12

Multicollinearity

So far we have considered the classical normal linear regression model and showed how it can be used to handle the twin problems of statistical inference, namely, estimation and hypothesis testing, as well as the problem of prediction. However, this model is based on several simplifying assumptions and hereafter we consider the theoretical and practical consequences of the violation of the classical assumptions of ordinary least squares.

Multicollinearity

The only additional assumption of the multiple regression model is that the independent variables are not perfectly correlated with each other . If this assumption is violated, a problem that potentially occurs which is called multicollinearity. It is a condition where the independent variables are not independent of one another. The term multicollinearity originally meant the existence of a "perfect" or exact linear relationship among some or all-explanatory variables of a regression model. One explanatory variables can be completely explained by a linear combination of other explanatory variables. In practice, perfect multicollinearity occurs only from an error in model specification. Perfect multicollinearity is an extreme situation. While perfect multicollinearity is often the result of model misspecification, near-perfect multicollinearity is a more common phenomenon. Near or imperfect multicollinearity refers to situations in which two or more of the the explanatory variables are "almost" linearly related. While this does not constitute a violation of the classical linear regression assumptions (and therefore the BLUE / Minimum Variance Unbiased Estimates properties), in this situation the separate effects of the explanatory variables cannot be estimated "precisely". This is a problem because our linear model, by including a separate term for each explanatory variable with its own parameter, requires that the individual effect of each explanatory variable on the response variable be estimated. Today multicollinearity is used in a broader sense. Broadly interpreted multicollinearity refers to the situation where there is either an exact or approximately exact relationship between the X variables. Another term used in econometric analysis is Orthogonality , which refers to no relationship between explanatory variables.

As an example of multicollinearity, suppose that we are investigating the relationship between wealth ,income and liquid assets and their effect on consumption levels. But these variables all share some information (that is they are not independent) they provide redundant information and may have serious negative consequences in a regression model. An analyst can perform a correlation analysis between each independent variable to determine to what degree the variables are related. The correlation coefficient measures the strength and direction of a relationship between two variables with ranges from -1 to 1. The closer the correlation coefficient is to -1 or 1 the stronger the relationship is between the variables. As noted in the correlation matrix below, wealth and income have a correlation of 1, so these variables contain identical information. Wealth and liquid assets also have a high correlation. The analyst should omit redundant variables from the regression model since their inclusion may have detrimental effects. The presence of multicollinearity can seriously damage the efforts to determine which explanatory variables are important and to measure the effect each has on the response variable .

	Wealth	Income	Liquid Assets
Wealth	1.00000		
Income	1.00000	1.00000	
Liquid Assets	0.86002	0.86002	1.00000

One way to indicate this idea visually is using a Ballantine diagram or Venn diagram for the two-independent variable model as illustrated in Figure 12.1. In a Venn Diagram, multicollinearity is shown by overlapping circles. In the following figure, the circle Y at the center represents the outcome variable and all surrounding ones represent the independent variables. The overlapping area denotes the variation explained. When there are too many variables, it is likely that Y is almost entirely covered by many inter-related Xs. The variance explained is very high but this model is over-specified and thus useless.

$$b_1 = \frac{(\Sigma x_1 y)(\Sigma x_2^2)-(\Sigma x_2 y)(\Sigma x_1 x_2)}{(\Sigma x_1^2)(\Sigma x_2^2)-(\Sigma x_1 x_2)^2}$$

Substituting kx_1 for x_2.

$$b_2 = \frac{\Sigma x_1 y(k^2 \Sigma x_2^2)-(k\,\Sigma x_1 y)(k\Sigma x_1^2)}{\Sigma x_1^2\,(k^2\Sigma x_1^2)-k^2(\Sigma x_1^2)^2}$$

$$= \frac{k^2\Sigma x_1 y\Sigma x_1^2 - k^2\Sigma x_1 y\Sigma x_1^2}{k^2\,(\Sigma x_1^2)^2 - k^2(\Sigma x_1^2)^2}$$

$$= \frac{0}{0}$$

Similarly b_2 can also be proved to be indeterminate.

$$b_2 = \frac{(\Sigma x_1 y)(\Sigma x_1^2)-(\Sigma x_1 y)(\Sigma x_1 x_2^2)}{(\Sigma x_1^2)(\Sigma x_2^2)-(\Sigma x_1 x_2)^2}$$

Substituting kx_1 for x_2

$$b_2 = \frac{(k\Sigma x_1 y)(k^2\Sigma x_1^2)-(k\,\Sigma x_1 y)(\Sigma x_1^2)}{\Sigma x_1^2\,(k^2\Sigma x_1^2)^2 - k^2(\Sigma x_1^2)^2}$$

$$= \frac{k(\Sigma x_1 y)(\Sigma x_1^2)-k(\Sigma x_1 y)(\Sigma x_1^2)}{k^2\,(\Sigma x_1^2)^2 - k^2(\Sigma x_1^2)^2}$$

$$= \frac{0}{0}$$

Therefore the parameters are indeterminate. There is no way of finding the values of b_1 and b_2 separately.

