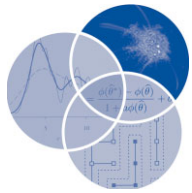
Focus Article



Multicollinearity

Aylin Alin*

Multicollinearity refers to the linear relation among two or more variables. It is a data problem which may cause serious difficulty with the reliability of the estimates of the model parameters. In this article, multicollinearity among the explanatory variables in the multiple linear regression model is considered. Its effects on the linear regression model and some multicollinearity diagnostics for this model are presented. © 2010 John Wiley & Sons, Inc. *WIREs Comp Stat* 2010 2 370–374

Multicollinearity refers to the linear relationship among two or more variables, which also means lack of orthogonality among them. This relation is also called *collinearity* or *ill conditioning* by some authors like Belsley[1] and Chatterjee and Hadi.[2] In more technical terms, *multicollinearity* occurs if k vectors lie in a subspace of dimension less than k. This is the definition of *exact multicollinearity* or *exact linear dependence*. It is not necessary for multicollinearity to be exact in order to cause a problem. It is enough to have k variables nearly dependent, which occurs if the angle between one variable and its orthogonal projections onto others is small. Through this article, multicollinearity will mean near dependence. Multicollinearity is a condition of deficient data, which frequently is encountered in observational studies in which the investigator does not interfere with the study. In many studies, multicollinearity has been confused with correlation. Correlation is the linear relationship between only two variables, whereas multicollinearity can exist between two variables or between one variable and linear combination of the others. Thus, correlation is a special case of multicollinearity. High correlation implies multicollinearity, but the converse is not true. One can have multicollinearity among explanatory variables, but still not have high correlation between pairs of these variables.

Multicollinearity creates difficulties when one builds a regression model between response variable $\mathbf{y}$ and explanatory variable $\mathbf{X}$. Through this article, the focus will be on the multicollinearity problem among the columns of $\mathbf{X}$ and the multiple linear regression between $\mathbf{y}$ and $\mathbf{X}$. The multiple linear regression model between $\mathbf{y}$ and $\mathbf{X}$ is given in Eq. (1).

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{y}$ is the $n \times 1$ vector of response variable, $\mathbf{X}$ is the matrix of explanatory variables with dimensions $n \times (k+1)$ with the first column consisting of the vector of ones, $\boldsymbol{\beta}$ is the $(k+1) \times 1$ vector for regression coefficients, and $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of errors, which is assumed to have mean $\mathbf{0}$ and variance covariance matrix $\sigma^2\mathbf{I}$. If the explanatory variables are centered, the model given in Eq. (1) does not include intercept and then there is no need to include the vector of ones in $\mathbf{X}$ matrix. In this case, the $\mathbf{X}$ matrix will have k columns whereas $\boldsymbol{\beta}$ will have k rows. The least squares estimates of $\boldsymbol{\beta}$ is obtained as follows:

$$b = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{2}$$

EFFECTS OF MULTICOLLINEARITY

If there is multicollinearity, the signs of b_i's for $i = 1, 2, \ldots, k$ may be wrong, i.e., different from the signs of correlation between the corresponding explanatory variable and response variable. However, the signs of the regression coefficients can be 'right' even when there is a high degree of multicollinearity and 'wrong' when there is essentially no multicollinearity (Ref 3, p. 131). The most serious effect of multicollinearity is that b_i's will have large standard errors ($\sigma^2_{b_i}$), i.e., large sampling variability. This makes these coefficients unreliable and, therefore, decreases their precision. The more the multicollinearity is exact, the less reliable are the estimates. In addition to having unreliable coefficients, which may vary from one sample to other, the inflated variances of coefficient estimates harm hypothesis testing, estimation, and forecasting. The regression coefficient b_i is interpreted

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: aylin.alin@deu.edu.tr

Department of Statistics, Dokuz Eylul University, Izmir, Turkey

DOI: 10.1002/wics.84

as the change in the expected value of $\mathbf{y}$ when $\mathbf{x}_i$ is increased by one unit under the condition that all remaining explanatory variables are held constant. The b_i is called *partial regression coefficient*. This interpretation is not applicable in case of multicollinearity, because it may not be possible in practice to vary one explanatory variable and hold other constant. Therefore, multicollinearity makes it hard to measure the marginal effect of explanatory variables.

