MULTICOLLINEARITY: CAUSES, EFFECTS AND REMEDIES

RANJIT KUMAR PAUL
M. Sc. (Agricultural Statistics), Roll No. 4405
I.A.S.R.I, Library Avenue, New Delhi-110012

Chairperson: Dr. L. M. Bhar

Abstract: If there is no linear relationship between the regressors, they are said to be orthogonal. Multicollinearity is a case of multiple regression in which the predictor variables are themselves highly correlated. If the goal is to understand how the various X variables impact Y, then multicollinearity is a big problem. Multicollinearity is a matter of degree, not a matter of presence or absence. In presence of multicollinearity the ordinary least squares estimators are imprecisely estimated. There are several methods available in literature for detection of multicollinearity. By observing correlation matrix, variance influence factor (VIF), eigenvalues of the correlation matrix, one can detect the presence of multicollinearity. The degree of the multicollinearity becomes more severe as $|\mathbf{X}'\mathbf{X}|$ approaches zero. Complete elimination of multicollinearity is not possible but the degree of multicollinearity can be reduced by adopting ridge regression, principal components regression, etc.

Key words: *Multicollinearity*, *Multiple Regression*, *Variance Influence Factor (VIF)*, *Ridge Regression*, *Principal Components*.

1. Introduction

The use of multiple regression model often depends on the estimates of the individual regression coefficients. Some examples of inferences that are frequently made include

1. Identifying the relative effects of the regressor variables,
2. Prediction and/or estimation, and
3. Selection of an appropriate set of variables for the model.

If there is no linear relationship between the regressors, they are said to be orthogonal. When the regressors are orthogonal, the inferences such as those illustrated above can be made relatively easily. Unfortunately, in most applications of regression, the regressors are not orthogonal. Sometimes the lack of orthogonality is not serious. However, in some situations the regressors are nearly perfectly linearly related, and in such cases the inferences based on the regression model can be misleading or erroneous. When there are near-linear dependencies among the regressors, the problem of multicollinearity is said to exist.

Multicollinearity is a case of multiple regression in which the predictor variables are themselves highly correlated. One of the purposes of a regression model is to find out to what extent the outcome (dependent variable) can be predicted by the independent variables. The strength of the prediction is indicated by R^2, also known as variance explained or strength of determination.

2. Multicollinearity

2.1 Multiple Regression

A regression model that involves more than one regressor variables is called a multiple regression model. Or in other words it is a linear relationship between a dependent variable and a group of independent variables. Multiple regression fits a model to predict a dependent (Y) variable from two or more independent (X) variables. Multiple linear regression models are often used as approximating functions. That is, true functional relationship between y and $x_1, x_2, \ldots, x_k$ is unknown, but over certain ranges of the regressor variables the linear regression model is an adequate approximation to the true unknown function. If the model fits the data well, the overall R^2 value will be high, and the corresponding P value will be low (P value is the observed significance level at which the null hypothesis is rejected). In addition to the overall P value, multiple regressions also report an individual P value for each independent variable. A low P value here means that this particular independent variable significantly improves the fit of the model. It is calculated by comparing the goodness-of-fit of the entire model to the goodness-of-fit when that independent variable is omitted. If the fit is much worse when that variable is omitted from the model, the P value will be low, telling that the variable has a significant impact on the model.

2.2 Multicollinearity

The term multicollinearity refers to a situation in which there is an exact (or nearly exact) linear relation among two or more of the input variables, [Hawking, 1983]. Exact relations usually arise by mistake or lack of understanding. For example, the user feels that x_1 and x_2 are important and that their sum is also important so $x_3 = x_1 + x_2$ is also included. Clearly, there is no additional information in x_3 and hence one of these variables should be selected.

Consider the following multiple regression models

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \ , \qquad \ldots(2.1)$$

where $\mathbf{y}$ is an $n \times 1$ vector of responses, $\mathbf{X}$ is an $n \times p$ matrix of the regressor variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors, with $\varepsilon_i \sim \text{IID } N(0, \sigma^2)$. It will be convenient to assume that the regressor variables are standardized. Consequently, $\mathbf{X}'\mathbf{X}$ is a $p \times p$ matrix of correlations between the regressors and $\mathbf{X}`\mathbf{y}$ is a $p \times 1$ vector of correlation between the regressors and the response.

Let the j^{th} column of $\mathbf{X}$ matrix be denoted by $\mathbf{X}_j$, so that $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_p]$. Thus $\mathbf{X}_j$ contains the n levels of the regressor variable. Formally multicollinearity can be defined as the linear dependence of the columns of $\mathbf{X}$. The vectors are linearly dependent if there is a set of constants $t_1, t_2, ..., t_p$, not all zero such that

$$\sum_{j=1}^{p} t_j \mathbf{X}_j = \mathbf{0} \qquad \ldots(2.2)$$

If Equation (2.2) holds exactly for a subset of the columns of $\mathbf{X}$, then the rank of the $\mathbf{X}'\mathbf{X}$ matrix is less than p and $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist. However suppose the Equation (2.2) is

approximately true for some subset of the columns of **X**. Then there will be a near linear dependency in $\mathbf{X}'\mathbf{X}$ and the problem of multicollinearity is said to exist. It is to be noted that the multicollinearity is a form of ill-conditioning in the $\mathbf{X}'\mathbf{X}$ matrix. Furthermore, the problem is one of the degrees, that is, every data set will suffer from multicollinearity to some extent unless the columns of **X** are orthogonal. The presence of multicollinearity can make the usual least-squares analysis of the regression model dramatically inadequate.

