Lecture 11: Multiple Regression
Multicollinearity

Matt Golder & Sona Golder

Pennsylvania State University

Notes

Multiple Regression

So far we have focused exclusively on simple regression for the purpose of learning the basic theory and procedures behind the linear regression model.

Obviously, we have severely limited our ability to test more complicated hypotheses.

We now turn to multiple regression and, in particular, the three-variable regression model.

The population regression function is:

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

Notes

Assumption

When we shift to multiple regression, we will keep the assumptions of the "classic normal linear regression model" (CNLRM) that we introduced for simple regression.

However, we now need to add the assumption of no perfect multicollinearity.

Assumption 10: (No Multicollinearity) When there is more than one independent variable, there are no perfect linear relationships between any of those variables.

Notes

Multicollinearity

Multicollinearity occurs when there is a linear relationship among one or more of the independent variables.

In the case where we have two independent variables, X_1 and X_2, multicollinearity occurs when $X_{1i} = a + bX_{2i}$, where a and b are constants.

When we have multicollinearity, the inclusion of both X_{1i} and X_{2i} in our model is problematic for estimation.

Intuitively, a problem arises because the inclusion of both X_1 and X_2 adds no more information to the model than the inclusion of just one of them.

Effectively, we are asking the regression model to estimate an additional parameter, but we are not supplying it with any additional information.

Notes

Multicollinearity

Consider what happens if $X_{2i} = 1 + 2X_{1i}$.

$$\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \\
Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2(1 + 2X_{1i}) + \epsilon_i \\
Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 + \beta_2 2X_{1i} + \epsilon_i \\
Y_i &= (\beta_0 + \beta_2) + (\beta_1 + 2\beta_2)X_{1i} + \epsilon_i \\
Y_i &= \beta_0' + \beta_1' X_{1i} + \epsilon_i
\end{aligned}$$

You can see that we're only getting two parameters (β_0' and β_1') back even though the original model specification had three (β_0, β_1, and β_2).

Notes

Multicollinearity

There are multiple values of β_0, β_1, and β_2 which will solve the two equations:

$$\begin{aligned}
\beta_1' &= \beta_1 + 2\beta_2 \\
\beta_0' &= \beta_0 + \beta_2
\end{aligned}$$

The bottom line is that we cannot estimate the separate influence of X_1 and X_2 on Y.

We often say that our model is not **identified**.

Notes

Multicollinearity

Our standard errors will also be infinite.

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

As we'll see shortly, the variance of, say, β_1 is

$$\text{Var}(\beta_1) = \frac{\sigma^2}{\sum (X_{1i} - \bar{X}_{1i})^2 (1 - r_{12}^2)} = \frac{\sigma^2}{\sum x_{1i}^2 (1 - r_{12})^2}$$

where r_{12} is the correlation between X_1 and X_2.

If X_1 and X_2 are linearly related, then $r_{12}^2 = 1$, and the denominator goes to zero (in the limit), and the variance goes to infinity.

Notes

Multicollinearity

What we have been discussing so far is really **perfect multicollinearity**.

In the case where we have two independent variables, this occurs when the correlation, r, between two variables is 1.

In the case where we have more than two independent variables, this occurs when your independent variables, or combinations of your independent variables, are not linearly independent, i.e., your matrix of independent variables is not of full rank.

- This is a problem because you cannot estimate all of your coefficients.
- Your standard errors will be infinite.
- Statistical packages will either not produce anything, drop some of your variables, or produce nonsensical results.

Notes

Multicollinearity

Sometimes people use the term multicollinearity to describe situations (i) where there is a *perfect* linear relationship between the independent variables **and** (ii) where there is a *nearly perfect* linear relationship between the independent variables (r_{12}^2 close to one).

The Gauss-Markov assumption only requires that there be no perfect multicollinearity.

So long as we don't have perfect multicollinearity

1. Our model is identified, i.e., we can estimate all of our coefficients.
2. Our coefficients will remain best linear unbiased estimates (BLUE)
3. Our standard errors will be correct and efficient.

Notes

Multicollinearity

In practice, scholars almost never face perfect multicollinearity.

However, they often encounter near-perfect multicollinearity.

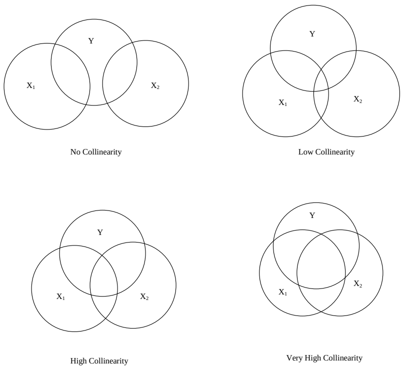
Although the standard errors are technically "correct" and will have minimum variance with near perfect multicollinearity, they will be very, very large.

The intuition is, again, that the independent variables are not providing much independent information in the model and so our coefficients are not estimated with a lot of certainty.

