Inference in Simple Regression 1

Class 6

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.

Distributions of OLS Estimators

Distribution of OLS Estimator

- Additional assumption: $e_i \sim N\left(0,\, \sigma^2\right)$ ***Normality assumption***

$$\hat{\beta}_2 = \beta_2 + \sum \omega_i e_i = \beta_2 + \underbrace{\omega_1 e_1 + \omega_2 e_2 + \cdots + \omega_n e_n}_{\textit{Linear Combination of } e_i}$$

  - Note: Any linear combination of independent *normal* random variables has a *normal* distribution

$$\rightarrow \hat{\beta}_2 - \beta_2 = \sum \omega_i e_i \sim N\left(0,\, \sigma^2 \sum \omega_i^2\right) \equiv N\left(0,\, \frac{\sigma^2}{\sum x_i^2}\right)$$

$$\therefore \hat{\beta}_2 \sim N\left(\beta_2,\, Var\left(\hat{\beta}_2\right)\right)$$

- Likewise, $\hat{\beta}_1 \sim N\left(\beta_1,\, Var\left(\hat{\beta}_1\right)\right)$

Distribution of OLS Estimator [cont'd]

$$\frac{\hat{\beta}_2 - \beta_2}{\sqrt{Var\left(\hat{\beta}_2\right)}} \sim N\left(0, 1\right), \quad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{Var\left(\hat{\beta}_1\right)}} \sim N\left(0, 1\right)$$

- But we don't know $Var\left(\hat{\beta}_2\right)$ and $Var\left(\hat{\beta}_1\right)$ because we don't know σ^2.
  - Let us use $\widehat{Var\left(\hat{\beta}_2\right)} = \frac{\hat{\sigma^2}}{\Sigma x_i^2}$ and $\widehat{Var\left(\hat{\beta}_1\right)} = \frac{\hat{\sigma^2}}{n}\frac{\Sigma X_i^2}{\Sigma x_i^2}$ where $\hat{\sigma^2} = \frac{\sum \hat{e}_i^2}{n-2}$.
  - Then, $\frac{\hat{\beta}_2 - \beta_2}{\sqrt{\widehat{Var\left(\hat{\beta}_2\right)}}}$ follows a normal? *Unfortunately, Not.*
    - *(optional)* We can show that it follows t distribution, $t\left(n-2\right)$.

Distribution of OLS Estimator [cont'd]

- When n is sufficiently large, we can apply **CLT**!!
  - By **CLT**, it can be shown that $\hat{\beta}_2$ is approximately normal (***without normality assumption***):

$$\frac{\hat{\beta}_2 - \beta_2}{\sqrt{Var\left(\hat{\beta}_2\right)}} \sim N\left(0, 1\right)$$

  - When n is large, we can expect the value of $\widehat{Var\left(\hat{\beta}_2\right)}$ and $Var\left(\hat{\beta}_2\right)$ are very close ($\widehat{Var\left(\hat{\beta}_2\right)} \longrightarrow Var\left(\hat{\beta}_2\right)$ since $\hat{\sigma^2} \longrightarrow \sigma^2$), and therefore we have

$$\frac{\hat{\beta}_2 - \beta_2}{\sqrt{\widehat{Var\left(\hat{\beta}_2\right)}}} \sim N\left(0, 1\right)$$

Hypothesis Testing

Hypothesis Testing

- Many economic problems require some basis for deciding whether a parameter is a specified value or not, or whether it is positive or negative.
- For example, we are interested in finding out whether the slope coefficient β_2 is zero or not.
  - Consider an econometric model with a dependent variable of food expenditure and an independent variable of income.
  - In this model, we are interested in figuring out whether the food expenditure is affected by your income, that is, whether the coefficient β_2 is zero or not.
- Then, the null hypothesis will be $H_0 : \beta_2 = 0$ against the alternative $H_1 : \beta_2 \neq 0$.
- Hypothesis testing procedures in econometrics compare our conjecture about the econometric model to the information contained in a sample of data.

Hypothesis Testing [cont'd]

- We want to test whether $\beta_2 = 0$ or not.
- **[Step 1]** Set the hypothesis
- **[Step 2]** Test statistic
- **[Step 3]** Set the rejection region
- **[Step 4]** Decision

Step 1: Set the Hypothesis

- Specify the null and alternative hypotheses.
- Null hypothesis

$$H_0 : \ \beta_2 = 0$$

- (Two-sided) Alternative hypothesis

$$H_1 : \ \beta_2 \neq 0$$

Step 2: Test Statistic

- **Distribution of estimator** (that is, distribution of $\hat{\beta}_2$)
  - Recall that

$$\frac{\hat{\beta}_2 - \beta_2}{\sqrt{Var\left(\hat{\beta}_2\right)}} \sim N(0,1)$$