In multiple linear regression, extra (sequential) sum of squares (SSR) measures the marginal contribution of an explanatory variable in reducing the error sum of squares (SSE). When there is multicollinearity among explanatory variables, the marginal contribution of any one variable in reducing the SSE depends on which other variables are already in the regression model. For instance, consider the data given by Wold.[4] Data consist of seven explanatory variables. When there is only $\mathbf{x}_1$ in the model, $\text{SSR}(\mathbf{x}_1) = 56.964$, but when $\mathbf{x}_2$ and $\mathbf{x}_3$ are in the model, the marginal contribution of $\mathbf{x}_1$ is $\text{SSR}(\mathbf{x}_1|\mathbf{x}_2, \mathbf{x}_3) = 1.308$, which is very small when compared with that of SSR ($\mathbf{x}_1$). The explanation for this is the multicollinearity problem among these variables, which will be shown numerically in the following section. The marginal contribution of $\mathbf{x}_1$ is small when $\mathbf{x}_2$ and $\mathbf{x}_3$ are already in the model, because $\mathbf{x}_2$ and $\mathbf{x}_3$ contain much of the same information as $\mathbf{x}_1$. Including a variable having multicollinearity with others already in the model contributes nothing except decreasing the degrees of freedom. Combining large sampling variability of the coefficient estimate and small marginal contribution to the reduction of SSE, the partial regression coefficients will be insignificant.

Multicollinearity does not affect the model significance given in Eq. (1). Actually, even though all variables in the model are insignificant, we can still have significant model. For the data mentioned above, the *P*-value for the analysis of variance performed for model significance is 0.007 even though the *P*-values for the significance of partial regression coefficients are 0.584, 0.067, and 0.615 for $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$, respectively. For model significance, we test the null hypothesis H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$. Hence, the reduced model does not contain any of the variables and it does not matter whether there is multicollinearity or not. However, the null hypothesis for the significance of *i*th partial regression coefficient is H_0: $\beta_i = 0$. The reduced model will include all the variables already in the model except $\mathbf{x}_i$. Unlike the model significance, the multicollinearity will matter this time. However, multicollinearity does not have to affect the precision of the mean response or prediction of new observations if these inferences are made within the range of observations (see the example in Ref 5, p. 293).

Because multicollinearity is not a modeling error, the problems, which may be caused by multicollinearity, should be investigated after the model has been satisfactorily specified. Belsley[1] (p. 26) emphasizes that some of the problems caused by multicollinearity such as the sensitivity of the regression coefficients to the addition or deletion of a variable or observation or having explanatory variables with low *t*-statistics may be the result of combination of building the wrong model and having influential data without any presence of multicollinearity.

DIAGNOSING MULTICOLLINEARITY

The effects of multicollinearity mentioned in the previous section can also be used as an indication of multicollinearity, but this would be informal. The mostly used diagnosis is to examine the correlation matrix of explanatory variables. As correlation and collinearity are not same, there can still be multicollinearity even when all correlations are low. Hence, using correlation matrix $\mathbf{R}$ is not enough. Determinant of $\mathbf{R}$ ($\det\mathbf{R}$) is another diagnostic for multicollinearity but it suffers from the same problem with pair-wise correlations. Even though determinant is very close to one, there still might be multicollinearity among the columns of $\mathbf{X}$. Furthermore, both pair-wise correlations and $\det\mathbf{R}$ do not reveal the number of coexisting relations and their structure.

