J. Korean Soc. Math. Educ. Ser. B: Pure Appl. Math.
https://doi.org/10.7468/jksmeb.2020.27.2.71
Volume 27, Number 2 (May 2020), Pages 71–81

ISSN(Print) 1226-0657
ISSN(Online) 2287-6081

ILL-CONDITIONING IN LINEAR REGRESSION MODELS AND ITS DIAGNOSTICS

Hamid Ghorbani

Abstract. Multicollinearity is a common problem in linear regression models when two or more regressors are highly correlated, which yields some serious problems for the ordinary least square estimates of the parameters as well as model validation and interpretation. In this paper, first the problem of multicollinearity and its subsequent effects on the linear regression along with some important measures for detecting multicollinearity is reviewed, then the role of eigenvalues and eigenvectors in detecting multicollinearity are bolded. At the end a real data set is evaluated for which the fitted linear regression models is investigated for multicollinearity diagnostics.

1. Introduction

Regression analysis is a commonly used statistical method in various disciplines, including engineering, finance, medicine, social sciences, etc., that allows examining the relationship between a dependent (target) variable and independent (or explanatory) variable(s). However, the efficiency of regression analysis highly depends on correlation structure among predictive variables. Multiple linear regression is the most common form of regression analysis, where we want to predict the value of dependent variable based on the value of two or more independent variables by fitting a linear equation to observed data. A basic assumption in multiple linear regression model is that the matrix of observations on explanatory observations is a full column rank variable matrix, i.e., the rank of this matrix is same as the number of explanatory variables. This implies that all the regressors are independent, i.e., there is no linear relationship among them [3]. Violation of this assumption leads to problems referred to as multicollinearity. This phenomenon of collinearity and near-collinearity was first described by [10]. If within the set of the explanatory variables,

Received by the editors July 22, 2019. Accepted November 04, 2019.
2010 *Mathematics Subject Classification.* Primary 62J20, Secondary 65F35.
Key words and phrases. diagnostic measures, ill-conditioned property, multicollinearity, regression analysis, singularity.

one or more linear relations exist, it is said that these variables are multicollinear which causes the existence of the multicollinearity problem in the regression model. In other words, multicollinearity depends on the sample correlations between the regressors, not on theoretical population quantities.

The reason why multicollinearity is important is that the coefficient estimates of the multiple regression may change erratically in response to small changes in the explanatory variables. Therefore, the ability to detect multicollinearity is important in linear regression analysis. Such detection involves two successive and related steps, firstly detecting the presence of multicollinearity problem, and secondly detecting its strength or severity.

The rest of this paper is organized as follows. First multicollinearity is introduced briefly, then its effects on parameter estimates and model prediction and its different diagnostic indices are discussed. At the end, a real data set is examined, by using the R statistical software [18], for detecting and relaxing the multicollinearity problem among the data.

2. MULTICOLLINEARITY

We will give a short summary of the problem of ill-conditioning in linear regression, that is, the problem of multicollinearity. This problem is highly common in practice when modelling the real data. A key goal of regression analysis is to explain the relationship between one dependent (or response) variable and a set of independent (or regressor) variables.

Assume that data for the dependent variable are arranged in the $n \times 1$ vector y and the data for the explanatory variables are in the $n \times p$ matrix X, known as design matrix. Consider now the multiple linear regression equation:

$$y = X\beta + \varepsilon, \tag{2.1}$$

where β is a $p \times 1$ vector of unknown parameters and ϵ 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.

In the case of perfect multicollinearity (in which one independent variable is an exact linear combination of the others) the design matrix X is not full rank. Therefore, the matrix $X^\mathsf{T}X$ becomes singular and as a result the ordinary least squares (OLS) estimator $\hat{\beta} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}y$ does not exist (if so, it is said that $X^\mathsf{T}X$ becomes ill-conditioned which is the antonym of well-conditioned), see [16].

In the case of near to perfect multicollinearity (or near-linear dependence), problems like large coefficients in absolute value, large variance or standard errors with wider confidence intervals (making the models less accurate and useful), and small t-ratios, are occurred.

