Package ‘rvif’

November 14, 2024

Type Package

Title Collinearity Detection using Redefined Variance Inflation Factor and Graphical Methods

Version 2.0

Date 2024-11-04

Author R. Salmerón [aut, cre],
C.B. García [aut]

Maintainer R. Salmerón <romansg@ugr.es>

Description The detection of troubling approximate collinearity in a multiple linear regression model is a classical problem in Econometrics. The objective of this package is to detect it using the variance inflation factor redefined and the scatterplot between the variance inflation factor and the coefficient of variation. For more details see Salmerón R., García C.B. and García J. (2018) <doi:10.1080/00949655.2018.1463376>, Salmerón, R., Rodríguez, A. and García C. (2020) <doi:10.1007/s00180-019-00922-x>, Salmerón, R., García, C.B, Rodríguez, A. and García, C. (2022) <doi:10.32614/RJ-2023-010>, Salmerón, R., García, C.B. and García, J. (2024) <doi:10.1007/s10614-024-10575-8> and Salmerón, R., García, C.B, García J. (2023, working paper) <doi:10.48550/arXiv.2005.02245>.

License GPL (>= 2)

Encoding UTF-8

URL http://colldetreat.r-forge.r-project.org/

Depends R (>= 3.5.0), multiColl

NeedsCompilation no

Repository CRAN

Date/Publication 2024-11-14 11:50:03 UTC

Contents

rvif-package 2
CDpf 3
CV_VIF 4
employees 6

euribor 7
RVIF 8
SLM1 9
SLM2 10
Theorem 11
Wissel 12

Index **14**

rvif-package *Multicollinearity detection using RVIF and graphical methods*

Description

The detection of troubling near multicollinearity in a multiple linear regression model is a classical problem in Econometrics. The purpose of this package is to detect it by using the Redefined Variance Inflation Factor (RVIF) and the scatterplot between the Variance Inflation Factor (VIF) and the Coefficient of Variation (CV).

In addition, the RVIF is used to determine whether the statistical analysis of the model is affected by the degree of multicollinearity of the model.

Details

This package contains three functions. The first, CV_VIF, returns the values of the Variance Inflation Factor (VIF) and the Coefficient of Variation (CV), as well as their representation in a scatterplot. Taking into account that the VIF is useful for detecting essential multicollinearity and the CV is useful for detecting non-essential multicollinearity, the scatterplot of both measures can provide interesting information for determining whether there is a troubling degree of multicollinearity, what kind of multicollinearity it is and what variables are causing the multicollinearity.

On the other hand, the funcion RVIF calculates the redefined VIF, the percentage of near multicollinearity due to each independent variable and, using the above function, the scatterplot between the CV and VIF.

Finally, Theorem determines whether the degree of multicollinearity in the regression model affects the statistical analysis of the model, i.e., whether the non-rejection of the null hypothesis in the individual significance tests is due to the linear relationships between the independent variables of the model.

Author(s)

Román Salmerón Gómez (University of Granada) and Catalina García García (University of Granada).

Maintainer: Román Salmerón Gómez (romansg@ugr.es)

References

R. Salmerón, C. García, and J. García. Variance inflation factor and condition number in multiple linear regression. Journal of Statistical Computation and Simulation, 88:2365-2384, 2018.

R. Salmerón, A. Rodríguez, and C. García. Diagnosis and quantification of the non-essential collinearity. Computational Statistics, 35:647-666, 2020.

Salmerón, R., García, C.B., Rodríguez, A. and García, C. Limitations in detecting multicollinearity due to scaling issues in the mcvis package. R Journal, 14(4), 264-279, 2022.

Salmerón, R., García, C.B. and García, J. A redefined Variance Inflation Factor: overcoming the limitations of the Variance Inflation Factor. Computational Economics (2024, online), doi: https://doi.org/10.1007/s10614-024-10575-8.

Overcoming the inconsistences of the variance inflation factor: a redefined VIF and a test to detect statistical troubling multicollinearity by Salmerón, R., García, C.B and García, J. (working paper, https://arxiv.org/pdf/2005.02245).

