

1 What is Multi Collinearity.

Note. It is not the same as “bivariate correlation” among x’s.

Don’t be silly and do Pearson correlations to check for multicollinearity. The word has “multi” for a reason! It is not *bi-collinearity*!

MC means that it is possible to calculate the value of one x using a linear formula that combines the other x’s. If a model has IV’s $X1_i$ and $X2_i$ and $X3_i$, MC would exist if you could predict $X3$ with this formula:

$$X3_i = k_1X1_i + k_2X2_i$$

If you make a mistake and put the same variable in a regression two times, what do you get? Perfect multicollinearity, the estimation process for the model should crash and complain to you.

If you put variables in a model that are similar, but not identical, then you have what in practice we call multicollinearity.

Brief Technical Comment

Recall that the OLS estimator is

$$\hat{b} = (X'X)^{-1}X'Y$$

where X is the matrix of independent variables and Y is the column of observed responses. Recall that $(X'X)$ is the matrix of cross products of the variables. (do a sketch to make sure).

Recall also that the estimated variance of b also has $(X'X)^{-1}$ in the formula:

$$V(\hat{b}) = s_e^2 * (X'X)^{-1}$$

This is a column of variances of the b’s. The symbol s_e^2 is the root MSE, the estimated variance of the error term.

If it is impossible to calculate this:

$$(X'X)^{-1}$$

then it is impossible to calculate either \hat{b} or $V(\hat{b})$.

Perfect multicollinearity means that $(X'X)^{-1}$ cannot be calculated. Another way of saying this is that $(X'X)$ cannot be “inverted.”

Here is an analogy with ordinary numbers. Suppose we have a number $X = 0$. Then the inverse, X^{-1} is undefined. X cannot be inverted.

Suppose instead $X = 0.0000000001$. Now the inverse of X does exist, but it is some HUGE number, $X^{-1} = \frac{1}{0.0000000001} = 10^9$.

What’s that mean? Recall that $(X'X)^{-1}$ is a matrix defined in the following way:

$$(X'X) * (X'X)^{-1} = I = \begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & \dots & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{matrix}$$

The “identity matrix” I is filled up with zeroes, except for the “main diagonal”. If X is an nxp matrix, then X' is pxn, and so the product $(X'X)$ is pxp. That is to say, it is square and has the number of columns equal to the number of variables in X.

If $(X'X)$ cannot be inverted, it means that there are 2 or more redundant rows in $(X'X)$. Some computer programs will give the error “the model is not full rank.” The term “rank” means the number of rows that are

not simple multiples of each other. How could two rows be redundant? Well, if you put the same variable in the model twice, that could happen. If “rank” of $(X'X)$ is smaller than p , then it means there are redundant rows.

Usually, in practice, we don't find perfect collinearity. Instead, we find moderate to severe multicollinearity. The cross product matrix $(X'X)$ can still be inverted, however, the values in $(X'X)^{-1}$ are HUGE.

2 Effects of MC:

2.1 MC raises the standard error of estimated coefficient.

Suppose $H_0 : b = 0$. MC makes t-statistics smaller, since $t = \frac{\hat{b}}{s.e(\hat{b})}$ (or whatever notation you use for standard error...) Find a book that gives the formula for the $s.e(\hat{b})$ for a model with a few independent variables. It should be easy to see that as the variables become more similar, then the $s.e(\hat{b})$ gets bigger.

2.2 Variance inflation factor (VIF) interpretation

s_e^2 is the estimated variance of the error term (MSE)

s_j is the standard deviation of variable j .

R_j^2 is the coefficient of determination from a regression in which the j 'th variable is the dependent variable and all other independent variables are included in r.h.s.

The “variance inflation factor” is defined $\frac{1}{1-R_j^2}$. You can see why in this expression for the variance of the estimated coefficient b_j :

$$V(\hat{b}_j) = \frac{s_e^2}{(n-1)s_j^2} \times \frac{1}{1-R_j^2}$$

2.3 Specification change causes unpredictable changes in coefficients. Estimate of a coefficient b_j should not be affected by putting other variables in and out of the model. But it is.

2.4 Another symptom: Really big R^2 but small t-statistics.

3 Diagnosis: How to find MC, and measure it.

simplest approach. Run all regressions, predicting each IV from other IV's. Then look at their r-squared values $R_1^2, R_2^2, \dots, R_m^2$. One rule of thumb is that if any $R_j^2 \geq R^2$, then multicollinearity is a problem.

4 Interpretation

4.1 \hat{b} 's are still unbiased.

4.2 If t's are "good", don't worry about it.

4.3 MC between two variables (or within a block of variables) need not affect estimates of other coefficients.

4.4 Excluding variables is very VERY very dangerous, because it causes bias.

5 Solutions

5.1 Do nothing: acknowledge problem

5.2 Get more data!

Gather more data and hope the X's are not so intercorrelated. This is the best and only truly meaningful solution.

That is impossible in many projects.

The sad fact is that there is no *magic bullet*.

5.3 Change specification:

5.3.1 use "index" variables.

I see some people begin with a set of variables that are almost the same, and then they combine them by adding them or calculating an average.

There's just about no methodological justification for doing this. But people do it anyway.

5.3.2 Principal Components: combine variables in a more sophisticated way.

If I were doing a project, and could not get more data, and I was forced to write a paper with intercorrelated, I would take this approach.

