Computational Economics

A guide to using the R Package “multiColl” for Detecting Multicollinearity

--Manuscript Draft--

Manuscript Number:	CSEM-D-19-00382R1
Full Title:	A guide to using the R Package “multiColl” for Detecting Multicollinearity
Article Type:	Software review
Keywords:	Multicollinearity; Detection; Intercept; Dummy; Software
Corresponding Author:	catalina garcia Universidad de Granada Granada, Andalucia SPAIN
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Universidad de Granada
Corresponding Author's Secondary Institution:	
First Author:	ROMAN SALMERON GOMEZ
First Author Secondary Information:	
Order of Authors:	ROMAN SALMERON GOMEZ
	catalina garcia
	JOSE GARCIA PEREZ
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	The detection of problematic collinearity in a linear regression model is treated in all the existing statistical software packages. However, such detection is not always done adequately. The main shortcomings relate to treatment of independent qualitative variables and completely ignoring the role of the intercept in the model (consequently, ignoring the nonessential collinearity). This paper presents the R package multiColl, which implements the usually applied measures for detecting near collinearity while overcoming the weaknesses observed in other existing packages.
Response to Reviewers:	It is with excitement that we submit a revised version (with 8 pages) of manuscript entitled “A guide to using the R Package “multicoll” for detecting multicollinearity” that we would like to submit for your consideration for publication in Computational Economics. Thank you for your time and consideration, Catalina García

Noname manuscript No.
(will be inserted by the editor)

A guide to using the R Package "multiColl" for Detecting Multicollinearity

Román Salmerón · Catalina García* · José García

Received: date / Accepted: date

Abstract The detection of problematic collinearity in a linear regression model is treated in all the existing statistical software packages. However, such detection is not always done adequately. The main shortcomings relate to treatment of independent qualitative variables and completely ignoring the role of the intercept in the model (consequently, ignoring the nonessential collinearity). This paper presents the **R** package **multiColl**, which implements the usually applied measures for detecting near collinearity while overcoming the weaknesses observed in other existing packages.

Keywords Multicollinearity · Detection · Intercept · Dummy · Software

1 Introduction

One of the initial assumptions to estimate a linear regression model by ordinary least squares (OLS) is that the matrix $\mathbf{X}'\mathbf{X}$ is well-conditioned. This condition is not verified when the exogenous variables are linearly dependent and, consequently the estimation of the coefficient has infinite solutions (perfect

Román Salmerón
Department of Quantitative Methods for the Economy and Business, University of Granada (Spain)
Tel.: +34-958248791
E-mail: romansg@ugr.es

Catalina García
Department of Quantitative Methods for the Economy and Business, University of Granada (Spain)
Tel.: +34-958248790
E-mail: cbgarcia@ugr.es

José García
Department of Economics and Business, University of Almería (Spain)
Tel.: +34-958248790
E-mail: jgarcia@ual.es

collinearity). Another case even more worrying is when the linear dependence is not perfect but approximate (near-collinearity) and the estimation by OLS will be unstable leading to inflated variances and covariances, inflated correlations and inflated prediction variance (see, for example, Farrar and Glauber [1], Gunst and Mason [5], Marquardt [10], Marquardt and Snee [11] or Willan and Watts [17]).

For an appropriate analysis of the diagnosis and treatment of multicollinearity, it is important to focus on the diverse causes that were distinguished by Marquardt and Snee [9] into the following types: i) nonessential multicollinearity that exists due to the relation between the intercept and the rest of the independent variables and ii) the essential multicollinearity that exists due to the relation between the independent variables apart from the intercept. This distinction is useful when applying the most common measures of approximate multicollinearity detection presented below, since it be shown that not all measures are useful for detecting both types of approximate multicollinearity and, see Simon and Lesage [15], "even when the intercept coefficient itself is not of interest, near linear relations involving the intercept column are important". These authors also stated that "diagnostics can only be fully effective if they are able to detect and report ill-conditioning arising from both types of collinearity". At the same time, the most appropriate solution for multicollinearity also varies depending on the type of approximate multicollinearity existing in the model.

Despite its relevance, the detection of problematic collinearity in a linear regression model is not always done adequately in different existing statistical software packages. One of the shortcomings is that the role of the intercept in the model is not considered and, consequently, the nonessential collinearity is ignored. Another usual limitation is related to the independent qualitative variables that are treated in some cases as quantitative showing misleading results.