`CDpf` *Cobb-Douglas data*

Description

Data used in Example 2 of Salmerón, García and García (2024) (sub-section 4.2) on data for the Cobb-Douglas production function.

Usage

```
data("CDpf")
```

Format

A data frame with 28 observations on the following 4 variables:

`P` Production (dependent variable).

`cte` Intercept.

`logK` Capital (in logarithm).

`logW` Work (in logarithm).

Details

This dataset is originally used by Olva Maldonado (2009).

References

Olva Maldonado, H. Análisis de la función de producción Cobb-Douglas y su aplicación en el sector productivo mexicano. Tesis, Universidad Autónoma de Chapingo, 2009.

Salmerón, R., García, C.B. and García, J. A redefined Variance Inflation Factor: overcoming the limitations of the Variance Inflation Factor. Computational Economics (2024, online), doi: https://doi.org/10.1007/s10614-024-10575-8.

Examples

```
data(CDpf)
head(CDpf, n=5)
y = CDpf[,1]
X = as.matrix(CDpf[,2:4])
Theorem(y, X)
```

CV_VIF *VIF, CV and a common scatter plot*

Description

This function provides the values for the Variance Inflation Factor (VIF) and the Coefficient of Variation (CV), as well as a common representation of both.

Usage

```
CV_VIF(X, size=NULL, top=82.64, limit=40, dummy=FALSE, pos=NULL, intercept=TRUE)
```

Arguments

`X` A numerical design matrix that should contain more than one regressor (including the intercept).

`size` A numerical vector containing the percentage of multicollinearity due to each variable. By default `size=NULL`.

`top` A real number that indicates the threshold from which the percentage of multicollinearity due to each variable is considered troubling. By default `top=82.64`.

`limit` A real number that indicates the lower limit of the vertical axis. By default `limit=40`.

`dummy` A logical value that indicates if there are dummy variables in the design matrix `X`. By default `dummy=FALSE`.

`pos` A numerical vector indicating the position of the dummy variables, if any, in the design matrix `X`. By default `pos=NULL`.

`intercept` A logical value used only by the function RVIF. By default `intercept=TRUE`.

Details

It is interesting to note the distinction between essential (near-linear relationship between at least two independent variables excluding the intercept) and non-essential multicollinearity (near-linear relationship between the intercept and at least one of the remaining independent variables), due to the VIF is not an appropriate measure to detect non-essential collinearity (only detects essential collinearity), while the CV is useful to detect only non-essential collinearity.

Then, this distinction between essential and non-essential multicollinearity and the limitations of each measure for detecting the different kinds of multicollinearity, can be very useful to detect if there is a troubling degree of multicollinearity, what kind of multicollinearity it is and what variables are causing the multicollinearity.

For this purpose, it is important to include in the figures the lines corresponding to the established thresholds for each measure (CV and VIF): dashed vertical line for 0.1002506 (CV) and dotted horizontal line for 10 (VIF). These lines determine four regions (see Example 1) which can be interpreted as follows: A, existence of troubling non-essential and non-troubling essential multicollinearity; B, existence of troubling essential and non-essential multicollinearity; C, existence of non-troubling non-essential and troubling essential multicollinearity; D: non-troubling degree of existing multicollinearity (essential and non-essential).

Value

`CV`	Coefficient of Variation of each independent variable.
`VIF`	Variance Inflation Factor of each independent variable.

Author(s)

R. Salmerón (<`romansg@ugr.es`>) and C. García (<`cbgarcia@ugr.es`>).

References

R. Salmerón, C. García, and J. García. Variance inflation factor and condition number in multiple linear regression. Journal of Statistical Computation and Simulation, 88:2365-2384, 2018.

R. Salmerón, A. Rodríguez, and C. García. Diagnosis and quantification of the non-essential collinearity. Computational Statistics, 35:647-666, 2020.

Salmerón, R., García, C.B., Rodríguez, A. and García, C. Limitations in detecting multicollinearity due to scaling issues in the mcvis package. R Journal, 14(4), 264-279, 2022.

