mctest: An R Package for Detection of Collinearity Among Regressors

by Muhammad Imdadullah, Muhammad Aslam, Saima Altaf

Abstract It is common for linear regression models to be plagued with the problem of multicollinearity when two or more regressors are highly correlated. This problem results in unstable estimates of regression coefficients and causes some serious problems in validation and interpretation of the model. Different diagnostic measures are used to detect multicollinearity among regressors. Many statistical software and R packages provide few diagnostic measures for the judgment of multicollinearity. Most widely used diagnostic measures in these software are: coefficient of determination (R^2), variance inflation factor/tolerance limit (VIF/TOL), eigenvalues, condition number (CN) and condition index (CI) etc. In this manuscript, we present an R package, **mctest**, that computes popular and widely used multicollinearity diagnostic measures. The package also indicates which regressors may be the reason of collinearity among regressors.

Brief introduction of collinearity

Consider the conventional multiple linear regression equation

$$y = X\beta + u,$$

where y is an $n \times 1$ vector of observation on response variable, X is known design matrix of order $n \times p$, β is an $p \times 1$ vector of unknown parameters and u is an $n \times 1$ vector of random errors with mean zero and variance $\sigma^2 I_n$, where I_n is an identity matrix of order n.

One of the important assumptions of the classical linear regression model is that there is no exact collinearity among the regressors otherwise, the issue is referred to as multicollinearity. Generally, the problem of multicollinearity may also refer to have not only exact linear relationship but also high correlations among some or all regressors of a regression model under study. Strictly speaking, multicollinearity is usually refers to the existence of more than one exact linear relationship among regressors, while collinearity refers to the existence of a single linear relationship among regressors. However, in general, the term multicollinearity may be referred to both the cases. Data collection method, constraints on the fitted regression model, model specification error, overdefined model, some common trend in time series data and naturally correlated data may be some potential sources of multicollinearity.

The problem of multicollinearity has potentially serious effect on the regression estimates such as implausible coefficient signs, impossible inversion of matrix $X'X$ as it becomes either singular (in the case of perfect multicollinearity) or near to singular (in the case of near to perfect multicollinearity), large magnitude of coefficients in absolute value, large variance or standard errors with wider confidence intervals and small t-ratios. The ordinary least squared (OLS) estimators and standard errors also become sensitive to small change in data when regressors are collinear to each other (see Belsley et al., 1980; Dorsett et al., 1983; Farrar and Glauber, 1967; Gunst and Mason, 1977; Johnston, 1963; Mason et al., 1975). On the basis of theoretical considerations, these indications signify the need for detection of multicollinearity among regressors (Belsley et al., 1980; Greene, 2002; Younger, 1979).

This paper presents the overview of existing collinearity diagnostic measures along with commonly used threshold values for the judgment of existence of collinearity among regressors. These diagnostic measures are being implemented in R with the proposed **mctest** package (Imdadullah and Aslam, 2016).

Collinearity diagnostic measures

Several numerical methods for the detection of collinearity are available in the existing literature proposed or discussed by various authors e.g., (see Belsley et al., 1980; Curto and Pinto, 2011; Koutsoyiannis, 1977; Kovács et al., 2005; Marquardt, 1970; Montgomery and Peck, 1982, etc.). Widely used and the most suggested collinearity diagnostic measures are values of pair-wise correlations, R-squared value (c.f. Gujarati and Porter (2008)), variance inflation factor (VIF), tolerance limit (TOL) (Kutner et al., 2004; Marquardt, 1970), eigenvalues (Kendall, 1957; Silvey, 1969), condition number (CN) and condition index (CI) (Belsley et al., 1980), Leamer's method (Greene, 2002), Kleins rule (Klein, 1962), three tests proposed by Farrar and Glauber (Farrar and Glauber, 1967), the Red indicator (Kovács et al., 2005), and Theil's measure (Theil, 1971). However, none of these can be regarded as the

best choice for the detection of collinearity (Kovács et al., 2005).

Following are the diagnostics that can be considered as the classical symptoms of harmfulness of multicollinearity. (i) If zero-order (pair-wise) correlation coefficient between two regressors is high (say >0.8) then multicollinearity may be a serious problem (Gujarati and Porter, 2008; Maddala, 1988). However, it is not sufficient and necessary condition for the detection of multicollinearity because a linear relation involves many of the regressors, therefore it may not be possible to detect such a relation with a simple correlation or pairs-wise plot (Chatterjee and Hadi, 2006; Judge et al., 1985). (ii) High R^2 (say >0.8) may indicate the problem of multicollinearity Gujarati and Porter (2008). In most of the cases, overall F-test rejects the null hypothesis of partial slopes for being zero, but some or all individual t-ratios of partial slopes may be non-significant. Therefore, a model having no multicollinearity problem should have high R^2 and larger (significant) t-ratios of partial slopes. (iii) High variance of regression coefficients' estimates and low t-ratios also suggest the existence of multicollinearity.

