

LINEAR MODEL

SP03 7 - 2

MULTICOLLINEARITY

Multicollinearity is actually related to *three* of the OLS assumptions:

1. No perfect linear relationship among the regressors
2. A greater number of observations than parameters (variables)
3. Sufficient variability in the values of the regressors

For the first of these, we say that there cannot be any set of λ s such that:

$$\lambda_0 X_0 + \lambda_1 X_1 + \dots + \lambda_k X_k = 0 \quad (1)$$

where the lambdas are constants not all zero.

If this is the case, then any variable X_j can be written as an exact function of the other variables:

$$X_j = (-\lambda_0/\lambda_j)X_0 - (\lambda_1/\lambda_j)X_1 + \dots$$

Multicollinearity is also the name we give to the problem of *nearly* perfect linear relationships among our regressors

- This is the more common problem...

Why is this a problem?

Mathematically: (scalar case...)

Consider: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$

where: $X_2 = \lambda X_1$

Then:
$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \\ &= \beta_0 + \beta_1 X_1 + \beta_2 (\lambda X_1) + u \\ &= \beta_0 + (\beta_1 + \beta_2 \lambda) X_1 + u \end{aligned}$$

So what are the values of β_1 and β_2 ?

A: There are no unique values for them:

- Its essentially one equation in two unknowns

- You can pick a value for one, and there will be an associated value for the other

BUT:

- We **can** estimate a linear combination of the parameters $(\beta_1 + \beta_2\lambda)$...
- This can be useful, as we'll see later...

MOREOVER...

- If the linear relationship is not perfect, then we can estimate the values of β_1 and β_2 just fine...
- I.e., the estimates are unbiased, efficient, etc...

What about the standard errors of the two?

- In the bivariate case, if the collinearity is perfect, they will be infinite, and inestimable...
- $\text{Var}(\beta_1) = \sigma^2 / \{\sum(X_{1i} - \bar{X}_1)^2(1 - r^2_{12})\}$ - variance of the errors over the product of the squared deviations of X_1 and (one minus the correlation between X_1 and X_2)

If the two variables are perfectly linearly related, then $r^2_{12} = 1$, the denominator goes to zero (in the limit), and the variance (and thus standard error) of the coefficients are infinite...

What about near-perfect multicollinearity?

- Then r^2_{12} is nearly equal to one
- The variance (and thus the s.e.'s) of the coefficients will be VERY LARGE

We'll come back to this in a minute...

Now think about this problem in matrix terms...

$$\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

If the X variables are perfectly related

- We can't invert $\mathbf{X}'\mathbf{X}$ (because its determinant will be zero)
- We therefore can't get a unique solution to the vector of coefficients β
- This is the same result as in scalar notation...

What about the standard errors?...

Recall that $\text{Var-Cov}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

Once again, because we can't invert $\mathbf{X}'\mathbf{X}$, the standard errors cannot be computed...

If the multicollinearity is near-perfect, $(\mathbf{X}'\mathbf{X})^{-1}$ will be nonzero and very large...

- Intuitively, because the X's are so highly related, their covariation is quite high
- Since the overall VCV is inversely proportional to the amount of (independent) variability in \mathbf{X} , small amounts of variability among the X's yield large standard errors...

Alternatively, note that in the matrix $\mathbf{X}'\mathbf{X}^{-1}$, the k th diagonal element is equal to

$$1 / \{(\mathbf{X}_k'\mathbf{X}_k)(1 - R_k^2)\}$$

where $\mathbf{X}_k'\mathbf{X}_k$ is the variance of the k th variable X_k and R_k^2 is the R-squared from the regression of X_k on all the other X's.

This means that:

- If the relationship is perfect, then the R-squared is one and the s.e.'s are inestimable
- If the relationship is near-perfect, then the denominator is small and the variance of β_k is LARGE...

THE IDEA OF MULTICOLLINEARITY...

Remember I said that multicollinearity is related to all three assumptions?

1. No perfect linear relationship among the regressors
2. A greater number of observations than parameters (variables)
3. Sufficient variability in the values of the regressors

Well, it is.

