1. Multicollinearity in Econometrics

Reviewing the initial assumption

- CLRM (Classical Linear Regression Model) assumption: Independent variables do not have exact linear relationship
  - If they do have it then there is a presence of multicollinearity, a phenomenon that independent variables in a model are correlated and their relationships can be expressed in form of a function
Example

Perfect Multicollinearity:
- $X_2 \quad X_3 \quad X_4$
- 10 50 52
- 15 75 75
- 18 90 97
- 24 120 129
- $X_2$ and $X_3$ have an exact linear relationship, i.e., $X_3 = 5X_2$

Example (continued)

Suppose that we are to estimate a consumption function. $Y = \text{consumption}$, $X_2 = \text{income}$ and $X_3 = \text{assets}$

$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3$

$X_3 = 5X_2$

$Y = \beta_1 + \beta_2 X_2 + \beta_3 5X_2$

$Y = \beta_1 + (\beta_2 + 5\beta_3)X_3$
Example (continued)

- We can estimate $(\beta_2 + 5\beta_3)$ but cannot estimate partial regression coefficients
  - There cannot be a unique solution to each regression coefficient (recall how to compute regression coefficient).
  - Thus regression coefficients cannot be determined
    - The standard errors of regression coefficients are extremely large

Multicollinearity

- Perfect multicollinearity rarely occurs in reality
  - Unless we fall into a dummy trap, which will be examined later

- Near multicollinearity often occurs in reality (when independent variables are highly correlated):
  - In this case we can estimate regression coefficients
    - However, the standard errors are very high and hence the estimated regression coefficients are not accurate, the statistics are statistically insignificant and the null hypothesis is very likely to be rejected
Multicollinearity

- A Case Study

2. Sources of Multicollinearity
Sources of Multicollinearity

Methods of collecting data

- Values of independent variables are correlated in the sample but are not correlated in the total population
- Example: Persons who have higher income are likely to possess more assets. This can be true for the sample but can be wrong for the total population
  - In the total population there will be observations of those who have high income but do not have a lot of assets and vice versa.

Form of function for the model:

- Example: multicollinearity arises when regressing independent variables in squared form (function form), especially when the independent variables start with small values
- Independent macro variables are observed in time series
  - Example: National import is dependent on its GDP and CPI (these indices are collected from time series data). Explain multicollinearity at macro level?
3. Consequences of Multicollinearity

Consequences

- Perfect multicollinearity
  - We cannot estimate the model
  - Computer Applications will give the following warnings
    - “Matrix singular”: an abnormal matrix that the application cannot perform the estimation of regression coefficients
    - “Exact collinearity encountered”: A case of perfect (exact) multicollinearity
Consequences

Consequences of near multicollinearity
(1) OLS estimators are still BLUE
  • Unbiased estimators: Mean value of repeated sample estimators are approximate to those of total population
  • Variance of estimators is still minimum but it is not necessarily small relative to the estimator value

Practical Consequences

(2) Standard errors are high
  ■ Confidence interval is large and t-statistics are less insignificant
  ■ Estimators are less accurate
  ■ We are likely not to have a reason to reject the null hypothesis, which may not be true
Practical Consequences

(3) Very high value of $R^2$ even though t-statistic is less significant
- Why high determinant coefficient?
  - Not many significant variations between independent variables because they really are correlated
- Easy to reject the null hypothesis of f-statistic and conclude that the model is not valid

Practical Consequences

(4) Estimators and standard errors are sensitive to changing data
- A small change in sample data makes the estimators alter drastically
- Because estimators contain strong relationships among independent variables
**Example**

- Examine results of estimating the consumption model:
  - \[ Y = 24.77 + 0.94X_2 - 0.04X_3 \]
  - \[ t \ (3.67) \ (1.14) \ (-0.53) \]
  - \[ R^2=0.96, \ F = 92.40 \]
  - \( X_2 \) : Income
  - \( X_3 \) : Assets
  - Very high \( R^2 \) explaining 96% of variation of the consumption function

**Example**

- No independent variables are significant (t-statistics are too low).
- One variable with incorrect sign.
- Very high f-statistic leading to rejection of null hypothesis and conclusion of valid model
  - Income and assets are strongly correlated and therefore impossible to accurately estimate marginal effect of income and assets on consumption
Example

Make a regression of $X_3$ on $X_2$

\[ X_3 = 7.54 + 10.19X_2 \]

\[ (0.26) \quad (62.04) \quad R^2 = .99 \]

We see a nearly perfect multicollinearity between $X_2$ and $X_3$

Regression of Consumption on Income:

\[ Y = 24.45 + 0.51X_2 \]

\[ (3.81) \quad (14.24) \quad R^2 = 0.96 \]