We classified other widely used collinearity diagnostics as overall and individual measures of collinearity. This classification is due to the fact that there are some diagnostic measures resulting in a single number, while other yield as many quantities as the number of regressors in the model. The overall diagnostic measures help to get an idea about only the existence of collinearity and they do not tell which regressor may be the reason of collinearity, while the individual measures point out the regressors causing collinearity. Since no specific collinearity diagnostic measure is superior and each of these measures has different collinearity detection criterion (threshold value) proposed by various authors in the textbooks and research articles, there is need to study multiple collinearity diagnostics. That is, there is no clear-cut criterion for evaluating multicollinearity in linear regression models. Similarly, some diagnostic measures are statistically criticized such as tests proposed by Farrar and Glauber (1967) while threshold values of many other diagnostic measures are subjective in nature as no unique or standard critical values exist for these measures. Moreover, different collinearity detection methods are not comparable with each other. That is why, many regression analysts often rely on more than one collinearity diagnostic measures.

Following is the list of overall and individual collinearity diagnostic measures along with short description, formula, detection criterion (threshold value) and reference for each measure. These diagnostic measures will assist the researchers in determining when and where some corrective action is necessary. According to Belsley et al. (1980), the investigations concerning the presence of multicollinearity have been based on judging the magnitudes of various diagnostic measures.

Overall collinearity diagnostic measures

- ***Determinant:***
  The matrix $X'X$ will be singular if it contains linearly dependent columns or rows. Therefore, determinant of normalized correlation matrix ($R = X'X$) without intercept can be used to indicate existence of collinearity among regressors. However, determinant does not provide information about interdependence among regressors, it only provides information about singularity (departure from orthogonality) of a correlation matrix. The determinant of $X'X$ on the scale is $0 \leq |X'X| \leq 1$ (Cooley and Lohnes, 1971). If $|X'X| \sim 0$, then collinearity exists among regressors (Asteriou and Hall, 2007).
- ***R-squared:***
  Coefficient of determination (R^2) from regression of all x on y. The R^2 is a monotonic non-decreasing function of number of regressors included in the model, that is, R^2 indicates how well the regression fits the data (Gujarati and Porter, 2008; Stock and Watson, 2010). On the other hand, higher the R^2 values, the more chances of regressors to be plagued with multicollinearity (Asteriou and Hall, 2007; Gujarati and Porter, 2008; Maddala, 1988), since R^2 is affected by regressors sharing their variances (Gujarati and Porter, 2008; Maddala, 1988).
- ***Farrar χ^2:***
  It is the Chi-square test for detecting the strength of collinearity over the complete set of regressors. $\chi^2 = -\left[n - 1 - \frac{1}{6(2p+5)}\right] \times \log_e [X'X] \sim \psi^2_{v=\frac{1}{2}p(p-1)}$. Collinearity exists among regressors if $\chi^2 > \chi^2_{\frac{1}{2}p(p-1)}$ (Farrar and Glauber, 1967).
- ***Condition index:***
  $CI_j = \sqrt{\frac{\max(\lambda_j)}{\lambda_j}}$ $j = 1, 2, \cdots, p$; $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p$. Collinearity exists if any of $CI_j > 10, 15$, or 30 (Belsley et al., 1980; Chatterjee and Hadi, 2006; Maddala, 1988).
- ***Sum of reciprocal of eigenvalues:***
  In an orthogonal system $\sum_{j=1}^{p} \frac{1}{\lambda_j} = p$, therefore, for a sample based correlation matrix R with

eigenvalues λ_j comparing p with $\sum_{j=1}^{p} \frac{1}{\lambda}$ can be used to indicate collinearity. If $\sum_{j=1}^{p} \frac{1}{\lambda_j}$ is (say) five times larger than the number of regressors used in the model then collinearity exists among regressors (Chatterjee and Price, 1977; Dillon and Goldstein, 1984).

- ***Theil's indicator:***
  Theil (1971) proposed a measure of collinearity based on an incremental contribution $(R^2 - R_j^2)$ to the squared multiple correlation, where R_j^2 is the R^2 from auxiliary regression of regressors. $m = R^2 - \sum_{i=1}^{p} (R^2 - R_{-i}^2)$. If $m = 0$ then all X's are mutually uncorrelated (no redundancy exists) as the incremental contribution all add up to R^2. However, if $m \sim 1$ then collinearity exists among regressors.

- ***Red indicator:***
  Kovács et al. (2005) presented a synthetic and new normalized indicator for diagnostic of collinearity by using eigenvalues or quantifying the average correlation of the data. $Red = \frac{\sqrt{\sum_{j=1}^{p} \left(\lambda_j - 1\right)^2}}{\sqrt{p-1}}$. If value of the Red indicator is zero ($Red = 0$) then it indicates the absence of redundancy and value near to 1 ($Red \sim 1$) indicates maximum redundancy ($Red \sim 1$).