2. What if you have more regressors than observations? (A special case of (1))

- Remember that multivariate OLS is really a solution to a system of k equations in k unknowns...
- X's are assumed to be *fixed*
- If there are fewer observations (X_i 's) than variables, there are not enough "fixed" values to allow us to solve for the unknowns

- Like the original discussion of two data points...needed two to get a slope, three for a variability, etc... (*degrees of freedom...*)

Likewise, if the number of observations is more than, but close to the number of parameters, your estimates:

- Will not be based on very much data, and so...
- Will not be very precise.

3. Sufficient variability in the regressors...

- This really gets at the heart of it...

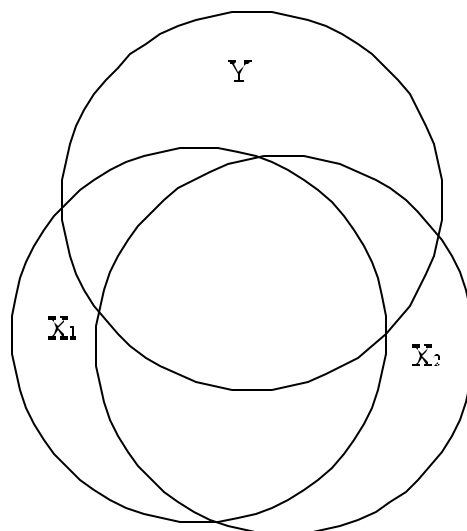
Gujarati's text talks about "micronumerosity" (i.e. too few observations)...

This is really what the near-perfect multicollinearity problem is about...

- If you have a lot of correlation among your independent variables, there is little *independent* variation in them
- That is, each one "explains" the other very, very well

Intuitively, this means that:

- It is difficult to separate the impact of X_1 on Y from that of X_2
- Because of this difficulty, it is more difficult to be sure that the estimate you get is in fact close to the "true" population parameter
- Hence, larger standard errors, and wider confidence intervals
- Venn diagram:



The “sufficient variability” issue is also a special case of near-perfect multicollinearity.

- Essentially, it means that you don't have enough data on “odd cases” to have confidence in your results...

Illustration:

- Effect of Judge's party, and that of appointing president, on decision making
- Two variables are highly correlated ($r = 0.90$ or so)
- In order to determine if my hypotheses are correct, I need sufficient data on judges who are NOT of the same party as their appointing president
- If I have this, then my estimates will be more precise, and I can be more confident in them, BUT
- If I have these data, it means that the two variables aren't so collinear after all...

So it's a bit of a vicious circle...

This suggests two things about the multicollinearity problem:

- It is a **sample** problem
 - * Isolated to the sample data you're using
 - * May not (indeed, if your theory is correct, probably does not) hold in the population
- It is a matter of **degree**
 - * You will always have some covariation among the regressors
 - * The important question is how much, and whether or not it matters...

What does all this mean in practical terms?

- If you have perfect multicollinearity, you'll know it
 - * Programs either bomb, or give complete nonsense results...
 - * Not generally a problem, unless you've created variables which are collinear (more on this when we do model specification...)
- Near-perfect multicollinearity is a stickier issue

Results of near-perfect multicollinearity:

- Estimates are still BLUE
- Estimates will have large standard errors

BUT

- This is not due to any statistical problem, but
- To the fact that its difficult to disentangle the separate effects of the independent variables...

Remember:

- **Multicollinearity is a problem of the *sample*** (not the population, or the data generally)
- **Multicollinearity is a matter of *degree***
 - * Can't "test for" MCLIN;
 - * Can only look to see if/how bad a problem it is.

So how do we detect it?

1) High R-squareds, but nonsignificant coefficients.

- Regression with R^2 high, but none of the variables showing significant effects
- Classic symptom of multicollinearity

BUT

- Don't necessarily need a high R^2 to have a MCLIN problem...
 - * You could have a model that only explains 10% of the variance in Y and still have massively collinear variables
 - * E.g. regression of shoe size and hat size on exam performance...
- This means that detection is a bit harder, since we tend to take insignificant t 's (along with low R-squareds) as signs of a crappy model...

2) High pairwise correlations among independent variables.

- This is an easy way to test for simple MCLIN: correlation matrix on the X's

BUT

- It is also not always effective...
- High pairwise correlations are a sufficient, but not necessary condition for multicollinearity
- More complex linear relationships may exist among the independent variables
- E.g. Suppose that we have four X's and $X_1 = 1 + 2X_2 - 0.5X_3 - 100X_4$ yields an $R^2=0.97...$
 - * The linear combination of X's is nearly perfect, but the individual correlations of X_1 and the other variables may not be that high...

