Package 'detectseparation'

March 25, 2020

Title Detect and Check for Separation and Infinite Maximum Likelihood Estimates

Version 0.1

Description Provides pre-fit and post-fit methods for detecting separation and infinite maximum likelihood estimates in generalized linear models with categorical responses. The pre-fit methods apply on binomial-response generalized liner models such as logit, probit and cloglog regression, and can be directly supplied as fitting methods to the glm() function. They solve the linear programming problems for the detection of separation developed in Konis (2007, <https://ora.ox.ac.uk/objects/uuid:8f9ee0d0-d78e-4101-9ab4-f9cbceed2a2a>) using 'ROI' <https://cran.r-project.org/package=ROI> or 'lpSolveAPI' <https://cran.r-project.org/package=lpSolveAPI>. The post-fit methods apply to models with categorical responses, including binomial-response generalized linear models and multinomial-response models, such as baseline category logits and adjacent category logits models; for example, the models implemented in the 'brglm2' <https://cran.r-project.org/package=brglm2> package. The post-fit methods successively refit the model with increasing number of iteratively reweighted least squares iterations, and monitor the ratio of the estimated standard error for each parameter to what it has been in the first iteration. According to the results in Lesaffre & Albert (1989, <https://www.jstor.org/stable/2345845>), divergence of those ratios indicates data separation.

URL https://github.com/ikosmidis/detectseparation

BugReports https://github.com/ikosmidis/detectseparation/issues

Imports ROI, ROI.plugin.lpsolve, lpSolveAPI, pkgload

Depends R (>= 3.3.0)

License GPL-3

Encoding UTF-8

LazyData true

RoxygenNote 7.0.2

Suggests testthat, knitr, rmarkdown, covr, brglm2, AER, ROI.plugin.ecos, ROI.plugin.glpk, ROI.plugin.neos, ROI.plugin.alabama

VignetteBuilder knitr

NeedsCompilation no

Author Ioannis Kosmidis [aut, cre] (<https://orcid.org/0000-0003-1556-0302>),
Dirk Schumacher [aut],
Kjell Konis [ctb]

Maintainer Ioannis Kosmidis <ioannis.kosmidis@warwick.ac.uk>

Repository CRAN

Date/Publication 2020-03-25 16:00:02 UTC

R topics documented:

`check_infinite_estimates` *Generic method for checking for infinite estimates*

Description

Generic method for checking for infinite estimates

Usage

```
check_infinite_estimates(object, ...)
```

Arguments

`object`	a fitted model object (e.g. the result of a `glm` call).
`...`	other options to be passed to the method.

See Also

check_infinite_estimates.glm

`check_infinite_estimates.glm` *A simple diagnostic of whether the maximum likelihood estimates are infinite*

Description

A simple diagnostic of whether the maximum likelihood estimates are infinite

Usage

```
## S3 method for class 'glm'
check_infinite_estimates(object, nsteps = 20, ...)
```

Arguments

`object`	the result of a `glm` call.
`nsteps`	starting from `maxit = 1`, the GLM is refitted for `maxit = 2`, `maxit = 3`, ..., `maxit = nsteps`. Default value is 30.
`...`	currently not used.

Details

`check_infinite_estimates` attempts to identify the occurence of infinite estimates in GLMs with binomial responses by successively refitting the model. At each iteration the maximum number of allowed IWLS iterations is fixed starting from 1 to `nsteps` (by setting `control = glm.control(maxit = j)`, where `j` takes values 1, ..., nsteps in `glm`). For each value of `maxit`, the estimated asymptotic standard errors are divided to the corresponding ones from `control = glm.control(maxit = 1)`. Then, based on the results in Lesaffre & Albert (1989), if the sequence of ratios in any column of the resultant matrix diverges, then complete or quasi-complete separation occurs and the maximum likelihood estimate for the corresponding parameter has value minus or plus infinity.

Value

An object of class `inf_check` that has a `plot` method.