Example

Variable Income in the above model is statistically significant but is insignificant in the previous model

Similarly, regression of Income $Y$ on Assets:

\[ Y = 24.41 + 0.05X_3 \]

\[ (3.55) \quad (13.29) \quad R^2 = 0.96 \]

Variable Assets now becomes statistically significant but it is insignificant in the previous model
4. Identifying Multicollinearity

Identifying Multicollinearity

(1) Very high $R^2$ and low t-statistic
(2) Strong linear relationship between independent variables

- Formulate a matrix of pairwise correlations and examine to identify the strength of the pairwise correlations between the independent variables
- Examine the economic importance of variables that are very likely to be correlated
Identifying Multicollinearity

(3) Make secondary regression
- Make regression of one independent on the rest of independent variables and observe the R² values of the secondary regressions
- Compute F-statistic
  - F = [R²/(k-1)] /[(1-R²)/(n-k)]
  - k independent variable in secondary regression
  - If F > F* then we can conclude that R² is statistically not zero, which means presence of multicollinearity in the model

Identifying Multicollinearity

(4) Variance inflation factor-VIF)
- VIF = 1/(1-rij²)
- rij² is correlation coefficient between two independent variables in the model
- As rij goes up VIF increases and so does multicollinearity
  - Rule of thumb: VIF >= 10 indication of multicollinearity between two independent variables in the model
5. Solutions to Multicollinearity

Rules of Thumb: Benign Neglect

- Ignore multicollinearity if $t > 2$
- Ignore it if $R^2$ of the model higher than $R^2$ of secondary regression
- Ignore it if the purpose of the model is to forecast rather than to test hypotheses
Solutions when multicollinearity is of particular concern

- Eliminating independent variables
  - Example: Drop Variable Assets from the consumption model
  - Assumption that there is no relationship between the dependent variable and the independent variable dropped from the model
    - If theory assures a relationship to the variable to be dropped then this leads to elimination of a significant variable and we make a specification error

Solutions

- Using extraneous information or new data
  - Try a new sample data or increase the sample size
    - If a large sample size still has multicollinearity present then it is still worth for larger sample size makes variance smaller and estimators more accurate than in case of smaller sample size
Solutions

Reformulating the model
- Econometric models have a variety of functions
- Reformulating a model means restructuring the model

Using priori information
- Using the results of previous econometric models with little or absence of multicollinearity
- Example: We may know the marginal effect of assets on consumption equal only 1/10 of that of income on consumption.

Solutions

Example: $\beta_3 = 0.10 \beta_2$
- Run the model with priori information
- $Y = \beta_1 + \beta_2 X_2 + 0.10 \beta_2 X_3 + e$
  - $Y = \beta_1 + \beta_2 X$ where $X = X_2 + 0.1X_3$
- When $\beta_2$ has been estimated, derive $\beta_3$ from the above priori relations
Solutions

- Use variables as changes from one time period to another (first differences)
  - Differences may reduce the severity of multicollinearity
  - Back to example of consumption model
    - Income and assets are highly correlated and multicollinearity is unavoidable

- We want to estimate
  - $Y_t = \beta_1 + \beta_2X_{2t} + \beta_3X_{3t} + \epsilon_t$
  - For $t-1$
    - $Y_{t-1} = \beta_1 + \beta_2X_{2t-1} + \beta_3X_{3t-1} + \epsilon_{t-1}$
  - Make differences of variables over time period
    - $Y_t - Y_{t-1} = \beta_2(X_{2t} - X_{2t-1}) + \beta_3(X_{3t} - X_{3t-1}) + \nu_t$
Solutions

- This may solve the problem of multicollinearity because multicollinearity arises from the independent variables themselves but not from their differences.
- However, there can be a violation of standard assumptions about random errors.

Solutions

- Combining cross-section and time series data
  - Example: Study of demand for cars when only time series data available
    - $\ln Y = \beta_1 + \beta_2 \ln Price + \beta_3 \ln Income + e$
      - $Y$ number of cars sold
      - Price and income are usually strongly correlated over time and hence there is surely a presence of multicollinearity when using the time series data.
Suppose cross-section data is available
   - We can estimate income elasticity using cross-section data while price elasticity must be estimated from the time series data

Estimating the regression model over time
   - $Y = \beta_1 + \beta_2 \ln P + e$
   - Then $Y = \ln Y - \beta_3 \ln \text{Income}$
   - $Y$ denotes the number of cars sold when the impact of income is eliminated
   - With the known $\beta_3$ we can estimate the price elasticity of demand for car with absence of multicollinearity
   - However, we have to assume that the elasticity computed from time series data and cross-section data are consistent