Individual collinearity diagnostic measures

- ***Klein's rule:***
  If R_j from the auxiliary regression is greater than the overall R^2 (obtained from the regression of y on all the regressors) then multicollinearity may be troublesome. The decision rule for detection of collinearity is, $R^2_{x_j.x_1,x_2,\cdots,x_p} > R^2_{y.x_1,x_2,\cdots,x_p}$ (Klein, 1962).

- ***VIF and TOL:***
  VIF measures how much variances of the estimated regression coefficients are increased over the case of no correlation among p regressors. The diagonal elements of $(X'X)^{-1}$ matrix are considered as very important in detecting multicollinearity. $VIF_j = (X'X)^{-1}_{jj} = \frac{1}{1-R_j^2}$ and $TOL_j = \frac{1}{VIF_j} = 1 - R_j^2$.

  The criticism on VIF is that $var(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2} VIF_j$ depends on σ^2, $\sum x_j^2$ and VIF, which shows that a high VIF can be counterbalanced by a low σ^2 or high $\sum x_j^2$. So a high VIF is neither a necessary nor a sufficient measure of multicollinearity (Gujarati and Porter, 2008). The value of $VIF > 3, 5, 10$ or value of $TOL \sim 0$ indicates existence of collinearity among regressors (Kutner et al., 2004; Marquardt, 1970).

- ***Eigenvalues:***
  Kendall (1957) and Silvey (1969) suggested the use of eigenvalues of $X'X$ (correlation matrix) to check the presence of multicollinearity and set the criteria that small eigenvalues (near to zero) are indication of high collinearity, however, they did not mentioned how much small it should be. One or more smaller eigenvalues of $X'X$ or its related correlation matrix indicate collinearity (Kendall, 1957; Silvey, 1969).

- ***CVIF:***
  Curto and Pinto (2011) proposed new measure of multicollinearity to evaluate the impact of the correlation among regressors in the variance of the OLSEs. $CVIF_j = VIF_j \times \frac{1-R^2}{1-R_0^2}$ where, $R_0^2 = R^2_{yx_1} + R^2_{yx_2} + \cdots + R^2_{yx_p}$. Collinearity exists if $CVIF_j \geq 10$ (Curto and Pinto, 2011).

- ***Leamer's method:***
  Leamer (in Greene (2002)) suggested a measure of the effect of multicollinearity for the jth variable; $C_j = \left\{ \frac{\left(\sum_i^n \left(X_{ij} - \overline{X}_j\right)^2\right)^{-1}}{(X'X)^{-1}_{jj}} \right\}^{\left(\frac{1}{2}\right)}$. This measure is the square root of the ratio of variances of estimated coefficients $(\hat{\beta}_j)$ when estimated without and with the other regressors. If X_j is uncorrelated with the other regressors C_j would be 1 otherwise will be equal to $(1 - R_j^2)^{\frac{1}{2}}$, i.e., $C_j \sim 0$ indicates existence of collinearity among regressors.

- ***F and R^2 relation:***
  The relationship of F-test and R^2 from regressing X_j on the other remaining regressors can

be used to detect multicollinearity. The relationship is described as: $F_j = \frac{\frac{R^2_{x_j,x_1,\cdots,x_p}}{p-2}}{\frac{1-R^2_{x_j,x_1,\cdots,x_p}}{n-p+1}} \sim F(p-2, n-p+1)$, where $F^* = F_{p-2,n-p+1}$. If $F_j > F^*$, then it means that the regressor X_j is collinear with other regressors and it should be dropped from the model (Gujarati and Porter, 2008).

- ***Farrar*** w_j**:**
  It is an F-test for locating the regressors which are collinear with others and it makes use of multiple correlation coefficients among regressors. $w_j = \frac{R_j^2}{1-R_j^2}\left(\frac{n-p}{p-1}\right) \sim F_{(n-p,\, p-1)}$. If $w_j > F_{(n-p,p-1)}$, there is indication of considerable collinearity (Farrar and Glauber, 1967).

There are few software and R packages that provide some collinearity diagnostic measures such as correlation matrix, VIF/TOL, eigenvalues/eigenvectors, and CN/CI. The design goal of our developed package **mctest** is primarily to provide a comprehensive suite of all the listed diagnostic measures. All R packages mentioned in Table 1 are compared with our **mctest** package regarding diagnostic measures in these packages. Other features in these packages and collinearity related measures available in different statistical software are also discussed.

	perturb	HH	car	fmsb	rms	faraway	usdm	mctest
Overall collinearity diagnostics								
$\lvert X'X\rvert$								✓
R-squared								✓
Farrar χ^2								✓
CN/CI	✓							✓
$\sum_{j=1}^{p} \frac{1}{\lambda_j}$								✓
Theil's indicator								✓
Red indicator								✓
Individual collinearity diagnostics								
Correlation matrix								✓
Var and t-ratios								✓
Klein's rule								✓
VIF		✓	✓	✓	✓	✓	✓	✓
TOL								✓
Eigenvalues								✓
CVIF								✓
Leamer's method								✓
Farrar W_i								✓
F and R^2 relation								✓