Case - 2 : The variances of the estimates become infinitely large

$$Var\,(b_1) = \frac{\hat{\sigma}_u^2\Sigma x_2^2}{(\Sigma x_1^2)(\Sigma x_2^2)-(\Sigma x_1 x_2)^2}$$

$$\text{Var}(b_1) = \frac{k\hat{\sigma}_u^2 \Sigma x_1^2}{\left(\Sigma x_1^2\right)\left(k^2 \Sigma x_2^2\right) - k^2\left(\Sigma x_1^2\right)^2} = \alpha$$

Similarly,

$$\text{Var}(b_2) = \frac{\hat{\sigma}_u^2 \Sigma x_1^2}{\left(\Sigma x_1^2\right)\left(k^2 \Sigma x_1^2\right) - k^2\left(\Sigma x_1^2\right)^2} = \alpha$$

$$\text{Var}(b_1) = \frac{\hat{\sigma}_u^2 \Sigma x_1^2}{\left(\Sigma x_1^2\right)\left(k^2 \Sigma x_1^2\right) - k^2\left(\Sigma x_1^2\right)^2} = \alpha$$

Note : $kx_1 = x_2$ and $\Sigma x_2^2 = (k\Sigma x_1)^2$

Thus, when two variables are perfectly correlated, OLS estimates (b_i's) can not be obtained and the variances of these estimates are infinitely large.

As an example, suppose we consider the following data on Y , X_1 and X_2. Obviously X_2 equals five times X_1 (Table 12.1).

TABLE 12.1 DATA ON Y , X1 AND X2

Consumption (Y)	Income(X_1)	Wealth (X_2)
20	1	5
25	2	10
30	2	10
33	3	15
36	4	20
38	5	25

The variables X_1 and X_2 are perfectly correlated . The variable X_2 is a linear function of X_1, ($X_2 = 5X_1$). When two variables are perfectly correlated, OLS estimates (b_i's) can not be obtained and the variances of these estimates are infinitely large. All the output is meaningless except for the degrees of freedom and the correlation matrix, which contains a one for the correlation of X_1 with X_2. Thus, when two or more independent variables are perfectly correlated or collinear, the method of Ordinary Least Square is not applicable.

Worksheet for the Multiple Linear Regression Model

Y	X_1	X_2	X_1Y	X_2Y	X_1X_2	X_{12}	X_{22}	Y_2
20	1	5	20	100	5	1	25	400
25	2	10	50	250	20	4	100	625
30	2	10	60	300	20	4	100	900

33	3	15	99	495	45	9	225	1089
36	4	20	144	720	80	16	400	1296
38	5	25	190	950	125	25	625	1444
$\Sigma Y =$ **182**	$\Sigma X_1 =$ **17**	$\Sigma X_2 =$ **85**	$\Sigma X_1Y =$ **563**	$\Sigma X_2Y =$ **2815**	$\Sigma X_1X_2 =$ **295**	$\Sigma X_1^2 =$ **59**	$\Sigma X_2^2 =$ **1475**	$\Sigma Y^2 =$ **5754**

Intermediate Results

N(or) n =	6	$\Sigma X_1Y =$	563			$\Sigma X_1^2 =$	59
$\Sigma Y =$	182	$\Sigma X_2Y =$	2815			$\Sigma X_2^2 =$	1475
$\Sigma X_1 =$	17	$\Sigma X_1X_2 =$	295			$\Sigma Y_2 =$	5754
$\Sigma X_2 =$	85	Mean of Y =	30.3333333	Mean of X_1 =	2.83333333	Mean of X_2 =	14.166667