In some cases, multiple regression results may seem paradoxical. Even though the overall P value is very low, all of the individual P values are high. This means that the model fits the data well, even though none of the X variables has a statistically significant impact on predicting Y. How is this possible? When two X variables are highly correlated, they both convey essentially the same information. In this case, neither may contribute significantly to the model after the other one is included. But together they contribute a lot. If both variables are removed from the model, the fit would be much worse. So the overall model fits the data well, but neither X variable makes a significant contribution when it is added to the model. When this happens, the X variables are collinear and the results show multicollinearity.

2.3 Why is Multicollinearity a Problem?

If the goal is simply to predict Y from a set of X variables, then multicollinearity is not a problem. The predictions will still be accurate, and the overall R^2 (or adjusted R^2) quantifies how well the model predicts the Y values.

If the goal is to understand how the various X variables impact Y, then multicollinearity is a big problem. One problem is that the individual P values can be misleading (a P value can be high, even though the variable is important). The second problem is that the confidence intervals on the regression coefficients will be very wide. The confidence intervals may even include zero, which means one can't even be confident whether an increase in the X value is associated with an increase, or a decrease, in Y. Because the confidence intervals are so wide, excluding a subject (or adding a new one) can change the coefficients dramatically and may even change their signs.

3. Sources of Multicollinearity

There are four primary sources of multicollinearity:

1. The data collection method employed
2. Constraints on the model or in the population.
3. Model specification.
4. An over defined model.

It is important to understand the differences among these sources of the multicollinearity, as the recommendations for analysis of the data and interpretation of the resulting model depend to some extent on the cause of the problem.

The data collection method can lead to multicollinearity problems when the analyst samples only a subspace of the region of the regressors defined in Equation (2.2).

Constraints of the model or in the population being sampled can cause multicollinearity. For example, suppose an electric utility is investigating the effect of family income (x_1)

per month in terms of thousands rupees and house size (x_2) in terms of square meters on residential electricity consumption. The levels of the two regressors variables obtained in the sample data are shown below (Figure 2.1). Note that the data lie approximately along a straight line, indicating a potential multicollinearity problem.

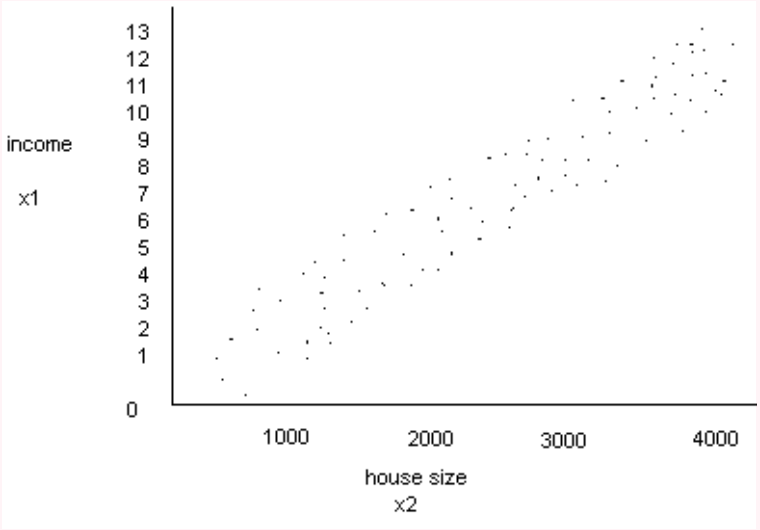

Figure 2.1

In this example a physical constraints in the population has caused this phenomenon, namely the family with the higher incomes generally have larger homes than families with lower incomes. When physical constraints such as this present, multicollinearity will exist regardless of the sampling method employed.

Multicollinearity may also be induced by the choice of model. We know that adding a polynomial term to a regression model causes ill conditioning of the $\mathbf{X}'\mathbf{X}$ matrix. Furthermore if the range of x is small, adding an x^2 term can result in significant multicollinearity.

An over defined model has more regressor variables than number of observations. These models are sometimes encountered in medical and behavioral research, where there may be only a small number of subjects (sample units) available, and information is collected for a large number of regressors on each subject. The usual approach to dealing with the multicollinearity in this context is to eliminate some of the regressor variables from consideration.

4. Effect of Multicollinearity

To assess multicollinearity, it should be noticed that how well each independent (X) variable is predicted from the other X variables. And what is the value of individual R^2 and a Variance Inflation Factor (VIF). When these R^2 and VIF values are high for any of the X variables, the fit is affected by multicollinearity.