Table 1: Comparison of collinearity related R packages

There are few statistical software (SAS (SAS 9.3, 2011), Stata (StataCorp, 2015), Minitab (Minitab, Inc., 2014), NCSS (NCSS , 2016), and StatGraphics (Statgraphics Centurion XVII, 2015) etc.), giving different collinearity diagnostic measures such as (R^2, eigenvalues, VIF, CN, and correlation matrix etc.). The R packages mentioned in Table 1 have some other functionalities related to collinearity. For example, **perturb** (Hendrickx, 2012) evaluates collinearity by adding random noise to selected variables and computes the CN and variance decomposition proportion to test the collinearity and to uncover its sources. The package **car** (Fox and Weisberg, 2011) computes the VIF and GVIF for linear and generalized linear models. The function `vif` of package **usdm** (Naimi, 2015) computes the VIF for a set of variables and excludes highly correlated variables from the set through a stepwise procedure. The package **rms** Harrell Jr (2016) computes VIF from the covariance matrix of parameter estimates from binary or ordinal regression models, Cox regression, accelerated failure time models, ordinary linear models, the Buckley-James model, generalized least squares for serially or spatially correlated observations, generalized linear models, and quantile regression. The packages, **HH** (Heiberger, 2016), **fmsb** (Nakazawa, 2015) and **faraway** (Faraway, 2016) present different statistical methods and an extensive use of graphical display.

There are some other R packages such as **VIF** (Lin, 2012), **leaps** (Lumley, 2009), **bestglm** (McLeod and Xu, 2014), **glmulti** (Calcagno, 2013), and **meifly** (Wickham, 2014) that are used for collinear datasets. These packages involved procedures to search for adequate predictors and for parsimonious

models (subset or all subset regression). The availability of different collinearity diagnostic measures in R packages, shown in Table 1 and in different statistical software, suggests that the package **mctest** is a useful addition on CRAN.

R implementation

In this section, we illustrate the use of our developed package **mctest**. The R package **mctest** mainly implements functions for the detection of collinearity among regressors by calling `omcdiag()` and `imcdiag()` functions. For the graphical representation of VIF values and eigenvalues, `mc.plot()` function can also be used. We try to build a simple interface to facilitate the usage of this package.

The functions, `omcdiag`, `imcdiag`, and `mctest` ensure that the number of regressors provided as x argument should be at least two. Similarly, the values of regressors and response variable (y) should contain only numbers provided that both have equal number of observations. All the other arguments are optional and have default threshold values for different collinearity diagnostic measures. Following is the list of functions available in **mctest**;

Function	Description
`omcdiag()`	Computation of overall collinearity measures.
`imcdiag()`	Computation of individual collinearity measures for each regressor.
`mctest()`	Calls overall and individual collinearity measures.
`mc.plot()`	Graphical representation of VIF and eigenvalues.

Table 2: Functions available in **mctest** package

Overall collinearity diagnostics

For overall collinearity diagnostic measures, the function `omcdiag()` has only two mandatory arguments: the vector of response variable y and the matrix of regressors x. The argument `na.rm` removes the missing values in dataset and is set to `TRUE`. Therefore, all calculations will be performed on newly created data after removing missing observations if any, otherwise, calculations will be performed on complete observation available in the provided dataset. The optional argument `Inter`, when it takes the value `TRUE`, allows to compute eigenvalues and condition index including intercept term in design matrix $X'X$, otherwise, without it. The other arguments `detr`, `red`, `theil`, `cn`, and `conf` are used as threshold values as collinearity detection criteria. If all these optional arguments are not used, the eigenvalues and CIs with intercept term will be computed and all these values will be compared with the default threshold values (can be provided by the user) for the indication of existence of collinearity by each of the diagnostic methods.

```
omcdiag(x,y,na.rm=TRUE,Inter=TRUE,detr=0.01,red=0.5,conf=0.95,theil=0.5,cn=30,...)
```

The results from each of overall collinearity diagnostic measures are displayed with an indication that whether existence of collinearity among regressors is correctly detected by diagnostic methods or not. The eigenvalues and CIs are also displayed for the confirmation of existence of collinearity.