This paper presents the **R** package **multiColl**, which implements the usually applied measures for detecting near collinearity while overcoming the weaknesses observed in other existing packages. The structure of the paper is as follows: Section 2 presents the function which allows to analyze the variations in the estimates of coefficients after small changes in the data and the obtention of the correlation matrix of the model's independent variables and its determinant, the variance inflation factors, the condition number with and without the intercept, the Stewart index and the coefficient of variation. Section 3 illustrates the detection of near multicollinearity in the simple linear regression as a special case not treated in many specialized statistical softwares. Finally, Section 4 summarizes the main contributions of this paper.

2 All detection measures: function *multiColLM*

This section presents the function *multiColLM* which allows to obtain estimates by OLS specifying the quantiles 2.5 and 97.5 of the percentage change

experienced by estimates of model coefficients with a perturbation of 1% in the observations and all the diagnostic measures enumerated in the introduction (see Salmerón et al. [12] for more details).

The dataset of Longley [8] will be used to illustrate this function. This dataset consists of a set of seven economical variables (GNP deflator, GNP, unenployed, armed forces, population, year and employed) observed yearly from 1947 to 1962. In order to illustrate the treatment of qualitative variable, a dummy variable is included that is equal to 1 if the number of persons in armed forces is higher than 200 and zero otherwise. The code and results are as follows:

```
> install.packages("multiColl")
> library(multiColl)
> help(multiColLM)
> attach(longley)
> longley.y = Employed
> dummy = ifelse(Armed.Forces>200, 1, 0, dummy)
> longley.X = cbind(array(1,length(Employed)),longley[,-7],dummy)
> multiColLM(longley.y, longley.X, dummy=TRUE, pos1=8, n=5000,
    mu=5, dv=5, tol=0.01, pos2 = 1:6)

$`Linear Model`

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.12623 -0.07790 -0.04562  0.04243  0.30007

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.710e+03  5.134e+02  -7.227 9.01e-05 ***
xGNP.deflator -3.864e-02  5.021e-02  -0.769  0.46372
xGNP          -6.434e-02  2.028e-02  -3.173  0.01313 *
xUnemployed   -2.344e-02  2.896e-03  -8.092 4.02e-05 ***
xArmed.Forces -2.229e-02  2.983e-03  -7.472 7.11e-05 ***
xPopulation    2.962e-01  1.518e-01   1.951  0.08683 .
xYear          1.934e+00  2.623e-01   7.372 7.82e-05 ***
xdummy         3.107e+00  7.061e-01   4.399  0.00229 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1749 on 8 degrees of freedom
Multiple R-squared:  0.9987,    Adjusted R-squared:  0.9975
F-statistic: 863.2 on 7 and 8 DF,  p-value: 7.151e-11
```

```
$Perturbation
     2.5%     97.5%
 99.94171 103.05738

$multiCol
$multiCol$`Coeficients of Variation`
[1] 0.102761096 0.248230907 0.283339359 0.258497138 0.057358090
    0.002358543

$multiCol$`Proportion of ones in the dummys variable`
[1] 75

$multiCol$`R and det(R)`
$multiCol$`R and det(R)`$`Correlation matrix`
             GNP.deflator GNP       Unemployed Armed.Forces
GNP.deflator 1.0000000    0.9915892 0.6206334  0.4647442
GNP          0.9915892    1.0000000 0.6042609  0.4464368
Unemployed   0.6206334    0.6042609 1.0000000  -0.1774206
Armed.Forces 0.4647442    0.4464368 -0.1774206 1.0000000
Population   0.9791634    0.9910901 0.6865515  0.3644163
Year         0.9911492    0.9952735 0.6682566  0.4172451

               Population Year
GNP.deflator   0.9791634  0.9911492
GNP            0.9910901  0.9952735
Unemployed     0.6865515  0.6682566
Armed.Forces   0.3644163  0.4172451
Population     1.0000000  0.9939528
Year           0.9939528  1.0000000

$multiCol$`R and det(R)`$`Correlation matrix's determinant`
[1] 1.579615e-08

$multiCol$`Variance Inflation Factors`
GNP.deflator GNP Unemployed Armed.Forces Population Year
 135.53244 1788.51348 33.61889 3.58893 399.15102 758.98060

$multiCol$CN
$multiCol$CN$`Condition Number without intercept`
[1] 1296.066

$multiCol$CN$`Condition Number with intercept`
[1] 46046.61

$multiCol$CN$`Increase (in percentage)`
```

```
[1] 97.18532

$multiCol$ki
$multiCol$ki$`Stewart index`
[1] 1.365233e+08 1.297238e+04 3.082410e+04 4.523829e+02
    5.729879e+01 1.215241e+05 1.367397e+08

$multiCol$ki$`Proportion of essential collinearity in i-th
    independent variable (without intercept)`
[1] 1.0447767212 5.8023217457 7.4315117497 6.2635357812
    0.3284541179 0.0005550549

$multiCol$ki$`Proportion of non-essential collinearity in i-th
    independent variable (without intercept)`
[1] 98.95522 94.19768 92.56849 93.73646 99.67155 99.99944
```