The presence of multicollinearity has a number of potentially serious effects on the least-squares estimates of the regression coefficients. Some of these effects may be easily

demonstrated. Suppose that there are only two regressor variables, x_1 and x_2. The model, assuming that x_1, x_2 and y are scaled unit length, is

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

and the least-squares normal equations are

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix}$$

where r_{12} is the simple correlation between x_1 and x_2 and r_{jy} is the simple correlation between x_j and y, j=1,2. Now the inverse of $(\mathbf{X}'\mathbf{X})$ is

$$\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \frac{1}{(1 - r_{12}^2)} & \frac{-r_{12}}{(1 - r_{12}^2)} \\ \frac{-r_{12}}{(1 - r_{12}^2)} & \frac{1}{(1 - r_{12}^2)} \end{bmatrix} \quad \ldots(4.1)$$

And the estimates of the regression coefficients are

$$\hat{\beta}_1 = \frac{r_{1y} - r_{12} r_{2y}}{(1 - r_{12}^2)}, \qquad \hat{\beta}_2 = \frac{r_{2y} - r_{12} r_{1y}}{(1 - r_{12}^2)} \quad \ldots(4.2)$$

If there is strong multicollinearity between x_1 and x_2, then the correlation coefficient r_{12} will be large. From Equation 4.1 we see that as $|r_{12}| \to 1$, $\text{var}(\hat{\beta}_j) = C_{jj}\sigma^2 \to \alpha$ and $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = C_{12}\sigma^2 \to \pm\alpha$ depending on whether $r_{12} \to +1$ or $r_{12} \to -1$. Therefore strong multicollinearity between x_1 and x_2 result in large variances and covariances for the least-squares estimators of the regression coefficients.

When there are more than two regressor variables, multicollinearity produces similar effects. It can be shown that the diagonal elements of the $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$ matrix are

$$C_{jj} = \frac{1}{1 - R_j^2}, \quad j=1,2,\ldots,p$$

where R_j^2 is the coefficient of multiple determinations from the regression of x_j on the remaining p-1 regressor variables.

In short the consequences of the multicollinearity can be listed as follows:

(i) With exact linear relationships among the explanatory variables, the condition of exact collinearity, or exact multicollinearity, exists and the least-squares estimator is not defined. That means $\mathbf{X}'\mathbf{X}$ is singular and estimation of coefficients and standard errors is not possible.

(ii) For variables that are highly related to one another (but not perfectly related), the OLS (Ordinary Least Squares) estimators have large variances and covariances, making precise estimation difficult.

(iii) Because of the consequences of point 2, confidence intervals tend to be much wider, leading to the acceptance of the null hypothesis more readily. This is due to the relatively large standard error. The standard error is based, in part, on the correlation between the variables in the model.

(iv) Although the t ratio of one or more of the coefficients is more likely to be insignificant with multicollinearity, the R^2 value for the model can still be relatively high.

(v) The OLS estimators and their standard errors can be sensitive to small changes in the data. In other words, the results will not be robust.

5. Multicollinearity Diagnostics

Multicollinearity is a matter of degree, not a matter of presence or absence. The higher the degree of multicollinearity, the greater the likelihood of the disturbing consequences of multicollinearity. Several techniques have been proposed for detecting multicollinearity.

5.1 Examination of Correlation Matrix

A very simple measure of multicollinearity is inspection of the off-diagonal elements r_{ij} in $\mathbf{X'X}$. If regressors x_i and x_j are nearly linearly dependent, then $|r_{ij}|$ will be near unity.

Example 5.1 [Acetylene Data taken from Marquardt and Snee, 1975]: The variables are: x_1: reactor temperature in ^{0}c, x_2: ratio of H2 to n-heptane, x_3: contact time in second, y: conversion of n-heptane to acetylene.

y	x_1	x_2	x_3
49	1300	7.5	0.012
50.2	1300	9	0.012
50.5	1300	11	0.0115
48.5	1300	13.5	0.013
47.5	1300	17	0.0135
44.5	1300	23	0.012
28	1200	5.3	0.04
31.5	1200	7.5	0.038
34.5	1200	11	0.032
35	1200	13.5	0.026
38	1200	17	0.034
38.5	1200	23	0.041
15	1100	5.3	0.084
17	1100	7.5	0.098
20.5	1100	11	0.092
29.5	1100	17	0.086

The model fitted is

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_1 x_2+\beta_5 x_1 x_3+\beta_6 x_2 x_3+\beta_7 x_1^2+\beta_8 x_2^2+\beta_9 x_3^2+\varepsilon$$

After standardizing all the variables, $\mathbf{X'X}$ is calculated. The $\mathbf{X'X}$ matrix reveals the high correlation between x_1 and x_3 since r_{13}= -0.958. Furthermore, there are other large correlation between $x_1 x_2$ and $x_1 x_3$, $x_1 x_3$ and x_1^2 and x_1^2 and x_3^2. This is not surprising as these variables are generated from the linear terms and they involve the highly correlated regressors x_1 and x_3. Thus, inspection of the correlation matrix indicates that there are several near linear dependencies in the data.

