

# 1 Multicollinearity (Soc 504)

Until now, we have assumed that the data are well-conditioned for linear regression. The next few sets of notes will consider what happens when the data are not so cooperative. The first problem we will discuss is multicollinearity.

## 1.1 Defining the Problem

Multicollinearity is a problem with being able to separate the effects of two (or more) variables on an outcome variable. If two variables are significantly alike, it becomes impossible to determine which of the variables accounts for variance in the dependent variable. As a rule of thumb, the problem primarily occurs when  $x$  variables are more highly correlated with each other than they are with the dependent variable.

Mathematically, the problem is that the  $X$  matrix is not of full rank. When this occurs, the  $X$  matrix (and hence the  $X^T X$  matrix) has determinant 0 and cannot be inverted. For example, take a general  $3 \times 3$  matrix  $A$ :

$$A = \begin{bmatrix} a & b & b \\ c & d & d \\ e & f & f \end{bmatrix}$$

The determinant of this matrix is:

$$adf + bde + bcf - (bde + adf + bcf) = 0$$

Recall from the notes on matrix algebra that the inverse can be found using the determinant function:

$$A^{-1} = \frac{1}{\det(A)} \text{adj}(A).$$

However, when  $\det(A) = 0$ , all of the elements of the inverse are clearly undefined. Generally, the problem is not severe enough (e.g., not every element of one column of  $X$  will be identical to another) to crash a program, but it will produce other symptoms.

To some extent, multicollinearity is a problem of not having enough information. Additional data points, for example, will tend to produce more variation across the columns of the  $X$  matrix and allow us to better differentiate the effects of two variables.

## 1.2 Detecting/Diagnosing Multicollinearity

In order to lay the foundation for discussing the detection of multicollinearity problems, I conducted a brief simulation, using the following generated variables ( $n = 100$  each):

$$\begin{aligned}
u &\sim N(0, 1) \\
x_2 &\sim N(0, 1) \\
e &\sim N(0, 1) \\
x_1 &= .9 \times u + (1 - .9^2)^{\frac{1}{2}}
\end{aligned}$$

The construction of the fourth variable gives us variables  $x_1$  and  $x_2$  that have a correlation of .9. I then created a series of  $y$  variables:

$$\begin{aligned}
y_1 &= 3 + x_1 + x_2 + (.01)e \\
y_2 &= 3 + x_1 + x_2 + (.1)e \\
y_3 &= 3 + x_1 + x_2 + (1)e \\
y_4 &= 3 + x_1 + x_2 + (5)e
\end{aligned}$$

Changing the error variance has the effect of altering the noise contained in  $y$ , reducing the relationship between each  $x$  and  $y$ . For example, the following are the correlations of each  $x$  and  $y$ :

	$X_1$	$X_2$
$X_1$	1	.91
$X_2$	.91	1
$Y_1$	.98	.98
$Y_2$	.98	.98
$Y_3$	.87	.87
$Y_4$	.37	.38

Notice that all of the correlations are high (which is atypical of sociological variables), but that the  $X$  variables are more highly correlated with each other than they are with  $Y_3$  and  $Y_4$ .

I conducted regression models of each  $y$  on  $x_1$  and  $x_2$  to examine the effect of the very high collinearity between  $x_1$  and  $x_2$ , given different levels of ‘noise’ ( $f(e)$ ) disrupting the correlation between each  $x$  and  $y$ . I then reduced the data to size  $n = 20$  and re-conducted these regressions. Here are the results:

	$y_1$	$y_2$	$y_3$	$y_4$
N=100				
$R^2$	1	1	.79	.15
F	1740033 ***	17463 ***	181.11 ***	8.51 ***
$b_0$	3.0 ***	3.0	3.04 ***	3.19 ***
$b_1$	1.0 ***	.99 ***	.93 **	.63
$b_2$	1.0 ***	1.01 ***	1.11 **	1.57
N=20				
$R^2$	1	1	.82	.27
F	333802 ***	3378 ***	38.63 **	3.08#
$b_0$	3.0 ***	2.98 ***	2.76*	1.82
$b_1$	1.01 ***	1.12 ***	2.22*	7.08
$b_2$	.99 ***	.90 ***	0	-4.0

Some classic symptoms of multicollinearity include: 1) having a significant F, but no significant t-ratios; 2) wildly changing coefficients when an additional (collinear) variable is included in a model; and 3) unreasonable coefficients.

This example highlights some of these classic symptoms. First, in the final model ( $n = 20$ ,  $y_4$ ), we have a significant  $F$  ( $p < .1$ ), but none of the coefficients is significant (based on the t-ratios). Second, if we were to reconduct this final model with only  $x_1$  included, the coefficient for  $x_1$  would be  $2.06 ***$ . However, as the model reported above indicates, the coefficient jumps to  $7.08$  when  $x_2$  is included. Finally, as the final model results indicate, the coefficients also appear to be unreasonable. If we examined the bivariate correlation between each  $x$  and  $y$ , we would find moderate and positive correlations—so, it may be unreasonable for us to find regression coefficients that are opposite and large as in this model.

To summarize the findings of the simulation, it appears that when there is relatively little ‘noise’ in  $y$ , collinearity between the  $x$  variables doesn’t appear to cause much of a problem. However, when there is considerable noise (as is typical in the social sciences), collinearity significantly influences the coefficients, and this effect is exacerbated when there is less information (e.g.,  $n$  is smaller). This highlights that to some extent multicollinearity is a problem of having too little information.