A matrix inheriting from class `inf_check`, with `nsteps` rows and `p` columns, where `p` is the number of model parameters. A `plot` method is provided for `inf_check` objects for the easy inspection of the ratios of the standard errors.

Note

For the definition of complete and quasi-complete separation, see Albert and Anderson (1984). Kosmidis and Firth (2019) prove that the reduced-bias estimator that results by the penalization of the logistic regression log-likelihood by Jeffreys prior takes always finite values, even when some of the maximum likelihood estimates are infinite. The reduced-bias estimates can be computed using the **brglm2** R package.

References

Lesaffre, E., & Albert, A. (1989). Partial Separation in Logistic Discrimination. *Journal of the Royal Statistical Society. Series B (Methodological)*, **51**, 109-116

Kosmidis I. and Firth D. (2019). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. arXiv:1812.01938. https://arxiv.org/abs/1812.01938v3

See Also

multinom, detect_separation, brmultinom

Examples

```
## endometrial data from Heinze & Schemper (2002) (see ?endometrial)
data("endometrial", package = "detectseparation")
endometrial_ml <- glm(HG ~ NV + PI + EH, data = endometrial,
                      family = binomial("probit"))
## clearly the maximum likelihood estimate for the coefficient of
## NV is infinite
(estimates <- check_infinite_estimates(endometrial_ml))
plot(estimates)



## Aligator data (Agresti, 2002, Table~7.1)
if (requireNamespace("brglm2", quietly = TRUE)) {
    data("alligators", package = "brglm2")
    all_ml <- brglm2::brmultinom(foodchoice ~ size + lake , weights = round(freq/3),
                         data = alligators, type = "ML", ref = 1)
    ## Clearly some estimated standard errors diverge as the number of
    ## Fisher scoring iterations increases
    plot(check_infinite_estimates(all_ml))
    ## Bias reduction the brglm2 R packages can be used to get finite estimates
    all_br <- brglm2::brmultinom(foodchoice ~ size + lake , weights = round(freq/3),
                         data = alligators, ref = 1)
    plot(check_infinite_estimates(all_br))
}
```

`detectseparation` *detectseparation: Methods for Detecting and Checking for Separation and Infinite Maximum Likelihood Estimates*

Description

detectseparation: Methods for Detecting and Checking for Separation and Infinite Maximum Likelihood Estimates

See Also

detect_separation, check_infinite_estimates

---

detect_separation — *Method for glm that tests for data separation and finds which parameters have infinite maximum likelihood estimates in generalized linear models with binomial responses*

Description

detect_separation is a method for glm that tests for the occurrence of complete or quasi-complete separation in datasets for binomial response generalized linear models, and finds which of the parameters will have infinite maximum likelihood estimates. detect_separation relies on the linear programming methods developed in Konis (2007).

Usage

```
detect_separation(
  x,
  y,
  weights = rep(1, nobs),
  start = NULL,
  etastart = NULL,
  mustart = NULL,
  offset = rep(0, nobs),
  family = gaussian(),
  control = list(),
  intercept = TRUE,
  singular.ok = TRUE
)

detectSeparation(
  x,
  y,
  weights = rep(1, nobs),
  start = NULL,
  etastart = NULL,
  mustart = NULL,
  offset = rep(0, nobs),
  family = gaussian(),
  control = list(),
  intercept = TRUE,
  singular.ok = TRUE
)
```

Arguments

x	x is a design matrix of dimension n * p.
y	y is a vector of observations of length n.
weights	an optional vector of 'prior weights' to be used in the fitting process. Should be NULL or a numeric vector.
start	currently not used.
etastart	currently not used.
mustart	currently not used.
offset	this can be used to specify an *a priori* known component to be included in the linear predictor during fitting. This should be NULL or a numeric vector of length equal to the number of cases. One or more offset terms can be included in the formula instead or as well, and if more than one is specified their sum is used. See model.offset.
family	a description of the error distribution and link function to be used in the model. For glm this can be a character string naming a family function, a family function or the result of a call to a family function. For glm.fit only the third option is supported. (See family for details of family functions.)
control	a list of parameters controlling separation detection. See detect_separation_control for details.
intercept	logical. Should an intercept be included in the *null* model?
singular.ok	logical. If FALSE, a singular model is an error.