(a) The OLS estimates (b_i's) can not be obtained. Recall

$$b_1 = \frac{(x_1y)(x_2^2)-(x_2y)(x_1x_2)}{(x_1^2)(x_2^2)-(x_1x_2)^2}$$

$$= \frac{(47.33333)(270.8333333)-(236.6666667)(54.16666667)}{(10.833333)(270.8333333)-(54.16666667)(54.166667)}$$

$$= \frac{12819.444-12819.444}{2934.0278-2934.0278}$$

$$= \frac{0}{0}$$

$$b_2 = \frac{(x_2y)(x_1^2)-(x_1y)(x_1x_2)}{(x_1^2)(x_2^2)-(x_1x_2)^2}$$

$$= \frac{(236.66667)(10.8333333)-(47.33333333)(54.16666667)}{(10.833333)(270.8333333)-(54.16666667)(54.166667)}$$

$$= \frac{2563.8889-2563.8889}{2934.0278-2934.0278}$$

$$= \frac{0}{0}$$

(b) The variances of the estimates become infinitely large. Recall

$$\text{Var } b_1 = \frac{\sigma_u^2 \Sigma x_2^2}{\Sigma x_1^2 . \Sigma x_2^2 - (\Sigma x_1 x_2)^2}$$

$$\text{Var } b_2 = \frac{\sigma u^2 \Sigma x_1^2}{\Sigma x_1^2 \Sigma x_2^2 - (\Sigma x_1 x_2)^2}$$

$$\text{Var } (b_1) = \frac{\hat{\sigma}_u^2 \Sigma x_2^2}{(2934.0278) - (2934.0278)}$$

$$\text{Var } (b_1) = \frac{k\hat{\sigma}_u^2 \Sigma x_1^2}{0} = \alpha$$

Similarly,

$$\text{Var } (b_2) = \frac{\hat{\sigma}_u^2 \Sigma x_1^2}{(2934.0278) - (2934.0278)}$$

$$\text{Var } (b_1) = \frac{\hat{\sigma}_u^2 \Sigma x_1^2}{0} = \alpha$$

CONSEQUENCES OF IMPERFECT MULTICOLLINEARITY

1. Imperfect or extreme or near multicollinearity is the more common problem and it arises when two or more of the explanatory variables are approximately linearly related. If collinearity is high but not perfect the estimation of the coefficients is possible but the following are some of the consequences .
2. Even extreme multicollinearity (so long as it is not perfect) does not violate OLS assumptions. OLS estimates are unbiased, consistent, and efficient and are BLUE (Best Linear Unbiased Estimators) in the presence of multicollinearity. However, they may be 'unstable'. By unstable we mean that they may be particularly sensitive to model specification, or to outliers in the data.
3. Standard errors of the regression coefficients will be high. Though in fact this is not necessarily the result of multicollinearity alone. The greater the multicollinearity, the greater the standard errors .The main consequence is that the variances (and standard errors) of some coefficient estimates will be higher than they would be in the absence of multicollinearity. When high multicollinearity is present, confidence intervals for coefficients tend to be very wide and t statistics tend to be very small. Coefficients will have to be larger in order to be statistically significant, i.e. it will be harder to reject the null when multicollinearity is present. In some cases, high R^2 and F test statistics, but low individual significance of the individual

coefficients. Thus the presence of a *high degree* of multicollinearity will result in the following combination:

High R^2 model will appear to *fit the data* well.

High calculated F value indicates the model explains a statistically significant portion of the variation in the dependent variable (The variables are jointly significant).

Low t values indicate the variables are not statistically significant .Coefficients may have very high standard errors and low levels of significance in spite of the fact that they are jointly highly significant.

This combination of result gives an indication that multicollinearity may be a problem.

4. Addition/deletion of an independent variable results in large changes of regression coefficients or signs. Signs and magnitudes of regression coefficients may be different from what are expected.
5. The overall fit of the equation (R^2 and adjusted R^2) will be largely unaffected.
6. The estimated coefficients of non-multicollinear variables will be largely unaffected.

DETECTION OF MULTICOLLINEARITY:

The easiest way to do this is to examine the correlations between each pair of explanatory variables. If two of the variables are highly correlated (e.g., they have a correlation less than -0.80 or greater than 0.80), then multicollinearity may be a problem. The correlation approach can only detect when pairs of variables are highly (linearly) related. High values of simple correlation coefficients may be considered to be sufficient, but not necessary for multicollinearity. It is possible for a group of independent variables, acting together, to cause multicollinearity. The form of multicollinearity can be much more complicated, involving a relationship between three or more variables, and, thus, will not necessarily be detected by the simple correlation approach. Hence, these relationships are measured by the partial correlation coefficients, which measure the correlation between two variables after holding the others constant. This will provide information on the existence of more complex relationships between or among the independent variables.

1. Multicollinearity can also be detected after the model has been fitted to the data by looking at the output for the linear regression. Very unreasonable estimates or extremely large estimated standard errors for some slope parameters can be an indication that multicollinearity is present. Additionally, if the linear model seems to fit the data well overall (e.g., the null hypothesis of no effect is rejected in the F-test for overall significance or, identically, R^2 is high) , but most of the coefficients are not significant according to their p-values, then multicollinearity might be the cause. In a regression model, the high R^2 is not a result of good independent predictors,

but a mis-specified model that carries mutually dependent and thus redundant predictors! Variance inflation factor (VIF) is common way for detecting multicollinearity.