3) High *partial* correlations among the X's.

- This is the natural way of getting around the problem with (2).
- Can be useful, though complicated as the number of X's increases...
- An easier way to get at the same thing is...

4) Auxilliary regressions of the X's.

- Remember that, in the multivariate context, $\text{Var}(\hat{\beta}_k) = 1 / (\mathbf{X}_k' \mathbf{X}_k)(1 - R_k^2)$?
- That "auxiliary" R^2 is very important (it can tell you a lot).
- SO: Regress each of the X's on all the other X's to see if there are any strong linear dependencies...

- Some general rules:
 - * Can do an F-test : $(R_k^2 / k-2) / [(1-R_k^2)/(N-k+1)]$ on the equation...
 - * *Klein's Rule*: If any of the R_k^2 are larger than the R^2 of the model, you have a problem...
 - * Tolerance and the "VIF"
 - The "Variance Inflation Factor" is just $1/(1-R_k^2)$: tells how much the variance of the β_k is "inflated by the multicollinearity"
 - If this is larger than, say, 10 (i.e. $R_k^2 > 0.90$), then you may have a problem
 - Tolerance is just $1 / \text{VIF}$; rescales it to $[0,1]$ — zero = perfect multicollinearity, 1 = none.

5) Eigenvalues.

- Not going to go into it at any length, but we can use the ratio of the maximum and minimum eigenvalues to test for multicollinearity
- Not really any better than auxiliary regressions

Example:

- Fictional data on **Y** and four covariate **X**'s, **N** = 100
- Regression: classic indices of multicollinearity
 - * High R^2
 - * Not especially significant t -values / imprecisely estimated β 's
- Correlation matrix of the variables
 - * Indicates that X_2 and X_3 are highly collinear
 - * X_4 isn't especially highly correlated with the others, BUT
 - * May not tell the whole story
- Auxiliary regressions:
 - * Indicates that $X_4 \approx 6X_2 - 3X_3$
 - * In other words, X_4 is highly collinear with the *combination* of X_2 and X_3

. su

Variable	Obs	Mean	Std. Dev.	Min	Max
Y	100	.4377146	12.36219	-25.39955	27.62726
X1	100	-.0688582	.9145279	-1.787513	1.983905
X2	100	-.0078136	1.007411	-2.528759	2.154691
X3	100	.0508892	2.363825	-5.782297	4.606578
X4	100	-.2246954	3.225809	-7.817698	9.417317

. reg Y X1 X2 X3 X4

Source	SS	df	MS	Number of obs =	100
Model	12310.5005	4	3077.62512	F(4, 95) =	103.71
Residual	2819.0427	95	29.6741336	Prob > F =	0.0000
Total	15129.5432	99	152.823668	R-squared =	0.8137
				Adj R-squared =	0.8058
				Root MSE =	5.4474

Y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
X1	2.409832	.5996684	4.019	0.000	1.219339	3.600324
X2	-.2063823	3.528496	-0.058	0.953	-7.211332	6.798568
X3	-4.097815	1.692719	-2.421	0.017	-7.458286	-.7373438
X4	.6159355	.5531322	1.114	0.268	-.4821708	1.714042
_cons	.9489711	.5475383	1.733	0.086	-.1380299	2.035972

. corr
(obs=100)

	Y	X1	X2	X3	X4
Y	1.0000				
X1	0.1840	1.0000			
X2	-0.7217	0.0186	1.0000		
X3	-0.8723	0.0015	0.8954	1.0000	
X4	0.5302	0.0448	-0.0413	-0.4604	1.0000

. reg X4 X1 X2 X3

Source	SS	df	MS	Number of obs =	100
Model	933.190133	3	311.063378	F(3, 96) =	307.89
Residual	96.9884764	96	1.01029663	Prob > F =	0.0000
Total	1030.17861	99	10.4058445	R-squared =	0.9059
				Adj R-squared =	0.9029
				Root MSE =	1.0051

X4	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
X1	.0472055	.1105438	0.427	0.670	-.1722223	.2666332
X2	5.984867	.2253359	26.560	0.000	5.537579	6.432155
X3	-2.91205	.0960169	-30.329	0.000	-3.102642	-2.721458
_cons	-.0264895	.1009937	-0.262	0.794	-.2269605	.1739815

WHAT DO WE DO ABOUT IT???

