Package 'DiSSMod'

April 15, 2019

Type Package

Title Fitting Sample Selection Models for Discrete Response Variables

Version 1.0.0

Date 2019-04-15

Author Sang Kyu Lee <lsk0816@gmail.com>, Adelchi Azzalini <adelchi.azzalini@unipd.it>, Hyoung-Moon Kim <hmk966a@gmail.com>

Maintainer Sang Kyu Lee <lsk0816@gmail.com>

Description Tools to fit sample selection models in case of discrete response variables, through a parametric formulation which represents a natural extension of the well-known Heckman selection model are provided in the package. The response variable can be of Bernoulli, Poisson or Negative Binomial type. The sample selection mechanism allows to choose among a Normal, Logistic or Gumbel distribution.

Depends R (>= 2.10)

Imports sfsmisc, matrixcalc, psych, MASS

License GPL-2

Encoding UTF-8

LazyData true

RoxygenNote 6.1.1

NeedsCompilation no

Repository CRAN

Date/Publication 2019-04-15 10:42:40 UTC

R topics documented:

coef.DiSSMod 2
confint.DiSSMod 3
CreditMDR 4
DiSSMod 5
DoctorRWM 8
plot.DiSSMod 10
summary.DiSSMod 11

Index **13**

`coef.DiSSMod` *Getting Coefficients of Discrete Sample Selection Model Fits*

Description

`coef` method for a class `"DiSSMod"`.

Usage

```
## S3 method for class 'DiSSMod'
coef(object, only = NULL, ...)
```

Arguments

`object`	an object of class "DiSSMod" made by the function `DiSSMod`.
`only`	a character value for choosing specific variable's coefficients. Initial value is
`...`	not used, but exists because of the compatibility. `NULL`, which shows all variable's coefficients. If `"response"` is written, only coefficients for response variables will be returned, and if `"selection"` is written, only coefficients for selection variables will be returned.

Details

It looks as similar as the generic function `coef`, but this case there are two equations. Therefore, there exist little differences.

Value

a numeric vector or a list is given.

See Also

See also `DiSSMod` and `coef`.

Examples

```
# example continued from DiSSMod
set.seed(45)
data(DoctorRWM, package = "DiSSMod")
n0 <- 600
set.n0 <- sample(1:nrow(DoctorRWM), n0)
reduce_DoctorRWM <- DoctorRWM[set.n0,]
result0 <- DiSSMod(response = as.numeric(DOCVIS > 0) ~ AGE + INCOME_SCALE + HHKIDS + EDUC + MARRIED,
                   selection = PUBLIC ~ AGE + EDUC + FEMALE,
                  data = reduce_DoctorRWM, resp.dist="bernoulli", select.dist = "normal",
                   alpha = seq(-5.5, -0.5, length.out = 21), standard = TRUE)
```

```
coef(result0)

data(CreditMDR, package = "DiSSMod")
n1 <- 600
set.n1 <- sample(1:nrow(CreditMDR), n1)
reduce_CreditMDR <- CreditMDR[set.n1,]
result1 <- DiSSMod(response = MAJORDRG ~ AGE + INCOME + EXP_INC,
                   selection = CARDHLDR ~ AGE + INCOME + OWNRENT + ADEPCNT + SELFEMPL,
                   data = reduce_CreditMDR, resp.dist="poi", select.dist = "logis",
                  alpha = seq(-0.3, 0.3,length.out = 21), standard = FALSE, verbose = 1)

coef(result1)
```

`confint.DiSSMod` *Getting Confidence Intervals for Parameters of Discrete Sample Selection Model Fits*

Description

`confint` method for a class `"DiSSMod"`.

Usage

```
## S3 method for class 'DiSSMod'
confint(object, parm, level = 0.95, ...)
```

Arguments

`object`	an object of class "DiSSMod" made by the function `DiSSMod`.
`parm`	not used, but it exists for compatibility reasons.
`level`	a numeric value between 0 and 1 for controlling the significance level of confidence interval; default value is 0.95.
`...`	not used, but it exists for compatibility reasons.

Value

a list, containing `level` and `confidence intervals` for parameters, is given.

See Also

See also `confint`, `DiSSMod` and `summary.DiSSMod`.