$$\mathbf{X'X}=\begin{array}{c} \begin{array}{ccccccccc} x_1 & x_2 & x_3 & x_1x_2 & x_1x_3 & x_2x_3 & x_1^2 & x_2^2 & x_3^2 \end{array} \\ \begin{pmatrix} 1 & 0.224 & -0.958 & -0.132 & 0.443 & 0.205 & -0.271 & 0.031 & -0.577 \\ 0.224 & 1 & -0.24 & 0.039 & 0.192 & -0.023 & -0.148 & 0.498 & -0.224 \\ -0.958 & -0.240 & 1 & 0.194 & -0.661 & -0.274 & 0.501 & -0.018 & 0.765 \\ -0.132 & 0.039 & 0.194 & 1 & -0.265 & -0.975 & 0.246 & 0.398 & 0.274 \\ 0.443 & 0.192 & -0.661 & -0.265 & 1 & 0.323 & -0.972 & 0.126 & -0.972 \\ 0.205 & -0.023 & -0.274 & -0.975 & 0.323 & 1 & -0.279 & -0.374 & 0.358 \\ -0.271 & -0.148 & 0.501 & 0.246 & -0.972 & -0.279 & 1 & -0.124 & 0.874 \\ 0.031 & 0.498 & -0.018 & 0.398 & 0.126 & -0.374 & -0.124 & 1 & -0.158 \\ -0.577 & -0.224 & 0.765 & 0.274 & -0.972 & 0.358 & 0.874 & -0.158 & 1 \end{pmatrix} \end{array}$$

The least square estimates of the regression coefficients and corresponding standard error are given in brackets.

β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
-2.630	0.205	-4.436	-0.392	-18.710	-0.159	-11.408	0.010	-6.664
(5.796)	(0.436)	(7.803)	(2.199)	(31.179)	(2.460)	(18.293)	(0.424)	(11.579)

Clearly the variances of the estimated regression coefficients are large that means the estimated regression coefficients are not stable.

5.2 Variance Inflation Factor (VIF)

The diagonal elements of the inverse of the $\mathbf{X'X}$ matrix are very useful for detecting multicollinearity. The j[th] diagonal element of C matrix can be written as $C_{jj}=\left(1-R_j^2\right)^{-1}$, where R_j^2 is the coefficient of determination obtained when x_j is regressed on the remaining p-1 regressors. If x_j is nearly orthogonal to the remaining p-1 regressors, R_j^2 is small and C_{jj} is close to unity, while if x_j is nearly linearly dependent on some subset of the remaining regressors, R_j^2 is near unity and C_{jj} is large. Since the variance of the jth regression coefficient is $C_{jj}\sigma^2$, we can view C_{jj} as the factor by which the variance of $\hat{\beta}_j$ is increased due to near linear dependences among the regressors. We call this as variance inflation factor or VIF,

$$VIF_j = C_{jj} = \left(1-R_j^2\right)^{-1}.$$

The terminology is due to Marquardt (1970). The VIF for each term in the model measures the combined effect of the dependences among the regressors on the variance of that term. One or more large VIF indicate multicollinearity. Practical experience indicates that if any of the VIFs exceeds 5 or 10, it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity.

The VIF for the acetylene data:

var	x_1	x_2	x_3	x_1^2	x_2^2	x_3^2	$x_1 x_2$	$x_1 x_3$	$x_2 x_3$
VIF	2856749	10956.1	2017163	2501945	65.73	12667.1	9802.9	1428092	240.36

Here the maximum VIF is 2856748.97. So it is clear that multicollinearity problem exists. Furthermore, the VIFs for several of the other cross product and squared variables involving x_1 and x_3 are large. Thus, the VIF can help identify which regressors are involved in the multicollinearity.

5.3 Eigensystem Analysis of X′X

The characteristic roots or eigenvalues of $\mathbf{X'X}$, say $\lambda_1, \lambda_2, ..., \lambda_p$, can be used to measure the extent of the multicollinearity in the data. If there are one or more near-linear dependences in the data, then one or more characteristic roots will be small. One or more small eigenvalues imply that there are near-linear dependences among the columns of $\mathbf{X}$. Some analysts prefer to examine the **condition number** of $\mathbf{X'X}$, defined as $\kappa = \lambda_{max}/\lambda_{min}$. This is just a measure of the spread in the eigenvalues spectrum of $\mathbf{X'X}$. Generally if the **condition number** is less than 100, there is no serious problem with multicollinearity. **Condition number** between 100 and 1000 imply moderate to strong multicollinearity, and if it exceeds 1000, severe multicollinearity is indicated.

The eigenvalues of $\mathbf{X'X}$ for the above data are $\lambda_1 = 4.2048$, $\lambda_2 = 2.1626$, $\lambda_3 = 1.1384$, $\lambda_4 = 1.0413$, $\lambda_5 = 0.3845$, $\lambda_6 = 0.0495$, $\lambda_7 = 0.0136$, $\lambda_8 = 0.0051$, $\lambda_9 = 0.0001$. There are four very small eigenvalues, a symptom of seriously ill conditioned data. The condition number $\kappa = \lambda_{max}/\lambda_{min} = 4.2048/0.0001 = 42048$ which exceeds 1000. This indicates presence of severe multicollinearity.

