BS2280 - Econometrics 1

Lecture 6 - Part 2: Multiple Regression Analysis II

Dr. Yichen Zhu

Structure of today's lecture

1. Review: Multiple Regression Model: F-Tests
2. Testing the Joint Significance of A Group of Variables
3. Testing Restrictions

Intended Learning Outcomes

- Testing the joint significance of a group of variables
- Testing restrictions

Multiple Regression Model: F-Tests

F-tests are extremely popular! We will use them to

1. Testing the overall significance or overall fit of a model
2. Testing the joint significance of a group of variables
3. Testing restrictions

Testing the Overall Significance or Overall Fit of a model

1. Testing the overall significance or overall fit of a model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i$$

- We will use F-Tests again to check the overall significance of the model
- There are two ways to write down the hypotheses:
  1. $H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0$; $H_1 :$ *at least one* $\beta \neq 0$
  2. $H_0 : R^2 = 0$; $H_1 : R^2 \neq 0$
- The F-test is identical to before
- Calculate F-statistic and compare with critical F-value
- Reject or do not reject null hypothesis

Testing the Joint Significance of A Group of Variables

- There is a more general F test of goodness of fit.
- This is a test of the joint explanatory power of a group of variables when they are added to a regression model.
- For example,

  Original model specification (Model 1): Y is a simple function of X_2

  $$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \tag{1}$$

  Modified model specification (Model 2): we add X_3 and X_4

  $$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \tag{2}$$

- **Question**: Adding these two variables X_3 and X_4 improves the explanatory power of the model or not

Testing the Joint Significance of A Group of Variables

- There is a more general F test of goodness of fit.
- This is a test of the joint explanatory power of a group of variables when they are added to a regression model.
- For example,

  Original model specification (Model 1): Y is a simple function of X_2

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \tag{1}$$

  Modified model specification (Model 2): we add X_3 and X_4

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \tag{2}$$

- **Question**: Adding these two variables X_3 and X_4 improves the explanatory power of the model or not

Testing the Joint Significance of A Group of Variables

- There is a more general F test of goodness of fit.
- This is a test of the joint explanatory power of a group of variables when they are added to a regression model.
- For example,

  Original model specification (Model 1): Y is a simple function of X_2

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \tag{1}$$

  Modified model specification (Model 2): we add X_3 and X_4

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \tag{2}$$

- **Question**: Adding these two variables X_3 and X_4 improves the explanatory power of the model or not

Testing the Joint Significance of A Group of Variables

- There is a more general F test of goodness of fit.
- This is a test of the joint explanatory power of a group of variables when they are added to a regression model.
- For example,

  Original model specification (Model 1): Y is a simple function of X_2

  $$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \tag{1}$$

  Modified model specification (Model 2): we add X_3 and X_4

  $$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \tag{2}$$

- **Question**: Adding these two variables X_3 and X_4 improves the explanatory power of the model or not

Null and Alternative Hypotheses

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \tag{3}$$

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \tag{4}$$

- Conduct hypothesis testing to determine whether the inclusion of these two variables X_3 and X_4 improves the explanatory power of the model

$$H_0 : \beta_3 = \beta_4 = 0 \tag{5}$$

$$H_1 : \beta_3 \neq 0 \; or \; \beta_4 \neq 0 \; or \; both \; \beta_3 \; and \; \beta_4 \neq 0 \tag{6}$$

- The null hypothesis is that neither X_3 nor X_4 belongs in the model. Adding X_3 and X_4 will not improve the explanatory power of the model.
- The alternative hypothesis is that at least one of them does, perhaps both.

Null and Alternative Hypotheses

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \tag{3}$$

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \tag{4}$$

- Conduct hypothesis testing to determine whether the inclusion of these two variables X_3 and X_4 improves the explanatory power of the model

$$H_0 : \beta_3 = \beta_4 = 0 \tag{5}$$

$$H_1 : \beta_3 \neq 0 \textit{ or } \beta_4 \neq 0 \textit{ or both } \beta_3 \textit{ and } \beta_4 \neq 0 \tag{6}$$

- The null hypothesis is that neither X_3 nor X_4 belongs in the model. Adding X_3 and X_4 will not improve the explanatory power of the model.
- The alternative hypothesis is that at least one of them does, perhaps both.

