Inferences in Multiple Regression

Class 9

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.

Hypothesis Testing of Individual Coefficient

Distribution of OLS Estimators

- We assume that sample size is sufficiently large.
  - In the case where sample size is small, we need normality assumption, $e_i \sim N\left(0, \sigma^2\right)$.
- When sample size is large, we have

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{Var\left(\hat{\beta}_k\right)}} \sim N\left(0, 1\right) \quad \text{for } k = 1, \cdots, K$$

- And we can expect that the values of $\widehat{Var\left(\hat{\beta}_k\right)}$ and $Var\left(\hat{\beta}_k\right)$ are very close, therefore we have

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{\widehat{Var\left(\hat{\beta}_k\right)}}} \sim N\left(0, 1\right) \quad \text{for } k = 1, \cdots, K$$

Hypothesis Testing of Individual Coefficient

- **[Step 1]** Set the hypothesis
  - Two-sided test: $H_0 : \beta_k = 0$ v.s. $H_1 : \beta_k \neq 0$
  - One-sided test: $H_0 : \beta_k = 0$ v.s. $H_1 : \beta_k < 0$ (or $\beta_k > 0$)
- **[Step 2]** Test statistic

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

- **[Step 3]** Set the rejection region
  - Choose significance level and find the corresponding critical value
  - Set the rejection region: $t > z_{\frac{\alpha}{2}}$ or $t < -z_{\frac{\alpha}{2}}$ (two-sided)
- **[Step 4]** Decision

Goodness of Fit

Recall: Coefficient of Determination, R^2

- Define, as before

$$TSS = \sum \left(Y_i - \bar{Y}\right)^2$$
$$ESS = \sum \left(\hat{Y}_i - \bar{Y}\right)^2$$
$$RSS = \sum \hat{e}_i^2$$

$\Longrightarrow$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

- R^2 is the proportion of the total variation in the dependent variable Y explained by explanatory variables included in the model ***jointly***.

$\bar{R}^2$ (Adjusted R^2)

- Problem of R^2: R^2 increases if more regressors are added in the regression *even if there is no economic justification.*
  - Mathematically, it is a fact that as variables are added, ESS goes up (or RSS goes down) and hence R^2 goes up.
  - Extremely, if you have the same number of regressors as the number of observations, then you have $R^2 = 1$!
- Alternative measure of goodness of fit is **Adjusted R^2 ($\bar{R}^2$or R^2_{adj})**:

$$\bar{R}^2 = 1 - \frac{\frac{RSS}{n-K}}{\frac{TSS}{n-1}}$$

  - $\bar{R}^2$ is not necessarily affected by adding more regressor.
  - In other words, R^2 is always increasing when regressor is added but $\bar{R}^2$ is not always.

$\bar{R}^2$ (Adjusted R^2) [cont'd]

$$\bar{R}^2 = 1 - \left(1 - R^2\right) \frac{n-1}{n-K}$$

- We can show that $\bar{R}^2 < R^2$ for $K > 1$.
  - Let's define $\alpha = \frac{n-1}{n-K} \rightarrow$ If $K > 1$ then $\alpha > 1$.

$$\begin{aligned} \bar{R}^2 - R^2 &= 1 - \left(1 - R^2\right)\alpha - R^2 \\ &= 1 - \alpha + R^2\alpha - R^2 \\ &= (1-\alpha) - (1-\alpha) R^2 \\ &= \underbrace{(1-\alpha)}_{<0} \underbrace{\left(1 - R^2\right)}_{>0} < 0 \end{aligned}$$

  - It implies that as the number of independent variables increases, $\bar{R}^2$ increases less than R^2.

$\bar{R}^2$ (Adjusted R^2) [cont'd]

- $\bar{R}^2$ can be negative, although R^2 is necessarily non-negative.
  - In case $\bar{R}^2$ turns out to be negative, its value is taken as zero.
- IMPORTANT NOTE: R^2 and $\bar{R}^2$ should not be used as a device for the selection of independent variables or a model selection.
- Also, it is crucial to note that in comparing two models on the basis of the coefficient of determination, whether adjusted or not, **the sample size n and the dependent variable must be same!**
  - For example, we cannot compare R^2s of the two models below:

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

Test of Linear Restriction

Restricted Linear Squares

- There are many cases where tests involving more than one parameter are appropriate in the multiple regression.
- Model

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

- Examples of **linear restrictions**

$$\begin{aligned} H_0 :& \beta_2 = \beta_3 = \cdots = \beta_K = 0 \\ H_0 :& \beta_2 = \beta_3 \\ H_0 :& \beta_2 + 2\beta_3 = 1 \\ H_0 :& \beta_3 = \beta_4 = \beta_5 = 0 \end{aligned}$$

