Multicollinearity

Class 11

Wonmun Shin

(wonmun.shin@sejong.ac.kr)

Department of Economics, Sejong University

* This lecture note is written based on Professor Chang Sik Kim's lecture notes.

Perfect Multicollinearity

Perfect Multicollinearity

- **Perfect multicollinearity**: Existence of exact linear relationship(s) among independent variables
  - For the K-variable regression involving explanatory variables X_{1i}, X_{2i}, $\cdots$, X_{Ki} (where $X_{1i} = 1$ for all i to allow for the intercept term), an exact linear relationship is said to exist if the following condition is satisfied:
  $$\lambda_1 X_{1i} + \lambda_2 X_{2i} + \cdots + \lambda_K X_{Ki} = 0$$
  where $\lambda_1, \lambda_2, \cdots, \lambda_K$ are constants such that not all of them are zero simultaneously.
  - Assume that $\lambda_2 \neq 0$, then
  $$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots - \frac{\lambda_K}{\lambda_2} X_{Ki}$$

Perfect Multicollinearity [cont'd]

- Consider the three variable model: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + e_i$
- Suppose $X_{2i} = 3X_{3i}$, which means there is exact linear relationship between X_{2i} and X_{3i}.
- Then,

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + e_i \\ &= \beta_1 + \beta_2 (3X_{3i}) + \beta_3 X_{3i} + e_i \\ &= \beta_1 + (3\beta_2 + \beta_3) X_{3i} + e_i \end{aligned}$$

- Now, we cannot estimate either β_2 and β_3 separately!!
  - We can estimate only the linear combination of two coefficients, *i.e.* $3\widehat{\beta_2 + \beta}_3$.
  - There is no unique value for $\hat{\beta}_2$ (or $\hat{\beta}_3$).
  - This problem is usually called as **Identification Problem**.

- Recall the OLS estimator in the three-variable regression model:

$$\hat{\beta}_2 = \frac{\sum x_{2i} y_i \sum x_{3i}^2 - \sum x_{3i} y_i \sum x_{2i} x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i} x_{3i}\right)^2}$$

  - Suppose that $X_{2i} = 3X_{3i} \implies$ indeterminate expression

$$\hat{\beta}_2 = \frac{3 \sum x_{3i} y_i \sum x_{3i}^2 - 3 \sum x_{3i} y_i \sum x_{3i}^2}{3^2 \sum x_{3i}^2 \sum x_{3i}^2 - \left(3 \sum x_{3i}^2\right)^2} = \frac{0}{0}$$

- Why do we obtain this result?
  - Meaning of $\hat{\beta}_2$: The rate of change in the average value of Y as X_2 changes by a unit, holding X_3 constant
  - But if X_2 and X_3 are perfectly collinear, there is no way X_3 can be kept constant.
  - As X_2 changes, so does X_3 by 3: There is no way of disentangling the separate influences of X_2 and X_3 from the given sample.

Near Multicollinearity

Near Multicollinearity

- Originally, multicollinearity meant the existence of a perfect linear relationship among regressors.
- Today, however, the term multicollinearity is used in a broader sense.
  - Includes the case of perfect multicollinearity as well as the case where the X variables are intercorrelated but not perfectly.

$$\lambda_1 X_{1i} + \lambda_2 X_{2i} + \cdots + \lambda_K X_{Ki} + v_i = 0$$

    where v_i is a stochastic error term.
  - Assume that $\lambda_2 \neq 0$, then

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots - \frac{\lambda_K}{\lambda_2} X_{Ki} - \frac{1}{\lambda_2} v_i$$

    which shows that X_2 is not an exact linear combination of other X's because it is also determined by the stochastic error term v_i.