Details

detect_separation is a wrapper to the separator_ROI function and separator_lpSolveAPI function (a modified version of the separator function from the **safeBinaryRegression** R package). detect_separation can be passed directly as a method to the glm function. See, examples.

The coefficients method extracts a vector of values for each of the model parameters under the following convention: 0 if the maximum likelihood estimate of the parameter is finite, and Inf or -Inf if the maximum likelihood estimate of the parameter if plus or minus infinity. This convention makes it easy to adjust the maximum likelihood estimates to their actual values by element-wise addition.

detectSeparation is an alias for detect_separation.

Value

A list that inherits from class detect_separation, glm and lm. A print method is provided for detect_separation objects.

Note

For the definition of complete and quasi-complete separation, see Albert and Anderson (1984). Kosmidis and Firth (2019) prove that the reduced-bias estimator that results by the penalization of the logistic regression log-likelihood by Jeffreys prior takes always finite values, even when some of

the maximum likelihood estimates are infinite. The reduced-bias estimates can be computed using the **brglm2** R package.

detect_separation was designed in 2017 by Ioannis Kosmidis for the **brglm2** R package, after correspondence with Kjell Konis, and a port of the separator function had been included in **brglm2** under the permission of Kjell Konis.

In 2020, detect_separation and check_infinite_estimates were moved outside **brglm2** into the dedicated **detectseparation** package. Dirk Schumacher authored the separator_ROI function, which depends on the **ROI** R package and is now the default implementation used for detecting separation.

Author(s)

Ioannis Kosmidis [aut, cre] <ioannis.kosmidis@warwick.ac.uk>, Dirk Schumacher [aut] <mail@dirk-schumacher.net> Kjell Konis [ctb] <kjell.konis@me.com>

References

Konis K. (2007). *Linear Programming Algorithms for Detecting Separated Data in Binary Logistic Regression Models*. DPhil. University of Oxford. https://ora.ox.ac.uk/objects/uuid:8f9ee0d0-d78e-4101-9ab4-f9cbceed2a2a

Konis K. (2013). safeBinaryRegression: Safe Binary Regression. R package version 0.1-3. https://CRAN.R-project.org/package=safeBinaryRegression

Kosmidis I. and Firth D. (2019). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. arXiv:1812.01938. https://arxiv.org/abs/1812.01938v3

See Also

glm.fit and glm, check_infinite_estimates, brglm_fit,

Examples

```
## endometrial data from Heinze & Schemper (2002) (see ?endometrial)
data("endometrial", package = "detectseparation")
endometrial_sep <- glm(HG ~ NV + PI + EH, data = endometrial,
                       family = binomial("logit"),
                       method = "detect_separation")
endometrial_sep
## The maximum likelihood estimate for NV is infinite
summary(update(endometrial_sep, method = "glm.fit"))


## Example inspired by unpublished microeconometrics lecture notes by
## Achim Zeileis https://eeecon.uibk.ac.at/~zeileis/
## The maximum likelihood estimate of sourhernyes is infinite
if (requireNamespace("AER", quietly = TRUE)) {
    data("MurderRates", package = "AER")
    murder_sep <- glm(I(executions > 0) ~ time + income +
                      noncauc + lfp + southern, data = MurderRates,
                      family = binomial(), method = "detect_separation")
```

```
    murder_sep
    ## which is also evident by the large estimated standard error for NV
    murder_glm <- update(murder_sep, method = "glm.fit")
    summary(murder_glm)
    ## and is also reveal by the divergence of the NV column of the
    ## result from the more computationally intensive check
    plot(check_infinite_estimates(murder_glm))
    ## Mean bias reduction via adjusted scores results in finite estimates
    if (requireNamespace("brglm2", quietly = TRUE))
        update(murder_glm, method = brglm2::brglm_fit)
}
```

`detect_separation_control` *Auxiliary function for the* `glm` *interface when* `method` *is* `detect_separation`.

