

POL 502
Lecture 8 Prof. Matthew Lebo
March 31st, 2005

Perfect multicollinearity is a violation of the classical assumption that no independent variable is a perfect linear function of one or more other independent variables.

Its very unlikely that we would make such a mistake, but multicollinearity can be a problem even when it is not perfect.

Why?

The coefficient β_k can be thought of as the impact on the dependent variable of a one-unit increase in the independent variable X_k , holding constant the other independent variable in the equation.

But if two of the independent variables are significantly related in a sample how do we hold one constant while moving the other? It becomes more difficult to distinguish the impact of one each variable.

So, the more correlated are two or more independent variables the more difficult it becomes to estimate the coefficients of the true model. And, if the variables are perfectly related to each other, regression is impossible with all of them included.

Perfect Multicollinearity

‘Perfect’ here means that the variation in on independent variable can be explained completely by the variation in the other.

For example:

$$X_{1i} = \alpha_0 + \alpha_1 X_{2i}$$

where α s are constants and the Xs are independent variables in:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i.$$

Note that there is no error term in the first equation because the relationship is a perfect one – graphing X_1 and X_2 gives us a regression line with all observations falling on the line.

For example, $X_1 = 32 + 1.8X_2$ where X_1 is the temperature in Fahrenheit and X_2 is Celsius. We wouldn’t want to include both in a model that explains how many people go to the beach in a day.

How can we interpret β_1 in that case?

Holding the temperature in Celsius constant it is the change in the number of beach-goers from a one degree change in Fahrenheit.

But obviously we can't hold Celsius constant while changing Fahrenheit.

If we were to try to estimate the two independent variable equation, MINITAB wouldn't let us.

We would not be able to calculate either $\hat{\beta}_1$ or $\hat{\beta}_2$ since it would require division by zero.

Also, the standard errors of each would be infinite.

To calculate the level of significance using t we are left with $t = \frac{?}{\infty}$.

Imperfect Multicollinearity

We are unlikely to make a mistake that would give us perfect multicollinearity, but very often we are confronted with data that have severe imperfect multicollinearity.

Its important to know how what the consequences are, how to figure out if we have it, and how to fix it.

We've discussed this in class before but now we'll go into some greater detail.

With imperfect multicollinearity we have some error in the relationship between the independent variables:

$$X_{1i} = \alpha_0 + \alpha_1 X_{2i} + u_i$$

But if this is a very close relationship – very small errors and high R^2 – we will encounter problems. (see figure 8.2).

Consequences of Multicollinearity

Remember that we want our OLS estimates to be BLUE – unbiased and having minimum variance.

If we have multicollinearity our estimates will remain unbiased. So long as we do not have perfect multicollinearity and we haven't broken other regression assumptions, our

estimates of the $\hat{\beta}$ s in repeated sampling will still be centered around the true population β .

That's the good news.

The bad news is that the variances and standard errors of our estimates will increase.

This is main problem with multicollinearity.

Note that this isn't a small problem – we are most concerned with hypothesis testing and even if our coefficients are unbiased, we are unable to get reliable t statistics if our standard errors are artificially increased.

This occurs because we are unable to separate the effects of the multicollinear variables.

So our estimated coefficients, while unbiased, now come from distributions with much larger variances and therefore standard errors.

Why? Here is equation 4.10, the standard error of a coefficient in a model with 2 independent variables:

$$SE(\hat{\beta}_1) = \sqrt{\frac{\sum e_i^2 / (n-3)}{\sum (X_{1i} - \bar{X}_1)^2 (1 - r_{12}^2)}}$$

The correlation between the two independent variables is r_{12} .

When r_{12} gets very high $(1 - r_{12}^2)$ will get very low causing $SE(\hat{\beta}_1)$ to get higher.

Figure 8.3 shows that with or without multicollinearity we can have the same mean in the sampling distribution of $\hat{\beta}$ but with severe multicollinearity we have a much wider distribution.

The end result of this is that we get lower t statistics.

With smaller correlations we get smaller $SE(\hat{\beta}_1)$ and therefore higher t statistics.

So, regardless of the size of our coefficient the $SE(\hat{\beta}_1)$ is larger with multicollinearity and this makes rejection of our null hypothesis more difficult.

What other problems are there?

Sensitivity to changes in specification.

This just means that you can get wildly different results as you add or remove variables from the model.

Some of you may have experienced this with your projects already. You have a variable that is significant, but when you remove another variable the effect goes away.

Or this could happen when adding or removing a few cases.

If an independent variable has absolutely no correlation with a set of other independent variables, its effect on the dependent variable will not change no matter which or how many of the other independent variables are included in a model.