Near Multicollinearity [cont'd]

X_{2i}	X_{3i}	X_{3i}^*
10	50	52
15	75	75
18	90	97
24	120	129
30	150	152

- $X_{3i} = 5X_{2i}$: There is perfect multicollinearity between X_2 and X_3. ($r_{23} = 1$)
- X_{3i}^* was created from X_3 by simply adding the random numbers to it .
  - Now there is no longer perfect multicollinearity between X_{2i} and X_{3i}^*.
  - However, the two variables are highly correlated! ($r_{23^*} = 0.9959$)
  - In this case, there is **Near (Perfect) Multicollinearity** (or **High but Imperfect Multicollinearity**).

Near Multicollinearity [cont'd]

- Practical example:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + e_i$$

  - Y_i: consumption expenditure
  - X_{2i}: income
  - X_{3i}: wealth
  - X_{3i} may have close positive relationship with X_{2i}!

Consequences of Multicollinearity

Consequences of Multicollinearity

1. Although BLUE, the OLS estimators may have ***large variances***.

   - (Near) multicollinearity does not violate the classical assumptions.
   - Gauss-Markov Theorem $\Rightarrow$ OLS estimators are BLUE!
   - However, they have large variances, making precise estimation difficult.
   - For the three-variable regression model, we can obtain

$$Var\left(\hat{\beta}_2\right) = \frac{\sigma^2}{\sum x_{2i}^2\left(1 - r_{23}^2\right)}$$

     - As r_{23} tends toward 1 or -1 (*i.e.* as collinearity increases), the variance of $\hat{\beta}_2$ increases.
     - Extremely, when $r_{23} = 1$ (or -1), $Var\left(\hat{\beta}_2\right)$ is infinite ($\rightarrow$ identification problem).

2. **Insignificant** t ratios
    - Because of large variance (**Consequence 1)**, the t ratio of one or more coefficients tends to be statistically insignificant.
    - Recall, to test $H_0 : \beta_2 = 0$, we use the t ratio.

$$t = \frac{\hat{\beta}_2}{s.e.\left(\hat{\beta}_2\right)}$$

    - As we have seen, in cases of high collinearity, the estimated standard errors increase dramatically.
    - Therefore, t value becomes smaller $\longrightarrow$ One will increasingly accept the null hypothesis!
    - Probability of Type II error increases (or low power of test).
        - cf. Type II error: Error of not rejecting H_0 when H_0 is false.
3. **Wider** confidence interval

Consequences of Multicollinearity [cont'd]

4. **Significant** F ratio but **few significant** t ratios
   - Although the t ratio of one or more coefficients is statistically insignificant, **R^2 (overall measure of goodness of fit) can be very high.**
   - Consider the K-variable linear regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + e_i$$

     - F-test of overall significance of coefficients: $H_0 : \beta_2 = \cdots = \beta_K = 0$

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - K}{K - 1} \sim F\left(K - 1, n - K\right)$$

     - F can be very large even if individual t ratio is insignificant!
   - Under multicollinearity, explanatory variables can be jointly significant even if each of them is individually insignificant.
     - Remember the distinction between a joint test and an individual test.
5. OLS estimators and their standard errors may be **sensitive** to small changes in the data.

Detection and Remedies of Multicollinearity

Detection of Multicollinearity

- How do we know that multicollinearity is present in any given situation?
  - Multicollinearity is essentially a sample phenomenon (arising out of non-experimental data collected in most social sciences).
  - Therefore, we do not have one unique method of detecting it or measuring its strength.

1. High R^2 (or significant F ratio) but few significant t-ratios.
   - If R^2 is high (usually, in excess of 0.8) and individual t test show that none or very few coefficients are significant, then there may be multicollinearity problem.
   - **Problem:** Although this diagnostic is sensible, its disadvantage is that it is too strong.
     - Note that multicollinearity is considered as "harmful" only when all of the influences of regressors on Y cannot be disentangled.