There are several classical tests for diagnosing collinearity problems, but we will focus on only one—the variance-inflation factor—perhaps the most common test. As Fox notes, the sampling distribution variance for OLS slope coefficients can be expressed as:

$$V(b_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma_e^2}{(n - 1)S_j^2}$$

In this formula,  $R_j^2$  is the explained variance we obtain when regressing  $x_j$  on the other  $x$  variables in the model, and  $S_j^2$  is the variance of  $x_j$ . Recall that the variance of  $b_j$  is used in constructing the  $t$ -ratios that we use to evaluate significance. This variance is increased if either  $\sigma_e^2$  is large,  $S_j^2$  is small, or  $R_j^2$  is large. The first term of the expression above is called

the variance inflation factor ( $VIF$ ). If  $x_j$  is highly correlated with the other  $x$  variables, then  $R_j^2$  will be large, making the denominator of the  $VIF$  small, and hence the  $VIF$  very large. This inflates the variance of  $b_j$ , making it difficult to obtain a significant  $t$ -ratio. To some extent, we can offset this problem if  $\sigma_e^2$  is very small (e.g., there is little noise in the dependent variable-or alternatively, that the  $x$ 's account for most of the variation in  $y$ ). We can also offset some of the problem if  $S_j^2$  is large. We've discussed this previously, in terms of gaining leverage on  $y$ , but here increasing the variance of  $x_j$  will also help generate more noise in the regression of  $x_j$  on the other  $x$ 's, and will thus tend to make  $R_j^2$  smaller.

What value of  $VIF$  should we use to determine whether collinearity is a problem? Typically, we often use 10 as our 'threshold' at which we consider it to be a problem, but this is simply a rule of thumb. The figure below shows what  $VIF$  is at different levels of correlation between  $x_j$  and the other variables.  $VIF = 10$  implies that the r-square for the regression must be .9.

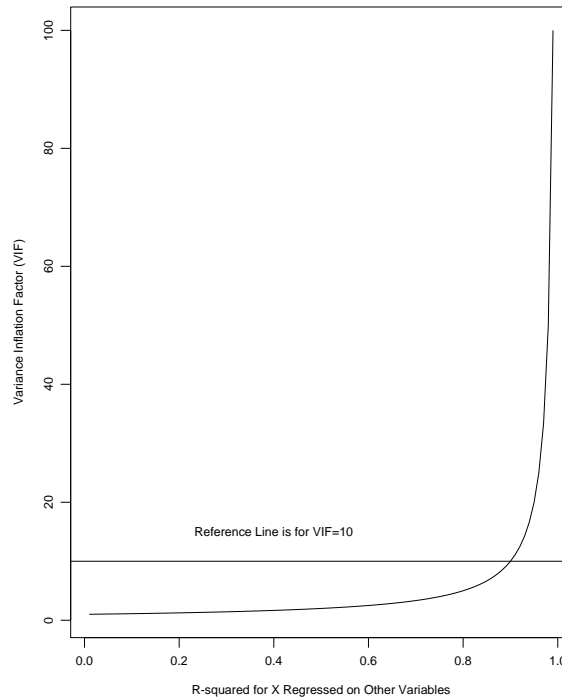


Figure. Variance Inflation Factor by Level of Relationship between X Variables.

### 1.3 Compensating for multicollinearity

There are several ways for dealing with multicollinearity when it is a problem. The first, and most obvious, solution is to eliminate some variables from the model. If two variables are highly collinear, then it means they contain highly redundant information. Thus, we can pick one variable to keep in the model and discard the other one.

If collinearity is a problem but you can't decide an appropriate variable to omit, you

can combine the offending  $x$  variables into a reduced set of variables. One such approach would be to conduct an exploratory factor analysis of the data, determine the number of unique factors the  $x$  variables contain, and generate either (factor) weighted or unweighted scales, based on the factors on which each variables ‘load.’ For example, suppose we have 10  $x$  variables that may be collinear. Suppose also that a factor analysis suggests that the variables really reflect only two underlying factors, and that variables 1 – 5 strongly correlate with the first factor, while variables 6 – 10 strongly correlate with the second factor. In that case, we could sum variables 1 – 5 to create a scale, and sum variables 6 – 10 to make a second scale. We can either sum the variables directly, or we can weight them based on their factor loadings. We then include these two scales in the regression model.

Another solution is to transform one of the offending  $x$  variables. We have already seen that multicollinearity becomes particularly problematic when two  $x$  variables have a stronger relationship with each other than they have with the dependent variable. Ideally, if we want to model the relationship between each  $x$  and  $y$ , we would like to see a strong relationship between the  $x$  variables and  $y$ . Transforming one or both  $x$  variables may yield a better relationship to  $y$ , and at the same time, it will eliminate the collinearity problem. Of course, be sure not to perform the same transformation on both  $x$  variables, or you will be back at square 1.

A final approach to remedying multicollinearity is to conduct ‘ridge regression.’ Ridge regression involves transforming all variables in the model and adding a biasing constant to the new  $(X^T X)$  matrix before solving the equation system for  $b$ . You can read about this technique more in-depth in the book.

## 1.4 Final Note

As a final note, we should discuss why collinearity is an issue. As we’ve discussed before, the only reason we conduct multiple regression is to determine the effect of  $x$  on  $y$ , net of other variables. If there is no relationship between  $x$  and the other variables, then multiple regression is unnecessary. Thus, to some extent, collinearity is the basis for conducting multiple regression. However, when collinearity is severe, it leads to unreasonable coefficient estimates, large standard errors, and consequently bad interpretation/inference. Ultimately there is a very thin line between collinearity being problematic and collinearity simply necessitating the use of multiple regression.