Null and Alternative Hypotheses

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \tag{3}$$

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \tag{4}$$

- Conduct hypothesis testing to determine whether the inclusion of these two variables X_3 and X_4 improves the explanatory power of the model

$$H_0 : \beta_3 = \beta_4 = 0 \tag{5}$$

$$H_1 : \beta_3 \neq 0 \; or \; \beta_4 \neq 0 \; or \; both \; \beta_3 \; and \; \beta_4 \neq 0 \tag{6}$$

- The null hypothesis is that neither X_3 nor X_4 belongs in the model. Adding X_3 and X_4 will not improve the explanatory power of the model.
- The alternative hypothesis is that at least one of them does, perhaps both.

F Test Statistic

$$F(cost\ in\ dof, dof\ remaining) = \frac{reduction\ in\ RSS / cost\ in\ dof}{RSS\ remaining / dof\ remaining} \tag{7}$$

Note: dof is degree of freedom

$$\begin{array}{cccccc} Y_i & = \beta_1 + \beta_2 & X_i & + & u_i \\ \text{Total variations, variations in } Y & & \text{Explained variations, variations in } X & & \text{Residual variations, variations in } u_i \\ \text{Total Sum of Squares, TSS} & & \text{Explained sum of squares, ESS} & & \text{Residual sum of squares, RSS} \\ TSS = \sum (Y_i - \bar{Y})^2 & & ESS = \sum (\hat{Y}_i - \bar{Y})^2 & & RSS = \sum (Y_i - \hat{Y}_i)^2 = \sum u_i^2 \end{array} \tag{8}$$

- When new X variables are added, RSS cannot rise. In general, it will **fall**. Because we would expect our model can explain more variations, the unexplained variations will get smaller.
- If the new X variables are irrelevant, they will fall only by a random amount, a pure coincidence.
- **Rationale of the test**: when we add new X variables in the model, is the fall in RSS bigger enough that we can reject the claim that it is just a pure coincidence

F Test Statistic

$$F(cost\ in\ dof,\ dof\ remaining) = \frac{reduction\ in\ RSS\,/\,cost\ in\ dof}{RSS\ remaining\,/\,dof\ remaining} \tag{7}$$

Note: dof is degree of freedom

$$\underset{\substack{\text{Total variations, variations in } Y \\ \text{Total Sum of Squares, TSS} \\ TSS = \sum(Y_i - \bar{Y})^2}}{Y_i} = \beta_1 + \beta_2 \underset{\substack{\text{Explained variations, variations in } X \\ \text{Explained sum of squares, ESS} \\ ESS = \sum(\hat{Y}_i - \bar{Y})^2}}{X_i} + \underset{\substack{\text{Residual variations, variations in } u_i \\ \text{Residual sum of squares, RSS} \\ RSS = \sum(Y_i - \hat{Y}_i)^2 = \sum u_i^2}}{u_i} \tag{8}$$

- When new X variables are added, RSS cannot rise. In general, it will **fall**. Because we would expect our model can explain more variations, the unexplained variations will get smaller.
- If the new X variables are irrelevant, they will fall only by a random amount, a pure coincidence.

- **Rationale of the test**: when we add new X variables in the model, is the fall in RSS bigger enough that we can reject the claim that it is just a pure coincidence

F Test Statistic

$$F(\textit{cost in dof}, \textit{dof remaining}) = \frac{\textit{reduction in RSS} / \textit{cost in dof}}{\textit{RSS remaining} / \textit{dof remaining}} \tag{7}$$

Note: dof is degree of freedom

$$\begin{array}{cccc} Y_i & = \beta_1 + \beta_2 & X_i & + & u_i \\ \text{Total variations, variations in } Y & & \text{Explained variations, variations in } X & & \text{Residual variations, variations in } u_i \\ \text{Total Sum of Squares, TSS} & & \text{Explained sum of squares, ESS} & & \text{Residual sum of squares, RSS} \\ TSS = \sum (Y_i - \bar{Y})^2 & & ESS = \sum (\hat{Y}_i - \bar{Y})^2 & & RSS = \sum (Y_i - \hat{Y}_i)^2 = \sum u_i^2 \end{array} \tag{8}$$

- When new X variables are added, RSS cannot rise. In general, it will **fall**. Because we would expect our model can explain more variations, the unexplained variations will get smaller.
- If the new X variables are irrelevant, they will fall only by a random amount, a pure coincidence.