Examples

```
# example continued from DiSSMod
set.seed(45)
data(DoctorRWM, package = "DiSSMod")
n0 <- 600
set.n0 <- sample(1:nrow(DoctorRWM), n0)
reduce_DoctorRWM <- DoctorRWM[set.n0,]
result0 <- DiSSMod(response = as.numeric(DOCVIS > 0) ~ AGE + INCOME_SCALE + HHKIDS + EDUC + MARRIED,
                   selection = PUBLIC ~ AGE + EDUC + FEMALE,
                  data = reduce_DoctorRWM, resp.dist="bernoulli", select.dist = "normal",
                   alpha = seq(-5.5, -0.5, length.out = 21), standard = TRUE)

confint(result0, level = 0.90)

data(CreditMDR, package = "DiSSMod")
n1 <- 600
set.n1 <- sample(1:nrow(CreditMDR), n1)
reduce_CreditMDR <- CreditMDR[set.n1,]
result1 <- DiSSMod(response = MAJORDRG ~ AGE + INCOME + EXP_INC,
                   selection = CARDHLDR ~ AGE + INCOME + OWNRENT + ADEPCNT + SELFEMPL,
                   data = reduce_CreditMDR, resp.dist="poi", select.dist = "logis",
                  alpha = seq(-0.3, 0.3,length.out = 21), standard = FALSE, verbose = 1)

confint(result1)
```

CreditMDR *Credit cards derogatory reports data*

Description

Data is originally from Greene (1992), used for studying statistical credit scoring methods.

Usage

```
data(CreditMDR)
```

Format

A data frame with 13444 observations of 8 variables as below;

MAJORDRG count of major derogatory reports (numeric)

CARDHLDR 1 for cardholders, 0 for denied applicants (categorical)

AGE age in years and twelfths of a year (numeric)

INCOME primary income, divided by 10,000 (numeric)

OWNRENT ownRent, individual owns (1) or rents (0) home (categorical)

ADEPCNT not classified yet (numeric)

SELFEMPL self employed; 1=yes, 0=no (categorical)

EXP_INC average expenditure for 12 months/average monthly income (numeric)

Source

Greene, W. H. (1992) *A Statistical Model for Credit Scoring*. Working Paper No. EC-92-29, Department of Economics, Stern School of Business, New York University, 1992.

http://pages.stern.nyu.edu/~wgreene/Text/Edition7/tablelist8new.htm

References

Greene, W. H. (2012) *Econometric Analysis, 7th Edition*. Pearson education.

Azzalini, A., Kim, H.-M. and Kim, H.-J. (2019) Sample selection models for discrete and other non-Gaussian response variables. *Statistical Methods & Applications*, **28**, 27–56. First online 30 March 2018. https://doi.org/10.1007/s10260-018-0427-1

`DiSSMod` *Fitting Sample Selection Models for Discrete Response Variables*

Description

Function `DiSSMod` fits sample selection models for discrete random variables, by suitably extending the formulation of the classical Heckman model to the case of a discrete response, but retaining the original conceptual framework. Maximum likelihood estimates are obtained by Newton-Raphson iteration combined with use of profile likelihood.

Usage

```
DiSSMod(response, selection, data, resp.dist, select.dist, alpha,
  trunc.num, standard = FALSE, verbose = 1, eps = 1e-07,
  itmax = 1000)
```

Arguments

`response`	a formula for the response equation.
`selection`	a formula for the selection equation.
`data`	a data frame and data has to be included with the form of `data.frame`.
`resp.dist`	a character for the distribution choice of the response variable, `"bernoulli"` for Bernoulli distribution, `"poisson"` for Poisson distribution, and `"negbinomial"` for Negative binomial distribution. Also, the character strings can be abbreviated and can be upper or lower case as preferred.
`select.dist`	a character for the distribution choice of the selection variable, `"gumbel"` for Gumbel distribution, `"normal"` for Normal distribution, and `"logistic"` for Logistic distribution. Also, the character strings can be abbreviated and can be upper or lower case as preferred.
`alpha`	a vector of $alpha$ values on which the profile log-likelihood function is evaluated; if the argument is missing, a set of values in the interval `(-10, 10)` is used for the initial search, followed by a second search on a revised interval which depends on the outcome from the first search.

trunc.num	an integer numeric constant used as the truncation point of an infine summation of probabilities involved when resp.dist equals "Poisson" or "NegBinomial"; if the argument is missing, a default choice is made, as described in the 'Details' section. Notice: this default choice of trunc.num may be subject to revision in some future version of the package, and the argument trunc.num itselt may possibly be replaced by some other ingredient.
standard	a logical value for the standardizing explanatory variables, if TRUE two types of values (standardized and not) will be returned.
verbose	an integer value for the level of printed details (values: 0|1|2); the default value is 1 which stands for shortly printed details. If the value is 2, more details are viewed such as values of the log likelihood functions and iteration numbers. If the value is 0, there is no printed detail.
eps	a numeric value for the estimating parameters, which is needed for the step of the optimization. If the sum of absolute differences between present step estimated parameters and former step estimated parameters is smaller than eps, we assume that estimated parameters are optimized.
itmax	an integer stands for maximum number for the iteration of optimizing the parameters.