2. Many econometrics texts outline methods designed to detect the presence, severity and form of multicollinearity and top econometricians have suggested variance inflation factors, auxiliary regressions (i.e., regressing one explanatory variable on another), computing the determinants of the (X'X) and its characteristic roots.

3. **Tolerance:** Multicollinearity is measured by the tolerance statistic, defined as $1 - R^2$ predicting each predictor using all other predictors (values close to 1 are better, values close to 0 are bad) The *tolerance* of X_h is the proportion that is *not* explained by the other variables, i.e. tolerance of X_h $= 1 - R_h^2$ It is the proportion of the variance of X_h that is not shared with the other variables in our analysis. If the tolerances are low (say .1 or .2) there are multicollinearity problems.. A tolerance close to 1 means there is little multicollinearity, whereas a value close to 0 suggests that multicollinearity may be a threat. The reciprocal of the tolerance is known as the *Variance Inflation Factor (VIF).*

$$VIF_h = 1/(1-R^2)$$

Auxiliary Regressions: The problem of multicollinearity may arise due to relationship between more than one independent variable. For example: $X_{1i} = b_0 + b_1 X_{2i} + b_2 X_{3i} + v_i$ To find these types of relationships, you can you can proceed by estimating separate regressions of each of your independent variables against all of the remaining independent variables. These regressions are called *auxiliary regressions*.If there are *k* independent variables $(X_1, X_2, \ldots, X_k$ run an OLS regression for each regressor as a function of all the other explanatory variables. For example, estimate auxiliary regression equations and calculate the VIF as discussed below

$$X_{1i} = \alpha_0 + \alpha_1 X_{2i} + \alpha_2 X_{3i} + \ldots + \alpha_{k-1} X_{ki} + v_i$$

$$X_{2i} = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{3i} + \ldots + \gamma_{k-1} X_{ki} + \varepsilon_i$$

$$\vdots$$

$$X_{ki} = \delta_0 + \delta_1 X_{1i} + \delta_2 X_{2i} + \ldots + \delta_{k-1} X_{k-1i} + \omega_i$$

Variance Inflation Factor

The variance inflation factor associated with X_h :

$$VIF(X_h) = \frac{1}{1 - R_h^2}$$

where R_h^2 is the R^2 value obtained for the regression of X on the other independent variables.

Relationship between R^2_h and VIF. Numerical example:

1. If $R^2_h = 0.1$, then the VIF $= 1.11$.
2. If $R^2_h = 0.25$, then the VIF $= 1.33$.
3. If $R^2_h = 0.5$, then the VIF $= 2$
4. If $R^2_h = 0.75$, then the VIF $= 4$
5. If $R^2_h = 0.9$, then the VIF $= 10$
6. If $R^2_h = 0.99$, then the VIF $= 100$.