Examples

```
## Example 1
plot(-2:20, -2:20, type = "n", xlab="Coefficient of Variation", ylab="Variance Inflation Factor")
abline(h=10, col="black", lwd=3, lty=2)
abline(v=0.1002506, col="black", lwd=3, lty=3)
text(-1.25, 2, "A", pos=3, col="red")
text(-1.25, 12, "B", pos=3, col="red")
text(10, 12, "C", pos=3, col="red")
text(10, 2, "D", pos=3, col="red")

## Example 2
library(multiColl)
set.seed(2022)
obs = 100
cte = rep(1, obs)
x2 = rnorm(obs, 5, 0.01)
x3 = rnorm(obs, 5, 10)
x4 = x3 + rnorm(obs, 5, 1)
x5 = rnorm(obs, -1, 30)
x = cbind(cte, x2, x3, x4, x5)
CV_VIF(x, size = c(1, 1, 1, 1))
```

`employees` *Spanish company employee data*

Description

Data used in example 3 of Salmerón, García and García (2024) (subsection 4.3) on number of employees of Spanish companies.

Usage

```
data("employees")
```

Format

A data frame with 15 observations on the following 5 variables:

`NE` Number of employees (dependent variable).

`cte` Intercept.

`FA` Fixed assets (in euros).

`OI` Operating income (in euros).

`S` Sales (in euros).

Details

This dataset is originally used by Salmerón, Rodríguez, García and García (2020).

References

Salmerón, R., Rodríguez, A., García, C.B. and García, J. The VIF and MSE in raise regression. Mathematics, 8(4), 2020.

Salmerón, R., García, C.B. and García, J. A redefined Variance Inflation Factor: overcoming the limitations of the Variance Inflation Factor. Computational Economics (2024, online), doi: https://doi.org/10.1007/s10614-024-10575-8.

Examples

```
data(employees)
head(employees, n=5)
y = employees[,1]
X = as.matrix(employees[,3:5])
Theorem(y, X)
```

euribor *Euribor data*

Description

Data used in example 1 of Salmerón, García and García (2024) (subsection 4.1) on euribor data.

Usage

```
data("euribor")
```

Format

A data frame with 47 observations on the following 5 variables:

`E` Euribor (dependent variable, in percentage).

`cte` Intercept.

`HIPC` Harmonized index of consumer prices (in percentage).

`BC` Balance of payments to net current account (millions of euros).

`GD` Goverment deficit to net nonfinancial accounts (millions of euros).

Details

This dataset is originally used by Salmerón, Rodríguez and García (2020).

References

Salmerón, R., Rodríguez, A. and García, C.B. Diagnosis and quantification of the non-essential collinearity. Computational Statistics, 35(2), 647-666, 2020.

Salmerón, R., García, C.B. and García, J. A redefined Variance Inflation Factor: overcoming the limitations of the Variance Inflation Factor. Computational Economics (2024, online), doi: https://doi.org/10.1007/s10614-024-10575-8.

Examples

```
data(euribor)
head(euribor, n=5)
y = euribor[,1]
X = as.matrix(euribor[,2:5])
Theorem(y, X)
```

RVIF *RVIF calculation*

Description

This function provides the values of the Redefined Variance Inflation Factor (RVIF) and the the percentage of near multicollinearity due to each independent variable.

Usage

```
RVIF(X, l_u=TRUE, l=40, intercept=TRUE, graf=TRUE)
```

Arguments

`X`	A numerical design matrix that should contain more than one regressor.
`l_u`	A logical value that indicates if the variables in the design matrix `X` are transformed to unit length. By default `l_u=TRUE`.
`l`	A real number that indicates the lower limit of the vertical axis of the scatter plot between the Variance Inflation Factor (VIF) and the Coefficient of Variation (CV). By default `l=40`.
`intercept`	A logical value that indicates if the design matrix `X` have intercept. By default `intercept=TRUE`.
`graf`	A logical value that indicates if the scatter plot between the VIF and CV is represented by using the CV_VIF function. By default `graf=TRUE`.

Details

The Redefined Variation Inflation Factor (RVIF) is capable to detect both kind of multicollinearity: the essential (near-linear relationship between at least two independent variables excluding the intercept) and non-essential (near-linear relationship between the intercept and at least one of the remaining independent variables). This measure also quantifies the percentage of near multicollinearity due to each independent variable.