Details

The specification of the two linear models regulating the response variable and the selection mechanism, as indicated in the 'Background' section, is accomplished by two arguments of formula type, denoted response and selection, respectively. Each formula is specified with the same syntax of similar arguments in standard functions such as lm and glm, with the restriction that the intercept term (which is automatically included) must not be removed.

The distributional assumptions associated to the response and selection components are specified by the arguments resp.dist and select.dist, respectively. Argument select.dist refers to the unobservable continuous variable of which we observe only the dichotomous outcome Yes-No.

In this respect, a remark is appropriate about the option "Gumbel" for select.dist. This choice is equivalent to the adoption of an Exponential distribution of the selection variables combined an exponential transformation of the linear predictor of the selection argument, as it is presented in Section 3.2 of Azzalini et al. (2019). Also, it corresponds to work with the log-transformation of an Exponential variable, which is essentially a Gumbel type of variable, up to a linear transformation with respect to its more commonly employed parameterization.

When resp.dist is "Poisson" or "NegBinomial" and trunc.num is missing, a default choice is made; this equals 1.5*m or 2*m in the two respective cases, where m denotes the maximum observed value of the response variable.

Function DiSSMOd calls lower level functions, nr.bin, nr.nbinom, nr.pois and the others for the actual numerical maximization of the log-likelihood via a Newton-Raphson iteration.

Notice that the automatic initialization of the alpha search interval, when this argument is missing, may change in future versions of the package.

Value

DiSSMod returns an object of class "DiSSMod", which is a list containing following components:

`call` a matched call.

`standard` a logical value, stands for standardization or not.

`st_loglik` a vector containing the differences between log likelihoods and maximized log likelihood.

`max_loglik` a maximized log likelihood value.

`mle_alpha` a maximized likelihood estimator of alpha.

`alpha` a vector containing grids of the alpha

`Nalpha` a vector containing proper alpha, which does not have `NA` value for corresponding log likelihood.

`num_NA` a number of `NA` values of log likelihoods.

`n_select` a number of selected response variables.

`n_all` a number of all response variables.

`estimate_response`
estimated values for the response model.

`std_error_response`
estimated standard errors for the response model.

`estimate_selection`
estimated values for the selection model.

`std_error_selection`
estimated standard errors for the selection model.

Background

Function `DiSSMod` fits sample selection models for discrete random variables, by suitably extending the formulation of the classical Heckman model to the case of a discrete response, but retaining the original conceptual framework. This logic involves the following key ingredients: (1) a linear model indicating which explanatory variables influence the response variable; (2) a linear model indicating which (possibly different) explanatory variables, besides the response variable itself, influence a 'selection variable', which is intrinsically continuous but we only observe a dichotomous outcome from it, of type Yes-No, which selects which are the observed response cases; (3) distributional assumptions on the response and the selection variable.

The data fitting method is maximum likelihood estimation (MLE), which operates in two steps: (i) for each given value of parameter $alpha$ which regulates the level of selection, MLE is performed for all the remaining parameters, using a Newton-Raphson iteration; (ii) a scan of the $alpha$ axis builds the profile log-likelihood function and its maximum point represents the overall MLE.

A detailed account of the underlying theory and the operational methodology is provided by Azzalini et al. (2019).

References

Azzalini, A., Kim, H.-M. and Kim, H.-J. (2019) Sample selection models for discrete and other non-Gaussian response variables. *Statistical Methods & Applications*, **28**, 27–56. First online 30 March 2018. https://doi.org/10.1007/s10260-018-0427-1

See Also

The functions summary.DiSSMod, coef.DiSSMod, confint.DiSSMod, plot.DiSSMod are used to obtain and print a summary, coefficients, confidence interval and plot of the results.

The generic function logLik is used to obtain maximum log likelihood of the result.

See also lm, glm and formula.