Example: `omcdiag()`

This section uses the Hald data (Hald, 1952) bundled in **mctest** package for checking of existence of collinearity among regressors using `omcdiag()` function. Different examples of `omcdiag()` with use of difference arguments are provided, however, results are shown only for the last command.

```
> library('mctest')
> head(Hald)
> x <- Hald[ , -1]
> y <- Hald[ ,  1]

> omcdiag (x, y, detr = 0.001, red = 0.6, conf = 0.99, theil = 0.6, cn = 15)
> omcdiag (x, y, Inter = FALSE)
> omcdiag (x, y)

Call:
```

```
omcdiag(x = x, y = y)

Overall Multicollinearity Diagnostics

                       MC Results detection
Determinant |X'X|:         0.0011         1
Farrar Chi-Square:        59.8700         1
Red Indicator:             0.5414         1
Sum of Lambda Inverse:   622.3006         1
Theil's Method:            0.9981         1
Condition Number:        249.5783         1

1 --> COLLINEARITY is detected
0 --> COLLINEARITY in not detected by the test

==================================
Eigenvalues with INTERCEPT
                 Intercept     X1     X2      X3       X4
Eigenvalues:        4.1197 0.5539 0.2887  0.0376   0.0001
Condition Indexes:  1.0000 2.7272 3.7775 10.4621 249.5783
```

Results from `omcdiag` shows that all of the overall collinearity diagnostic measures correctly detected the presence of multicollinearity among regressors. Similarly, eigenvalues and CIs also indicate regressors are collinear since, some eigenvalues are small enough and at least one of the CIs is greater than 30.

Individual collinearity diagnostics

For the individual collinearity diagnostic measures, `imcdiag()` also has two mandatory arguments like `omcdiag()` or `mctest()` has. The optional argument `method`, when it takes value "VIF", "TOL", "Wi", "Fi", "Klein", "conf", "CVIF", or "Leamer", will compute only provided method with an indication of whether regressor(s) is(are) possible reason of collinearity or not. The other optional arguments (such as `vif`, `tol` `conf`, `cvif`, and `leamer`) are threshold values to compare with diagnostic measures of VIF, TOL, confidence level for the Farrar-Glauber test of Wi, Fi, CVIF, and Leamer's method, respectively for possible detection of collinearity among regressors. The `corr` argument is set to `FALSE`, if it takes value as `TRUE`, the correlation matrix will also be produced along with collinearity diagnostic measures with the indication of which pair of regressors are collinear. The computed value of certain diagnostic measure, provided to `method` argument, is displayed with an indication of whether diagnostic measure correctly detected the existence of collinearity or not. The `all` argument is set to `FALSE`, if it takes value as `TRUE`, the individual collinearity diagnostics will be returned in form of 0 or 1. From `"lm"` function, non-significant *t*-values are also displayed for further subjective judgment and confirmation of the existence of collinearity among regressors.

```
imcdiag(x,y,method = NULL,na.rm = TRUE,corr = FALSE,vif = 10,tol = 0.1,
        conf = 0.95,cvif = 10,leamer = 0.1,all = FALSE,...)
```

Example: `imcdiag()`

Different examples of `imcdiag()` function with use of different arguments are provided, however, results are shown only for the last command.

```
> imcdiag(x, y, corr = TRUE)
> imcdiag(x, y)

Call:
imcdiag(x = x, y = y)

Individual Multicollinearity Diagnostics

        VIF    TOL       Wi        Fi Leamer    CVIF Klein
X1  38.4962 0.0260 112.4886  187.4811 0.1612 -0.5846     0
X2 254.4232 0.0039 760.2695 1267.1158 0.0627 -3.8635     1
X3  46.8684 0.0213 137.6052  229.3419 0.1461 -0.7117     0
X4 282.5129 0.0035 844.5386 1407.5643 0.0595 -4.2900     1
```

```
1 --> COLLINEARITY is detected
0 --> COLLINEARITY is not detected by the test

X1 , X2 , X3 , X4 , coefficient(s) are non-significant may be due to multicollinearity

* use method argument to check which regressors may be the reason of collinearity
```

Each column of output from `imcdiag(x,y)` indicates that the computed values of individual collinearity diagnostic measures for each regressor. The last column results in either 0 (no collinearity due X_j) or 1 (collinearity due to X_j) due to Klein's rule.

To get certain individual collinearity diagnostic with custom threshold can be obtained by using `method` argument. The first column of output contains the value of diagnostic measure. In the second column, 1 and 0 denotes the detection and non-detection of collinearity, respectively, for each of the regressor. The use of `switch` statement is made to fulfill the purpose of obtaining diagnostic values and the indication of collinearity detection for certain collinearity diagnostics provided as value to argument `method`. Some examples of obtaining certain individual collinearity diagnostic measures are;

```
> imcdiag(x, y, method = "VIF", vif = 5)
> imcdiag(x, y, method = "VIF", vif = 10, corr = TRUE)
> imcdiag(x, y, method = "CVIF", cvif = 10)

Call:
imcdiag(x = x, y = y, method = "CVIF", cvif = 10)

Individual Multicollinearity Diagnostics

      CVIF detection
X1 -0.5846         0
X2 -3.8635         0
X3 -0.7117         0
X4 -4.2900         0

NOTE:  CVIF Method Failed to detect multicollinearity

0 --> COLLINEARITY in not detected by the test
```