Description

Typically only used internally by `detect_separation` but may be used to construct a `control` argument.

Usage

```
detect_separation_control(
  implementation = c("ROI", "lpSolveAPI"),
  solver = "lpsolve",
  linear_program = c("primal", "dual"),
  purpose = c("find", "test"),
  tolerance = 1e-04,
  solver_control = list()
)

detectSeparationControl(
  implementation = c("ROI", "lpSolveAPI"),
  solver = "lpsolve",
  linear_program = c("primal", "dual"),
  purpose = c("find", "test"),
  tolerance = 1e-04,
  solver_control = list()
)
```

Arguments

`implementation`	should the implementation using `ROI` or the implementation using `lpSolveAPI` be used? Default is `ROI`.

solver	should the linear program be solved using the "lpsolve" (using the **ROI.plugin.lpsolve** package; default) or another solver? Alternative solvers are "glpk", "cbc", "clp", "cplex", "ecos", "gurobi", "scs", "symphony". If **ROI.plugin.[solver]** is not installed then the user will be prompted to install it before continuing.
linear_program	should detect_separation solve the "primal" (default) or "dual" linear program for separation detection? Only relevant if implementation = "lpSolveAPI".
purpose	should detect_separation simply "test" for separation or also "find" (default) which parameters are infinite? Only relevant if implementation = "lpSolveAPI".
tolerance	maximum absolute variable value from the linear program, before separation is declared. Default is 1e-04.
solver_control	a list with additional control parameters for the "solver". This is solver specific, so consult the corresponding documentation. Default is list() unless solver is "alabama" when the default is list(start = rep(0,p)), where p is the number of parameters.

Value

A list with the supplied linear_program, solver, solver_control, purpose, tolerance, implementation, and the matched separator function (according to the value of implementation).

---

endometrial *Histology grade and risk factors for 79 cases of endometrial cancer*

Description

Histology grade and risk factors for 79 cases of endometrial cancer

Usage

```
endometrial
```

Format

A data frame with 79 rows and 4 variables:

NV neovasculization with coding 0 for absent and 1 for present

PI pulsality index of arteria uterina

EH endometrium height

HG histology grade with coding 0 for low grade and 1 for high grade

Source

The packaged data set was downloaded in .dat format from http://www.stat.ufl.edu/~aa/glm/data. The latter link provides the data sets used in Agresti (2015).

The endometrial data set was first analyzed in Heinze and Schemper (2002), and was originally provided by Dr E. Asseryanis from the Medical University of Vienna.

References

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Wiley Series in Probability and Statistics. Wiley

Heinze, G., & Schemper, M. (2002). A Solution to the Problem of Separation in Logistic Regression. *Statistics in Medicine*, **21**, 2409–2419

See Also

brglm_fit

`lizards` *Habitat preferences of lizards*

Description

The lizards data frame has 23 rows and 6 columns. Variables `grahami` and `opalinus` are counts of two lizard species at two different perch heights, two different perch diameters, in sun and in shade, at three times of day.

Usage

```
lizards
```

Format

An object of class `data.frame` with 23 rows and 6 columns.

Details

- grahami. count of grahami lizards
- opalinus. count of opalinus lizards
- height. a factor with levels `<5ft`, `>=5ft`
- diameter. a factor with levels `<=2in`, `>2in`
- light. a factor with levels `sunny`, `shady`
- time. a factor with levels `early`, `midday`, `late`

Source

McCullagh, P. and Nelder, J. A. (1989) _Generalized Linear Models_ (2nd Edition). London: Chapman and Hall.

Originally from

Schoener, T. W. (1970) Nonsynchronous spatial overlap of lizards in patchy habitats. _Ecology_ *51*, 408-418.

See Also

brglm_fit

Index