In practice the case of perfect multicollinearity is rare, so it is more useful to speak about near to perfect multicollinearity or multicollinearity problem's severity. It is worth mentioning that although multicollinearity makes it hard to interpret the regression coefficients and it reduces the power of the model to identify independent variables that are statistically significant but it does not influence the predictions, the precision of the predictions, and the goodness-of-fit statistics. If ones primary goal is to make predictions, and he/she does not need to understand the role of each independent variable (like what happens in Machine Learning), there is no need to detect multicollinearity and eliminate its sources, see [16], where it is mentioned that:

"The fact that some or all predictor variables are correlated among themselves does not, in general, inhibit our ability to obtain a good fit nor does it tend to affect inferences about mean responses or predictions of new observations."

In general, there are five (primer) sources of multicollinearity [1]:

- The data collection method employed,
- Constraints on the model or in the population,
- Existence of identities or definitional relationships,
- Imprecise formulation of model,
- An over defined model.

3. Effects of Multicollinearity

One issue with multicollinearity in data might be that the coefficient of determination, R^2, will be high so that the regression looks good as a whole (note that multicollinearity does not affect the value of R^2) but some variables are statistically insignificant when they should be significant without multicollinearity. To show this, note that the variance of j-th non-intercept parameter in linear multiple regression model can be expressed as [9]:

$$\mathrm{v}ar(\hat{\beta}_j) = \frac{\sigma^2}{(n-1)s_j^2} \times \left(\frac{1}{1 - R^2_{j|others}} \right),$$

where s_j^2 is the sample variance of j-th regressor, and $R^2_{j|others}$ is the squared coefficient of determination from the regression of j-th regressor on other regressors. It is obvious from the second part of the above equation that, the value of $R^2_{j|.}$ approaches one if we face severe multicollinearity in the model, which yields that the v$ar(\hat{\beta}_j)$ approaches ∞. In such situation, t test statistics, which are the ratio of the coefficients estimates to their standard errors, become smaller. In this case, we might not be able to trust the p-values which identify some independent variables, statistically insignificant. Under extreme (not perfect) multicollinearity, as long as the OLS assumptions do not violate, the OLS estimator $\hat{\beta}$ remain unbiased but it is less accurate, i.e., the addition or deletion of just a few sample observations can substantially change the estimated coefficients. Due to multicollinearity the confidence interval for coefficients estimators will be wider (reflecting greater uncertainty in the estimates) because for example 95 percent confidence bounds are coefficients estimates plus or minus approximately two standard errors, [15]. Multicollinearity also results in the opposite signs of the estimated coefficient, [5]. See [11] where important consequences of multicollinearity in the linear regression are discussed using numeral examples.

Despite the above mentioned effects, multicollinearity does not influence the predictions, the precision of the predictions, and the goodness-of-fit statistics. If the researcher primary goal is just to make predictions, and understanding the (partial) role of each regressor on response does not matter, one can ignore multicollinearity, safety. To explain this, consider for example the following simple regression model,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon, \tag{3.1}$$

assume $x_3 = 2x_2 + x_1$, which indicates a perfect collinearity and a typical OLS solution will not exist because $(X^\mathsf{T} X)^{-1}$ has a singularity. However, let's plug one equation into another:

$$y = \beta_0 + (\beta_1 + \beta_3)x_1 + (\beta_2 + 2\beta_3)x_2 + \varepsilon = \beta_0 + b_1 x_1 + b_2 x_2 + \varepsilon,$$

where $b_1 = \beta_1 + \beta_3$ and $b_2 = \beta_2 + 2\beta_3$. So, clearly we can estimate β_0, b_1 and b_2 by usual OLS method, i.e., there is a solution, and replacing back to estimate the original β_1, β_2 and β_3 parameters, but the only problem is their non-uniqueness. This means, we can choose any $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_3$, which would give us $\hat{b_1} = \hat{\beta}_1 + \hat{\beta}_3$ and $\hat{b_2} = \hat{\beta}_2 + 2\hat{\beta}_3$, i.e., we have infinite number of triple vectors $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ that correspond to a unique solution of $(\hat{b_1}, \hat{b_2})$ pairs. Obviously, any of these triples is as