One thing that you'll often hear is that "If you have multicollinearity, and your coefficients are significant, *stop right there.*"

- B/c your estimates are still "Good enough" to give you the results you want.
- This is not bad practical advise, but its only *mediocre* statistical advise.
 - * Since the estimates are BLUE, you're OK.
 - * Moreover, since you're dealing with a *sample* problem, it may be the case that your efficient, significant (but imprecise) estimates are the "best you can do"
BUT
 - * We also want the most precise estimates available (not just for big *t*-stats, but in general terms, for precision)
 - * Thus, *even if* your estimates are significant, I suggest at least considering some of the possible "remedies" to improve your estimates.

What not to do...

1) A priori restrictions on coefficients.

- Possible for economists, not likely for us...
- Typically requires too strong a theory.
- We generally want to *estimate* these values...

2) Dropping one or more variables.

- DO NOT DO NOT DO NOT DO THIS!!! (SPECIFICATION ERROR)
- Its very tempting... (when you do, your results will, likely, be significant)
- Taking good (BLUE) estimates and exchanging them for biased, inefficient, crappy ones
- Let THEORY be your guide...
 - * If your theory suggests that one of the variables can be dropped, that's one thing
 - * If not, DO NOT DO IT

What you can do...

1) ADD DATA.

- This is always good...
 - * At the very least, will decrease σ^2 and give you more precise estimates
 - * Moreover, if the data yield more "odd" observations, it will also reduce the MCLIN and improve them even further

NOTE

- * This does NOT mean go collect ONLY data which "go against" the MCLIN pattern in the independent variables (e.g. only Prez/judge party disagreeing jurists)

- Ways to do this...
 - * Just go do it...
 - * “Pool” data across different cross-sections, or time periods
 - Can yield very good results, if you’re careful
 - More on this when we do model specification

2) Transform the Independent Variables.

- Lots of ways to do this...
- Combine related variables into an “index”
 - * If they’re all comparable, add them up, or take a proportion
 - * Use factor analysis to generate factor scores
 - * This MUST be guided by theory: DON’T index/factor variables that are theoretically distinct, even if / just because they’re highly correlated
- Take “First differences” to eliminate time - related multicollinearity
 - * E.g. trending variables are often collinear
 - * Differencing allows you to estimate the same parameters, but without the MCLIN
 - BUT
 - * Can cause some other problems (see autocorrelation, next week)
- “Center” the variables (i.e. subtract the means)
 - * This can be useful if the MCLIN is due to interaction terms
 - * “Centered” variables are less highly correlated with their interaction terms than are non-centered ones, BUT
 - * Note that this *doesn’t* eliminate the problem, and can be misleading as well, if you’re not careful with your interpretation
 - * More on this when we do model specification/interactions

3) Ridge Regression.

- Rarely used...
- Trades off some (known) bias in the coefficients for more efficiency
- E.g. Hoerl and Kennard (1970 *Technometrics*): ORR (“Ordinary Ridge Regression”)
- Where:
 - OLS: $\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$
 - ORR: $\beta_{\text{orr}} = (\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}$
- We’re introducing a bias factor equal to a constant κ into the regression

- Has two effects:
 - * Biases the estimate of β ; since $E(\hat{\beta})$ isn't equal to β anymore
BUT
 - * Decreases the variance of the estimates for β_k : $\sigma^2 / (\mathbf{X}_k' \mathbf{X}_k + \kappa)(1 - R_k^2)$
 - Increases the denominator, decreasing the variance and increasing the overall precision of the estimates
- The larger the κ , the greater the bias, but the smaller the variance...
BUT
- Since we know κ , we know how far "off" our estimate is.
- This is typically done with standardized variables (rescaled to [0,1]) using $\kappa \approx .05$.

Conclusions...

- Perfect MCLIN is not a problem, because you won't get any estimates at all
- Near-perfect MCLIN is different from the other regression violations we'll talk about...
 - * Is a data problem, not an estimator problem, and
 - * One of degree, not kind
- Its always a good idea to do some diagnostics for this, and possibly put them in a footnote if your theory/operationalization would suggest it is a problem...
- When you can, collect more data!

After break: *Measurement Error.*