- **Rationale of the test**: when we add new X variables in the model, is the fall in RSS bigger enough that we can reject the claim that it is just a pure coincidence

F Test Statistic

Original model specification (Model 1): $Y_i = \beta_1 + \beta_2 X_{2i} + u_i \quad RSS_1$
Modified model specification (Model 2): $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \quad RSS_2$

$$F(cost\ in\ dof, dof\ remaining) = \frac{reduction\ in\ RSS / cost\ in\ dof}{RSS\ remaining / dof\ remaining} = \frac{(RSS_1 - RSS_2)/cost\ in\ dof}{RSS_2 / dof\ remaining} \tag{9}$$

Note:

- **reduction in** ***RSS***: reduction in *RSS* after the group of new X variables is added, $RSS_1 - RSS_2$
- **cost in dof**: reduction in the number of degrees of freedom after the group of new X variables is added. It is equal to the number of new variables added (here we add X_3 and X_4, so it is 2)
- ***RSS*** **remaining**: *RSS* of the model after the group of new X variables is added, RSS_2
- **dof remaining**: number of degrees of freedom remaining after the group of new X variables is added, $n - k$

F Test Statistic

Original model specification (Model 1): $Y_i = \beta_1 + \beta_2 X_{2i} + u_i \quad RSS_1$
Modified model specification (Model 2): $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \quad RSS_2$

$$F(cost\ in\ dof, dof\ remaining) = \frac{reduction\ in\ RSS/cost\ \ in\ dof}{RSS\ remaining/dof\ remaining} = \frac{(RSS_1 - RSS_2)/cost\ \ in\ dof}{RSS_2/dof\ remaining} \tag{9}$$

Note:

- **reduction in** ***RSS***: reduction in *RSS* after the group of new X variables is added, $RSS_1 - RSS_2$
- **cost in dof**: reduction in the number of degrees of freedom after the group of new X variables is added. It is equal to the number of new variables added (here we add X_3 and X_4, so it is 2)
- *RSS* **remaining**: *RSS* of the model after the group of new X variables is added, RSS_2
- **dof remaining**: number of degrees of freedom remaining after the group of new X variables is added, $n - k$

F Test Statistic

Original model specification (Model 1): $Y_i = \beta_1 + \beta_2 X_{2i} + u_i \quad RSS_1$
Modified model specification (Model 2): $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \quad RSS_2$

$$F(cost\ in\ dof, dof\ remaining) = \frac{reduction\ in\ RSS / cost\ in\ dof}{RSS\ remaining / dof\ remaining} = \frac{(RSS_1 - RSS_2)/cost\ in\ dof}{RSS_2 / dof\ remaining} \tag{9}$$

Note:

- **reduction in** ***RSS***: reduction in *RSS* after the group of new X variables is added, $RSS_1 - RSS_2$
- **cost in dof**: reduction in the number of degrees of freedom after the group of new X variables is added. It is equal to the number of new variables added (here we add X_3 and X_4, so it is 2)
- *RSS* **remaining**: *RSS* of the model after the group of new X variables is added, RSS_2
- **dof remaining**: number of degrees of freedom remaining after the group of new X variables is added, $n - k$

F Test Statistic

Original model specification (Model 1): $Y_i = \beta_1 + \beta_2 X_{2i} + u_i \quad RSS_1$
Modified model specification (Model 2): $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \quad RSS_2$

$$F(cost\ in\ dof, dof\ remaining) = \frac{reduction\ in\ RSS / cost\ in\ dof}{RSS\ remaining / dof\ remaining} = \frac{(RSS_1 - RSS_2)/cost\ in\ dof}{RSS_2/dof\ remaining} \tag{9}$$