Also, the overall fit of the equation and the estimation of non-multicollinear variables will be largely unaffected.

So we may be results that show a high R^2 value but that none of our variables are significant – this is a telltale sign of multicollinearity.

Example of Multicollinearity

As an example, we can look at some of the variables in the Country Health data set.

How is the infant mortality rate affected by the life expectancy rates of men and women?

Well, probably spuriously, but lets ignore that for now.

Here is a regression:

Regression Analysis: infmr versus lifeexpm, lifeexpf

The regression equation is
infmr = 303 + 0.187 lifeexpm - 3.86 lifeexpf

Predictor	Coef	SE Coef	T	P
Constant	302.867	6.707	45.16	0.000
lifeexpm	0.1866	0.6615	0.28	0.778
lifeexpf	-3.8595	0.5835	-6.61	0.000

S = 11.3087 R-Sq = 93.3% R-Sq(adj) = 93.1%

We have an R^2 value of 93.3 but we have only one significant variable with a t statistic that – though high – is not high enough to coincide with the R^2 value.

So the coefficient estimates here are unbiased and the R^2 value is unaffected, but the standard errors, t statistics and P values are all based on inflated variance in our estimates.

Maybe this tells us that the infant mortality rate has more to do with women's life expectancy than with men's, but its pretty hard to rely on a model with so much multicollinearity.

Detection of Multicollinearity

What are the correlations?

Correlations: lifeexpm, lifeexpf, birthrat, deathrat, infmr, fertrate

	lifeexpm	lifeexpf	birthrat	deathrat	infmr
lifeexpf	0.988 0.000				
birthrat	-0.846 0.000	-0.870 0.000			
deathrat	-0.826 0.000	-0.799 0.000	0.568 0.000		
infmr	-0.953 0.000	-0.966 0.000	0.862 0.000	0.780 0.000	
fertrate	-0.795 0.000	-0.821 0.000	0.954 0.000	0.547 0.000	0.820 0.000

This is a correlation matrix.

Simply a grid of the Pearson's correlations between pairs of independent variables.

What is too high?

We can choose some arbitrary number, such as an absolute value higher than .8 or .6.

Or, we can use Klein tests as discussed before.

We test to see what the R^2 value is when we regress one independent variable on all the others.

If this value is higher than in our model of Y, we conclude that the independent variables are determining each other excessively.

For example, we begin with:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

We then run regressions of:

$$X_{1i} = \beta_{0a} + \beta_{2a} X_{2i} + \beta_{3a} X_{3i} + \varepsilon_{ai}$$

and $X_{2i} = \beta_{0b} + \beta_{1b}X_{1i} + \beta_{3b}X_{3i} + \varepsilon_{bi}$

and $X_{3i} = \beta_{0c} + \beta_{1c}X_{1i} + \beta_{2c}X_{2i} + \varepsilon_{ci}$.

We can compare the R^2 of these models and if it is higher than when Y is the dependent variable we have too much multicollinearity.

What to do?

Maybe do nothing.

If your t-statistics are still significant then you are able to reject your null hypotheses, even with multicollinearity. So the multicollinearity isn't really causing a problem.

If a variable belongs in an equation theoretically, dropping it out leaves an omitted variable. So we are creating a new problem. This might be ok if the new problem is smaller than the old one, but we should be careful about removing variables that we believe belong there.

If this is the choice you make, you still need to explain the correlation between the variables.

Or, drop one of the redundant variables.

This is certainly easy if the theoretical reasons behind including one independent variable are the same as another.

For instance, if we think that life expectancy is important, we shouldn't have two independent variables in a model measuring it.

When we drop out male life expectancy we get:

Regression Analysis: infmr versus lifeexpf

The regression equation is
`infmr = 304 - 3.70 lifeexpf`

Predictor	Coef	SE Coef	T	P
Constant	303.637	6.103	49.75	0.000
lifeexpf	-3.69694	0.09075	-40.74	0.000

S = 11.2652 R-Sq = 93.3% R-Sq(adj) = 93.2%

Notice that the R^2 is exactly the same and that the adjust R^2 is a little bit higher.

The t statistic is now makes much more sense – if a single variable is explaining so much of the dependent variable then we should expect an enormous t .

Also, note that the coefficient hasn't changed that much – we were getting a good estimate the first time.

Another alternative is to combine variables.

Instead of including both male and female life expectancy we can compute their average.

This is easy enough since they are on the same scale.

Combining variables can be a handy trick in many cases.

We can create indices of variables, such as pollution in our environmental data set.

When things aren't measured on the same scale, we can standardize them all and then combine them.

Another option to deal with multicollinearity are increasing the sample size.

Next week we will begin dummy variables and interaction terms.

Section 7.4 of the text is especially important.