Value

`RVIF`	Redefined Variance Inflation Factor of each independent variable.
`%`	Percentage of near multicollinearity due to each independent variable.
`Graph`	Scatter plot of VIF and CV.

Author(s)

R. Salmerón (<`romansg@ugr.es`>) and C. García (<`cbgarcia@ugr.es`>).

References

R. Salmerón, C. García, and J. García. Variance inflation factor and condition number in multiple linear regression. Journal of Statistical Computation and Simulation, 88:2365-2384, 2018.

R. Salmerón, A. Rodríguez, and C. García. Diagnosis and quantification of the non-essential collinearity. Computational Statistics, 35:647-666, 2020.

Salmerón, R., García, C.B. y García, J. A redefined Variance Inflation Factor: overcoming the limitations of the Variance Inflation Factor. Computational Economics (2024, online), doi: https://doi.org/10.1007/s10614-024-10575-8.

See Also

`CV_VIF`

Examples

```
library(multiColl)
set.seed(2022)
obs = 100
cte = rep(1, obs)
x2 = rnorm(obs, 5, 0.01)
x3 = rnorm(obs, 5, 10)
x4 = x3 + rnorm(obs, 5, 1)
x5 = rnorm(obs, -1, 30)
x = cbind(cte, x2, x3, x4, x5)
RVIF(x)
```

SLM1 *First simulated data for the simple linear regression model*

Description

First data used in example 4 of Salmerón, García and García (2024) (subsection 4.4) on special case of simple linear model.

Usage

```
data("SLM1")
```

Format

A data frame with 50 observations on the following 3 variables:

`y1` Dependent variable simulated as y = 3 + 4*V + u with u is normally distributed with a mean equal to 0 and a variance equal to 2.

`cte` Intercept.

`V` Simulated from a normal distribution with mean equal to 10 and variance equal to 100.

References

Salmerón, R., García, C.B. and García, J. A redefined Variance Inflation Factor: overcoming the limitations of the Variance Inflation Factor. Computational Economics (2024, online), doi: https://doi.org/10.1007/s10614-024-10575-8.

Examples

```
data(SLM1)
head(SLM1, n=5)
y = SLM1[,1]
X = as.matrix(SLM1[,2:3])
Theorem(y, X)
```

SLM2 *Second simulated data for the simple linear regression model*

Description

Second data used in example 4 of Salmerón, García and García (2024) (subsection 4.4) on special case of simple linear model.

Usage

```
data("SLM2")
```

Format

A data frame with 50 observations on the following 3 variables:

`y2` Dependent variable simulated as y = 3 + 4*Z + u where u is normally distributed with a mean equal to 0 and a variance equal to 2.

`cte` Intercept.

`Z` Simulated from a normal distribution with mean equal to 10 and variance equal to 0.1.

References

Salmerón, R., García, C.B. and García, J. A redefined Variance Inflation Factor: overcoming the limitations of the Variance Inflation Factor. Computational Economics (2024, online), doi: https://doi.org/10.1007/s10614-024-10575-8.

Examples

```
data(SLM2)
head(SLM2, n=5)
y = SLM2[,1]
X = as.matrix(SLM2[,2:3])
Theorem(y, X)
```

`Theorem` *Theorem*

Description

Given a multiple linear regression model with n observations and k independent variables, the degree of near-multicollinearity affects its statistical analysis (with a level of significance of afa%) if there is a variable i, with i = 1,...,k, that verifies that the null hypothesis is not rejected in the original model and is rejected in the orthogonal model of reference.

Usage

```
Theorem(y, X, alfa = 0.05)
```

Arguments

`y`	A numerical vector representing the dependent variable of the model.
`X`	A numerical design matrix that should contain more than one regressor (intercept included).
`alfa`	Level of significance (by default, 5%).

Details

This function compares the individual inference of the original model with that of the orthonormal model taken as reference.

Thus, if the null hypothesis is rejected in the individual significance tests in the model where there are no linear relationships between the independent variables (orthonormal) and is not rejected in the original model, the reason for the non-rejection is due to the existing linear relationships between the independent variables (multicollinearity) of the original model.

The second model is obtained from the first model by performing a QR decomposition which allows to eliminate the initial linear relationships.