Note:

- **reduction in** ***RSS***: reduction in *RSS* after the group of new X variables is added, $RSS_1 - RSS_2$
- **cost in dof**: reduction in the number of degrees of freedom after the group of new X variables is added. It is equal to the number of new variables added (here we add X_3 and X_4, so it is 2)
- *RSS* **remaining**: *RSS* of the model after the group of new X variables is added, RSS_2
- **dof remaining**: number of degrees of freedom remaining after the group of new X variables is added, $n - k$

Example

Original model specification (Model 1):

$$S_i = \beta_1 + \beta_2 ASVABC_i + u_i \tag{10}$$

Modified model specification (Model 2):

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SM_i + \beta_4 SF_i + u_i \tag{11}$$

Note:
S_i: schooling years of the i^{th} respondent
$ASVABC_i$: the ability score of the i^{th} respondent
SM_i: the highest grade completed by the mother of the i^{th} respondent
SF_i: the highest grade completed by the father of the i^{th} respondent

Step 1. State the null and alternative hypotheses

Null Hypothesis	$H_0 : \beta_3 = \beta_4 = 0$
Alternative Hypothesis	$H_1 : \beta_3 \neq 0$ *or* $\beta_4 \neq 0$ *or both* β_3 *and* $\beta_4 \neq 0$

Step 2. Select the significance level. Significance level $\alpha = 5\%$

Example

Step 3. Select and calculate the test statistics

$$F(cost\ in\ dof, dof\ remaining) = \frac{reduction\ in\ RSS/cost\ in\ dof}{RSS\ remaining/dof\ remaining} = \frac{(RSS_1 - RSS_2)/cost\ in\ dof}{RSS_2/dof\ remaining} \tag{12}$$

Example

Original model specification (Model 1):

$$S_i = \beta_1 + \beta_2 ASVABC_i + u_i \tag{13}$$

Note:
S_i: schooling years of the i^{th} respondent
$ASVABC_i$: the ability score of the i^{th} respondent

- We estimate first a simple model

```
Call:
lm(formula = S ~ ASVABC, data = EAWE21)

Coefficients:
(Intercept)       ASVABC
     14.437        1.581
```

- Then we use the `anova` command again to get information on RSS

```
> anova(educfit1)
Analysis of Variance Table

Response: S
           Df Sum Sq Mean Sq F value    Pr(>F)
ASVABC      1   1007 1007.00  182.56 < 2.2e-16 ***
Residuals 498   2747    5.52
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$RSS_1 = 2747$

Example

Modified model specification (Model 2):

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SM_i + \beta_4 SF_i + u_i \tag{14}$$

Note:
S_i: schooling years of the i^{th} respondent
$ASVABC_i$: the ability score of the i^{th} respondent
SM_i: the highest grade completed by the mother of the i^{th} respondent
SF_i: the highest grade completed by the father of the i^{th} respondent

- **Now we estimate the modified model:**

```
Call:
lm(formula = S ~ ASVABC + SM + SF, data = EAWE21)

Coefficients:
(Intercept)       ASVABC           SM           SF
   10.59674      1.24253      0.09135      0.20289
```

- **And use the `anova` command again to get the RSS**

```
Analysis of Variance Table

Response: S
           Df  Sum Sq Mean Sq F value       Pr(>F)
ASVABC      1 1007.00 1007.00 198.283   < 2.2e-16 ***
SM          1  112.38  112.38  22.128 0.000003312 ***
SF          1  115.68  115.68  22.778 0.000002396 ***
Residuals 496 2518.97    5.08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$RSS_2 = 2518.97 \quad n - k = 500 - 4 = 496$

Example

Null Hypothesis	$H_0 : \beta_3 = \beta_4 = 0$
Alternative Hypothesis	$H_1 : \beta_3 \neq 0 \text{ or } \beta_4 \neq 0 \text{ or both } \beta_3 \text{ and } \beta_4 \neq 0$