In this case, it is observed that small (1% variation) changes in the data imply major changes in the estimates of model coefficients. Apart from the numerical instability, there is no contradiction in the individual and global significance tests. In addition, it is possible to conclude the following:

- The coefficients of variation allow detecting that there are two variables, **Population** and **Year**, that have little variability, especially the latter variable. The measures indicate that these two variables can cause a problematic nonessential near multicollinearity in the model (see Salmerón et al. [14]).
- The existence of coefficients of simple linear correlation higher than $\sqrt{0.9} = 0.9486833$ indicates[1] that there are linear relations between pairs of variables (see García et al. [4]). The relevant variables are **GNP.deflator**, **GNP**, **Population** and **Year**.
- The determinant 0.00000001579615 of the correlation matrix is almost zero[2], indicating a problematic degree of essential multicollinearity.
- The preceding conclusion is confirmed by the calculation of variance inflation factors, 135.53244, 1788.51348, 33.61889, 3.58893, 399.15102 and 758.9806, as various factors have values higher than 10.
- The condition number (with and without the intercept) also indicates that the multicollinearity is problematic. In addition, because the CN increases

[1] This value differs from the threshold equal to 0.7 provided by Halkos and Tsilika [6] to indicate a problem of near collinearity.

[2] García et al. [4] show that values of the determinant of the correlation matrix lower than $0.1013 + 0.00008626 \cdot n - 0.01384 \cdot k$ indicate the presence of problematic near essential multicollinearity. Once again, this value differs from the threshold provided by Field [2], who claims that when the value of the determinant of the correlation matrix is less than 0.00001 there is severe multicollinearity. In the example presented by Halkos and Tsilika [6], the conclusion is that collinearity is not detected since the value of the determinant of the matrix of correlations (0.00663839) is higher than the threshold (0.00001). However, taking into account the paper presented by García et al. [4], the threshold will be 0.01964016 and, consequently, severe near collinearity is detected.

after changing from not accounting for the intercept to accounting for it, a strong essential multicollinearity is indicated.
- Note that the variable dummy is not perturbed and is not considered in the calculation of the coefficients of variation, the matrix of correlations and its determinant, the condition number and the index of Stewart.
- Finally, although there is no threshold for the Stewart index, high values indicate a problematic degree of collinearity. The value obtained for the index associated with the intercept also indicates a strong nonessential multicollinearity. Note also that the percentage of nonessential multicollinearity for variables **Population** and **Year** exceeds 99%.

Thus, to mitigate the detected problem it will be recommended to center the variables **Population** and **Year**, as it is known that this operation resolves the problem. It is also known that this approach to nonessential multicollinearity is not useful for treating essential multicollinearity. Note that the measures coefficients of simple linear correlation, the determinant of the correlation matrix and variance inflations factors indicate that the near essential multicollinearity is worrying. To mitigate this kind of collinearity, it will be necessary to apply other techniques, such as ridge regression (Hoerl and Kennard [7]), raise regression (Salmerón et al. [13]), residualization (York [18] or García et al. [3]) or LASSO (Tibshirani [16]).