Value

The function returns the value of the RVIF, the thresholds established as worroying and whether or not the individual significance analysis is affected by multicollinearity (at the significance level used).

Author(s)

Román Salmerón Gómez (University of Granada) and Catalina García García (University of Granada).

Maintainer: Román Salmerón Gómez (romansg@ugr.es)

References

Salmerón, R., García, C.B. and García, J. A redefined Variance Inflation Factor: overcoming the limitations of the Variance Inflation Factor. Computational Economics (2024, online), doi: https://doi.org/10.1007/s10614-024-10575-8.

Overcoming the inconsistences of the variance inflation factor: a redefined VIF and a test to detect statistical troubling multicollinearity by Salmerón, R., García, C.B and García, J. (working paper, https://arxiv.org/pdf/2005.02245).

See Also

RVIF

Examples

```
## Example 1
set.seed(2024)
obs = 100
cte = rep(1, obs)
x2 = rnorm(obs, 5, 0.01)  # related to intercept: non essential
x3 = rnorm(obs, 5, 10)
x4 = x3 + rnorm(obs, 5, 0.5) # related to x3: essential
x5 = rnorm(obs, -1, 3)
x6 = rnorm(obs, 15, 0.5)
y = 4 + 5*x2 - 9*x3 -2*x4 + 2*x5 + 7*x6 + rnorm(obs, 0, 2)
X = cbind(cte, x2, x3, x4, x5, x6)
Theorem(y, X)

## Example 2
obs = 25 # by decreasing the number of observations affected to x4
cte = rep(1, obs)
x2 = rnorm(obs, 5, 0.01)  # related to intercept: non essential
x3 = rnorm(obs, 5, 10)
x4 = x3 + rnorm(obs, 5, 0.5) # related to x3: essential
x5 = rnorm(obs, -1, 3)
x6 = rnorm(obs, 15, 0.5)
y = 4 + 5*x2 - 9*x3 -2*x4 + 2*x5 + 7*x6 + rnorm(obs, 0, 2)
X = cbind(cte, x2, x3, x4, x5, x6)
Theorem(y, X)

## Example 3
y = 4 - 9*x3 - 2*x5 + rnorm(obs, 0, 2)
X = cbind(cte, x3, x5) # independently generated
Theorem(y, X)
```

Wissel *Wissel data*

Description

Wissel data on outstanding mortgage debt.

Usage

```
data("Wissel")
```

Format

A data frame with 17 observations on the following 6 variables:

`t` Year.

`D` Outstanding mortgage debt (dependent variable).

`cte` Intercept.

`C` Personal consumption (trillions of dollars).

`I` Personal income (trillions of dollars).

`CP` Outstanding consumer credit (trillions of dollars).

References

Wissel, J. (2009). A new biased estimator for multivariate regression models with highly collinear variables. Ph.D. thesis, Erlangung des naturwissenschaftlichen Doktorgrades der Bayerischen Julius-Maximilians-Universität Würzburg.

Examples

```
data(Wissel)
head(Wissel, n=5)
y = Wissel[,2]
X = as.matrix(Wissel[,3:6])
Theorem(y, X)
```

Index

∗ **Coefficient of Variation**
  CV_VIF, 4
  RVIF, 8
  rvif-package, 2
∗ **Individual significance tests**
  Theorem, 11
∗ **Orthonormal model**
  Theorem, 11
∗ **Scatter plot**
  CV_VIF, 4
∗ **Variance inflation factor redefined**
  Theorem, 11
∗ **Variance inflation factor**
  CV_VIF, 4
∗ **Variation Inflation Factor Redefined**
  RVIF, 8
  rvif-package, 2
∗ **Variation Inflation Factor**
  RVIF, 8
  rvif-package, 2
∗ **datasets**
  CDpf, 3
  employees, 6
  euribor, 7
  SLM1, 9
  SLM2, 10
  Wissel, 12

CDpf, 3
CV_VIF, 4, *9*

employees, 6
euribor, 7

RVIF, 8, *12*
rvif *(*rvif-package*)*, 2
rvif-package, 2

SLM1, 9
SLM2, 10

Theorem, 11

Wissel, 12