Step 3. Select and calculate the test statistics

$$F(\text{cost in dof}, \text{dof remaining}) = \frac{\text{reduction in } RSS / \text{cost in dof}}{RSS \text{ remaining} / \text{dof remaining}} = \frac{(RSS_1 - RSS_2)/\text{cost in dof}}{RSS_2/\text{dof remaining}} \tag{15}$$

- **reduction in** *RSS*: reduction in *RSS* after the group of new X variables is added, $RSS_1 - RSS_2 = 2747 - 2518.97$
- **cost in dof**: reduction in the number of degrees of freedom after the group of new X variables is added. It is equal to the number of new variables added (here we add X_3 and X_4, so $\text{cost in dof} = 2$)
- *RSS* **remaining**: *RSS* of the model after the group of new X variables is added, $RSS_2 = 2518.97$
- **dof remaining**: number of degrees of freedom remaining after the group of new X variables is added, $n - k = 500 - 4 = 496$
-

$$F(\text{cost in dof}, \text{dof remaining}) = \frac{\text{reduction in } RSS / \text{cost in dof}}{RSS \text{ remaining} / \text{dof remaining}} = \frac{(2747 - 2518.97)/2}{2518.97/496} = 22.45 \tag{16}$$

Example

Null Hypothesis	$H_0 : \beta_3 = \beta_4 = 0$
Alternative Hypothesis	$H_1 : \beta_3 \neq 0 \text{ or } \beta_4 \neq 0 \text{ or both } \beta_3 \text{ and } \beta_4 \neq 0$

Step 3. Select and calculate the test statistics

$$F(\text{cost in dof}, \text{dof remaining}) = \frac{\text{reduction in } RSS / \text{cost in dof}}{RSS \text{ remaining} / \text{dof remaining}} = \frac{(RSS_1 - RSS_2)/\text{cost in dof}}{RSS_2/\text{dof remaining}} \tag{15}$$

- **reduction in** *RSS*: reduction in *RSS* after the group of new X variables is added, $RSS_1 - RSS_2 = 2747 - 2518.97$
- **cost in dof**: reduction in the number of degrees of freedom after the group of new X variables is added. It is equal to the number of new variables added (here we add X_3 and X_4, so $\text{cost in dof} = 2$)
- *RSS* **remaining**: *RSS* of the model after the group of new X variables is added, $RSS_2 = 2518.97$
- **dof remaining**: number of degrees of freedom remaining after the group of new X variables is added, $n - k = 500 - 4 = 496$
-

$$F(\text{cost in dof}, \text{dof remaining}) = \frac{\text{reduction in } RSS / \text{cost in dof}}{RSS \text{ remaining} / \text{dof remaining}} = \frac{(2747 - 2518.97)/2}{2518.97/496} = 22.45 \tag{16}$$

Example

Null Hypothesis	$H_0 : \beta_3 = \beta_4 = 0$
Alternative Hypothesis	$H_1 : \beta_3 \neq 0 \text{ or } \beta_4 \neq 0 \text{ or both } \beta_3 \text{ and } \beta_4 \neq 0$

Step 3. Select and calculate the test statistics

$$F(\text{cost in dof}, \text{dof remaining}) = \frac{\text{reduction in } RSS / \text{cost in dof}}{RSS \text{ remaining} / \text{dof remaining}} = \frac{(RSS_1 - RSS_2)/\text{cost in dof}}{RSS_2/\text{dof remaining}} \tag{15}$$

- **reduction in** *RSS*: reduction in *RSS* after the group of new *X* variables is added, $RSS_1 - RSS_2 = 2747 - 2518.97$
- **cost in dof**: reduction in the number of degrees of freedom after the group of new *X* variables is added. It is equal to the number of new variables added (here we add X_3 and X_4, so $\text{cost in dof} = 2$)
- *RSS* **remaining**: *RSS* of the model after the group of new *X* variables is added, $RSS_2 = 2518.97$
- **dof remaining**: number of degrees of freedom remaining after the group of new *X* variables is added, $n - k = 500 - 4 = 496$
-

$$F(\text{cost in dof}, \text{dof remaining}) = \frac{\text{reduction in } RSS / \text{cost in dof}}{RSS \text{ remaining} / \text{dof remaining}} = \frac{(2747 - 2518.97)/2}{2518.97/496} = 22.45 \tag{16}$$

