Package 'devianLM'

September 24, 2025

Type Package

Title Detecting Extremal Values in a Normal Linear Model

Version 1.0.3

Date 2025-09-12

Description Provides a method to detect values poorly explained by a Gaussian linear model. The procedure is based on the maximum of the absolute value of the studentized residuals, which is a parameter-free statistic. This approach generalizes several procedures used to detect abnormal values during longitudinal monitoring of biological markers. For methodological details, see: Berthelot G., Saulière G., Dedecker J. (2025). ``DEViaN-LM An R Package for Detecting Abnormal Values in the Gaussian Linear Model". HAL Id: hal-05230549. <https://hal.science/hal-05230549>.

License GPL-3

Encoding UTF-8

Imports Rcpp

LinkingTo Rcpp, RcppArmadillo

Suggests testthat (>= 3.0.0)

Config/testthat/edition 3

RoxygenNote 7.3.1

Depends R (>= 2.10)

LazyData true

SystemRequirements OpenMP

NeedsCompilation yes

Author Guillaume Sauliere [aut, cre] (ORCID: <https://orcid.org/0000-0001-8263-6456>), Geoffroy Berthelot [aut], Jérôme Dedecker [aut]

Maintainer Guillaume Sauliere <guillaumesauliere@hotmail.com>

Repository CRAN

Date/Publication 2025-09-24 08:10:02 UTC

Contents

---

devianLM-package *Detection of Poorly Explained Values in Gaussian Linear Models*

Description

The **devianLM** package provides tools to detect values that are poorly explained by a Gaussian linear model. The method is based on the maximum absolute value of studentized residuals, a statistic that is independent of the model parameters. This approach generalizes several procedures used to detect abnormal values, such as during the longitudinal monitoring of certain biological markers.

Details

The package offers two main functions:

- `get_devianlm_threshold`: Computes the detection threshold via Monte Carlo simulations.
- `devianlm_stats`: Fits a Gaussian linear model and flags potential outliers based on the computed threshold.

These methods are particularly useful for regression diagnostics, quality control, and longitudinal monitoring in applied statistics.

Author(s)

Guillaume Saulière <guillaumesauliere@hotmail.com> \ Geoffroy Berthelot <geoffroy.berthelot@insep.fr> \ Jérôme Dedecker <jerome.dedecker@u-paris.fr> \

Examples

```
set.seed(123)
x <- as.matrix(rnorm(50))
y <- 2 * x + rnorm(50)

# Small n_sims for quick example
result <- devianlm_stats(y, x, n_sims = 100)
```

`devianlm_stats` *Identify outliers using devianLM method*

Description

Identify outliers using devianLM method

Usage

```
devianlm_stats(
  y,
  x,
  threshold = NULL,
  n_sims = 50000,
  nthreads = detectCores() - 1,
  alpha = 0.95,
  ...
)
```

Arguments

`y`	a numeric variable
`x`	either a numeric variable or several numeric variables (explanatory variables) concatenated in a data frame.
`threshold`	numeric or NULL; if NULL, computed using devianlm_cpp()
`n_sims`	optional value which is the number of simulations, is set to 50.000 by default.
`nthreads`	optional value which is the number of CPU cores to use, is set to "number of CPU cores - 1" by default.
`alpha`	quantile of interest, is set to 0.95 by default.
`...`	additional arguments for get_devianlm_threshold()

Value

devianlm returns an object of class *list* with the following components:

reg_residuals Numeric vector. The studentized residuals from the linear model.

outliers Integer vector. The indices (positions in the original data) of observations identified as outliers based on the threshold.

threshold Numeric value. The cutoff applied to the absolute value of the studentized residuals to flag outliers. If not provided, it is estimated using `get_devianlm_threshold()`.

is_outliers Integer vector. A binary vector (0 or 1) of the same length as `reg_residuals`, indicating whether each observation is considered an outlier (1) or not (0).