good as any other for prediction of y. Moreover, all these triples are as good as the unique $(\hat{b_1}, \hat{b_2})$ coefficients for the purpose of forecasting. Indeed, multicollinearity does not affect the so-called predictive power of the model with producing bad predictions. The only problem with multicollinearity is the inference, where for example we are interested to know how x_1 impacts y, where the typical analysis of $\hat{b_1}$ coefficient and its variance will be problematic.

In sum, detection of multicollinearity among regressors is of great importance. In the following, we mainly focus on some multicollinearity diagnostic indexes that help us to detect existence of multicollinearity problem among regressors. Almost all of this indexes has been gathered by [13] and has been mainly recalled in the next section.

4. Multicollinearity Diagnostics

Having defined multicollinearity in a near to perfect rather than a complete sense, we must face the increased problem of detection. Indeed, the researcher whose set of independent variables are perfectly correlated may be more fortunate than one whose variables are nearly so. The former case are soon discovered by the mechanical inability to derive $\hat{\beta}$ while the latter's problems may never be fully understood [20].

In the following the list of some important multicollinearity diagnostic measures along with short description are given.

- Determinant, the matrix $X^{\mathsf{T}}X$ will be singular, or its determinant is zero, if it contains linearly dependent columns. On the other hand, since the correlation matrix R of the scaled X is equal to $X^{\mathsf{T}}X$, the determinant of R can be used as an indicator of multicollinearity existence among regressors. When determinant of R equals one (or it approaches zero), the columns of original X are mutually orthogonal (or are linearly dependent). In the later case, multicollinearity becomes most severe, see [2]. The use of the determinant to detect collinearity has been criticized by [19]. The determinant is excessively sensitive to scaling. For example, the matrix cI_n, whose determinant is c^n and can be made arbitrarily small, has a simple inverse $c^{-1}I_n$, for $c \neq 0$.
- R-squared, the 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. Therefore, higher value of R^2 indicates the more chances of multicollinearity among regressors.

- Farrar χ^2, it is the Chi-square test for detecting the strength of multicollinearity over the complete set of regressors and is defined as $\chi^2 = -(n-1-\frac{1}{6}(2p+5))\log(|X^{\mathsf{T}}X|) \sim \chi^2(-\frac{1}{2}p(p-1))$. The multicollinearity exists among regressors if $\chi^2 > \chi^2(-\frac{1}{2}p(p-1))$, [8]. Alternatively, [12] suggested the modified statistic $\chi^2 = -(n-1-\frac{1}{6}(2p+5))\log(1-|X^{\mathsf{T}}X|)$.
- The variance inflation factor (VIF), is a measures traditionally applied to detect the presence of multicollinearity, and is defined as:

$$\mathrm{V}IF_j = \frac{1}{1-R_j^2},$$

where R_j^2 is the coefficient of determination of a regression of explanatory j on all the other regressors. A $\mathrm{V}IF$ of 5 or 10 and above indicates a multicollinearity problem, [17]. Assuming the standardized regressors (i.e., the columns of design matrix X, to be centred and scaled for unit length), the covariance matrix of the estimated b has the simple form,

$$\mathrm{V}ar(\hat{b}) = \frac{\sigma^2}{n-1}R_X^{-1},$$

where R_X^{-1} is the correlation matrix among the standardized regressors. Then, the diagonal elements of R_X^{-1} are just the $\mathrm{V}IF_j$, [6].
- Condition indexes, the eigenvalues of $X^{\mathsf{T}}X$, say $\lambda_1, \lambda_2, ..., \lambda_p$ can be used to measure the multicollinearity in the data. If there are one or more near-linear dependencies in the data, then one or more of the characteristic roots will be small. One or more small eigenvalues imply that there are near-linear dependencies among the columns of X.