Restricted Linear Squares [cont'd]

- An example of restricted regression

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

$$H_0 : \beta_3 = 0$$

  - Estimation **under** restriction:

$$\min \sum \hat{e}_i^2 = \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} \right)^2 \text{ w.r.t. } \hat{\beta}_1, \hat{\beta}_2$$

$$RSS_R = \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} \right)^2 = \text{Restricted } RSS$$

  - Estimation **without** restriction:

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

$$RSS_U = \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} \right)^2 = \text{Unrestricted } RSS$$

Restricted Linear Squares [cont'd]

- $RSS_U \leq RSS_R$ always!
  - This is because the restricted regression model is just a special form of the unrestricted model (that is, when $\beta_3 = 0$).
  - The restricted regression can be obtained by minimizing

    $$\sum \hat{e}_i^2 = \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - 0 \cdot X_{3i} \right)^2$$

    which is the case restricting $\hat{\beta}_3$ to be zero on the unrestricted model.
  - Therefore, if you allow for $\hat{\beta}_3$ to be non-zero, you can only decrease the value of RSS.

Test of Linear Restriction

- We want to test whether the linear restriction of interest is true or not.
  - Note that H_0 (e.g. $H_0 : \beta_3 = 0$) is the linear restriction of interest.

$$\begin{cases} \text{If } H_0 \text{ is true, then } RSS_U \approx RSS_R \\ \text{If } H_0 \text{ is false, then } RSS_U < RSS_R \end{cases}$$

  - Hence, we will consider $RSS_R - RSS_U$ as the basis of the test statistic.
  - We reject H_0 if $RSS_R - RSS_U$ is large!

Test of Linear Restriction [cont'd]

- Review of F distribution

$$F \sim F\left(d_1, d_2\right)$$

  - Random variable F has the F distribution with degrees of freedom $\left(d_1, d_2\right)$
  - The pdf is positively skewed and located over the range of positive numbers.
  - If $V_1 \sim \chi^2\left(d_1\right)$, $V_2 \sim \chi^2\left(d_2\right)$, and V_1 and V_2 are independent, then

$$F = \frac{V_1 / d_1}{V_2 / d_2} \sim F\left(d_1, d_2\right)$$

Test of Linear Restriction [cont'd]

- When sample size is large (or under normality assumption when sample size is small), it can be shown that

$$V_1 = \frac{RSS_R - RSS_U}{\sigma^2} \sim \chi^2(J)$$

$$V_2 = \frac{RSS_U}{\sigma^2} \sim \chi^2(n-K)$$

  - J: the number of restrictions
  - n: sample size
  - K: the number of explanatory variables (including constant) in the unrestricted model

- We can also show that V_1 and V_2 are independent.

- Therefore, we have

$$F = \frac{V_1/J}{V_2/(n-K)} \sim F(J, n-K)$$

Test of Linear Restriction [cont'd]

- **Question:** Why do we need V_2 when our interest is only $RSS_R - RSS_U$?
- **Answer:** We do not know σ^2!

$$F = \frac{V_1 / J}{V_2 / (n-K)} = \frac{\frac{RSS_R - RSS_U}{\sigma^2} / J}{\frac{RSS_U}{\sigma^2} / (n-K)}$$
$$= \frac{(RSS_R - RSS_U) / J}{RSS_U / (n-K)} \sim F(J, n-K)$$

- Alternative form:

$$F = \frac{\left(R_U^2 - R_R^2\right) / J}{\left(1 - R_U^2\right) / (n-K)} \sim F(J, n-K)$$

  - $R_U^2 = 1 - \frac{RSS_U}{TSS}$: Unrestricted R^2
  - $R_R^2 = 1 - \frac{RSS_R}{TSS}$: Restricted R^2
  - Note that $R_U^2 \geq R_R^2$ always!

Test of Linear Restriction [cont'd]

- **[Step 1]** Set the hypothesis
  - H_0 is the linear restriction, and H_1 is the restriction is false.
  - e.g. $H_0 : \beta_3 = 0$ (1 restriction), $H_0 : \beta_2 = \beta_3 = 0$ (2 restrictions, **joint test**)
- **[Step 2]** Test statistic: F statistic

$$F = \frac{(RSS_R - RSS_U) / J}{RSS_U / (n - K)} \sim F(J, n - K)$$

- **[Step 3]** Set the rejection region
  - Choose significance level and find the corresponding critical value *(from F distribution)*
  - Set the rejection region: $F > F_\alpha (J, n - K)$
- **[Step 4]** Decision
  - If we reject H_0, we conclude that the restriction is not valid.