Examples

```
set.seed(123)
y <- salary$hourly_earnings_log
x <- cbind(salary$age, salary$educational_attainment, salary$children_number)

test_salary <- devianlm_stats(y, x, n_sims = 100, alpha = 0.95)

plot(test_salary$reg_residuals,
  pch = 16, cex = .8,
 ylim = c(-1 * max(abs(test_salary$reg_residuals)), max(abs(test_salary$reg_residuals))),
  xlab = "", ylab = "Studentized residuals",
  col = ifelse(test_salary$is_outliers, "red", "black"))

# Ajouter les lignes de seuil
abline(h = c(-test_salary$threshold, test_salary$threshold), col = "chartreuse2", lwd = 2)
```

`get_devianlm_threshold` *get_devianlm_threshold : Compute threshold using Monte Carlo simulations*

Description

This package determines whether the maximum of the absolute values of the studentized residuals of a Gaussian regression is abnormally high. The distribution of the maximum of the absolute of the studentized residuals (depending on the design matrix) is computed via Monte-Carlo simulations (with n_sims simulations).

Usage

```
get_devianlm_threshold(
  x,
  n_sims = 50000,
  nthreads = detectCores() - 1,
  alpha = 0.95
)
```

Arguments

`x`	either a numeric variable or several numeric variables (explanatory variables) concatenated in a data frame.
`n_sims`	optional value which is the number of simulations, is set to 50.000 by default.
`nthreads`	optional value which is the number of CPU cores to use, is set to "number of CPU cores - 1" by default.
`alpha`	quantile of interest, is set to 0.95 by default.

Value

Numeric value.

`threshold` The quantile of order 1-alpha of the distribution of the maximum of the absolute of the studentized residuals (depending on the design matrix) is computed via Monte-Carlo simulations (with n_sims simulations).

`salary` *Salary dataset*

Description

A random sample from the 2012 Current Population Survey (CPS). It is the primary source of labor force statistics for the US population.

- age. age of the individual (0–85)
- sex. sex of the individual ("F" = Female, "M" = Male)
- region. region ("NE" = Northeast, "W" = West, "S" = South, "NW" = Northwest)
- marital_status. marital status of the individual ("NM" = Never married, "M" = Married, "D" = Divorced, "S" = Separated, "W" = Widowed)
- hourly_earnings. how much does the individual earn per hour (00–9999)
- educational_attainment. educational attainment of the individual (0 = Children, 31 = Less than 1st grade, 32 = 1st,2nd,3rd,or 4th grade, 33 = 5th or 6th grade, 34 = 7th and 8th grade 35 = 9th grade, 36 = 10th grade, 37 = 11th grade, 38 = 12th grade no diploma, 39 = High school graduate - high school diploma or equivalent, 40 = Some college but no degree, 41 = Associate degree in college - occupation/vocation program, 42 = Associate degree in college - academic program 43 = Bachelor's degree (for example: BA,AB,BS), 44 = Master's degree (for example: MA,MS,MENG,MED,MSW, MBA), 45 = Professional school degree (for example:MD,DDS,DVM,LLB,JD) 46 = Doctorate degree (for example: PHD,EDD))
- persons_number. number of persons in household (0–16)
- children_number. number of children in household (0–9)
- family_income. family income from basic CPS income screener question (-1 = Not in universe, 01 = Less than $5,000, 02 = $5,000 to $7,499, 03 = $7,500 to $9,999 04 = $10,000 to $12,499, 05 = $12,500 to $14,999, 06 = $15,000 to $19,999, 07 = $20,000 to $24,999 08 = $25,000 to $29,999, 09 = $30,000 to $34,999, 10 = $35,000 to $39,999, 11 = $40,000 to $49,999 12 = $50,000 to $59,999, 13 = $60,000 to $74,999, 14 = $75,000 to $99,999, 15 = $100,000 to $149,999)
- hourly_earnings_log. log(hourly_earnings)

Usage

```
salary
```

Format

A data frame with 599 rows and 10 variables

See Also

Original data are available from <https://www.ilo.org/surveyLib/index.php/catalog/7379>.

The data dictionary is available from <https://www2.census.gov/programs-surveys/cps/datasets/2022/march/asec2022_ddl_p

Index