  The condition indexes of the $X^{\mathsf{T}}X$ matrix are defined as,

$$CI_j = \frac{\max(\lambda_i)}{\lambda_j}, j = 1, 2, .., p,$$

the number of condition indexes that are "large" is a useful measure of the number of near-linear dependences in $X^{\mathsf{T}}X$. What is to be considered large has been determined empirically by [3], by which weak dependencies are associated with condition indexes around 5-10, whereas moderate to strong dependencies are associated with condition indexes of 30-100.

Some analysts prefer to examine the condition number of $X^{\mathsf{T}}X$, defined as,

$$\kappa = \frac{\max(\lambda_j)}{\min(\lambda_j)}, j = 1, 2, .., p,$$

generally, condition number less than 100, between 100 and 900 and greater than 900, indicates no serious, moderate to strong and severe multicollinearity problem, respectively, see chapter 5 of [3].

- Kleins rule, Klein [14] argued that it is not necessarily a problem unless the inter-correlation is high relative to the overall degree of multiple correlation i.e., $R_y^2 < R_j^2$, where R_j^2 is from a regression of explanatory variable j on all of the others and R_y^2 is from a regression of response variable y on all explanatory variable.
- The corrected VIF (CVIF), for evaluating the impact of the correlation among regressors in the variance of the OLS, [7] defined a new index of multicollinearity, namely the corrected VIF (CVIF):

$$\mathrm{C}VIF_j = VIF_j \times \frac{1 - R^2}{1 - R_0^2},$$

where $R_0^2 = R_{yx_1}^2 + R_{yx_2}^2 + ... + R_{yx_p}^2$. Collinearity exists if $\mathrm{C}VIF_j \geq 10$.

5. Eigensystem of $X^{\mathsf{T}}X$

As we have seen, the collinearity diagnostics are all functions of the eigenvalues and eigenvectors (eigensystem) of the $X^{\mathsf{T}}X$ matrix in the regression model[1)]. Perfect linear relation in the data matrix $X_{n\times p}$, yields linear system of equations $Xv = 0$ that allow a solution for v to be obtained. It can be shown that for each exact linear dependency among the columns of X there is one zero eigen value of $X^{\mathsf{T}}X$. However, when there is no perfect multicollinearity but near to perfect multicollinearity in X we need to find one or more non-zero vectors v such that $Xv = a$ with $a \neq 0$ but close to zero or equivalently the length of a, $\|a\|$ should be small. Since finding the set v values which makes $\|a\| = \sqrt{v^{\mathsf{T}}X^{\mathsf{T}}Xv}$ small has non unique solution without considering any condition on v's, we search for v within those vectors which have unit length, i.e $\|v\| = 1$.

To find desired vector v and the corresponding minimum length $\|a\|$, consider the

[1)] or equivalently, the eigensystem of the correlation matrix of the predictors in the regression model.

singular-value decomposition (SVD) of the matrix $X = UDV^{\mathsf{T}}$, where X is an arbitrary $n \times p$ matrix, $n \geq p$, U is $n \times p$, V is $p \times p$ satisfying $U^{\mathsf{T}}U = V^{\mathsf{T}}V = I_p$ and $D = diag\{\lambda_1, ..., \lambda_p\}$ is a $p \times p$ non-negative and diagonal matrix of singular values (or eigenvalues) of X with $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. Now

$$\|a\|^2 = \|Xv\|^2 = v^{\mathsf{T}}X^{\mathsf{T}}Xv = v^{\mathsf{T}}VD^{\mathsf{T}}U^{\mathsf{T}}UDV^{\mathsf{T}}v = v^{\mathsf{T}}VD^2V^{\mathsf{T}}v,$$

according to [3] the minimum length $\|a\|$ and the corresponding vector v which produce this minimum length a are the positive square root of the smallest eigenvalues and corresponding eigenvector of $X^{\mathsf{T}}X$. Therefore the length of $\|a\|$ in the linear system $Xv = a$ that address the multicollinearity problem in columns of X is defined in terms of the square roots of the eigenvalues of $X^{\mathsf{T}}X$, which are the singular values of X.