  - When n is sufficiently large, we can expect that the values of $\widehat{Var}\left(\hat{\beta}_2\right)$ and $Var\left(\hat{\beta}_2\right)$ are very close, and therefore we have

$$\frac{\hat{\beta}_2 - \beta_2}{\sqrt{\widehat{Var}\left(\hat{\beta}_2\right)}} \sim N(0,1)$$

- Under the null, compute a test statistic: ***t-statistic***

$$t = \frac{\hat{\beta}_2}{\sqrt{\widehat{Var}\left(\hat{\beta}_2\right)}} \sim N(0,1)$$

Step 2: Test Statistic [cont'd]

- Decomposition of t-statistic:

$$t = \overbrace{\frac{\hat{\beta}_2 - \beta_2}{\sqrt{\widehat{Var}\left(\hat{\beta}_2\right)}}}^{(A)} + \overbrace{\frac{\beta_2}{\sqrt{\widehat{Var}\left(\hat{\beta}_2\right)}}}^{(B)}$$

  - (A): this part follows $N(0, 1)$.
  - (B): this is zero under H_0, and this is positive or negative under H_1.
  - $\Rightarrow$ t will have zero mean if H_0 is correct, and t will have positive mean or negative mean if H_1 is correct.
  - $\Rightarrow$ $\hat{\beta}_2$ close to zero implies it is likely that H_0 is correct, and $\hat{\beta}_2$ far from zero implies it is likely that H_1 is correct.
  - $\Rightarrow$ We would **reject** H_0 in favor of H_1 if t is *larger (for positive values) or smaller (for negative values)* than the number, called **critical value**.

- Then, how can we choose the critical value?

Step 3: Set the Rejection Region

- Note that

$$\begin{cases} \text{Critical Value} \uparrow (\textit{in absolute value}) & \Rightarrow \begin{cases} \text{Type I error} \downarrow \\ \text{Type II error} \uparrow \end{cases} \\ \text{Critical Value} \downarrow (\textit{in absolute value}) & \Rightarrow \begin{cases} \text{Type I error} \uparrow \\ \text{Type II error} \downarrow \end{cases} \end{cases}$$

- Fix a significance level (α)
  - Significance level is the probability of Type I error and size of the test.
  - Usually, $\alpha = 0.01, 0.05, 0.10$.
- Find the critical value corresponding to the chosen α.

Step 3: Set the Rejection Region [cont'd]

- How can we obtain the **critical value**?
  - Recall that our test statistic (t-statistic) follows standard normal distribution if H_0 is correct.
  - From the standard normal distribution table (or using software package), find $z_{\frac{\alpha}{2}}$.
    - As we consider two-sided alternative, we should find $z_{\frac{\alpha}{2}}$ but not z_α.
    - For example, if you choose $\alpha = 5\%$, then you should find $z_{0.025}$.
  - $z_{\frac{\alpha}{2}}$ is the critical value!
- Set the **rejection region**: Reject $H_0 : \beta_2 = 0$ in favor of $H_1 : \beta_2 \neq 0$ if

$$t > z_{\frac{\alpha}{2}} \text{ or } t < -z_{\frac{\alpha}{2}}$$

Step 4: Decision

- We would reject H_0 in favor of H_1 if t is larger (for positive values) or smaller (for negative values) than critical value.
  - When we reject $H_0 : \beta_2 = 0$, we say that **the estimate $\hat{\beta}_2$ is (statistically) significant** at the significance level of α.
  - If t is not on the rejection region, we cannot reject H_0. However, it does not mean that H_0 is true (due to Type II error).
    - For this reason, economists use *"not reject H_0"* instead of *"accept H_0"*.
  - Also, obviously, the result of test depends on your choice of the significance level.
    - The higher α, the greater chance the null hypothesis will be rejected.
    - Therefore, the choice of α is important.

Step 4: Decision [cont'd]

- ***p*-value** (probability value): The smallest significance level at which a null hypothesis can be rejected
  - *p*-value is very useful and convenient when you decide whether the estimate is significant or not. <u>(You can calculate it by yourself, but please rely on computer!)</u>
  - You can directly compare the *p*-value of the estimate with α.
  - If *p*-value is smaller than α, the estimate is significant!