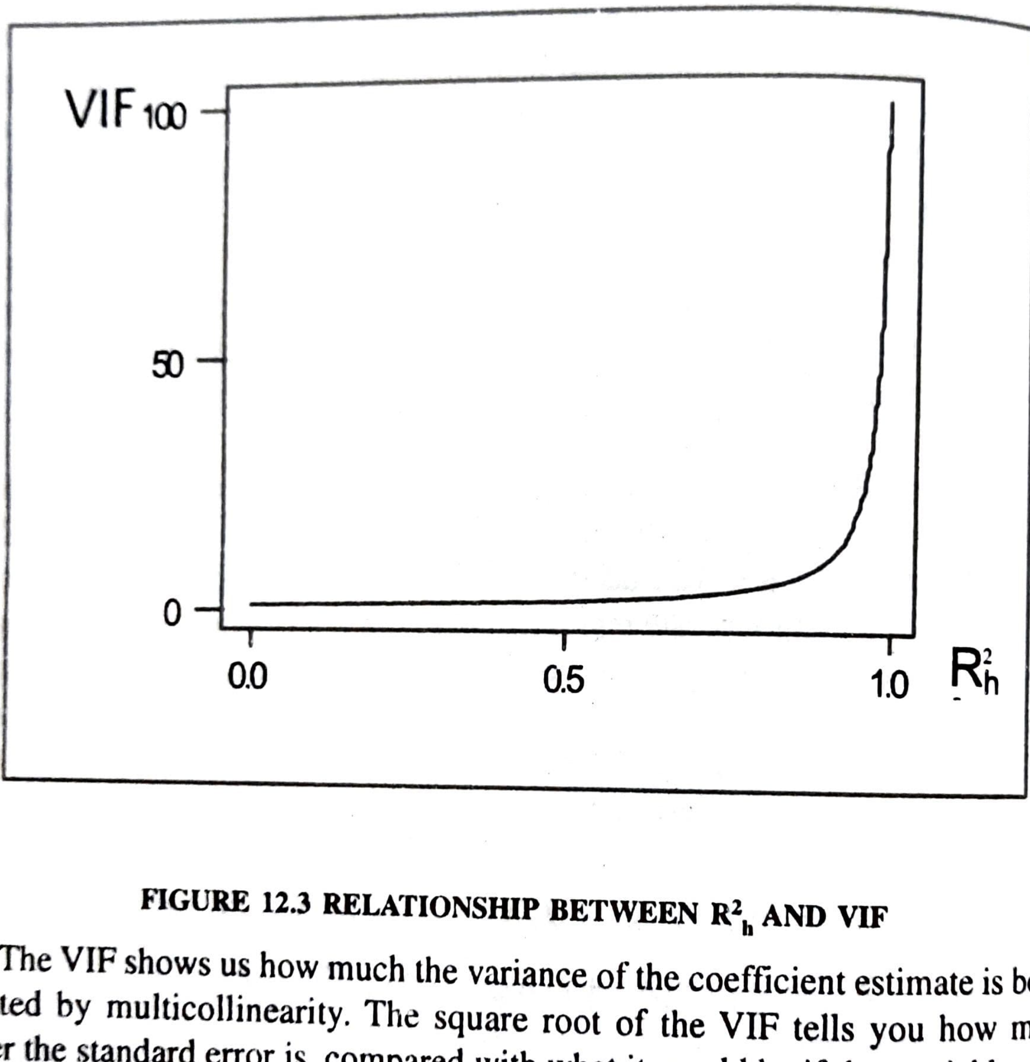

FIGURE 12.3 RELATIONSHIP BETWEEN R^2_h AND VIF

The VIF shows us how much the variance of the coefficient estimate is being inflated by multicollinearity. The square root of the VIF tells you how much larger the standard error is ,compared with what it would be if that variable were uncorrelated with the other X variables in the equation. For example, if the VIF for a variable were 9, its standard error would be three times as large as it would be if its VIF was 1. In such a case, the coefficient would have to be 3times as large to be statistically significant. High VIFs suggest the presence of a multicollinearity problem . The higher a VIF, the higher the variance of the estimated coefficient of that explanatory variable. When $R_{2i} = 1$, there is a perfect multicollinearity and the VIF is infinity; when $R_{2i} = 0$, there is no multicollinearity and the VIF is one. VIF > 30 usually indicates a severe multicollinearity problem.

A common rule of thumb is that a problem exists when any *VIF* exceeds 10 ome suggested that 5) or when the sum of all *VIFs* exceeds 10. But VIFs above) are not unusual. The Longley data (Barone *et al*, 1976) has several VIFs ceeding 100 (and one above 1,000)). Such VIFs almost certainly indicate ill-nditioned data and/or inappropriate model specification.

5. Step by step regression method may be used to test for the presence of multicollinearity . If introduction of a variable in a model increases the standard error sharply, so as to make the coefficient insignificant while R^2 remains nearly same, multicollinearity is present The first method is the computation of a correlation matrix of the independent regression variables. The correlation matrix allows us to identify those explanatory variables that are highly correlated with one another and this causes the problem of multicollinerity. Collinearity is often suspected when R^2 is high (between 0.8 and 1.0) and none or very few estimated coefficients are individually significant on the basis of the students t-Test. To measure the ill-effect of multicollinearity we use the variance inflation factor (VIF).
6. Klein Suggests that multicollinearity should not be considered serious unless the simple correlation coefficient between any two explanatory variables is greater then or equal to the multiple correlation coefficient (R2). *Klein's Rule* says that multicollinearity becomes a problem only if

$$r^2_{x_i x_j} \geq R^2_{y.x_1,x_2......x_k}$$

where r^2 is the square of the simple correlation coefficient between any two explanatory variables ($X_i X_j$) and R^2 is the multiple correlation coefficient [L.R. Klein, Introduction to Economics, Prentice Hall International, Landon, pp.64 & 101].
7. Check to see how stable coefficients are when different samples are used. For example, you might randomly divide your sample in two. If coefficients differ dramatically, multicollinearity may be a problem.
8. The presence and degree of multicollinearity are more precisely determined by an examination of the characteristic roots and vectors of the X'X matrix (Judge, et al., 1985). Collinearity is present when one or more characteristics roots are "small." This measure was developed in detail by Belsley, Kuh, and Welsch (1980), who suggested that a more precise method of defining "small" involves the formation of condition indices" and a corresponding matrix of cross variances between variables and eigenvalues. The condition index refers to a vector consisting of the square root of the ratio of the largest eigen value to each individual eigenvalue. Elements in the cross variance matrix are calculated as the proportion of the variance of each variable associated with each single characteristic root. In case of linear dependence between the variables the eigenvalues of all different eigenvectors will differ much from each other, such that the ratio