5.4 Other Diagnostics

There are several other techniques that are occasionally useful in diagnosing multicollinearity. The determinant of $\mathbf{X'X}$ can be used as an index of multicollinearity. Since the $\mathbf{X'X}$ matrix is in correlation form, the possible range of values of the determinant is $0 \le |\mathbf{X'X}| \le 1$. If $|\mathbf{X'X}| = 1$, the regressors are orthogonal, while if $|\mathbf{X'X}| = 0$, there is an exact linear dependence among the regressors. The degree of the multicollinearity becomes more severe as $|\mathbf{X'X}|$ approaches zero. While this measure of multicollinearity is easy to apply, it does not provide any information on the source of the multicollinearity.

The F statistics for significance of regression and the individual t statistics can sometimes indicate the presence of multicollinearity. Specifically, if the overall F statistic is significant but the individual t statistics are all non significant, multicollinearity is present.

Unfortunately, many data sets that have significant multicollinearity will not exhibit this behavior, and so the usefulness of this measure is questionable.

The sign and magnitude of the regression coefficients will sometimes provide an indication that multicollinearity is present. In particular if adding or removing a regressor produces large changes in the estimates of the regression coefficients, multicollinearity is indicated. If the deletion of one or more data points results in large changes in the regression coefficients, there may be multicollinearity present. Finally if the signs or magnitude of the regression coefficients in the regression model are contrary to prior expectation, we should be alert to possible multicollinearity.

6. Remedies of Multicollinearity

6.1 Model Respecification

Multicollinearity is often caused by the choice of model, such as when two highly correlated regressors are used in the regression equation. In these situations some respecification of the regression equation may lessen the impact of multicollinearity. One approach to model respecification is to redefine the regressors. For example, if x_1, x_2, and x_3 are nearly linearly dependent, it may be possible to find some function such as $x = (x_1 + x_2)/x_3$ or $x = x_1 x_2 x_3$ that preserves the information content in the original regressors but reduces the ill conditioning.

Another widely used approach to model respecification is variable elimination. That is, if x_1, x_2, and x_3 are nearly linearly dependent, eliminating one regressor may be helpful in combating multicollinearity. Variable elimination is often a highly effective technique. However, it may not provide a satisfactory solution if the regressors dropped from the model have significant explanatory power relative to the response y, that is eliminating regressors to reduce multicollinearity may damage the predictive power of the model. Care must be exercised in variables selection because many of the selection procedures are seriously distorted by the multicollinearity, and there is no assurance that the final model will exhibit any lesser degree of multicollinearity than was present in the original data.

6.2 Use Additional or New Data

Since multicollinearity is a sample feature, it is possible that in another sample involving the same variables collinearity may not be so serious as in the first sample. Sometimes simply increasing the size of the sample may attenuate the collinearity problem. If one uses more data, or increase the sample size, the effects of multicollinearity on the standard errors (SE) will decrease. This is because the standard errors are based on both the correlation between variables and the sample size. The larger the sample size, the smaller is the SE.

Unfortunately, collecting additional data is not always possible because of economic constraints or because the process being studied is no longer available for sampling. Even when the additional data are available it may be inappropriate to use if the new data extend the range of the regressor variables far beyond the analyst's region of interest. Of course collecting additional data is not a viable solution to the multicollinearity problem when the multicollinearity is due to constraints on the model or in the population.

6.3 Ridge Regression

When the method of least-squares is applied to nonorthogonal data, very poor estimates of the regression coefficients can be obtained. The problem with the method of least squares is the requirement that $\hat{\boldsymbol{\beta}}$ be an unbiased estimator of $\boldsymbol{\beta}$. The Gauss-Markov property assures that the least-squares estimator has minimum variance in the class of unbiased linear estimators.

One way to alleviate this problem is to drop the requirement that the estimator of $\boldsymbol{\beta}$ be unbiased. Suppose that a biased estimator of $\boldsymbol{\beta}$ is found say $\hat{\boldsymbol{\beta}}^*$ that has smaller variance than the unbiased estimator $\hat{\boldsymbol{\beta}}$. The mean square error of $\hat{\boldsymbol{\beta}}^*$ is defined as

$$\text{MSE}\left(\hat{\boldsymbol{\beta}}^*\right) = \text{E}\left(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}\right)^2 = \text{V}\left(\hat{\boldsymbol{\beta}}^*\right) + \left[\text{E}\left(\hat{\boldsymbol{\beta}}^*\right) - \boldsymbol{\beta}\right]^2$$

$$\text{or} \quad \text{MSE}\left(\hat{\boldsymbol{\beta}}^*\right) = \text{V}\left(\hat{\boldsymbol{\beta}}^*\right) + \left(\text{bias in } \hat{\boldsymbol{\beta}}^*\right)^2$$

by allowing a small amount of bias in $\hat{\boldsymbol{\beta}}^*$, the variance of $\hat{\boldsymbol{\beta}}^*$ can be made small such that the MSE of $\hat{\boldsymbol{\beta}}^*$ is less than the variance of the unbiased estimator $\boldsymbol{\beta}$.