Examples

```
set.seed(45)
data(DoctorRWM, package = "DiSSMod")
n0 <- 600
set.n0 <- sample(1:nrow(DoctorRWM), n0)
reduce_DoctorRWM <- DoctorRWM[set.n0,]
result0 <- DiSSMod(response = as.numeric(DOCVIS > 0) ~ AGE + INCOME_SCALE + HHKIDS + EDUC + MARRIED,
                   selection = PUBLIC ~ AGE + EDUC + FEMALE,
                  data = reduce_DoctorRWM, resp.dist="bernoulli", select.dist = "normal",
                   alpha = seq(-5.5, -0.5, length.out = 21), standard = TRUE)

print(result0)

data(CreditMDR, package = "DiSSMod")
n1 <- 600
set.n1 <- sample(1:nrow(CreditMDR), n1)
reduce_CreditMDR <- CreditMDR[set.n1,]
result1 <- DiSSMod(response = MAJORDRG ~ AGE + INCOME + EXP_INC,
                   selection = CARDHLDR ~ AGE + INCOME + OWNRENT + ADEPCNT + SELFEMPL,
                   data = reduce_CreditMDR, resp.dist="poi", select.dist = "logis",
                  alpha = seq(-0.3, 0.3,length.out = 21), standard = FALSE, verbose = 1)

print(result1)
```

DoctorRWM *German doctor first visits data*

Description

Data is from Riphahn, Wambach and Million (2003), used for studying longitudinal analysis concerning the usage of the German health insurance system. The original data contain a few years data for patients, but we have only for first year.

Usage

```
data(DoctorRWM)
```

Format

A data frame with 7293 observations of 26 variables as below;

ID identification number (numeric)
FEMALE female or not (categorical)
YEAR year (categorical)
AGE age (numeric)
HSAT health satisfaction coded 0 (low) to 10 (high) (numeric)
HANDDUM person is handicappe or not (categorical)
HANDPER percentage degree of handicap (numeric)
HHNINC monthly household net income (numeric)
HHKIDS child (ren) below age 16 in household (numeric)
EDUC years of schooling (numeric)
MARRIED person is married or not (categorical)
HAUPTS level of schooling (categorical)
REALS level of schooling (categorical)
FACHHS level of schooling (categorical)
ABITUR level of schooling (categorical)
UNIV level of schooling (categorical)
WORKING employed or not (categorical)
BLUEC person is blue collar worker or not (categorical)
WHITEC person is white collar worker or not (categorical)
SELF person is self-employed or not (categorical)
BEAMT civil servant or not (categorical)
DOCVIS number of doctor visits in last 3 months (numeric)
HOSPVIS number of hospital visits last year (numeric)
PUBLIC person is insured in public health insurance or not (categorical)
ADDON person is insured in add-on insurance or not (categorical)
INCOME_SCALE scaled income; original income/1000 (numeric)

Source

Riphahn, R. T., Wambach, A. and Million, A. (2003) Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation, *Journal of Applied Econometrics*, **18**, 4, 387–405. Published online 8 October 2002. https://doi.org/10.1002/jae.680

http://qed.econ.queensu.ca/jae/2003-v18.4/riphahn-wambach-million/

References

Greene, W. H. (2012) *Econometric Analysis, 7th Edition*. Pearson education.

Azzalini, A., Kim, H.-M. and Kim, H.-J. (2019) Sample selection models for discrete and other non-Gaussian response variables. *Statistical Methods & Applications*, **28**, 27–56. First online 30 March 2018. https://doi.org/10.1007/s10260-018-0427-1

plot.DiSSMod *Relative Log Likelihood Plot for Discrete Sample Selection Model Fits*

Description

plot method for a class "DiSSMod".

Usage

```
## S3 method for class 'DiSSMod'
plot(x, ...)
```

Arguments

x	an object of class "DiSSMod" made by the function DiSSMod.
...	additional control argument is as follows.

- level: an option for controlling the significance level of confidence interval. It has to be given in probability between 0 and 1. Initial level is set to $1 - \alpha = 0.95$.

Details

Function plot draws a convex line due to the values of twice relative log likelihoods by using the profile likelihood approach with following the grids of alpha. If confidence interval created from the function confint exists between the maximum and minimum value of the alpha, there will be two points drawn with the color red. Also, the Maximum Likelihood Estimator (MLE) of alpha can be seen easily, if it exists between the maximum and minimum value of the alpha.

See Also

See also DiSSMod and plot.