$$\text{Condition number} = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} = \sqrt{k}$$

becomes quite large. Collinearity exists when the condition index is large — around 5-10 for weak dependencies and 30-100 for moderate to strong relationships — and when the associated row vector in the cross variance matrix contains two or more large values — usually values greater than 0.50 (Judge, et al., 1985). If the square root of k (e.q. the condition number) is much larger than (approx.) 30 this could be, according to many authors, a sign of harmful multicollinearity.

Solutions to the Problem of Multicollinearity :

1. If multicollinearity doest not seriously affect the estimates of the coefficients, one may tolerate its presence in the regression model.
2. Drop redundant variable and this applies to the case that two or more variables in an equation are measuring essentially the same thing. It shall make no statistical difference in which variable is dropped.
3. A solution to the perfect multicollinearity is to drop one or more collinear variables, but one has to be careful about the interpretation of the coefficients. But, if the variable really belongs in the model, this can lead to specification error, which can be even worse than multicollinearity. Eliminating variables to "solve" multicollinearity problems results in estimators that are biased, inconsistent, and inefficient.
4. Step by step regression method may be used to eliminate the variable introduction of which in the model does not increase R^2 but increases standard error of significantly. One of the best solutions to the problem of multicollinearity is to delete collinear variables from the regression model. We have a regression of Y on X_1, X_2, X_3 and X_4 and we find that X_1 is highly correlated with X_4. By comparing the R^2 and adjusted R^2 of different regression with and without one of the variables, we can decide which of the two independent variables to drop from regression. We want to maintain a high R^2 and therefore should drop a variable of R^2 is not reduced much when the variable is dropped from the equation. When the adjusted R^2 increase when a variable is removed, we certainly want to drop the variable. For example that the R^2 of the regression with all four independent variables is 0.94, the R_2 when X_1 is removed is 0.87, and the R^2 of the regression of Y on X_1, X_2 and X_3 (X_4 deleted) is 0.92. In this case, we should drop the variable X_4 and not X_1.
5. Multicollinearity often occurs in small samples and many included variables (low degrees of freedom). With a few observations, variables just happen to be closely related. Repeated observations will lesson such chance occurrences. If it is not possible to get more data, one may have to conclude that the data available does not permit one to reliably determine

the individual effects of each variable. But what is important is not the number of observations but the informational content.

6. The fundamental problem is that if two variables are highly correlated, then it becomes difficult to identify the independent impact of the two variables. One possible solution to the problem is to increase the sample size, which should introduce greater variation to allow the independent effects to be disentangled. A larger data set will allow more accurate estimates and will reduce somewhat the variance of the estimated coefficients .In time series contexts, this may not be all that feasible as additional data may take some time to become available.
7. If the purpose of the analysis is to predict future values of the dependent variable ,and we are not interested in the values of the individual parameters, then we may regress the dependent variables, even when a set of independent variables is highly (but less than perfectly) correlated. However, if the goal is explanation, measures other than increasing the sample size are needed.
8. Collinearity is problematic when one's purpose is explanation rather than mere prediction. Collinearity makes it more difficult to achieve significance of the collinear parameters. But if such estimates are statistically significant, they are as reliable as any other variables in a model. And even if they are not significant,, the sum of the coefficient is likley to be reliable. In this case, increasing the sample size is a viable remedy for collinearity when prediction instead of explanation is the goal .
9. Multicollinearity often occurs when several variables seem to be moving together, particularly in the case of nominal time - series data. For instance, both investment spending and government spending tend to move together as the price level rises. Translation of nominal variables into real magnitudes through the use of a price index may alleviate this joint movement. A common remedy to the multicollinearity problem is deflating time series (mostly prices, or price indexes) by some time series measuring e.g. consumption prices. Thus, instead of working with nominal quantities it is preferred to use real quantities.
10. Variables may be highly correlated through time, but not across space, or vice versa. If data sources are available, multicollinearity can sometimes be lessened by using cross section data instead of time series, or in pooling time series and cross section observations.
11. Use other statistical methods rather than OLS approach . As stated previously, the multicollinear relationship between explanatory variables can often be very complex in nature, therefore not lending itself to this simple approach. If this is the case, then a fairly sophisticated solution would be to use one of the *biased regression techniques* such as ***ridge regression*** or ***principal components regression***; these regression techniques produce (biased) parameter estimates that are typically smaller in

magnitude than the corresponding least squares estimates.