Variance inflation factor (VIF) given in Eq. (3) is very popular as a multicollinearity diagnostic.

$$\text{VIF}_i = \frac{1}{1 - R_i^2} \text{ for } i = 1, 2, \ldots, k \tag{3}$$

R_i^2 is the coefficient of multiple determination of $\mathbf{x}_i$ on the remaining explanatory variables. Diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}$ are equal to VIF values when $\mathbf{X}$ is standardized. VIF measures the increase in $\sigma_{b_i}^2$ because of multicollinearity relative to the variance that would result if there was no multicollinearity. The larger the VIF, the more $\sigma_{b_i}^2$ is inflated. A large value of VIF_i indicates the involvement of $\mathbf{x}_i$ in at least some linear dependency, but does not reveal which $\mathbf{x}_i$. The threshold value to deviate small from large is generally taken as 10. The mean of VIF values is also used as an index for multicollinearity. The mean considerably larger than one is an indication of multicollinearity. For the data including seven explanatory variables given by Wold[4], the VIF values are 307.1, 24.1, 25.7,

 © 2010 John Wiley & Sons, Inc.

TABLE 1 | Eigenvalues, Eigenvectors and Condition Indexes for the Data Given in Ref 15

λ	0.027	0.133	0.421	0.555	1.512	40.047	83.305
	−0.659	0.128	0.384	**0.397**	−0.088	−0.210	0.439
	−0.012	−0.001	0.234	**−0.805**	0.217	−0.276	0.417
	−0.052	0.052	**0.653**	−0.252	−0.516	0.241	−0.425
υ	0.102	**−0.798**	0.182	0.089	0.007	0.432	0.354
	−0.125	0.197	**−0.477**	−0.216	−0.640	0.325	0.396
	0.561	−0.112	0.114	0.208	−0.472	−0.593	0.207
	0.472	**0.542**	0.313	0.182	0.215	0.422	0.358
η	55.947	25.032	14.071	12.248	7.423	1.442	1.000

TABLE 2 | Variance Decomposition Proportions for the Data Given in Ref 15

η	$\sigma^2_{b_1}$	$\sigma^2_{b_2}$	$\sigma^2_{b_3}$	$\sigma^2_{b_4}$	$\sigma^2_{b_5}$	$\sigma^2_{b_6}$	$\sigma^2_{b_7}$
55.947	**0.955**	0.004	0.072	0.074	0.329	**0.970**	**0.768**
25.032	0.007	0.000	0.014	**0.907**	0.164	0.008	**0.202**
14.071	0.021	0.097	**0.709**	0.015	**0.305**	0.003	0.021
12.248	0.017	**0.872**	0.080	0.003	0.047	0.006	0.005

95.1, 32, 219.1, and 196.4 for $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$, $\mathbf{x}_5$, $\mathbf{x}_6$, and $\mathbf{x}_7$, respectively. Very large VIF values are indicators of multicollinearity problem for this data. Moreover, the mean of these values is 128.5, which is much larger than one. Despite its popularity, VIF is still unable to distinguish the number of coexisting linear dependencies and their structure.

Examining the eigenvalues (λ_s) and eigenvectors (υ_s) of $\mathbf{R}$ or $\mathbf{X}'\mathbf{X}$ yields another method to diagnose multicollinearity for $s = 1, 2, \ldots, k$. This method, also called a *principal component approach*, reveals not only the number of linear dependencies but also the structural relations within these relations. Small λ values indicate the presence of multicollinearity and there will be one small λ value for each linear dependency. The problem is to determine what small is. Chatterjee and Hadi[2] (p. 243) suggest that if one λ value is much smaller than the others (and near zero), then multicollinearity is present. The elements of υ corresponding small λ value help to diagnose the variables involved in the corresponding linear relation. The nonzero elements of υ indicate the variables which are in the linear relation, whereas zero elements indicate which are not. However, Belsley[1] (p. 36) argues this criterion. He claims that it is possible for an element of an eigenvector to be arbitrarily small even though the corresponding variable still belongs to the linear relation. Examining the eigenvalues of $\mathbf{X}'\mathbf{X}$ is equivalent to examining singular values of $\mathbf{X}(\mu_s)$ where $\mu_s = (\lambda_s)^{1/2}$ (see Ref 6, pp. 91–103, 334–336 and Ref 1, pp. 41–44 for more information about eigenvalues, eigenvectors, and singular values). The sum of reciprocals of λ_s for $s = 1, 2, \ldots, k$ is also used as multicollinearity diagnostic. The rule of thumb is that if this sum is larger than $5k$ then multicollinearity is present.