Example

Null Hypothesis	$H_0 : \beta_3 = \beta_4 = 0$
Alternative Hypothesis	$H_1 : \beta_3 \neq 0 \text{ or } \beta_4 \neq 0 \text{ or both } \beta_3 \text{ and } \beta_4 \neq 0$

Step 3. Select and calculate the test statistics

$$F(\text{cost in dof}, \text{dof remaining}) = \frac{\text{reduction in RSS}/\text{cost in dof}}{\text{RSS remaining}/\text{dof remaining}} = \frac{(RSS_1 - RSS_2)/\text{cost in dof}}{RSS_2/\text{dof remaining}} \tag{15}$$

- **reduction in** *RSS*: reduction in *RSS* after the group of new *X* variables is added, $RSS_1 - RSS_2 = 2747 - 2518.97$
- **cost in dof**: reduction in the number of degrees of freedom after the group of new *X* variables is added. It is equal to the number of new variables added (here we add X_3 and X_4, so $\text{cost in dof} = 2$)
- *RSS* **remaining**: *RSS* of the model after the group of new *X* variables is added, $RSS_2 = 2518.97$
- **dof remaining**: number of degrees of freedom remaining after the group of new *X* variables is added, $n - k = 500 - 4 = 496$
-

$$F(\text{cost in dof}, \text{dof remaining}) = \frac{\text{reduction in RSS}/\text{cost in dof}}{\text{RSS remaining}/\text{dof remaining}} = \frac{(2747 - 2518.97)/2}{2518.97/496} = 22.45 \tag{16}$$

Example

Null Hypothesis	$H_0 : \beta_3 = \beta_4 = 0$
Alternative Hypothesis	$H_1 : \beta_3 \neq 0 \text{ or } \beta_4 \neq 0 \text{ or both } \beta_3 \text{ and } \beta_4 \neq 0$

Step 3. Select and calculate the test statistics

$$F(\textit{cost in dof}, \textit{dof remaining}) = \frac{\textit{reduction in RSS}/\textit{cost in dof}}{\textit{RSS remaining}/\textit{dof remaining}} = \frac{(RSS_1 - RSS_2)/\textit{cost in dof}}{RSS_2/\textit{dof remaining}} \tag{15}$$

- **reduction in** *RSS*: reduction in *RSS* after the group of new X variables is added, $RSS_1 - RSS_2 = 2747 - 2518.97$
- **cost in dof**: reduction in the number of degrees of freedom after the group of new X variables is added. It is equal to the number of new variables added (here we add X_3 and X_4, so $\textit{cost in dof} = 2$)
- *RSS* **remaining**: *RSS* of the model after the group of new X variables is added, $RSS_2 = 2518.97$
- **dof remaining**: number of degrees of freedom remaining after the group of new X variables is added, $n - k = 500 - 4 = 496$
-

$$F(\textit{cost in dof}, \textit{dof remaining}) = \frac{\textit{reduction in RSS}/\textit{cost in dof}}{\textit{RSS remaining}/\textit{dof remaining}} = \frac{(2747 - 2518.97)/2}{2518.97/496} = 22.45 \tag{16}$$

Example

Step 4. Set the decision rule.
$\textit{cost in dof} = \textit{number of new variables added} = 2$
$k = 4, n = 500, \textit{dof remaining} = n - k = 500 - 4 = 496$

$F_{crit,5\%}(\textit{cost in dof}, \textit{dof remaining}) = F_{crit,5\%}(2, 496) = 3.01$

Step 5. Make statistical decisions.
$F = 22.45 > F_{crit,5\%}(2, 496) = 3.01$

We can reject the null $H_0 : \beta_3 = \beta_4 = 0$.
We conclude that the parental education variables *SM* and *SF* do have significant joint explanatory power.

