Simple Regression 1

Class 4

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.

Regression Problem: Basic Concepts

Regression Problem

- Y: Dependent (or Explained) variable, Regressand
- X: Independent (or Explanatory) variable, Regressor
- How can we explain the average behavior of the dependent variable using independent variable(s)?
- Examples
  - Y: Consumption / X: Income, Family size, Age, $\cdots$
  - Y: Demand / X: Prices, Income, Preferences, $\cdots$
  - Y: Sales / X: Expenditure of advertisement, number of employees, $\cdots$

Basic Concepts

1. (Conditional) Expectation of Dependent Variable; Population Regression Function (PRF)

$$E\left(Y_i \mid X_i\right) = \beta_1 + \beta_2 X_i$$

- Model specification: simple linear model
  - This is an assumption based on the economic theory (*i.e.* **Economic Model**)
- Unknown parameters β_1, β_2 characterize economic behavior.
- For the econometric analysis with corresponding data, we need to specify an **Econometric Model**.

2. Error Term (or Disturbance Term)

- Decompose the dependent variable Y_i into two parts:
  - Systematic component of Y_i: (conditional) expectation of Y_i, *i.e.* $E(Y_i \mid X_i) = \beta_1 + \beta_2 X_i$
  - Random component of Y_i: difference between Y_i and its expectation

$$\begin{aligned} e_i &= Y_i - E(Y_i \mid X_i) \\ &= Y_i - \beta_1 - \beta_2 X_i \end{aligned}$$

- Note: e_i is a **random variable**!
- Note: $E(e_i \mid X_i) = 0$

$$\begin{aligned} E(e_i \mid X_i) &= E(Y_i \mid X_i) - E[E(Y_i \mid X_i) \mid X_i] \\ &= E(Y_i \mid X_i) - E(Y_i \mid X_i) \\ &= 0 \end{aligned}$$

3. Linear Regression Model

$$Y_i = \beta_1 + \beta_2 X_i + e_i$$

- Econometric Model: simple (linear) regression model
- The dependent variable Y_i is explained by a systematic component $(\beta_1 + \beta_2 X_i)$ and by a random term e_i.
- The error term e_i represents:
  - Omitted variables: the influence of variables that are not explicitly included in the model
  - Measurement errors: the error comes from measurement of data
  - Idiosyncratic individual errors

Basic Concepts [cont'd]

4. Fitted Value of Dependent Variable

- Let $\hat{\beta}_1$, $\hat{\beta}_2$ be estimators of β_1 and β_2.

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

  - $\hat{Y}_i$: Fitted value of Y_i; **Sample Regression Function (SRF)**
  - **Note:** β_1 and β_2 are parameters, *i.e.* *unknown but fixed constant*. However, estimators $\hat{\beta}_1, \hat{\beta}_2$ are *random variables*.

- **Residuals ($\hat{e}_i$)**
  - Totally different from error term e_i
  - e_i is unknown (or unobservable), whereas $\hat{e}_i$ can be calculated by

$$\hat{e}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$$

  - Note: both e_i and $\hat{e}_i$ are random variables.

5. Estimated Regression Model

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{e}_i$$

Meaning of "Linear" Regression

- ***Linear*** regression model indicates models that are *linear in both variables and parameters*.
- Models that are nonlinear in variables but linear in parameters:

$$Y_i = \beta_1 + \beta_2 X_i^2 + e_i$$

$$Y_i = \beta_1 + \beta_2 \frac{1}{X_i} + e_i$$

- Models that are nonlinear in parameters but linear in variables:

$$Y_i = \beta_1 + \beta_2^2 X_i + e_i$$

$$Y_i = \beta_1 + \sqrt{\beta_2} X_i + e_i$$

- Models that are nonlinear in both variables and parameters:

$$Y_i = \beta_1 X_i^{\beta_2} + e_i$$

$$Y_i = \beta_1 + \frac{1}{\beta_2 X_i} + e_i$$

Simple Regression vs. Multiple Regression

- Simple regression model

$$Y_i = \beta_1 + \beta_2 X_i + e_i$$

- Multiple regression model

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

Estimation of Parameters: OLS

Ordinary Least Squares (OLS) Estimation

- Regression model: $Y_i = \beta_1 + \beta_2 X_i + e_i$
- Estimated model: $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{e}_i$
- We want to choose $\hat{\beta}_1$, $\hat{\beta}_2$ so that $\hat{e}_i$ are small. That is, the OLS estimator $\hat{\beta}_1$, $\hat{\beta}_2$ minimizes the following **criterion function**:

$$\sum \hat{e}_i^2 = \sum \left(Y_i - \hat{Y}_i\right)^2 = \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)^2$$