The relevance of this is when we want to find least square estimator of the parameter β in the linear regression model $y = X\beta + \varepsilon$ by solving the linear system $(X^{\mathsf{T}}X)^{-1}\hat{\beta} = X^{\mathsf{T}}y$ with variance-covariance matrix $\sigma^2(X^{\mathsf{T}}X)^{-1}$, where multicollinearity in data matrix X yield ill-conditioning problem in matrix $X^{\mathsf{T}}X$ causing $(X^{\mathsf{T}}X)^{-1}\hat{\beta} = X^{\mathsf{T}}y$ do not have a unique solution and its variance-covariance matrix to be numerically unstable.

Therefore the eigensystem of the matrix $X^{\mathsf{T}}X$ has been used for many years when dealing with multicollinearity problem in the linear regression model.

Note that the SVD of the matrix X is mathematically equivalent to determining the eigensystem of the matrix $X^{\mathsf{T}}X$, since if $X = UDV^{\mathsf{T}}$, the column vectors of V and the squares of singular values (diagonal elements of D) are the eigenvectors and the eigenvalues of the of the matrix $X^{\mathsf{T}}X$, respectively[2)].

Despite this mathematical equivalency between two methods when calculating the eigensystem of $X^{\mathsf{T}}X$, [4] gives some reasons to explain why the use of SVD of an ill-conditioned matrix X is preferred regarding numerical stability of numerical algorithm developed for computing SVD. The ill-conditioning in X is reflected in the size of singular values. The extent of ill-conditioning is described by how small is j-th singular values relative to maximum of all singular values. When anyone singular values of the data matrix X, λ_i is small relative to $\lambda_{\max}$ it may be interpreted as indicative of a near dependency between columns of X. The ratios $\frac{\lambda_{\max}}{\lambda_j}$ are the condition indexes of matrix X [3)] and according to [4] condition indexes less than

[2)]The columns of U are the eigenvectors of XX^{T}, as well.

[3)]The condition indexes are at least equal to one and their maximum is the condition number.

10 indicates weak dependencies weak dependencies, whereas moderate to strong relations are associated with values between 30 to 100.
One issue with the condition index is the role of measurement unit of explanatory variables (columns of X) on its value. This situation is known as artificial ill-conditioning. Before computing the singular values of data matrix X, each column should be scaled to have unit length but centering the columns must be avoided because it masks the role of the constant term in any underlying near-dependencies, [3]. This scaling ensures that the units of measurement of the original variables do not influence the value of condition indexes which are referred to scaled condition indexes of X.

6. Illustrative Example

Table 1. The estimated parameter using OLS method and variance inflation factors for PCD data.

Coefficients	Unscaled estimate	Scaled estimate	Std. Error	t value	$Pr(>\|t\|)$	VIF
Intercept	62.40	0.179	0.201	0.89	0.40	0
X_1	1.55	0.150	0.072	2.08	0.07	38
X_2	0.510	0.150	0.378	0.77	0.50	254
X_3	0.102	0.014	0.104	0.14	0.90	47
X_4	249.58	-0.144	0.250	-0.20	0.84	283

Table 2. Condition indexes.

		Scaled condition	Variance Decomposition Proportions				
No	Eigenvalues	index	intercept	X_1	X_2	X_3	X_4
1	4.120	1.00	0.0000	0.0004	0.0000	0.0002	0.0000
2	0.554	2.73	0.0000	0.0100	0.0000	0.0027	0.0001
3	0.299	3.78	0.0000	0.0006	0.0003	0.0016	0.0017
4	0.038	10.46	0.0001	0.0574	0.0028	0.0460	0.0009
5	0.0001	249.58	0.9999	0.9316	0.9969	0.9499	0.9973

As a leading example, we use the Portland Cement Dataset (PCD) originally due to [21]. This dataset contains 13 observations on the following five variables: Y (The heat evolved after 180 days of caring in calories per gram), X_1 (Tricalcium Aluminate), X_1(Tricalcium Silicate), X_3 (Tetracalcium Aluminoferrite)and X_4 (Dicalcium Silicate). For these data, the question is how well the heat evolved can be

explained by by the other variables using the following regression model

$$Y = \beta_0 + \Sigma_{i=1}^{4} \beta_i X_i \ + \ \varepsilon$$

To permit direct comparison of the variable coefficients, all variables were rescaled to have unit length.