The Gujarati textbook mentions principal components, but it does not describe this approach on the grounds that it requires matrix algebra that is outside the scope of the book. That may be true, but let me tell you this: I know plenty of matrix algebra, and still I don't understand most books about principal components. There is something that is fundamentally non-econometric about principal components (see below).

What is a principal component? It is an "underlying variable" (unmeasured variable) that is related to some variables, say X_1, X_2, X_3, X_4 . Suppose further that those variables are in "deviations form," meaning they are "centered" about their means. (So $\tilde{X}_1 = x_1 - \bar{x}$, if x_1 were the "original" data.) Now, suppose there were, say 2 columns of numbers, Z_1 and Z_2 , that can be used to predict the X's with a linear formula, like so (where u_1, u_2, u_3, u_4 are columns of errors, and the coefficients a_{ij} are scalars):

$$\begin{bmatrix} X_1 & X_2 & X_3 & X_4 \end{bmatrix} = \begin{bmatrix} Z_1 & Z_2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{bmatrix} + \begin{bmatrix} u_1 & u_2 & u_3 & u_4 \end{bmatrix}$$

The columns Z_1 and Z_2 are the so-called principal components.

You don't always have just 2 principal components. If you have 4 X variables, you could have 4 principal components. But that would not simplify anything, so we want to focus our attention on the principal components that "really count." In this example, I use just the first the first 2 eigenvectors (2 rows of a) because I'm assuming there are only 2 meaningful principal components.

I suppose a "whirl" of matrix calculations would usually result here, but it is not necessary. Here's what is important.

1. By design, the columns Z_1 and Z_2 are uncorrelated. So if we remove the X 's from the regression model, and we use the Z 's instead, then our "inputs" are not intercorrelated any more.

2. By design, the columns Z_1 and Z_2 give the “best possible” linear prediction about the values of the X ’s.
3. Sometimes you can give the principal component’s a substantive interpretation. You do that by interpreting the matrix of a ’s.

$$\begin{bmatrix} X_1 & X_2 & X_3 & X_4 \end{bmatrix} = \begin{bmatrix} Z_1 & Z_2 \end{bmatrix} \begin{bmatrix} .5 & .5 & 0.0 & 0.0 \\ 0.0 & 0.0 & .5 & .5 \end{bmatrix} + \begin{bmatrix} u_1 & u_2 & u_3 & u_4 \end{bmatrix}$$

You’d probably never see such a sharp separation of the rows of a in practice, but this example makes it very clear. It shows that the principal component Z_1 is giving the predictions for X_1 and X_2 , and the principal component Z_2 is predicting X_3 and X_4 .

So people would make the interpretation that Z_1 is the essence of variables X_1 and X_2 .

What’s that in practice? Well, suppose the variables are like this:

X_1	number of children in public school
X_2	number of teachers in public schools
X_3	number of employees in city government
X_4	number of desks owned by city

In the context of this example, clearly the first 2 go together, the last 2 go together, and the matrix of a ’s is telling you so. The first component is a description of the scale of the public schools (since teachers and students go together) and the second is the level of city employment.

4. The coefficients in the matrix a can be calculated from the X ’s: they are the *eigenvectors* of the correlation matrix of the X ’s. Who knows what that means? who cares? What is important: the *eigenvector* has a “magical” property that you can multiply it by the Z ’s and get the best predictions about the X ’s.

The psychologists and applied statistics folk like it, the econometricians do not. People who do medical research seem to like it.

Here’s why econometrics folk don’t favor this. In the words of William Greene, *Econometric Analysis*, 5th ed (p. 58)

The problem here is that if the original model in the form $y = X\beta + \epsilon$ were correct, then it is unclear what one is estimating when one regresses y on some set of linear combinations of the columns of X . Algebraically, it is simple; at least for the principal components case, in which we regress y on $Z = XC_L$ to obtain d , it follows that $E(d) = \delta = C_L C_L' \beta$. In an economic context, if β has an interpretation, then it is unlikely that δ will. (How do we interpret the price elasticity plus minus twice the income elasticity?)

Many of the details are discussed in Julian Faraway’s online text, “Practical Regression and Anova with R” <http://www.stat.lsa.umich>

5.4 Use “Ridge Regression” or another estimator that is biased, but has lower variance.

Some of these approaches take a Bayesian perspective of positing an original belief (value and associated uncertainty) and then the new estimate is added to update that belief. Adding information adds efficiency.

See Faraway.

The idea is that instead of using the OLS estimator, we use a scalar value, λ , known as the “ridge constant,” to create an adjusted estimator:

$$\hat{b} = (X'X + \lambda I)^{-1} X'Y$$

One adjusts the ridge coefficient. If a small value of λ is used, then, of course, the estimates are not far from the \hat{b}^{ols} .

This estimator is known to be biased, but it is also known to have much lower variance than the OLS estimator. People who are Bayesians have a much easier time incorporating this kind of estimator into their work. I recall liking the discussion of this in Greenberg and Webster’s *Advanced Econometrics: A Bridge to the Literature*.

Greene's text says ridge regression is rarely used because econometricians don't like bias. There's a hint in his tone that he thinks their views are too narrow.

How would you decide about using a biased estimator? Suppose your goal is to have an estimator with the smallest squared-error, in the sense you want the smallest value

$$E[(\hat{b} - b)^2]$$

That is the squared difference between the estimate and the true value.

$$E[(\hat{b} - b)^2] = (E[\hat{b} - b])^2 + E[(\hat{b} - E(\hat{b}))^2]$$

which is

$$E[(\hat{b} - b)^2] = \text{bias of estimator}^2 + \text{variance of estimator}$$