Condition number given in Eq. (4) is another measure to diagnose the overall multicollinearity. $\kappa \gg 1$ indicates evidence of strong multicollinearity.

$$\kappa = \sqrt{\frac{\lambda_{\max} \text{ of } \mathbf{X}'\mathbf{X}}{\lambda_{\min} \text{ of } \mathbf{X}'\mathbf{X}}} = \frac{\mu_{\max} \text{ of } \mathbf{X}}{\mu_{\min} \text{ of } \mathbf{X}} \tag{4}$$

κ is the largest among the condition indexes $\eta_s = \mu_{\max}$ of $\mathbf{X}/\mu_s$ of $\mathbf{X}$ defined by Belsley[1] (p. 55). η_s around 5–10 indicates weak multicollinearity and 30–100 indicates moderate to strong multicollinearity.

It is mentioned that multicollinearity causes inflation at $\sigma^2_{b_i}$. Variance decomposition proportions are the multicollinearity diagnostics developed based on this idea. These proportions are defined as

$$\pi_{si} = \frac{c_{is}}{\sum_{s=1}^{k} c_{is}}, \text{ where } c_{is} = \frac{\upsilon_{is}^2}{\lambda_s} \tag{5}$$

υ_{is} is the ith element in υ_s. π_{si} is the proportion of VIF of the ith regression coefficient associated with the multicollinearity represented with λ_s. In other

 © 2010 John Wiley & Sons, Inc.

words, it shows the extent to which $\sigma^2_{b_i}$ is inflated by multicollinearity corresponding to small λ_s or equivalently large η_s. As eigenvector approach, variance decomposition proportions enable us to determine which variables are involved in which linear relations.

Table 1 includes the eigenvalues, eigenvectors, and condition indexes for the data given by Wold.[4] These values have been obtained using variables centered and scaled to have zero mean and unit length to make the calculations easier and the comparisons more meaningful. The λ values and condition indexes suggest that there are four coexisting linear dependencies in these data from weak to stronger. The elements of ν given in bold represent the variables involved in these dependencies. It is clear that $\mathbf{x}_1$, $\mathbf{x}_6$, and $\mathbf{x}_7$ are involved in the strongest linear relation. Variance proportions for the four linear dependencies presented with Table 2 reveal that this dominating linear relation accounts for 96, 97, and 77% of the inflation at $\sigma^2_{b_1}$, $\sigma^2_{b_6}$, and $\sigma^2_{b_7}$, respectively. These are all given in bold in Table 2. The values in the eigenvector corresponding to the second smallest eigenvalue reveal that $\mathbf{x}_4$ and $\mathbf{x}_7$ are included in this relation and the variance proportions show that 91 and 20% of the inflation at $\sigma^2_{b_4}$ and $\sigma^2_{b_7}$, respectively are accounted by this relation.

There are other diagnostic procedures, which are less popular such as the ones proposed by Farrar and Glauber[7] and Haitovsky.[8] These procedures are based on statistical hypothesis testing. However, they have been criticized by some authors, e.g., Kumar,[9] Wichers[10] and Parker and Smith.[11]

It should be noted that multicollinearity may not be harmful. Harmful effects of multicollinearity could be counteracted by a sufficiently small error variance σ^2. The diagnostics given in this text only include information in $\mathbf{X}$. We need different diagnostics incorporating the size of σ^2. Belsley[1] suggested signal to noise ratios to detect harmful multicollinearity.