Example

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + e_i \tag{1}$$

1. $H_0 : \beta_2 = \beta_3 = 0 \quad \text{v.s.} \quad H_1 : \beta_2 \neq 0 \text{ or } \beta_3 \neq 0$
   - Under H_0, we have
   $$Y_i = \beta_1 + e_i \tag{2}$$
   - RSS from (1) is RSS_U, and RSS from (2) is RSS_R
   - $J = 2,\ K = 3 \Longrightarrow$ We can compute F statistic.

2. $H_0 : \beta_2 + 2\beta_3 = 1 \quad \text{v.s.} \quad H_1 : \beta_2 + 2\beta_3 \neq 1$
   - Under H_0, $\beta_2 = 1 - 2\beta_3$ so we have
   $$\begin{aligned} Y_i &= \beta_1 + (1 - 2\beta_3) X_{2i} + \beta_3 X_{3i} + e_i \\ \longrightarrow \quad Y_i - X_{2i} &= \beta_1 + \beta_3 (X_{3i} - 2X_{2i}) + e_i \end{aligned} \tag{3}$$
   - RSS from (1) is RSS_U, and RSS from (3) is RSS_R
   - $J = 1,\ K = 3 \Longrightarrow$ We can compute F statistic.
   - Note that we cannot use R_U^2 and R_R^2 in this case because the dependent variables are not same.

Test of Overall Significance

Test of Overall Significance

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

- Consider a joint test of *the relevance of* ***all*** *the included explanatory variables* as

$$\begin{cases} H_0: & \beta_2 = \beta_3 = \cdots = \beta_K = 0 \\ H_1: & \text{At least one of the } \beta_k \text{ is non-zero.} \end{cases}$$

- Then, if the null is true, none of the regressors influence Y, and thus the model is not constructed well at all!

Test of Overall Significance [cont'd]

- Under H_0, the model becomes

$$Y_i = \beta_1 + e_i \quad \text{(Restricted model)}$$

$\Longrightarrow$

$$\begin{aligned} RSS_R &= \sum \left(Y_i - \hat{\beta}_1\right)^2 \\ &= \sum \left(Y_i - \bar{Y}\right)^2 = TSS \end{aligned}$$

- What is RSS from the unrestricted model?

$$\begin{aligned} RSS_U &= \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_K X_{Ki}\right)^2 \\ &= \text{the usual } RSS \end{aligned}$$

- Note that J (# of restrictions) is $K - 1$.

Test of Overall Significance [cont'd]

$$F = \frac{\left(RSS_R - RSS_U\right)/J}{RSS_U/\left(n-K\right)} = \frac{\left(TSS - RSS\right)/\left(K-1\right)}{RSS/\left(n-K\right)}$$
$$= \frac{ESS/\left(K-1\right)}{RSS/\left(n-K\right)} \sim F\left(K-1, n-K\right)$$

- Moreover,

$$F = \frac{ESS/\left(K-1\right)}{RSS/\left(n-K\right)} = \frac{ESS}{RSS} \cdot \frac{n-K}{K-1}$$
$$= \frac{ESS/TSS}{RSS/TSS} \cdot \frac{n-K}{K-1}$$
$$= \frac{R^2}{1-R^2} \cdot \frac{n-K}{K-1} \sim F\left(K-1, n-K\right)$$

- Therefore, if you want to test the overall significance of a model, F statistic can be reduced into a simpler form.
- We conclude the model has overall significance if $F > F_\alpha\left(K-1, n-K\right)$.

Joint vs. Individual Test

Single Coefficient Testing

- As we discussed before, you can test a significance of a single coefficient in the multiple regression using t-test.
- We can get the same result with F-test because F-test statistic is exactly same as the square of t-test statistic.
  - *[Optional]* $F(1, n-K) = [t(n-K)]^2$
- Therefore, the p-values for the two tests are **identical** under two-sided alternative, meaning that the same conclusion will be drawn whichever test is used.
- However, you cannot have the same equivalence with one-sided test for the single coefficient since F-test is not appropriate when the alternative is an inequality.

Joint vs. Individual Test

- Consider the following two different testings:

$$H_0: \ \beta_2 = \beta_3 = 0 \tag{4}$$

$$\begin{cases} H_0: \ \beta_2 = 0 \\ H_0: \ \beta_3 = 0 \end{cases} \tag{5}$$

- Testing with (4) is **"Joint test"**.
  - It involves F-test and allows the correlation between two parameters.
  - It is related to *confidence ellipse*.
- Testing with (5) is two **"Individual test"**.
  - It does not consider the possibility of $\beta_2 = 0$ when we perform the test about $H_0 : \beta_3 = 0$.
  - It is related to *confidence interval*.
- Therefore, it is possible that we can get conflicting results from these two tests.