3 The special case of the simple linear model

As commented in the introduction, the detection of near multicollinearity in the simple linear regression is a special case not treated in many specialized statistical softwares, see Salmerón et al. [14]. This section illustrates how to detect the multicollinearity in a model that explains the variable *Employed* as a function of GNP where the unique multicollinearity possible to exist is the nonessential one. The code and the results are as follows:

```
> longley.X = cbind(rep(1, length(longley.y)), GNP)
> multiColLM(longley.y, longley.X, n=5000, mu=5, dv=5, tol=0.01,
    pos2 = 1)

$‘Linear Model‘

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.77958 -0.55440 -0.00944  0.34361  1.44594

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 51.843590   0.681372   76.09  < 2e-16 ***
x            0.034752   0.001706   20.37 8.36e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6566 on 14 degrees of freedom
Multiple R-squared:  0.9674,    Adjusted R-squared:  0.965
F-statistic: 415.1 on 1 and 14 DF,  p-value: 8.363e-12

$Perturbation
       2.5%       97.5%
0.008504537 0.542901482

$multiCol
$multiCol$‘Coeficient of Variation‘
[1] 0.2482309

$multiCol$‘Variance Inflation Factor‘
[1] 1

$multiCol$‘Condition Number‘
[1] 8.179275

$multiCol$‘Stewart index‘
[1] 17.22887 17.22887
```

Note that there are situations where small changes in the data imply relevant changes in the results. However, the coefficient of variation and condition number indicate that there is not worrying near multicollinearity.

4 Conclusions

This paper presents the package **multiColl** for detection of multicollinearity in a multiple linear regression. With this package, it is possible to obtain the correlation matrix of the model's independent variables and its determinant, the variance inflation factors, the condition number (with and without the intercept), the Stewart index, the coefficient of variation and the variations in the estimates of coefficients after small changes in the data. Interested reader can also consult Salmerón et al. [12] where some of these functions are developed, the help file generated by **R** and available in `https://cran.r-project.org/web/packages/multiColl/multiColl.pdf` or using the command *help* in the console of **R**.

The main contribution relative to other existing packages is the treatment of qualitative independence and the intercept. Indeed, this package allows the diagnosis of (nonessential) multicollinearity in a simple linear model that is

completely ignored in other existing **R** packages. It is also important to distinguish between essential and nonessential multicollinearity since not only the diagnostic measures but also the mitigation methodologies differ in each case.

References

1. Farrar, D.E., Glauber, R.R.: Multicollinearity in regression analysis: the problem revisited. The Review of Economic and Statistics pp. 92–107 (1967)
2. Field, A.: Discovering statistics using SPSS for Windows (3rd ed). Sage Publications, Los Angeles (2019)
3. García, C., Salmerón, R., García, C., García, J.: Residualization: justification, properties and application. Journal of Applied Statistics (in review) (2019)
4. García, C., Salmerón, R., García, C.: A choice of the ridge factor from the correlation matrix determinant. Journal of Statistical Computation and Simulation **2**(89), 211–231 (2018). URL `https://www.tandfonline.com/doi/abs/10.1080/00949655.2018.1543423?journalCode=gscs20`
5. Gunst, R.F., Mason, R.L.: Advantages of examining multicollinearities in regression analysis. Biometrics pp. 249–260 (1977)
6. Halkos, G., Tsilika, K.: Programming correlation criteria with free cas software. Computational Economics **52**(1), 299–311 (2018)
7. Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12**(1), 55—-67 (1970)
8. Longley, J.: An appraisal of least-squares programs from the point of view of the user. Journal of the American Statistical Association **62**, 819—-841 (1967)
9. Marquardt, D., Snee, R.: Ridge regression in practice. The American Statistician **1**(29), 3–20 (1975). URL `https://www.tandfonline.com/doi/abs/10.1080/00031305.1975.10479105`
10. Marquardt, D.W.: Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. Technometrics **12**(3), 591–612 (1970)
11. Marquardt, D.W., Snee, S.R.: Ridge regression in practice. The American Statistician **29**(1), 3–20 (1975)
12. Salmerón, R., García, C., García, J.: "multicoll": An r package to detect multicollinearity. arXiv preprint arXiv:1910.14590 (2019)
13. Salmerón, R., García, C., García, J., López, M.: The raise estimators. estimation, inference and properties. Communications in Statistics - Theory and Methods **46**(13), 6446—-6462 (2017)
14. Salmerón, R., Rodríguez, A., García, C.: Diagnosis and quantification of the nonessential collinearity. Computational Statistics (2019). URL `https://doi.org/10.1007/s00180-019-00922-x`
15. Simon, D., Lesage, J.: The impact of collinearity involving the intercept term on the numerical acauracy of regression. Computer Science in Economics and Management **1**, 137–152 (1988)
16. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) **58**(1), 267—-288 (1996)
17. Willan, A.R., Watts, D.G.: Meaningful multicollinearity measures. Technometrics **20**(4), 407–412 (1978)
18. York, R.: Residualization is not the answer: Rethinking how to address multicollinearity. Social science research **41**(6), 1379–1386 (2012)