  - $\sum \left(Y_i - \hat{Y}_i\right)^2$ is the sum of squared distance between the actual Y_i and the fitted Y_i

Ordinary Least Squares (OLS) Estimation [cont'd]

- Minimization problem

$$\min_{\hat{\beta}_1, \hat{\beta}_2} \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i \right)^2$$

  - (FOC wrt $\hat{\beta}_1$) $\quad -2 \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i \right) = 0$
  - (FOC wrt $\hat{\beta}_2$) $\quad -2 \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i \right) X_i = 0$

- $\therefore$ We can get the **normal equations**:

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i \tag{1}$$

$$\sum Y_i X_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2 \tag{2}$$

Ordinary Least Squares (OLS) Estimation [cont'd]

- From (1)

$$\begin{aligned}\frac{1}{n}\sum Y_i &= \hat{\beta}_1 + \hat{\beta}_2 \frac{1}{n}\sum X_i \\ \rightarrow \bar{Y} &= \hat{\beta}_1 + \hat{\beta}_2 \bar{X} \\ \rightarrow \hat{\beta}_1 &= \bar{Y} - \hat{\beta}_2 \bar{X}\end{aligned}$$

- From (2)

$$\begin{aligned}\sum Y_i X_i &= \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2 \\ &= (\bar{Y} - \hat{\beta}_2 \bar{X}) \sum X_i + \hat{\beta}_2 \sum X_i^2 \\ &= \bar{Y} \sum X_i - \hat{\beta}_2 \bar{X} \sum X_i + \hat{\beta}_2 \sum X_i^2\end{aligned}$$

Ordinary Least Squares (OLS) Estimation [cont'd]

$$\rightarrow n\sum Y_i X_i - n\cdot\left(\frac{1}{n}\sum Y_i\right)\sum X_i = n\hat{\beta}_2\sum X_i^2 - n\hat{\beta}_2\cdot\left(\frac{1}{n}\sum X_i\right)\sum X_i$$

$$\rightarrow n\sum Y_i X_i - \sum Y_i\sum X_i = \hat{\beta}_2\left[n\sum X_i^2 - \left(\sum X_i\right)^2\right]$$

$$\rightarrow \hat{\beta}_2 = \frac{n\sum Y_i X_i - \sum Y_i \Sigma X_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} = \frac{\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum\left(X_i - \bar{X}\right)^2} = \frac{\sum x_i y_i}{\sum x_i^2}$$

- $x_i = X_i - \bar{X}$, $y_i = Y_i - \bar{Y}$

OLS Estimators: $\hat{\beta}_1$ and $\hat{\beta}_2$

$$\hat{\beta}_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

Properties

1. $\bar{\hat{e}} = \frac{1}{n}\sum \hat{e}_i = 0$
   - In other words, the sum of residuals $(\sum \hat{e}_i)$ is zero.
   - This property comes directly from the FOC with respect to $\hat{\beta}_1$.
2. $\sum \hat{e}_i X_i = 0$
   - This property comes directly from the FOC with respect to $\hat{\beta}_2$.
3. $\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$
   - SRF (obtained by OLS) passes through $(\bar{X}, \bar{Y})$.

Interpretation of Estimation Results

- **Example 1**: TOEIC

$$\hat{Y}_i = 380.48 + 1.64 X_i$$

  - Y_i: RC score of an individual i, X_i: LC score of an individual i
  - If the LC score goes up by 1 point, on the average the RC score goes up by 1.64 points.

- **Example 2**: Purchasing Power Parity (PPP)
  - PPP theory tells that US price $\uparrow \rightarrow$ Demand for Korean goods $\uparrow \rightarrow$ Demand for KRW $\uparrow \rightarrow$ Appreciation of KRW $\rightarrow$ KRW/USD exchange rate $\downarrow$

$$\hat{Y}_i = 6.68 - 4.32 X_i$$

  - Y_i: KRW/USD exchange rate at time i, X_i: (US CPI)/(Korea CPI) at time i
  - Every unit increase in the relative price, on average, the exchange rate declines by 4.32 unit during the observed time period.