Notes

Multicollinearity

Figure: Different Degrees of Collinearity



Notes

Multicollinearity

Multicollinearity is actually related to three of our regression assumptions:

1. Assumption 7: The number of observations, N, must be greater than the number of parameters to be estimated.
2. Assumption 8: There must be sufficient variation in the values of the X variables.
3. Assumption 10: No perfect linear relationship among the independent variables.

Notes

Multicollinearity

Assumption 7: The number of observations, N, must be greater than the number of parameters to be estimated.

Having a greater number of observations than parameters is actually just a special case of having a perfect linear relationship among the regressors.

Multiple regression is really about solving a system of k equations and k unknowns where the X are assumed to be fixed.

If there are fewer observations than variables, then there are not enough "fixed" values to allow us to solve for the unknowns.

Notes

Multicollinearity

Assumption 8: There must be sufficient variation in the values of the X variables.

The central problem with multicollinearity, though, is that there is not enough variability in the values of the independent variables.

If there is a lot of correlation among the independent variables, then there is little independent variation in them.

Intuitively, this means that it is difficult to separate the impact of X_1 on Y from that of X_2.

As a result, it is difficult to be sure that the estimate you get is in fact close to the "true" population parameter, hence the large standard errors and wide confidence intervals.

Put differently, you don't have enough data on "odd cases" to have confidence in your results.

Notes

Multicollinearity

Example (Zorn): Suppose we are interested in the effect of a Supreme Court judge's political party affiliation and the party of the appointing president on court rulings.

These two variables are obviously highly correlated.

In order to separate the effects of the judge's party and that of the appointing president we need sufficient data on judges who are **not** of the same party as the appointing president.

If I have these data, then my estimates will be more precise and I can be more confident in them.

But, of course, if I have these data, then it means that my two variables aren't that collinear in the first place.

This leads to a bit of a vicious circle.

Notes

Multicollinearity

This all suggests two things about multicollinearity problems:

1. **It is a sample problem.**
   - It is isolated to the sample data you're using and probably does not hold in the population (especially if your theory is correct).
2. **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.

While perfect multicollinearity causes statistical problems, this is not the case for near-perfect multicollinearity.

Notes

Multicollinearity

So, how do you detect near-perfect multicollinearity?

A classic sign of near-perfect multicollinearity is when you have a high R^2 but none of the variables show significant effects.

A problem, though, is that you can still have near-perfect multicollinearity without having a high R^2.

Notes

Multicollinearity

So, how do you detect near-perfect multicollinearity?

You are likely to have near-perfect multicollinearity if you have high pairwise correlations among the independent variables.

A problem, though, is that high pairwise correlations are a sufficient, but not a necessary, condition for near-perfect multicollinearity.

For example, you can have more complex linear relationships among the independent variables where a linear combination of the Xs is nearly perfect but the individual pairwise correlations are not that high.

Notes

Multicollinearity

So, how do you detect near-perfect multicollinearity?

You can regress each of the Xs on all of the other Xs to see if there are any strong linear dependencies.

These are sometimes referred to as **auxiliary regressions**.

If any of the R^2s from these models is greater than the R^2 from the main model, then you may have a problem.

Notes

Multicollinearity

So, how do you detect near-perfect multicollinearity?

You can also look at what are known as **Variance Inflation Factors** (VIF).

These are measures of the how much the variance of your coefficients is "inflated" by multicollinearity.

The **variance-inflating factor** (VIF) is defined as

$$VIF = \frac{1}{1 - r_{23}^2}$$

Why?

Notes

Multicollinearity

We will see shortly that:

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum x_{1i}^2 (1 - r_{12})^2}$$

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{12})^2}$$

$$\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = \frac{-r_{12}\sigma^2}{(1 - r_{12}^2)\sqrt{\sum x_{1i}^2}\sqrt{\sum x_{2i}^2}}$$

Notes

Multicollinearity

So, we can rewrite the variances in terms of VIF:

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum x_{1i}^2}\text{VIF}$$

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2}\text{VIF}$$

As r_{12}^2 approaches 1, VIF approaches infinity, and the variance of the estimator becomes infinite.

If there is no collinearity between X_1 and X_2, then VIF $= 1$.

Notes

Multicollinearity

When you have k independent variables, we have:

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R^2 when we regress X_j on the remaining independent variables.

To obtain VIF scores after a regression in Stata , type

```
. regress y x1 x2
. estat vif
```

As a rule of thumb, VIF > 10 indicates high collinearity.

Notes

Multicollinearity

What, if anything, should you do if you have near-perfect multicollinearity?

To some extent, if you have near-perfect multicollinearity, but your estimates are still significant, you can stop right there.

You should **not** drop one or more variables, no matter how tempting.

You would be exchanging BLUE estimates for biased and inefficient ones.

Notes

Multicollinearity

You can try to add data since this will decrease s^2 and give you more precise estimates.

If the new data gives you more “odd” cases, this will reduce the level of multicollinearity and improve the precision of your estimates even more.

But, you should not selectively search for “odd” cases.

Remember, though, that near-perfect multicollinearity is not a statistical problem.

You just don’t have enough variation in your data, and you may not be able to do anything about this.

Notes

Notes

Notes