust variance estimator (with working independence) will be used even if the user specifies type `"M"`.

Details

The model is evaluated with numerical minimization of the negative log-likelihood (the `BFGS` is used). The probit model and error correlation parameters are weakly identified and hence the error variance is set at unity. The data must be complete (no missing values) and numeric, with the exception of factors, which may be used on the right hand side of equations.

Value

`sampselect` prints a summary of the coefficient estimates, standard errors, Wald-based confidence intervals, and p-values for the outcome model and the selection use probit model.

alpha
: estimate of the selection probit model parameters.

beta
: estimate of the outcome model parameters.

sigma
: estimate of the standard deviation of the outcome error.

rho
: estimate of the correlation between the errors.

vcov
: entire estimated variance-covariance matrix, provided if the user wishes to perform any more specific hypothesis tests.

init
: initial value ultimately used, whether specified by the user or generated through the default approach.

fitted
: vector of fitted outcome values.

call
: the matched call.

out.form
: the formula used for the outcome model.

prob.form
: the formula used for the selection probit model.

sterr
: the choice of the variance estimate procedure (either model-based or robust).

labels
: labels for predictors to be passed into output.

Author(s)

Andrew J. Spieker, Ph.D.

References

Heckman JJ. Dummy endogenous variables in a simultaneous equation system. *Econometrica* 46(4), 931-959.

Maddala GS. *Limited-dependent and qualitative variables in econometrics.* Cambridgeshire: Cambridge University Press; 1983.

Examples

```
#- Generate Data -#
require(mvtnorm)
set.seed(1)
N <- 2000
X1 <- rnorm(N, 0, 1);
X2 <- rnorm(N, 0, 1);
X3 <- rnorm(N, 0, 1);
errors <- rmvnorm(N, sigma = 50*matrix(c(1, 0.5, 0.5, 1), nrow = 2))
Y <- 50 + X1 + X2 + errors[,1]
Z <- rep(0, N)
Z[(-5 + X1 + X3 + errors[,2]) > 0] <- 1
Y[Z == 1] <- 0

#- Estimate Model -#
sampselect(Y ~ X1 + X2, probit = Z ~ X1 + X3)

#- Estimate Model with Model-Based Variance -#
sampselect(Y ~ X1 + X2, probit = Z ~ X1 + X3, se = "M")
```

Index