Most of the authors suggest that effects of multicollinearity may be removed by having more data or introducing prior information. Harvey[12] proposes that multicollinearity may be reduced by imposing appropriate restrictions, though it cannot, in general, be eliminated. Soper[13] proposes a technique to reduce multicollinearity in linear regression, which is also discussed by Kennedy.[14] There are other methods alternative to linear regression when multicollinearity is present in $\mathbf{X}$ such as ridge regression, principal component regression, and partial least squares regression. Even though these methods give biased estimates, these estimates have smaller variance compared with ordinary least squares estimates. See Refs 15–18 for more information about these methods.

CONCLUSION

Multicollinearity problem among the explanatory variables, its effects and diagnostics to determine multicollinearity have been presented. As in most of the text books and articles, multicollinearity in the linear regression has been studied in this article. However, it is also a problem for other generalized linear models such as the logistic regression, which is also a very popular technique, especially for medical studies. See Ryan[3] and Hosmer and Lemeshow[19] for the detection of multicollinearity in logistic regression.

The multicollinearity diagnostics are available in most of the statistical software. The correlation matrix and VIF values may be obtained easily using Minitab, SAS, and SPSS. SAS also provides eigenvalues, eigenvectors, and variance decomposition proportions. Moreover, MATLAB code for a function calculating eigenvalues, eigenvectors, condition indexes, and variance decomposition proportions will be provided in the electronic version of this text along with options for using original $\mathbf{X}$ matrix, scaled but not centered $\mathbf{X}$ matrix, or both centered and scaled $\mathbf{X}$ matrix.

REFERENCES

1. Belsley DA. *Conditioning diagnostics collinearity and weak data in regression*. New York: John Wiley & Sons; 1991.
2. Chatterjee S, Hadi AS. *Regression Analysis by Example*. 4th ed. New York: John Wiley & Sons; 2006.
3. Ryan TP. *Modern Regression Methods*. New York: John Wiley & Sons; 1997.
4. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst* 2001, 58:109–130.
5. Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. *Applied Linear Statistical Models*. 4th ed. Chicago: Irwin; 1990.
6. Seber GAF. *A Matrix Handbook for Statisticians*. New York: John Wiley & Sons; 2008.
7. Farrar DE, Glauber RR. Multicollinearity in regression analysis: the problem revisited. *Rev Econ Stat* 1967, 49(1):92–107.
8. Haitovsky Y. Multicollinearity in regression analysis: comment. *Rev Econ Stat* 1969, 51(4):486–489.

 © 2010 John Wiley & Sons, Inc.

9. Kumar TK. Multicollinearity in regression analysis. *Rev Econ Stat* 1975, 57(3):365–366.

10. Wichers CR. The detection of multicollinearity: a comment. *Rev Econ Stat* 1975, 57(3):366–368.

11. Parker RN, Smith MD. High correlations or multicollinearity, and what to do about either: reply to light. *Soc Forces* 1984, 62(3):804–807.

12. Harvey AC. Some comments on multicollinearity in regression. *Appl Stat* 1977, 26(2):188–191.

13. Soper JC. Second generation research in economic education: problems of specification and interdependence. *J Econ Educ* 1976, 8(2):40–48.

14. Kennedy PE. Eliminating problems caused by multicollinearity: a warning. *J Econ Educ* 1982, 13(1):62–64.

15. Geladi P, Kowalski BR. Partial least squares regression: a tutorial. *Anal Chim Acta* 1986, 185:1–17.

16. Höskuldsson A. PLS regression methods. *J Chemom* 1988, 2:211–228.

17. Frank IE, Friedman JH. A statistical view of some chemometrics regression tools. *Technometrics* 1993, 35(2):109–148.

18. McDonald GC. Ridge regression. *WIREs Comput Stat* 2009, 1:93–100.

19. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. New York: John Wiley & Sons; 1989.

 © 2010 John Wiley & Sons, Inc.