Multiple Regression

Class 8

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.

Introduction to Multiple Regression

General Model

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

- There may be more than one explanatory variable that may influence the dependent variable.
- Examples
  - Y: Consumption / X: Income, Wealth, Family size, Age, $\cdots$
  - Y: Aggregate money demand / X: aggregate income, interest rate, $\cdots$
  - Y: Demand for a commodity / X: own price, other commodity prices, income $\cdots$

Multiple Linear Regression Model

Meaning of Coefficient

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

- β_2: the change in the mean value of Y per unit change in X_2, holding the value of $X_3, \cdots, X_K$ constant.

  $\Longrightarrow$ Regression coefficient on the independent variable: **partial effect** of the independent variable on the mean of the dependent variable when other independent variables included in the model are held constant.

$$\beta_2 = \frac{\partial Y}{\partial X_2}, \cdots, \beta_K = \frac{\partial Y}{\partial X_K}$$

Classical Assumptions

- **(B1)** $X_{2i}, \cdots, X_{Ki}$: non-random
- **(B2)** $E\left(e_i\right) = 0$
- **(B3)** $Var\left(e_i\right) = \sigma^2$ *Homoskedasticity*
- **(B4)** $Cov\left(e_i,\ e_j\right) = 0$ for $i \neq j$ *No Autocorrelation*
- **(B5)** No exact linear relationship in regressors *No (Perfect) Multicollinearity*

Explanation for (B5)

- 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,

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

- Now, we cannot estimate either β_2 and β_3 separately!!
  - We can estimate only the linear combination of two coefficients, *i.e.* $3\widehat{\beta_2 + \beta_3}$.
  - This problem is usually called as **Identification Problem**.

OLS Estimation in Multiple Regression Model

Three-Variable Regression Model

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

- To find the OLS estimators, let's first write the (stochastic) SRF corresponding to the above equation:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{e}_i$$

  where $\hat{e}_i$ is the residual term (the sample counterpart of the error term e_i)

- The OLS procedure consists of choosing $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ so that $\sum \hat{e}_i^2$ (RSS, residual sum of squares) is as small as possible:

$$\min \sum \hat{e}_i^2 = \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} \right)^2$$

Three-Variable Regression Model [cont'd]

- From the FOCs, we can derive the following normal equations:

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_{2i} + \hat{\beta}_3 \sum X_{3i}$$
$$\sum Y_i X_{2i} = \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{2i}X_{3i}$$
$$\sum Y_i X_{3i} = \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{2i}X_{3i} + \hat{\beta}_3 \sum X_{3i}^2$$

- We can obtain the OLS estimators:

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3$$
$$\hat{\beta}_2 = \frac{\sum x_{2i}y_i \sum x_{3i}^2 - \sum x_{3i}y_i \sum x_{2i}x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i}x_{3i}\right)^2}$$
$$\hat{\beta}_3 = \frac{\sum x_{3i}y_i \sum x_{2i}^2 - \sum x_{2i}y_i \sum x_{2i}x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i}x_{3i}\right)^2}$$

Three-Variable Regression Model [cont'd]

- Note: Equations for $\hat{\beta}_2$ and $\hat{\beta}_3$
  - Two equations are symmetrical in nature because one can be obtained from the other by interchanging the roles of X_{2i} and X_{3i}.
  - The denominators of two equations are identical.
- Note: The three-variable case is a natural extension of the two-variable case.
- Also, we can derive the formulae for variances (+ covariances) of the OLS estimators .
- How about the OLS estimator of σ^2?

$$\hat{\sigma^2} = \frac{\sum \hat{e}_i^2}{n-3}$$

General Multiple Regression Model

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

- To find the OLS estimators, let's first write the (stochastic) SRF corresponding to the above equation:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_K X_{Ki} + \hat{e}_i$$

where $\hat{e}_i$ is the residual term (the sample counterpart of the error term e_i)

- The OLS procedure consists of choosing $\hat{\beta}_1$, $\hat{\beta}_2, \cdots$, $\hat{\beta}_K$ so that $\sum \hat{e}_i^2$ (RSS, residual sum or squares) is as small as possible:

$$\min \sum \hat{e}_i^2 = \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_K X_{Ki} \right)^2$$

General Multiple Regression Model [cont'd]

- We will get K first order conditions and should solve the system of equations! $\Longrightarrow$ **Complicated!**
- **No need to memorize** all formulae for OLS in a multivariate regression.
  - They can be obtained by any statistical packages such as Gretl.
  - In a matrix notation, there is a simple formula (which is beyond this course).
    - We can write $Y = X\beta + e$ where Y is $(n \times 1)$ vector, X is $(n \times K)$ matrix, β is $(K \times 1)$ vector, and e is $(n \times 1)$ vector.
    - Minimize $\hat{e}'\hat{e} = \left(Y - X\hat{\beta}\right)'\left(Y - X\hat{\beta}\right) \Rightarrow \hat{\beta} = \left(X'X\right)^{-1} X'Y$
    - Consider the classical assumptions in the matrix form. For example, $Var\left(e\right) = \sigma^2 I_n$
    - We can show that $\hat{\beta} = \beta + \left(X'X\right)^{-1} X'e \Rightarrow Var\left(\hat{\beta}\right) = \sigma^2 \left(X'X\right)^{-1}$

General Multiple Regression Model [cont'd]

- The estimated regression will be

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_K X_{Ki} + \hat{e}_i$$

- OLS estimator of σ^2:

$$\hat{\sigma^2} = \frac{\sum \hat{e}_i^2}{n - K}$$

where K is the number of regressors (including constant) in the multiple regression.

General Multiple Regression Model [cont'd]

- Most properties of the OLS estimators obtained in the simple regression will hold in the same way in multiple regression model.
  - Unbiasedness of OLS estimators
  - **Gauss-Markov Theorem**: Under the assumptions (B2) - (B5), the OLS estimators are BLUE (Best Linear Unbiased estimators).
  - Variance of OLS estimators
    - Same as simple regression case: $\sigma^2 \uparrow \rightarrow Var\left(\hat{\beta}_k\right) \uparrow$ // n, $(X_{ki} - \bar{X}_k) \uparrow \rightarrow Var\left(\hat{\beta}_k\right) \downarrow$
    - The larger the correlation between the regressors, the larger variance of OLS estimators **(Multicollinearity problem)**.