12. The problem of multicollinearity may be solved by using an alternative to the least squares method called ridge regression. **Ridge regression** is an effective counter measure because it allows better interpretation of the regression coefficients by imposing some bias on the regression coefficients.. Although the coefficient estimators produced by ridge regression are biased, in some cases, it may be worthwhile to tolerate some bias in the regression estimators in exchange for a reduction in the high variance of the estimators that results from multicollinearity.
13. Use information from prior research to place justifiable constraints on parameters and imposing restrictions on coefficients . For example, we might know from prior research that $b_1 = 3\, b_2$. Another form of extraneous information that can be used is a constraint, or restriction, on the parameters being estimated. For example, in the estimation of a Cobb-Douglas production function it is possible to restrict it to be homogenous of degree 1. Suppose that a researcher estimated.

$$Y = b_0 X_1^{b1} X_2^{b2} \qquad \text{....(12.1)}$$

where Y is the output, X_1 is input 1 and X_2 is input 2, with a sample where X_1 and X_2 are "highly" correlated. Introduce the constraint

$$b_1 + b_2 = 1 \qquad \text{....(12.2)}$$

which states that (12.1) is homogeneous of degree one. With this information, we can substitute $b_1 = (1 - b_2)$ into (12.2) to obtain

$$Y = b_0 X_1^{(1-b_2)} X_2^{b2} \qquad \text{....(12.3)}$$

Taking logs, we now get

$$Y^* = C_0 + (1 - b_2) X_1^* + b_2 X_2^* \qquad \text{....(12.4)}$$

where the asterisk denotes the natural logs of the original variables and $C_0 = L_n b_0, (Ln = \log)$. This yields

$$Y^* = C_0 + X_1^* + b_2 \left(X_2^* - X_1^*\right) \qquad \text{....(12.5)}$$

$$\left(Y^* - X_1^*\right) = C_0 + b_2 \left(X_2^* - X_1^*\right) \qquad \text{....(12.6)}$$

Let $Q = \left[Y^* - X_1^*\right]$ and $I = \left[X_2^* - X_1^*\right]$ and obtain an estimate of b_2 from

$$Q = C_0 + b_2 I \qquad \text{....(12.7)}$$

And obtain the estimate of b_1 from (12.2)

14. If the three X's are all indicators of the same concept, create some sort of composite scale and use it instead.
15. Do nothing.
    (a) It is possible to have severe multicollinearity, but yet have each

individual t-statistic be significant.

(b) If we drop nearly multicollinear variables, we might have created omitted variable bias (that is a more serious problem).

16. A final way of avoiding multicollinearity is through the use of instrumental variables, which are discussed in the later chapter.

An answer favored by many experts would be "nothing," a view nicely expressed by Blanchard (1987):"Multicollinearity is God's will, not a problem with OLS or statistical techniques in general." Other possible actions are summarized in Table 12.2.

TABLE 12.2 . POSSIBLE ACTIONS AND THEIR CONSEQUENCES

Action	Consequence
Drop one or more predictors	If the model was correctly specified, you now have a specification bias
Transform one or more variables	Predictive power may be lost and you may introduce specification bias
Use factor analysis, principal components, or ridge regression	Results may be impossible to interpret in the context of the original problem
Enlarge the sample	Reduces standard errors but may not be possible

The following figure 12.4 provides some idea about the model building when the explanatory variables are highly correlated with each other .