Example: Food Expenditure and Income

$$\widehat{\text{food\_exp}} = \underset{(43.410)}{83.4160} + \underset{(2.0933)}{10.2096}\,\text{income}$$

(standard errors in parentheses)

- Using sample, the estimated coefficient is 10.21, *i.e.* $\hat{\beta}_2 = 10.21$.
- **[Step 1]** Set the hypothesis
  - We are interested in whether income affects food expenditure or not, *i.e.* $\beta_2 = 0$ or not.
  - $H_0 : \beta_2 = 0$ v.s. $H_1 : \beta_2 \neq 0$
- **[Step 2]** Test statistic

$$t = \frac{\hat{\beta}_2}{\sqrt{\widehat{Var}\left(\hat{\beta}_2\right)}} = \frac{10.21}{2.09} = 4.88$$

Example: Food Expenditure and Income [cont'd]

$$\underset{}{\widehat{\text{food\_exp}}} = \underset{(43.410)}{83.4160} + \underset{(2.0933)}{10.2096}\,\text{income}$$

(standard errors in parentheses)

- **[Step 3]** Set the rejection region
  - Choose $\alpha = 0.05$.
  - We can find $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$, which is the critical value.
  - Rejection region: $t > 1.96$ or $t < -1.96$
- **[Step 4]** Decision
  - Since $t = 4.88$ is larger than the critical value, 1.96, we reject the null $H_0 : \beta_2 = 0$.
  - In other words, $\hat{\beta}_2 = 10.21$ is significant at 5% significance level.

Example: Food Expenditure and Income [cont'd]

Model 1: OLS, using observations 1–40
Dependent variable: food_exp

	Coefficient	Std. Error	t-ratio	p-value
const	83.4160	43.4102	1.922	0.0622
income	10.2096	2.09326	4.877	0.0000

- When you use statistics package (eg. Gretl, R, Stata, etc.), you can earn the resulting table such as the above one.
  - t-ratio presents t-statistic under the null hypothesis that the coefficient is zero.
  - $p-$value is very useful measure: p-value for $\hat{\beta}_2$ is close to zero, which means we reject the null $H_0 : \beta_2 = 0$ at 5% significance level as well as at extremely low significance level!
- **2-t rule of thumb**
  - If the number of degrees of freedom is more than 20, and if the significance level is set at 5%, then the coefficient is significant if $|t| > 2$ (or if the estimated coefficient value is more than twice as large as its standard error).

One-Sided Test

- Consider the null hypothesis against one-sided alternative as

$$H_0 : \beta_2 = 0 \quad v.s. \quad H_1 : \beta_2 > 0$$

  - The one-sided test is conducted when we strongly believe that the effect of independent variable on dependent variable is one direction.

- Same process as the two-sided test, except obtaining the critical value and setting the rejection region.
  - If you choose $\alpha = 0.05$, then you should find $z_\alpha = z_{0.05}$, but not $z_{\frac{\alpha}{2}}$.
  - The rejection region will be: $t > z_\alpha$

Confidence Interval

Confidence Interval

- Estimate $\hat{\beta}_2$ does not give any information about how far it can be from the true parameter β_2.
- Therefore, we need to combine the estimate and its variance to present the reasonable range of estimate by constructing **confidence interval** for β_2.
- Actually, it is closely associated with the hypothesis testing because the hypothesis test also uses the information of the estimate and its variance.
  - We know that

$$\frac{\hat{\beta}_2 - \beta_2}{\sqrt{\widehat{Var\left(\hat{\beta}_2\right)}}} \sim N\left(0, 1\right)$$

Confidence Interval [cont'd]

- For a given α (significance level), there exists $z_{\frac{\alpha}{2}}$ such that

$$
\begin{aligned}
1-\alpha &= P\left(-z_{\frac{\alpha}{2}} < \frac{\hat{\beta}_2-\beta_2}{\sqrt{\widehat{Var}\left(\hat{\beta}_2\right)}} < z_{\frac{\alpha}{2}}\right) \\
&= P\left(\hat{\beta}_2 - z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}\left(\hat{\beta}_2\right)} < \beta_2 < \hat{\beta}_2 + z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}\left(\hat{\beta}_2\right)}\right)
\end{aligned}
$$

- Therefore, $100\left(1-\alpha\right)\%$ confidence interval for β_2 is:

$$
\left[\hat{\beta}_2 \pm z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}\left(\hat{\beta}_2\right)}\right]
$$

Confidence Interval [cont'd]

- Interpretation of confidence interval
  - When the 95% ($\alpha = 0.05$) confidence interval of β_2 is $[A, B]$, it means that 95% of sample realizations of $[A, B]$ would contain the unknown population parameter β_2.
  - That is, if we sample repeatedly and calculate the intervals by the realization of random variables A and B, then 95% of the intervals will contain the true β_2.
- Note: if the $100\,(1-\alpha)\,\%$ confidence interval of β_2 contains 0 (zero), we can say that the estimated coefficient $\hat{\beta}_2$ is not significant at $100\alpha\%$ significance level.