Lecture 14: Multicollinearity

1. Multicollinearity

The sixth classical assumption is that no independent variable is a perfect linear combination of the other variables. This situation is called *perfect multicollinearity*. The correlation coefficient $r_{12}$ between two perfectly correlated variables will be one or minus one.

When an explanatory variable is not a perfect linear combination of other variables, but its nearly so, this is called *near multicollinearity* or *imperfect multicollinearity*. When two variables are very strongly correlated, the correlation coefficient will be near one in absolute value.

In the case of perfect multicollinearity, it is theoretically impossible to distinguish the effects of the collinear independent variables on the independent variable. For example, suppose that we have two variables $X_1$ and $X_2$ that are identically equal (i.e, $X_1 = X_2$). Then, although $Y$ might vary with both $X_1$ and $X_2$, we cannot possibly attribute the correlation to either $X_1$ or $X_2$ separately. Each observation with a large value of $X_1$ also has a large value of $X_2$. In order to separately distinguish the effects of $X_1$ and $X_2$, it is necessary to have enough independent variation in these two variables so that the sample contains some observations with, say, high values of $X_1$ and low values of $X_2$. Essentially, we cannot “hold $X_2$ constant” and look at the effect $X_1$ on $Y$ because $X_2$ is always moving exactly in the same way as $X_1$.

In the case of imperfect multicollinearity, it is theoretically possible to separately distinguish the effects of two $X_1$ and $X_2$, but doing so will require a larger sample size than if the two variables were not as strongly correlated. A larger sample size will make it more likely that our sample contains observations where $X_1$ and $X_2$ are different enough to separate their effects on $Y$.

Perfect multicollinearity commonly arises in cases where categorical dummy variables are used in a regression, but when one category is not excluded. For example, both MALE$_i$ and FEMALE$_i$ are included in the regression. In such cases, the sum of the variables is always equal to one, indicating that any one dummy variable is a perfect linear combination of the remaining variables.

Imperfect multicollinearity commonly arises when two variables are included that measure nearly the same thing. This is difficult to detect, but if you suspect that there might be imperfect multicollinearity, the sample correlation coefficient can be a useful indicator. For two variables $X_1$ and $X_2$, this is defined as

$$r_{12} = \frac{\sum_{i=1}^{n} (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{(n-1)s_1 s_2}$$

where $s_1$ is the sample standard deviation of $X_1$ defined as

$$s_1 = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_{1i} - \bar{X}_1)^2}.$$ 

and $s_2$ is defined analogously. For two variables $X_j$ and $X_k$, if $r_{jk}$ is close to 1 or -1, then the variables are highly (positively or negatively) correlated, which makes it difficult to separately distinguish the effects of $X_j$ and $X_k$ on $Y$. 


1.1. Consequences of Multicollinearity

The consequences of multicollinearity are:

1. Estimates will remain unbiased.
2. The standard errors of the estimated coefficients increases.
3. The \(t\)-statistics for the estimated coefficients will fall.
4. Estimates become more sensitive to model specification issues.
5. The overall fit of the equation and the coefficients of the other variables will be mostly unaffected.

1.2. Examples of Perfect Multicollinearity

**Example 1.** In our education expenditure example from before, suppose that in addition to total population \( (\text{POP}_i) \) we control for subsets of the population: children under age 18 \((\text{CHILDREN}_i)\), adults aged 18 to 65 \((\text{ADULTS}_i)\), and senior citizens, age 65 and older \((\text{SENIORS}_i)\). These variables are perfectly collinear since

\[
\text{POP}_i = \text{CHILDREN}_i + \text{ADULTS}_i + \text{SENIORS}_i.
\]

Similarly,

\[
\text{CHILDREN}_i = \text{POP}_i - \text{ADULTS}_i - \text{SENIORS}_i
\]

and so on.

**Example 2.** Suppose we regress housing expenditure on

1. wage income,
2. investment income,
3. other income,
4. total income.

Since total income equals the sum of income in each category (i.e., it is a perfect linear combination of the other variables), this regression would suffer from perfect multicollinearity.

In addition to the pure additive example above, perfect multicollinearity can arise in more complex situations. Two variables \( A \) and \( B \) are perfect linear combinations of each other if we can write

\[
B = a_0 + a_1 \cdot A
\]

for any numbers \( a_0 \) and \( a_1 \). So, if \( B = 2 - \frac{1}{3} A \), then \( A \) and \( B \) are collinear.

**Example 3.** Suppose we regress housing expenditure, \( Y_i \), on both annual income in dollars, \( X_{1i} \), and annual income in thousands of dollars \( X_{2i} \). This is not permitted because the variables \( X_{1i} \) and \( X_{2i} \) are perfect linear combinations of each other: \( X_{1i} = 1000 \cdot X_{2i} \).
1.3. Multicollinearity and Indicator Variables

In the case of indicator variables, we have to be careful to exclude at least one categorical indicator variable. For instance, we cannot include both indicators MALE\textsubscript{i} and FEMALE\textsubscript{i} or USED\textsubscript{i} and NEW\textsubscript{i} because

\[ \text{USED}_i = 1 - \text{NEW}_i. \]

For indicator variables with multiple categories, let’s say we have data on the sales prices of new American cars, PRICE\textsubscript{i}, produced by the three major manufacturers, with indicators named CHRYSLER\textsubscript{i}, FORD\textsubscript{i}, and GM\textsubscript{i}. Since each car model in our dataset must be manufactured by either Chrysler, Ford, or General Motors, and since each car is manufactured by exactly one manufacturer, these indicator variables must sum to one:

\[ \text{CHRYSLER}_i + \text{FORD}_i + \text{GM}_i = 1. \]

Given this relationship, we know that these variables are multicollinear and we must omit one of them from the regression. In the case of categorical dummy variables like these, the omitted variable is the baseline category, and the coefficients on the included dummies are to be interpreted relative to the baseline.

So we cannot regress PRICE\textsubscript{i} on all three indicators, but we can run the following regression and treat CHRYSLER\textsubscript{i} as the baseline:

\[ \text{PRICE}_i = \beta_0 + \beta_1 \text{FORD}_i + \beta_2 \text{GM}_i + \epsilon_i. \]

In this regression, the included variables are no longer multicollinear. However, the coefficients \( \beta_1 \) and \( \beta_2 \) have to be interpreted relative to the “excluded” group, Chrysler. So, \( \beta_1 \) represents the average price difference between cars produced by Ford and Chrysler. Similarly, \( \beta_2 \) represents the average price difference between cars produced by General Motors and Chrysler.

1.4. Examples of Near Multicollinearity

**Example 4.** Suppose we are interested in estimating the demand for gasoline using state-level data on petroleum consumption (PCON\textsubscript{i}), the gasoline tax rate in each state (TAX\textsubscript{i}), the number of miles of urban highways in each state (UHM\textsubscript{i}), and the number of registered automobiles in the state (REG\textsubscript{i}).

If we regress PCON\textsubscript{i} on TAX\textsubscript{i}, UHM\textsubscript{i}, and REG\textsubscript{i}, then we might find that the coefficient on either UHM\textsubscript{i} or REG\textsubscript{i} is insignificant or has an unexpected sign. These two variables are both essentially measuring of the size of the state, and are highly correlated.

1.5. Remedies for Multicollinearity

1. Accept the decreased precision and do nothing.
2. Drop one of the offending variables.
3. Collect more data.