Example

- We can use *anova* command in R studio to undertake F-test for testing groups of explanatory variables:
- Use both models as arguments of *anova* command

Original model (Model 1)

Modified model (Model 2)

```
> anova(educfit1,educfit2)
Analysis of Variance Table

Model 1: S ~ ASVABC
Model 2: S ~ ASVABC + SM + SF
  Res.Df  RSS Df Sum of Sq      F    Pr(>F)
1    498 2747
2    496 2519  2    228.06 22.453 4.627e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Use F test: We get the same F statistic = 22.45 > $F_{crit,5\%}(2,496) = 3.01$
  We can reject the null $H_0 : \beta_3 = \beta_4 = 0$.
- Use p-value: p-value = 4.627e-10 < 1%, variable is very significant (i.e. at the 1% level), reject null $H_0 : \beta_3 = \beta_4 = 0$.
  We conclude that the parental education variables *SM* and *SF* do have significant joint explanatory power.

Student Task

- Using the regression results from the following slides, undertake an F-test to identify whether adding per capita income (*pcinc*) and population (*pop*) as independent variables improves the explanatory power of the model significantly.
- Original model specification (Model 1):

$$officers_i = \beta_1 + \beta_2 crimes_i + u_i \tag{17}$$

Modified model specification (Model 2):

$$officers_i = \beta_1 + \beta_2 crimes_i + \beta_3 pcinc_i + \beta_4 pop_i + u_i \tag{18}$$

- Hypothesis test

Null Hypothesis	$H_0 : \beta_3 = \beta_4 = 0$
Alternative Hypothesis	$H_1 : \beta_3 \neq 0$ *or* $\beta_4 \neq 0$ *or both* β_3 *and* $\beta_4 \neq 0$

- Sample size $n = 45$
- $F_{crit,5\%}(2, 42) = 3.23$

Student Task

```
> anova(crimefit1)
Analysis of Variance Table

Response: officers
          Df   Sum Sq  Mean Sq F value    Pr(>F)
crimes     1 19508997 19508997  218.37 < 2.2e-16 ***
Residuals 44  3930948    89340
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

----------------------------------------------------------------

> anova(crimefit2)
Analysis of Variance Table

Response: officers
          Df   Sum Sq  Mean Sq  F value  Pr(>F)
crimes     1 19508997 19508997 267.0256 < 2e-16 ***
pcinc      1   338322   338322   4.6307 0.03720 *
pop        1   524088   524088   7.1734 0.01052 *
Residuals 42  3068537    73060
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$F(\text{cost in dof}, \text{dof remaining}) = \frac{\text{reduction in RSS}/\text{cost in dof}}{\text{RSS remaining}/\text{dof remaining}} = \frac{(RSS_1 - RSS_2)/\text{cost in dof}}{RSS_2/\text{dof remaining}} = ? \tag{19}$$

Testing Restrictions

- We can also test restrictions
- For example:

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SM_i + \beta_4 SF_i + u_i \tag{20}$$

  Is it possible that the coefficients of *SM* and *SF* are the same?
- Restriction is $\beta_3 = \beta_4$

Null Hypothesis	$H_0 : \beta_3 = \beta_4$
Alternative Hypothesis	$H_1 : \beta_3 \neq \beta_4$

Testing Restrictions

- To test this restriction, we estimate two models:
- **The unrestricted model**:

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SM_i + \beta_4 SF_i + u_i \tag{21}$$

Add restriction $\beta_3 = \beta_4$

- **The restricted model**:

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_c (SM_i + SF_i) + u_i \tag{22}$$

- The rationale of the test is the same as when testing the joint significance of a group of variables.

F Test Statistic

The unrestricted model: $S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SM_i + \beta_4 SF_i + u_i \quad RSS_{UR}$

The restricted model: $S_i = \beta_1 + \beta_2 ASVABC_i + \beta_c (SM_i + SF_i) + u_i \quad RSS_R$

$$F(\text{cost in dof}, \text{dof of unrestricted model}) = \frac{\text{reduction in RSS} / \text{cost in dof}}{\text{RSS remaining} / \text{dof of unrestricted model}}$$
$$= \frac{(RSS_R - RSS_{UR}) / \text{number of restrictions}}{RSS_{UR} / (n - k_{UR})}$$

Note:

- **cost in dof**: reduction in the number of degrees of freedom after the group of new X variables is added. It is equal to the number of restrictions (here we just have one restriction $\beta_3 = \beta_4$, so it is 1)
- **dof of unrestricted model**: number of degrees of freedom of the unrestricted model, $n - k_{UR}$

F Test Statistic

The unrestricted model: $S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SM_i + \beta_4 SF_i + u_i \quad RSS_{UR}$

The restricted model: $S_i = \beta_1 + \beta_2 ASVABC_i + \beta_c (SM_i + SF_i) + u_i \quad RSS_R$

$$F(\textit{cost in dof}, \textit{dof of unrestricted model}) = \frac{\textit{reduction in RSS} / \textit{cost in dof}}{\textit{RSS remaining} / \textit{dof of unrestricted model}}$$
$$= \frac{(RSS_R - RSS_{UR}) / \textit{number of restrictions}}{RSS_{UR}/(n - k_{UR})}$$

Note:

- **cost in dof**: reduction in the number of degrees of freedom after the group of new X variables is added. It is equal to the number of restrictions (here we just have one restriction $\beta_3 = \beta_4$, so it is 1)
- **dof of unrestricted model**: number of degrees of freedom of the unrestricted model, $n - k_{UR}$

Example

The unrestricted model: $S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SM_i + \beta_4 SF_i + u_i \quad RSS_{UR}$

The restricted model: $S_i = \beta_1 + \beta_2 ASVABC_i + \beta_c (SM_i + SF_i) + u_i \quad RSS_R$

Note:
S_i: schooling years of the i^{th} respondent
$ASVABC_i$: the ability score of the i^{th} respondent
SM_i: the highest grade completed by the mother of the i^{th} respondent
SF_i: the highest grade completed by the father of the i^{th} respondent

Step 1. State the null and alternative hypotheses

Null Hypothesis	$H_0 : \beta_3 = \beta_4$
Alternative Hypothesis	$H_1 : \beta_3 \neq \beta_4$

Step 2. Select the significance level. Significance level $\alpha = 5\%$

Example

Step 3. Select and calculate the test statistics

Unrestricted Model

```
Response: S
           Df  Sum Sq Mean Sq F value       Pr(>F)
ASVABC      1 1007.00 1007.00 198.283   < 2.2e-16 ***
SM          1  112.38  112.38  22.128 0.000003312 ***
SF          1  115.68  115.68  22.778 0.000002396 ***
Residuals 496 2518.97    5.08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Restricted Model

```
Response: S
           Df  Sum Sq Mean Sq F value    Pr(>F)
ASVABC      1 1007.00 1007.00 197.814 < 2.2e-16 ***
SMSF        1  216.99  216.99  42.625  1.64e-10 ***
Residuals 497 2530.04    5.09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$RSS_{UR} = 2518.97$ $\quad RSS_R = 2530.04$ $\quad$ one restriction $\beta_3 = \beta_4$ $\quad n - k_{UR} = 500 - 4 = 496$

$$F(cost\ in\ dof, dof\ of\ unrestricted\ model) = \frac{(RSS_R - RSS_{UR})/number\ of\ restrictions}{RSS_{UR}/(n - k_{UR})} = \frac{(2530.04 - 2518.97)/1}{2518.97/(500-4)} = 2.17$$

Example

Step 4. Set the decision rule.

$\textit{cost in dof} = \textit{number of restrictions} = 1$

For unrestricted model: $k_{UR} = 4$, $n = 500$,

$\textit{dof of unrestricted model} = n - k_{UR} = 500 - 4 = 496$

$F_{crit,5\%}(\textit{cost in dof}, \textit{dof of unrestricted model}) = F_{crit,5\%}(1, 496) = 3.86$

Step 5. Make statistical decisions.

$F = 2.17 < F_{crit,5\%}(1, 496) = 3.86$

We cannot reject the null $H_0 : \beta_3 = \beta_4$.

We conclude that the impact of mother (SM) and father's (SF) educational attainment on child's educational attainment could be the same.

What to do next:

- Attempt homework 6
- Read chapter 3.3 and 3.5 of Dougherty