If argument `all` in `imcdiag` or `mctest` is set to `TRUE`, a matrix of either 0 or 1 will be displayed. Few examples for use of `all` argument are;

```
> imcdiag(x, y, all = TRUE)
> imcdiag(x, y, all = TRUE, vif = 15, conf = 0.99, )
> imcdiag(x, y, method = "VIF", all = TRUE)
> mctest(x, y, all = TRUE, type="i")

Call:
imcdiag(x = x, y = y, all = TRUE)

All Individual Multicollinearity Diagnostics in 0 or 1

  VIF TOL Wi Fi Leamer CVIF Klein
X1   1   1  1  1      0    0     0
X2   1   1  1  1      1    0     1
X3   1   1  1  1      0    0     0
X4   1   1  1  1      1    0     1

1 --> COLLINEARITY is detected
0 --> COLLINEARITY in not detected by the test

X1 , X2 , X3 , X4 , coefficient(s) are non-significant may be due to multicollinearity

R-square of y on all x: 0.9824

* use method argument to check which regressors may be the reason of collinearity
```

```
   mctest(x,y,type = c("o","i","b"),na.rm = TRUE,Inter = TRUE,method = NULL,
            corr = FALSE,detr = 0.01,red = 0.5,cn = 30,vif = 10,tol = 0.1,conf = 0.95,
```

```
        cvif = 10,leamer = 0.1,all = FALSE,... )
```

The `mctest()` function also has two mandatory arguments: the vector of response y and the matrix of regressors as x. The argument `type` is optional for computation of overall (from `omcdiag`) by setting `type="o"`, individual (from `imcdiag`) by setting `type="i"` or both overall and individual collinearity diagnostics by setting `type="b"`, if `type` argument is not used overall collinearity measures will be computed and displayed.

Collinearity diagnostic plots: VIF and eigenvalues plot

The `mc.plot` function can also be used to draw the plots of VIF values and eigenvalues to graphically judge the existence of collinearity among regressors. The VIF values and eigenvalues are also drawn for each regressor along the y-axis. A horizontal red dashed line equal to either default threshold or may be provided by the user of **mctest**, for both VIF and eigenvalues.

```
> mc.plot(x, y)
> mc.plot(x, y, vif = 10, ev = 0.1)
```

The argument, `vif = 10` and `ev = 0.1` are user provided thresholds for VIF and eigenvalues, respectively and will be shown as horizontal red dashed line.

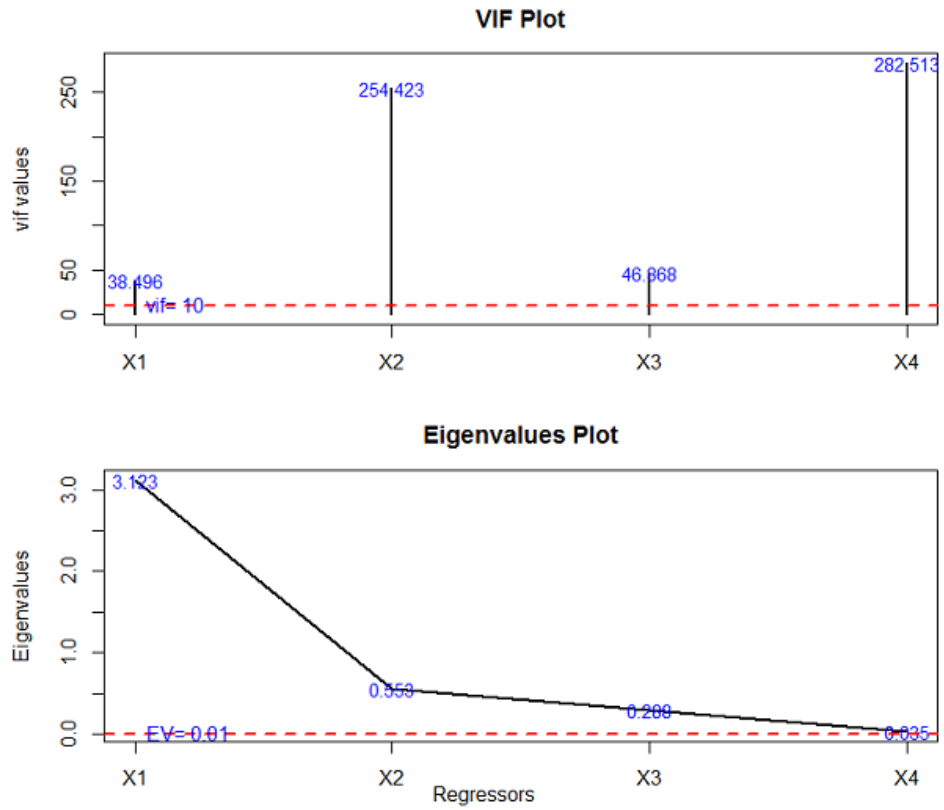

Figure 1: The VIF and Eigenvalues Plots.

From VIF plot, the VIF values of each regressor greater than 30 indicates the existence of multicollinearity among regressors. Similarly, the eigenvalues plot indicates that few regressors have relatively smaller eigenvalues than others, indicating the existence of collinearity. Note that the graphical output (shown in Figure 1) from `mc.plot()` and numerical results from, for example, `mctest()` are all equivalent. Only difference exits in the way of their representation.