Detection of Multicollinearity [cont'd]

2. High pairwise correlation(s) among explanatory variables
    - If the pairwise correlation coefficient between two regressors is high (in excess of 0.8, in absolute value), then multicollinearity is a serious problem.
    - **Problem:** Multicollinearity can exist even though the pairwise correlations are comparatively low (less than 0.5, in absolute value).
        - High pairwise correlations are a sufficient condition but not a necessary condition for the existence of multicollinearity.

3. Auxiliary regressions
    - One way of finding out which X variable is related to other X variables is to regress each X_{ki} on the remaining X variables.
    - Each one of these regression is called an **auxiliary regression** (auxiliary to the main regression of Y on X's).

Remedies

1. **Do nothing**
   - Multicollinearity is essentially a data problem, and sometimes we have no choice over data available.
   - Also, it is not that all coefficients in a regression model are insignificant.
     - Moreover, significant F ratio means the model is overall significant.
     - Depending on the objective of analysis, multicollinearity is not that problematic.

2. **A Priori information (about parameters)**

- **Example 1:** Suppose we know that $\beta_3 = 0.1\beta_2$ (given a priori information).

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + e_i \\ &= \beta_1 + \beta_2 \left(X_{2i} + 0.1 X_{3i} \right) + e_i \end{aligned}$$

  - Once we obtain $\hat{\beta}_2$, we can estimate $\hat{\beta}_3$ from the relationship between β_2 and β_3.

- **Example 2:** Cobb-Douglas production function

$$\begin{aligned} Y_i &= \beta_1 L_i^{\beta_2} K_i^{\beta_3} e^{e_i} \\ \Longrightarrow \quad \ln Y_i &= \ln \beta_1 + \beta_2 \ln L_i + \beta_3 \ln K_i + e_i \end{aligned}$$

  - Suppose we know CRTS: $\beta_2 + \beta_3 = 1$

$$\begin{aligned} \ln Y_i &= \ln \beta_1 + \beta_2 \ln L_i + (1 - \beta_2) \ln K_i + e_i \\ \Longrightarrow \quad \ln \left(Y_i / K_i \right) &= \ln \beta_1 + \beta_2 \ln \left(L_i / K_i \right) + e_i \end{aligned}$$

Remedies [cont'd]

3. **Dropping a variable(s)**

- The simplest remedy: dropping one of the collinear variables
- But, we may be committing a **specification error** of omitting relevant variables ($\rightarrow$ ***biased*** and ***inconsistent*** estimator)

4. **Transformation of variables: First difference**

- One reason for high multicollinearity : Variables tent to move in the same direction over time (eg. income and wealth)
- Although the variables may be highly correlated, there is no a priori reason to believe their first differences will also be highly correlated.
- But, *it might introduce autocorrelation of errors*.
- Example:

$$\underbrace{Y_t - Y_{t-1}}_{\triangle Y_t} = \beta_2 \underbrace{(X_{2t} - X_{2t-1})}_{\triangle X_{2t}} + \beta_3 \underbrace{(X_{3t} - X_{3t-1})}_{\triangle X_{3t}} + \underbrace{(e_t - e_{t-1})}_{\nu_t}$$

5. **Transformation of variables: Ratio transformation**

- Example : Y is consumption, X_2 is GDP, X_3 is population
  - GDP and population grow over time. → They are likely to be correlated.
  - One solution is to express the model on a per capita basis:

$$\frac{Y_t}{X_{3t}} = \beta_1 \left(\frac{1}{X_{3t}} \right) + \beta_2 \left(\frac{X_{2t}}{X_{3t}} \right) + \beta_3 + \left(\frac{e_t}{X_{3t}} \right)$$

- But, *it might introduce heteroskedastic errors*.

6. **Additional or new data**

- Again, multicollinearity is a sample feature.
- Simply increasing the size of the sample may attenuate the collinearity problem.

$$n \Uparrow \;\rightarrow\; Var\left(\hat{\beta}_2\right) = \frac{\sigma^2}{\sum x_{2i}^2 \left(1 - r_{23}^2\right)} \Downarrow$$