A number of procedures have been developed for obtaining biased estimators of regression coefficients. One of these procedures is ridge regression, originally proposed by Hoerl and Kennard (1970). Specifically the ridge estimator is defined as the solution to $\left(\mathbf{X}'\mathbf{X} + \text{k}\mathbf{I}\right)\hat{\boldsymbol{\beta}}_\text{R} = \mathbf{X}'\mathbf{y}$ or $\hat{\boldsymbol{\beta}}_\text{R} = \left(\mathbf{X}'\mathbf{X} + \text{k}\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{y}$ where $\text{k} \geq 0$ is a constant selected by the analyst. The procedure is called ridge regression. It is to be noted that when k=0 then the ridge estimator is the least-square estimator. The ridge estimator is a linear transformation of the least-squares estimator since

$$\begin{aligned}\hat{\boldsymbol{\beta}}_\text{R} &= \left(\mathbf{X}'\mathbf{X} + \text{k}\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{y} \\ &= \left(\mathbf{X}'\mathbf{X} + \text{k}\mathbf{I}\right)^{-1}\left(\mathbf{X}'\mathbf{X}\right)\hat{\boldsymbol{\beta}} \\ &= \mathbf{Z}_\text{k}\,\hat{\boldsymbol{\beta}}\end{aligned}$$

Therefore, since $\text{E}\left(\hat{\boldsymbol{\beta}}_\text{R}\right) = \text{E}\left(\mathbf{Z}_\text{k}\,\hat{\boldsymbol{\beta}}\right) = \mathbf{Z}_\text{k}\,\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_\text{R}$ is a biased estimator of $\boldsymbol{\beta}$. The constant k is usually referred to the biasing parameter. The covariance matrix of $\hat{\boldsymbol{\beta}}_\text{R}$ is

$$\text{V}\left(\hat{\boldsymbol{\beta}}_\text{R}\right) = \sigma^2\left(\mathbf{X}'\mathbf{X} + \text{k}\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X} + \text{k}\mathbf{I}\right)^{-1}$$

The total mean square error of the ridge estimator is

$$\begin{aligned}\text{MSE}\left(\hat{\boldsymbol{\beta}}_\text{R}\right) &= \text{V}\left(\hat{\boldsymbol{\beta}}_\text{R}\right) + \left(\text{bias in } \hat{\boldsymbol{\beta}}_\text{R}\right)^2 \\ &= \sigma^2\,\text{Tr}\left[\left(\mathbf{X}'\mathbf{X} + \text{k}\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X} + \text{k}\mathbf{I}\right)^{-1}\right] + \text{k}^2\,\boldsymbol{\beta}'\left(\mathbf{X}'\mathbf{X} + \text{k}\mathbf{I}\right)^{-2}\boldsymbol{\beta} \\ &= \sigma^2\sum_{\text{j}=1}^{\text{p}}\frac{\lambda_\text{j}}{\left(\lambda_\text{j} + \text{k}\right)^2} + \text{k}^2\,\boldsymbol{\beta}'\left(\mathbf{X}'\mathbf{X} + \text{k}\mathbf{I}\right)^{-2}\boldsymbol{\beta}\end{aligned}$$

where $\lambda_1, \lambda_2, ..., \lambda_\text{p}$ are the eigenvalues of $\mathbf{X}'\mathbf{X}$. If k > 0, note that the bias in $\hat{\boldsymbol{\beta}}_\text{R}$ increases with k. However, the variance decreases as *k* increases.

In using ridge regression we would like to choose a value of k such that the reduction in the variance term is greater than the increase in the squared bias. If this can be done, the

mean square error of the ridge estimator $\hat{\boldsymbol{\beta}}_R$ will be less than the variance of the least-square estimator $\hat{\boldsymbol{\beta}}$.

Hoerl and Kennard (1976) proved that there exists a non zero value of k for which the MSE of $\hat{\boldsymbol{\beta}}_R$ is less than the vatiance of the least-squares estimatror $\hat{\boldsymbol{\beta}}$. The residual sum of squares is

$$SS_{Res} = \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_R\right)' \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_R\right)$$

$$= \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)' \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) + \left(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}\right)' \mathbf{X}'\mathbf{X}\left(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}\right)$$

Since the first term in the right hand side of the above equation is the residual sum of squares for the least-squares estimates $\hat{\boldsymbol{\beta}}$, it is clear that as k increases, the residual sum of squares increases. Consequently, because the total sum of squares is fixed, R^2 decreases as *k* increases. Therefore, the ridge estimate will not necessarily provide the best fit to the data, but this should not be more concerned since the interest is in obtaining a stable set of parameter estimates.

