Package 'devianLM'

May 8, 2026

Type Package

Title Detecting Extremal Values in a Normal Linear Model

Version 1.1.0

Date 2026-04-21

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 7.3.3

Depends R (>= 2.10)

LazyData true

SystemRequirements OpenMP (optional, for parallel execution)

NeedsCompilation yes

Repository CRAN

Date/Publication 2026-04-30 09:40:02 UTC

Author Guillaume Sauliere [aut] (ORCID: <https://orcid.org/0000-0001-8263-6456>), Geoffroy Berthelot [aut, cre] (ORCID: <https://orcid.org/0000-0003-4036-6114>), Jérôme Dedecker [aut] (ORCID: <https://orcid.org/0000-0002-8838-0356>)

Maintainer Geoffroy Berthelot <geoffroy.berthelot@insep.fr>

Contents

`devianLM-package` *devianLM: Outlier detection via studentized residuals*

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 (conditional on the design matrix), which is a parameter-free statistic. This approach generalizes several procedures used to detect abnormal values during longitudinal monitoring of biological markers.

Details

The main R functions are:

- `get_devianlm_threshold`: computes the threshold via Monte-Carlo simulations.
- `devianlm_stats`: identifies outliers based on this threshold. It is also possible to use a custom threshold value, specified by the user.

Monte Carlo simulations are parallelized using OpenMP when available.

Author(s)

Maintainer: Geoffroy Berthelot <geoffroy.berthelot@insep.fr> (ORCID)

Authors:

- Guillaume Sauliere <guillaumesauliere@hotmail.com> (ORCID)
- Jérôme Dedecker <jerome.dedecker@u-paris.fr> (ORCID)

References

Sauliere, G., Berthelot, G., and Dedecker, J. (2025) *DEViaN-LM An R Package for Detecting Abnormal Values in the Gaussian Linear Model*. https://doi.org/10.48550/arXiv.2509.02202 doi:10.48550/arXiv.2509.02202

Berthelot G., Gelein B., Meinadier E. Orhant E., Dedecker J., (2025) *Z-scores-based methods and their application to biological monitoring: An extended analysis of professional soccer players and cyclists athletes*. https://arxiv.org/abs/2510.01810 doi:10.48550/arXiv.2510.01810

devianlm_stats *Identify outliers using devianLM method*

Description

This function determines whether the maximum of the absolute values of the studentized residuals of a Gaussian regression is abnormally high. Outliers are detected by comparing the absolute values of the studentized residuals to a threshold (depending on the design matrix), which can be supplied or estimated via `n_sims` Monte-Carlo simulations.

Usage

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

Arguments

`y`	Numeric. Response vector.
`x`	either a numeric variable or several numeric variables (explanatory variables) concatenated in a data frame. `devianLM` does not add an intercept automatically; include a column of ones in `x` if an intercept is desired.
`threshold`	Numeric or NULL. If NULL, the threshold value is computed using `get_devianlm_threshold()`.
`n_sims`	Integer. Optional value which is the number of Monte-Carlo simulations. Default is 50,000.
`verbose`	Logical. If TRUE (default), informative messages are printed during execution (e.g., when ties are detected and handled).
`nthreads`	Integer. Optional value which is the number of threads to use. Default is `parallel::detectCores() - 1`.
`quant`	Numeric. Order of the quantile of interest. Default is 0.95 (this corresponds to a risk level of 0.05).
`...`	Additional arguments passed to `get_devianlm_threshold()`.

Details

When ties are present in `y`, a small random perturbation is added to avoid numerical issues. The "Ties were detected in the data, they have been randomly broken" message is displayed when this occurs.

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$hourly_earnings_log
x <- cbind(1, salary$age, salary$educational_attainment, salary$children_number)

test_salary <- devianlm_stats(y, x, n_sims = 100, quant = 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"))

# Add the thresholds lines:
abline(h = c(-test_salary$threshold, test_salary$threshold), col = "chartreuse2", lwd = 2)
```

`get_devianlm_threshold` *Estimate threshold value via Monte-Carlo simulations.*

Description

Estimates the threshold for the maximum absolute studentized residual in a Gaussian linear model, conditional on the design matrix and using `n_sims` Monte-Carlo simulation for the quantile of order `quant`.

Usage

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

Arguments

x	either a numeric variable or several numeric variables (explanatory variables) concatenated in a data frame. **Note:** devianLM does not add an intercept automatically; include a column of ones in x if an intercept is desired.
n_sims	Integer. Optional value which is the number of Monte-Carlo simulations. Default is 50,000.
nthreads	Integer. Optional value which is the number of threads to use. Default is parallel::detectCores() - 1.
quant	Numeric. Order of the quantile of interest. Default is 0.95 (this corresponds to a risk level of 0.05).

Details

Monte-Carlo simulations are parallelized using OpenMP when available.

Value

Numeric value.

threshold	The quantile of order quant of the distribution of the maximum of the absolute values 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://webapps.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_pub_full.pdf.

Index