Dealing with multicollinearity

Complete elimination of multicollinearity is not possible, but the degree of collinearity can be reduced. Depending on the severity of the collinearity problem, there are two schools of thought (a) do nothing or (b) follow some rules of thumb. According to the first school of thought, Blanchard (1967) suggested to do nothing with the regressors or model, since multicollinearity is essentially a data deficiency

problem and sometimes there is no choice over the data available for empirical analysis. Regarding the second approach, some rules to alleviate the problem of multicollinearity are: (i) Drop one of the highly collinear regressor. If model has two or more regressors with high VIF, drop one from the model, because it supplies redundant information. Dropping one of the correlated regressor usually does not drastically reduce the R^2. However, omission of relevant regressor(s) from the model, may result in a specification error. Hence, the remedy may be worse than the disease in some situations, because, multicollinearity may prevent the precise estimation of parameters of the regression model. Therefore, omitting some regressor(s) may seriously mislead to the true values of the parameters (Gujarati and Porter, 2008, pg. 344). (ii) Use an appropriate experimental design and increase the sample if possible. However, obtaining additional or better data is not always easy. (iii) Transform the regressors (iv) Use some alternative methods to the OLS such as principal component regression and ridge regression etc. to control variance and instability of the OLS estimates. (v) Use stepwise regression, best subset regression or specialized knowledge of the dataset to remove the redundant regressors. (vi) Combine the redundant variables, if possible.

Summary

Strong linear relationship among regressors i.e. , the issue of multicollinearity results in unstable estimated regression coefficients and other inadequate statistical measures. Therefore, its severity should be tested. An R package, **mctest** has been designed with the goal of providing the most widely used and discussed collinearity diagnostic related statistics. Two main functions `omcdiag()` and `imcdiag` facilitate the users to get information about the existence of collinearity among regressors and also to get idea about which regressor may be the reason of multicollinearity. A function, `mc.plot()` can also be used to detect existence of collinearity among regressors by drawing a graph of VIF and eigenvalues. For further details about use of the said package and related functions, interested readers are referred to the documentation of the package.

Bibliography

D. Asteriou and S. G. Hall. *Applied Econometrics: A Modern Approach using Eviews and Microfit*. Palgrave Macmillan, New York, 2007. [p2]

D. A. Belsley, E. Kuh, and R. E. Welsch. *Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons, New York, 1980. chap. 3. [p1, 2]

O. J. Blanchard. Commnet. *Journal of Business and Economic Statistics*, 5:449–451, 1967. [p8]

V. Calcagno. *glmulti: Model Selection and Multimodel Inference Made Easy*, 2013. URL `https://CRAN.R-project.org/package=glmulti`. R package version 1.0.7. [p4]

S. Chatterjee and A. S. Hadi. *Regression Analysis by Example*. Wiley and Sons, 4th edition, 2006. [p2]

S. Chatterjee and B. Price. *Regression Analysis by Examples*. John Wiley & Sons, New York, 1977. [p3]

W. W. A. Cooley and P. R. A. Lohnes. *Multivariate Data Analysis*. John Wiley & Sons Australia, Limited, 1971. ISBN 9780471170600. [p2]

J. D. Curto and J. C. Pinto. The corrected VIF (CVIF). *Journal of Applied Statistics*, 38(7):1499–1507, 2011. [p1, 3]

W. R. Dillon and M. Goldstein. *Multivariate Analysis: Methods and Applications*. John Wiley & Sons, Inc., 1984. [p3]

D. Dorsett, R. F. Gunst, and E. C. J. Gartland. Multicollinear effects of weighted least squares regression. *Statistics & Probability Letters*, 1(4):207–211, 1983. [p1]

J. Faraway. *faraway: Functions and Datasets for Books by Julian Faraway*, 2016. URL `https://CRAN.R-project.org/package=faraway`. R package version 1.0.7. [p4]

D. E. Farrar and R. R. Glauber. Multicollinearity in Regression Analysis: The Problem Revisted. *The Review of Economics and Statistics*, 49:92–107, 1967. [p1, 2, 4]

J. Fox and S. Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, second edition, 2011. URL `http://socserv.socsci.mcmaster.ca/jfox/Books/Companion`. [p4]

W. H. Greene. *Econometric Analysis*. Prentic Hall, New Jersey, 5th edition, 2002. [p1, 3]

D. N. Gujarati and D. C. Porter. *Basic Econometrics*. McGraw Hill, 5 edition, 2008. [p1, 2, 3, 4, 9]

R. F. Gunst and R. L. Mason. Advantages of examining multicollinearities in regression analysis. *Biometrics*, 33:249–260, 1977. [p1]