Hoerl and Kennard (1976) have suggested that an appropriate value of k may be determined by inspection of the ridge trace. The ridge trace is a plot of the elements of $\hat{\boldsymbol{\beta}}_R$ versus k for values of k usually in the interval 0 – 1. If the multicollinearity is severe, the instability in the regression coefficients will be obvious from the ridge trace. As k is increased, some of the ridge estimates will vary dramatically. At some value of k, the ridge estimates $\hat{\boldsymbol{\beta}}_R$ will stabilize. The objective is to select a reasonably small value of k at which the ridge estimates $\hat{\boldsymbol{\beta}}_R$ are stable. Generally this will produce a set of estimates with smaller MSE than the least-squares estimates.

Several author have proposed several procedures for choosing the value of k. Hoerl, Kennard, and Baldwin (1975) have suggested that an appropriate choice of k is $\kappa = \dfrac{p\,\hat{\sigma}^2}{\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}}$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are found by least squares solution.

6.4 Principal Component Regression

Biased estimators of regression coefficients can also be obtained by using a procedure known as principal components regression. Consider the following model,

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

Let $\mathbf{X}'\mathbf{X} = \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}'$, where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_p)$ is a $p \times p$ diagonal matrix of the eigenvalues of $\mathbf{X}'\mathbf{X}$ and $\mathbf{T}$ is a $p \times p$ orthogonal matrix whose columns are the eigenvectors associated with $\lambda_1, \lambda_2, ..., \lambda_p$. Then the above model can be written as

$$\mathbf{y} = \mathbf{X}\mathbf{T}\mathbf{T}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \mathbf{T}\mathbf{T}' = \mathbf{I}$$

$$= \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

where $\mathbf{Z} = \mathbf{X}\mathbf{T}$, $\alpha = \mathbf{T}'\boldsymbol{\beta}$,

$$\mathbf{Z}'\mathbf{Z} = \mathbf{T}'\mathbf{X}'\mathbf{X}\mathbf{T} = \mathbf{T}'\mathbf{T}\boldsymbol{\Lambda}\mathbf{T}'\mathbf{T} = \boldsymbol{\Lambda}.$$

The columns of **Z**, which define a new set of orthogonal regressors, such as $\mathbf{Z}=\left[\mathbf{Z}_1,\mathbf{Z}_2,...,\mathbf{Z}_p\right]$ are referred to as principle components.

The least square estimator of $\boldsymbol{\alpha}$ is

$$\hat{\boldsymbol{\alpha}}=\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{y}=\boldsymbol{\Lambda}^{-1}\mathbf{Z}'\mathbf{y}$$

and the covariance matrix of $\hat{\boldsymbol{\alpha}}$ is

$$\mathrm{V}\left(\hat{\boldsymbol{\alpha}}\right)=\sigma^2\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}=\sigma^2\boldsymbol{\Lambda}^{-1}\ .$$

Thus a small eigenvalues of $\mathbf{X}'\mathbf{X}$ means that the variance of the corresponding regression coefficient will be large. Since $\mathbf{Z}'\mathbf{Z}=\sum_{i=1}^{p}\sum_{j=1}^{p}\mathbf{Z}_i\mathbf{Z}_j'=\boldsymbol{\Lambda}$. We often refer to the eigenvalue λ_j as the variance of the jth principle component. If all λ_j equal to unity, the original regressors are orthogonal, while if a λ_j is exactly equal to zero, this implies a perfect linear relationship between the original regressors. One or more λ_j near to zero implies that multicollinearity is present.

The principle components regression approach combats multicollinearity by using less than the full set of principle components in the model. To obtain the principle components estimator, assume that the regressors are arranged in order of decreasing eigenvalues, $\lambda_1\geq\lambda_1\geq...\geq\lambda_p>0$ suppose that the last s of these eigenvalues are approximately equal to zero. In principal components regression the principal components corresponding to near zero eigenvalues are removed from the analysis and least squares applied to the remaining components. That is, $\hat{\boldsymbol{\alpha}}_{pc}=\mathbf{B}\hat{\boldsymbol{\alpha}}$ where $b_1=b_2=...=b_{p-s}=1$ and $b_{p-s+1}=b_{p-s+2}=\cdots=b_p=0$ thus the principle components estimator is

$$\hat{\boldsymbol{\alpha}}_{pc}=\left[\hat{\alpha}_1\ \ \hat{\alpha}_2\ .\ .\ .\ \hat{\alpha}_{p-s}\ \ 0\ \ 0\ .\ .\ .\ 0\right]'$$

Principal components regression for the acetylene data:

The linear transformation **Z=XT** that transforms the original regressors into an orthogonal set of variables (the principal components).

