CHAPTER 8: MULTICOLLINEARITY

Perfect multicollinearity is the violation of **Assumption 6** (no explanatory variable is a perfect linear function of any other explanatory variables).

Perfect (or Exact) Multicollinearity

If two or more independent variables have an exact linear relationship between them then we have perfect multicollinearity.

Examples: including the same information twice (weight in pounds and weight in kilograms), not using dummy variables correctly (falling into the dummy variable trap), etc.

Here is an example of perfect multicollinearity in a model with two explanatory variables:

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

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

Consequence: OLS cannot generate estimates of regression coefficients (error message).

Why? OLS cannot estimate the marginal effect of X_1 on Y while holding X_2 constant because X_2 moves exactly when X_1 moves!

Solution: Easy - Drop one of the variables!

Imperfect (or Near) Multicollinearity

When we use the word multicollinearity we are usually talking about severe imperfect multicollinearity.

When explanatory variables are approximately linearly related, we have

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

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

The Consequences of Multicollinearity

1. Imperfect multicollinearity does not violate Assumption 6. Therefore the Gauss-Markov Theorem tells us that the OLS estimators are BLUE.

   So then why do we care about multicollinearity?

2. The variances and the standard errors of the regression coefficient estimates will increase. This means lower *t*-statistics.

3. The overall fit of the regression equation will be largely unaffected by multicollinearity. This also means that forecasting and prediction will be largely unaffected.

4. Regression coefficients will be sensitive to specifications. Regression coefficients can change substantially when variables are added or dropped.

The Detection of Multicollinearity

High Correlation Coefficients

Pairwise correlations among independent variables might be high (in absolute value). Rule of thumb: If the correlation > 0.8 then severe multicollinearity may be present.

High R^2 with low *t*-Statistic Values

Possible for individual regression coefficients to be insignificant but for the overall fit of the equation to be high.

High Variance Inflation Factors (VIFs)

A VIF measures the extent to which multicollinearity has increased the variance of an estimated coefficient. It looks at the extent to which an explanatory variable can be explained by all the other explanatory variables in the equation.

Suppose our regression is equation includes *k* explanatory variables:

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

In this equation there are *k* VIFs:

Step 1: Run the OLS regression for each *X* variable. For example for X_{1i}:

$$X_{1i} = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \ldots + \alpha_k X_{ki} + \upsilon_i$$

Step 2: Calculate the VIF for $\hat{\beta}_i$:

$$\text{VIF}(\hat{\beta}_i) = \frac{1}{1 - R_i^2}$$

R_i^2 is the R^2 for the auxiliary regression in Step 1.

Step 3: Analyze the degree of multicollinearity by evaluating each VIF($\hat{\beta}_i$).
Rule of thumb: If VIF($\hat{\beta}_i$) > 5 then severe multicollinearity may be present.

Remedies for Multicollinearity

No single solution exists that will eliminate multicollinearity. Certain approaches may be useful:

1. Do Nothing

Live with what you have.

2. Drop a Redundant Variable

If a variable is redundant, it should have never been included in the model in the first place. So dropping it actually is just correcting for a specification error. Use economic theory to guide your choice of which variable to drop.

3. Transform the Multicollinear Variables

Sometimes you can reduce multicollinearity by re-specifying the model, for instance, create a combination of the multicollinear variables. As an example, rather than including the variables GDP and population in the model, include GDP/population (GDP per capita) instead.

4. Increase the Sample Size

Increasing the sample size improves the precision of an estimator and reduces the adverse effects of multicollinearity. Usually adding data though is not feasible.

Example

How would perfect multicollinearity arise in our previous election example?

We've already seen one case:

$$\text{vote share}_i = \beta_0 + \beta_1 \text{ spending}_i + \beta_2 \text{ incumbency}_i + \beta_3 \text{ male}_i + \beta_4 \text{ Liberal}_i + \beta_5 \text{ Conservative}_i + \beta_6 \text{ BQ}_i + \beta_7 \text{ NDP}_i + \varepsilon_i$$

What else?

Campaign spending is really the sum of five separate expenditures:

- Advertising
- Election surveys
- Office expenses
- Salaries
- Other

CHAPTER 8: MULTICOLLINEARITY

What if a researcher were interested in the individual effect of each of these expenditures?

$$\text{vote share}_i = \beta_0 + \beta_1 \text{spending}_i + \beta_2 \text{ advertising} + \beta_3 \text{ surveys} + \beta_4 \text{ office} + \beta_5 \text{ salaries} + \beta_6 \text{ other} + \beta_7 \text{ incumbency}_i + \beta_8 \text{ male}_i + \beta_9 \text{ Liberal}_i + \beta_{10} \text{ Conservative}_i + \beta_{11} \text{ BQ}_i + \beta_{12} \text{ NDP}_i + \varepsilon_i$$

Even if you correct the model there still may be an imperfect multicollinearity between the components of campaign expenditures.