Examples

```
# example continued from DiSSMod
set.seed(45)
data(DoctorRWM, package = "DiSSMod")
n0 <- 600
set.n0 <- sample(1:nrow(DoctorRWM), n0)
reduce_DoctorRWM <- DoctorRWM[set.n0,]
result0 <- DiSSMod(response = as.numeric(DOCVIS > 0) ~ AGE + INCOME_SCALE + HHKIDS + EDUC + MARRIED,
                   selection = PUBLIC ~ AGE + EDUC + FEMALE,
                  data = reduce_DoctorRWM, resp.dist="bernoulli", select.dist = "normal",
                   alpha = seq(-5.5, -0.5, length.out = 21), standard = TRUE)

plot(result0, level = 0.90)

data(CreditMDR, package = "DiSSMod")
```

```
n1 <- 600
set.n1 <- sample(1:nrow(CreditMDR), n1)
reduce_CreditMDR <- CreditMDR[set.n1,]
result1 <- DiSSMod(response = MAJORDRG ~ AGE + INCOME + EXP_INC,
                   selection = CARDHLDR ~ AGE + INCOME + OWNRENT + ADEPCNT + SELFEMPL,
                   data = reduce_CreditMDR, resp.dist="poi", select.dist = "logis",
                  alpha = seq(-0.3, 0.3,length.out = 21), standard = FALSE, verbose = 1)

plot(result1)
```

---

summary.DiSSMod *Summarizing Discrete Sample Selection Model Fits*

Description

summary method for a class "DiSSMod".

Usage

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

## S3 method for class 'summary.DiSSMod'
print(x, digits = max(3, getOption("digits") -
  3), ...)
```

Arguments

object	an object of class "DiSSMod" made by the function DiSSMod.
...	additional control argument is as follows.
	• level: an option for controlling the significance level of confidence interval. It has to be given in probability between 0 and 1. Initial level is set to $1 - \alpha = 0.95$.
x	an object of class "summary.DiSSMod".
digits	a numeric number of significant digits.

Details

If standard equals TRUE, summary also additionally returns summary statistics of standardized results. Otherwise, it just returns summary statistics as similar statistics as the generic function summary.

Value

The function summary.DiSSMod returns a list of summary statistics of the fitted discrete sample selection model given in object.

The components, which are not duplicated from the object, are as follows:

z.value_response
: Z statistics (normal distribution) for coefficients of response equation.

z.value_selection
: Z statistics (normal distribution) for coefficients of selection equation.

CI_alpha
: confidence interval of the parameter alpha.

level
: a numeric value between 0 and 1 for controlling the significance level of confidence interval. Initial level is set to $1 - \alpha = 0.95$.

See Also

See also DiSSMod and summary.

Examples

```
# example continued from DiSSMod
set.seed(45)
data(DoctorRWM, package = "DiSSMod")
n0 <- 600
set.n0 <- sample(1:nrow(DoctorRWM), n0)
reduce_DoctorRWM <- DoctorRWM[set.n0,]
result0 <- DiSSMod(response = as.numeric(DOCVIS > 0) ~ AGE + INCOME_SCALE + HHKIDS + EDUC + MARRIED,
                   selection = PUBLIC ~ AGE + EDUC + FEMALE,
                  data = reduce_DoctorRWM, resp.dist="bernoulli", select.dist = "normal",
                   alpha = seq(-5.5, -0.5, length.out = 21), standard = TRUE)

summary(result0, level = 0.90)

data(CreditMDR, package = "DiSSMod")
n1 <- 600
set.n1 <- sample(1:nrow(CreditMDR), n1)
reduce_CreditMDR <- CreditMDR[set.n1,]
result1 <- DiSSMod(response = MAJORDRG ~ AGE + INCOME + EXP_INC,
                    selection = CARDHLDR ~ AGE + INCOME + OWNRENT + ADEPCNT + SELFEMPL,
                    data = reduce_CreditMDR, resp.dist="poi", select.dist = "logis",
                   alpha = seq(-0.3, 0.3,length.out = 21), standard = FALSE, verbose = 1)

summary(result1)
```

Index

∗Topic **datasets**
CreditMDR, 4
DoctorRWM, 8

coef, *2*
coef.DiSSMod, 2, *8*
confint, *3*
confint.DiSSMod, 3, *8*
CreditMDR, 4

DiSSMod, *2, 3*, 5, *10*, *12*
DoctorRWM, 8

formula, *8*

glm, *8*

lm, *8*
logLik, *8*

plot, *10*
plot.DiSSMod, *8*, 10
print.summary.DiSSMod
*(*summary.DiSSMod*)*, 11

summary, *12*
summary.DiSSMod, *3*, *8*, 11