Eigenvalue								
4.0628	2.2865	1.1653	1.1118	0.4528	0.3249	0.0154	0.0025	-0.422
Eigenvectors or T matrix								
-0.355	0.032	0.374	-0.528	0.140	-0.070	-0.467	0.453	-0.079
-0.146	0.265	0.602	0.387	-0.526	0.344	-0.017	0.017	-0.014
0.432	-0.020	-0.233	0.398	-0.111	-0.056	-0.404	0.632	0.149
0.198	0.546	-0.095	-0.234	0.310	0.584	0.047	0.036	0.398
-0.467	0.025	-0.247	0.201	0.149	0.204	0.510	0.535	-0.266
-0.153	-0.596	0.295	0.236	0.350	0.173	0.046	0.018	0.569
0.419	-0.023	0.347	-0.318	-0.169	-0.290	0.595	0.324	0.176
-0.045	0.477	0.274	0.398	0.521	-0.513	0.022	-0.047	0.016
0.455	-0.215	0.303	0.054	0.388	0.337	-0.015	-0.005	-0.622

The **T** matrix indicates that the relationship between z_1 and standardized regressors is

$$z_1 = -.35549\,x_1 - .14831\,x_2 + .432164\,x_3 + .198493\,x_1\,x_2 - .46653\,x_1\,x_3$$

$$-.15317\,x_2\,x_3 + .418952\,{x_1}^2 - .04508\,{x_2}^2 + .455406\,{x_3}^2$$

the relationship between the remaining principal components $z_1, z_2, \ldots, z_9$ and the standardized regressors are determined similarly.

The principal components estimator reduces the effect of multicollinearity by using a subset of the principal components in the model. Since there are four small eigenvalues for the acetylene data, this implies that there are four principal components that should be deleted. We will exclude z_6 , z_7 , z_8 , z_9 and consider regressions involving only the first five principal components. The estimate of

$$\hat{\boldsymbol{\alpha}}_{\text{pc}} = \begin{bmatrix} -0.3598 & -0.12087 & 0.29767 & -0.00948 & 0.130699 \\ (0.14274) & (0.24649) & (0.23690) & (0.29272) & (0.49574) \end{bmatrix},$$

The value in the bracket indicates the corresponding standard error.

The original $\hat{\boldsymbol{\beta}}$ can be obtained by the reverse transformation $\hat{\boldsymbol{\beta}} = \mathbf{T}\,\hat{\boldsymbol{\alpha}}_{pc}$ and the variance covariance matrix of $\hat{\boldsymbol{\beta}}$ will be $\mathrm{V}(\hat{\boldsymbol{\beta}}) = \mathbf{T}\,\mathrm{V}(\boldsymbol{\alpha}_{\text{pc}})\mathbf{T}'$. Following table shows the $\hat{\boldsymbol{\beta}}$ and their corresponding standard error in bracket.

$\hat{\boldsymbol{\beta}}$	0.259	0.127	-0.241	-0.123	0.109	0.258	-0.064	0.105	0.002
S. E.	0.198	0.326	0.153	0.218	0.130	0.249	0.162	0.313	0.222

7. Conclusions

Multicollinearity does not affect the properties of the OLS estimators. The estimators remain unbiased and efficient. But the fact is that when multicollinearity is present in the data then the OLS estimators are imprecisely estimated. If the goal is simply to predict Y from a set of X variables, then multicollinearity is not a problem. The predictions will still be accurate and the overall R^2 (or adjusted R^2) quantifies how well the model predicts the Y values. If the goal is to understand how the various X variables impact Y, then multicollinearity is a big problem.

There are several methods available in literature for detection of multicollinearity like examination of correlation matrix, calculating the variance inflation factor (VIF), by the eigensystem analysis *etc*. Complete elimination of multicollinearity is not possible but we can reduce the degree of multicollinearity present in the data. Several remedial measures are employed to tackle the problem of multicollinearity such as collecting the additional data or new data, respecification of the model, ridge regression, by using data reduction technique like principal component analysis.

References

Hawking, R. R. and Pendleton, O. J. (1983). the regression dilemma, *Commun*. *Stat*.- *Theo. Meth*, **12**, 497-527.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55-67.

Hoerl, A. E. and Kennard, R. W. and Baldwin, K. F. (1975). Ridge regression: some simulations, *Commun*. *Stat*.-*Theo. Meth*., **4**, 105-123.

Hoerl, A. E. and Kennard, R. W. (1976). Ridge regression: Iterative estimation of biasing parameter, *Commun*. *Stat*.- *Theo. Meth*, **5**, 77-88.

Marquardt, D. W. (1970) Generalized inverse, ridge regression, biased linear estimation, and non linear estimation, *Technometrics*, **12**, 591-612.

Marquardt, D. W. and Snee, R. D. (1975) Ridge regression in practice. *Amer*. *Statis*t., **29**, 3-19.

Some Additional References

Draper, N. R., Smith, H. (2003). *Applied regression analysis*, 3rd edition, Wiley, New York.

Gujrati, D. N. (2004). *Basic econometrics* 4th edition, Tata McGraw-Hill, New Delhi.

Johnston, J. and Dinardo, J. (1997) *Econometric methods*, 4th edition, McGraw-Hill, Singapore.

Montgomery, D. C., Peck, E. A., Vining, G. G. (2001). *Introduction to linear regression analysis*, 3rd edition, Wiley, New York.

Wetherill, G. B., Duncombe, P., Kenward, M., Kollerstrom, J. (1986). *Regression analysis with application*, 1st edition, Chapman and Hall, New York.