A. Hald. *Statistical Theory with Engineering Applications*. John Wiley and Sons, New York, 1952. [p5]

F. E. Harrell Jr. *rms: Regression Modeling Strategies*, 2016. URL https://CRAN.R-project.org/package=rms. R package version 4.5-0. [p4]

R. M. Heiberger. *HH: Statistical Analysis and Data Display: Heiberger and Holland*, 2016. URL http://CRAN.R-project.org/package=HH. R package version 3.1-32. [p4]

J. Hendrickx. *perturb: Tools for Evaluating Collinearity*, 2012. URL https://CRAN.R-project.org/package=perturb. R package version 2.05. [p4]

M. Imdadullah and D. M. Aslam. *mctest: Multicollinearity Diagnostic Measures*, 2016. URL https://CRAN.R-project.org/package=mctest. R package version 1.1. [p1]

J. Johnston. *Econometric Methods*. McGraw Hill, New York, 1963. [p1]

G. Judge, W. Griffiths, H. Lutkepohl, and T. Lee. *The Theory and Practice of Econometrics*. Wiley, 1985. [p2]

M. G. Kendall. *A Course in Multivariate Analysis*. Griffin, London, 1957. pp. 70–75. [p1, 3]

L. R. Klein. *An Introduction to Econometrics*. Prentic-Hall, Englewood, Cliffs, N. J., 1962. pp. 101. [p1, 3]

A. Koutsoyiannis. *Theory of Econometrics*. Macmillan Education Limited, 1977. [p1]

P. Kovács, T. Petres, and Tóth. A new measure of multicollinearity in linear regression models. *International Statistical Review / Revue Internationale de Statistique*, 73(3):405–412, 2005. [p1, 2, 3]

M. H. Kutner, C. J. Nachtsheim, and J. Neter. *Applied Linear Regression Models*. McGraw Hill Irwin, 4th edition, 2004. [p1, 3]

D. Lin. *VIF: VIF regression: A Fast Regression Algorithm for Large Data*, 2012. URL https://CRAN.R-project.org/package=VIF. R package version 1.0. [p4]

T. Lumley. *leaps: Regression Subset Selection using Fortran code by Alan Miller Including Exhaustive Search*, 2009. URL https://CRAN.R-project.org/package=leaps. R package version 2.9. [p4]

G. S. Maddala. *Introduction to Econometrics*. Macmillan, New York, 1988. [p2]

D. W. Marquardt. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12(3):591–612, 1970. [p1, 3]

R. L. Mason, R. F. Gunst, and J. T. Webster. Regression analysis and problems of multicollinearity. *Communications in Statistics*, 4(3):277–292, 1975. [p1]

A. McLeod and C. Xu. *bestglm: Best Bubset GLM*, 2014. URL https://CRAN.R-project.org/package=bestglm. R package version 0.34. [p4]

Minitab, Inc. Minitab Statistical Software, Release 17 for Windows, 2014. State College, Pennsylvania. [p4]

D. Montgomery and E. A. Peck. *Introduction to Linear Regression Analysis*. John Wiley and Sons, New York, 1982. [p1]

B. Naimi. *usdm: Uncertainty Analysis for Species Distribution Models*, 2015. URL https://CRAN.R-project.org/package=usdm. R package version 1.1-15. [p4]

M. Nakazawa. *fmsb: Functions for Medical Statistics Book with Some Demographic Data*, 2015. URL https://CRAN.R-project.org/package=fmsb. R package version 0.5.2. [p4]

NCSS . NCSS 11 Statistical Software, 2016. URL ncss.com/software/ncss. NCSS, LLC Paysville, Utah, USA. [p4]

SAS 9.3. SAS Institute Inc., 2011. Cary, NC, USA. [p4]

S. D. Silvey. Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society, Series B (Methodological)*, 31(3):539–552, 1969. [p1, 3]

StataCorp. Stata statistical software: Release 14, 2015. College Station, Texas 77845 USA. [p4]

Statgraphics Centurion XVII. Statpoint Technologies, Inc., 2015. Warrenton, Virginia. [p4]

J. H. Stock and M. W. Watson. *Introduction to Econometrics*. Pearson Addison-Wesley, 3rd edition, 2010. [p2]

H. Theil. *Principles of Econometrics*. John Wiley & Sons, New York, 1971. [p1, 3]

H. Wickham. *meifly: Interactive Model Exploration using GGobi*, 2014. URL https://CRAN.R-project.org/package=meifly. R package version 0.3. [p4]

M. S. Younger. *A Handbook for Linear Regression*. MA: Duxbury Resource Center, North Scituate, 1979. [p1]

Muhammad Imdadullah
Ph.D scholar (Statistics)
Department of Statistics
Bahauddin Zakariya University, Pakistan
mimdadasad@gmail.com

Muhammad Aslam
Department of Statistics
Bahauddin Zakariya University, Pakistan
aslamasadi@bzu.edu.pk

Saima Altaf
Department of Statistics
Bahaudding Zakariya University, Pakistan
drsaimaaltaf27@gmail.com