The model fits very well, with $R^2 = 0.98$; however, the partial t-tests for parameters shown in Table 1 indicate that there is no significant predictor at 0.05 significant level in the presence of other variables. These two facts together are amongst the signs of multicollinearity problem in data. Table 1 also shows the variance inflation factors. By the rules of thumb described later, all predictors have potentially serious problems of collinearity. The condition indices and coefficient variance decomposition proportions are given in Table 2.

REFERENCES

1. M.P. Allen: The problem of multicollinearity. In: Understanding Regression Analysis. Springer, Boston, (1997), 176-180.
2. D. Asteriou & S.G. Hall: Applied Econometrics: A Modern Approach Using Eviews and Microfit. Palgrave Macmillan Pub., New York, 2007.
3. D.A. Belsley: Conditioning Diagnostics: Collinearity and Weak Data Regression. John Wiley & Sons, New York, 1991.
4. D.A. Belsley, E. Kuh & R.E. Welsch: Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley & Sons, New York, 1980.
5. N. Billington: The location of foreign direct investment: an empirical analysis. Applied Economics **31** (1999), no. 1, 65-76.
6. S. Chatterjee & A. S. Hadi: Regression Analysis by Example (5th ed.). John Wiley & Sons, New York, 2012.
7. J.D. Curto & J.D. Pinto: The corrected VIF (CVIF). Journal of Applied Statistics **38** (2011), no. 7, 495-505.
8. D.E. Farrar & R.R. Glauber: Multicollinearity in regression analysis: the problem revisited. Review of Economics and Statistics **49** (1967), no. 1, 92107.
9. M. Friendly & E. Kwan: Where's Waldo? Visualizing Collinearity Diagnostics. The American Statistician **63** (2009), no. 1, 56-65.
10. R. Frisch: Statistical Confluence Analysis by Means of Complete Regression Systems. Universitetets Okonomiske Institutt., Oslo, 1934.
11. J. Groß: Linear Regression. Springer-Verlag, Berlin, Heidelberg, 2003.
12. Y. Haitovsky: Multicollinearity in Regression Analysis: Comment. The Review of Economics and Statistics **51** (2002), no. 4, 486-89.

13. M. Imdadullah, M. Aslam & S. Altaf: mctest: an R package for detection of collinearity among regressors. The R Journal **8** (2016), no. 2, 495-505.
14. R. Klein: An Introduction to Econometrics. Prentic-Hall Pub., Englewood, Cliffs, N. J., 1962.
15. C. Mela & P. Kopalle: The Impact of Collinearity on Regression Analysis: The Asymmetric Effect of Negative and Positive Correlations. Applied Economics **34** (2002), 667-77.
16. D.C. Montgomery, E.A. Peck & G.G. Vining: Introduction to Linear Regression Analysis (5th ed.), John Wiley & Sons, (2012).
17. R.M. OBrien: A caution regarding rules of thumb for variance inflation factors. Quality & Quantity **41** (2007), no. 5, 673-679.
18. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, (2019), url=https://www.R-project.org.
19. G.W. Stewart: Collinearity and Least Squares Regression. Statist. Sci. **2** (1987), no. 1, 68-100.
20. E.C. Willis & D.R. Perlack: Multicollinearity: Effects, Symptoms, and Remedies. Journal of the Northeastern Agricultural Economics Council. **7** (1978), 55-61.
21. H. Woods, H. Steinour & H.R. Starke: Effect of Composition of Portland Cement on Heat Evolved during hardening. Industrial & Engineering Chemistry **24** (1932), no. 11, 1207-1214.

Faculty of Mathematical Sciences, University of Kashan, Kashan, I. R. Iran
Email address: hamidghorbani@kashanu.ac.ir