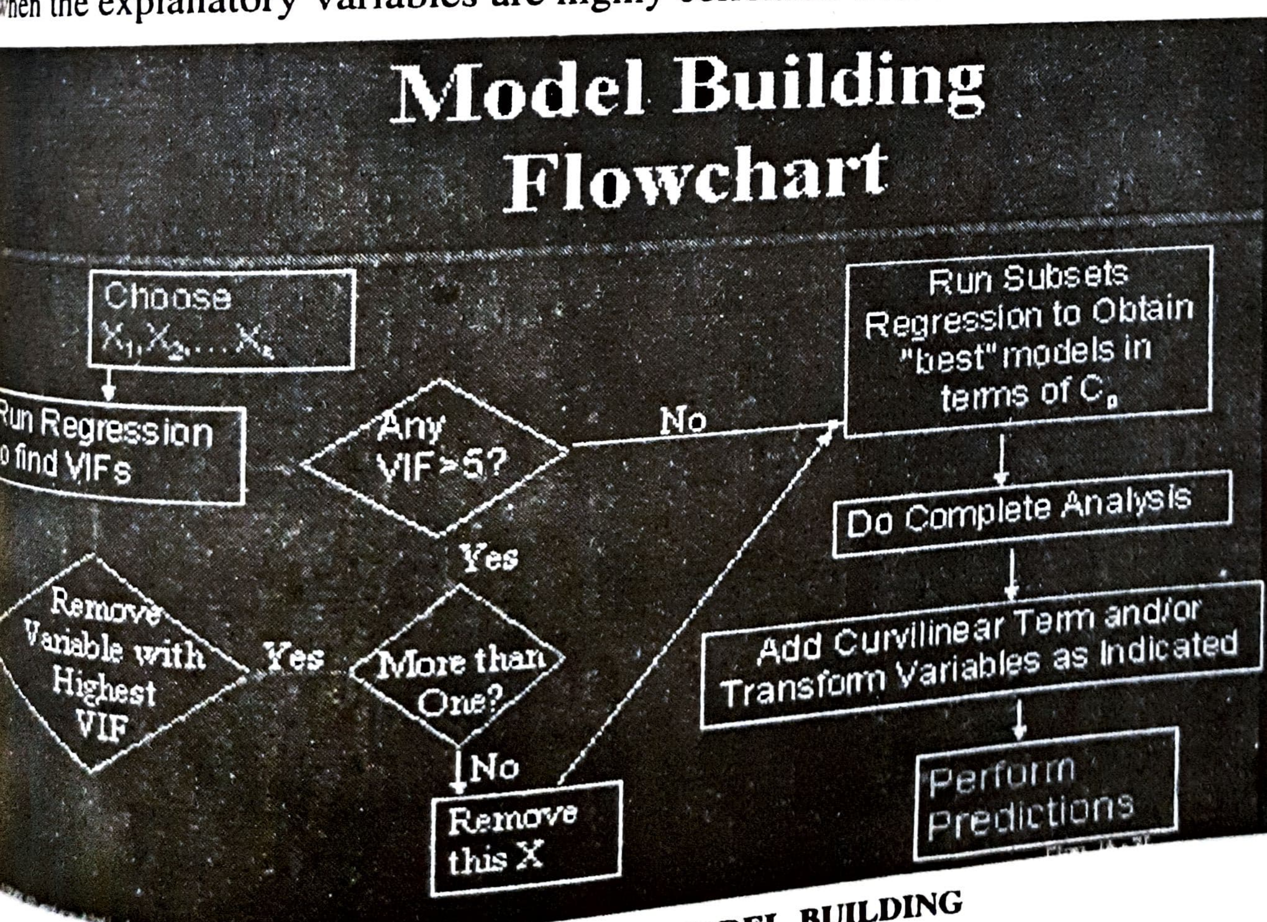

FIGURE 12.4. VIF AND MODEL BUILDING

where Cp is the Mallows Cp criterion to select the regressors.

EXERCISES

1. What is meant by multicollinearity ?.
2. What is the difference between perfect multi-collinearity and near multi-collinearity
3. Define multicollinearity and orthogonality.
4. Distinguish between multicollinearity and orthogonality.
5. How do we correct for perfect multi-collinearity.
6. How do we detect near multi-collinearity.
7. Answer the following questions about multicollinearity:
   (a) Define what it is.
   (b) Why does it occur?
   (c) What are its consequences?
   (d) How can it be detected?
   (e) Is it true that multicollinearity is always a bad thing and nothing can be done about it? Give reasons for your answer together with illustrative examples
8. What are the problems with perfect multicollinearity ?
9. What are the problems with imperfect multicollinearity?
10. Give the sources of multicollinearity
11. Explain the various consequences of multicollinearity
12. Discuss the problem of collinearity
13. Give solutions to remove multicollinearity
14. Explain Multcollinearity .
15. What are the consequences and solutions for multicollinearity?
16. Give the sources of multicollinearity.
17. Bring out the test of multicollinearity.
18. Examine the meaning of multicollinearity. Also eExplain the sources and consequences of multicollinearity.
19. Define multicollinearity, Describe the sources and solutions for it.
20. Discuss the effects of multicollinearity for $Y.X_1X_2$, $R^2 = 0.943$, n=15.

$$Y = 1.48 - 0.65 X_1 \quad R^2 = 0.78$$
$$(SE=0.12)$$
$$Y = 1.21 + 0.128 X_2, \quad R^2 = 0.941$$
$$(SE=0.011)$$
$$Y = -1.92 + 0.19X1 + 0.16 X_2 \quad R^2 = 0.943$$
$$SE = \quad (0.19) \quad (0a.03)$$