BS2280 - Econometrics 1

Lecture 6 - Part 1: Multiple Regression Analysis II

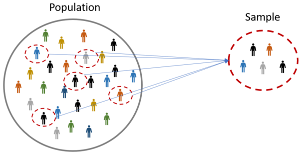
Dr. Yichen Zhu

Structure of today's lecture

Intended Learning Outcomes

- Understand OLS assumptions for the multiple regression model
- Testing the overall fit of the model

Review: Simple Regression Model OLS Assumptions



Population	Sample
$Y_i = \beta_1 + \beta_2 X_i + u_i$	$Y_i = \hat{Y}_i + \hat{u}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$
parameters β_1 and β_2	coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$
u_i disturbance term	$\hat{u}_i$ residual

OLS estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ have certain desirable properties, but these properties rely on a set of assumptions we need to make!!!

Simple Regression Model: OLS Assumptions

- **Assumption 1**. Model is linear in parameters and correctly specified.

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{1}$$

- **Assumption 2**. There is some variation in the X variable. X cannot be constant.
- **Assumption 3**. Disturbance term has zero expectation.
  $E(u_i) = 0$ for all i or $E(u_i|X_1, X_2, ..., X_n) = 0$ for all i
- **Assumption 4**. The disturbance term is homoscedastic. We assume that the error term has a constant variance, $E(u_i^2) = \sigma_u^2$ for all i
- **Assumption 5**. Values of disturbance term have independent distributions. We assume that the error terms are absolutely **independent** of each other.
  u_i is independently distributed from u_j for all $i \neq j$
- **Assumption 6**. The disturbance term has a normal distribution. Once we assume that the error term has a normal distribution, we can assume that $\hat{\beta}_1$ and $\hat{\beta}_2$ will also have a normal distribution, allowing us to carry out hypothesis tests on them.

Simple Regression Model: OLS Assumptions

- **Assumption 1**. Model is linear in parameters and correctly specified.

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{1}$$

- **Assumption 2**. There is some variation in the X variable. X cannot be constant.
- **Assumption 3**. Disturbance term has zero expectation.
  $E(u_i) = 0$ for all i or $E(u_i|X_1, X_2, ..., X_n) = 0$ for all i
- **Assumption 4**. The disturbance term is homoscedastic. We assume that the error term has a constant variance, $E(u_i^2) = \sigma_u^2$ for all i
- **Assumption 5**. Values of disturbance term have independent distributions. We assume that the error terms are absolutely **independent** of each other.
  u_i is independently distributed from u_j for all $i \neq j$
- **Assumption 6**. The disturbance term has a normal distribution. Once we assume that the error term has a normal distribution, we can assume that $\hat{\beta}_1$ and $\hat{\beta}_2$ will also have a normal distribution, allowing us to carry out hypothesis tests on them.

Simple Regression Model: OLS Assumptions

- **Assumption 1**. Model is linear in parameters and correctly specified.

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{1}$$

- **Assumption 2**. There is some variation in the X variable. X cannot be constant.
- **Assumption 3**. Disturbance term has zero expectation.
  $E(u_i) = 0$ for all i or $E(u_i|X_1, X_2, ..., X_n) = 0$ for all i
- **Assumption 4**. The disturbance term is homoscedastic. We assume that the error term has a constant variance, $E(u_i^2) = \sigma_u^2$ for all i
- **Assumption 5**. Values of disturbance term have independent distributions. We assume that the error terms are absolutely **independent** of each other.
  u_i is independently distributed from u_j for all $i \neq j$
- **Assumption 6**. The disturbance term has a normal distribution. Once we assume that the error term has a normal distribution, we can assume that $\hat{\beta}_1$ and $\hat{\beta}_2$ will also have a normal distribution, allowing us to carry out hypothesis tests on them.

Simple Regression Model: OLS Assumptions

- **Assumption 1**. Model is linear in parameters and correctly specified.

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{1}$$

- **Assumption 2**. There is some variation in the X variable. X cannot be constant.
- **Assumption 3**. Disturbance term has zero expectation.
  $E(u_i) = 0$ for all i or $E(u_i|X_1, X_2, ..., X_n) = 0$ for all i
- **Assumption 4**. The disturbance term is homoscedastic. We assume that the error term has a constant variance, $E(u_i^2) = \sigma_u^2$ for all i
- **Assumption 5**. Values of disturbance term have independent distributions. We assume that the error terms are absolutely **independent** of each other.
  u_i is independently distributed from u_j for all $i \neq j$
- **Assumption 6**. The disturbance term has a normal distribution. Once we assume that the error term has a normal distribution, we can assume that $\hat{\beta}_1$ and $\hat{\beta}_2$ will also have a normal distribution, allowing us to carry out hypothesis tests on them.

Simple Regression Model: OLS Assumptions

- **Assumption 1**. Model is linear in parameters and correctly specified.

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{1}$$

- **Assumption 2**. There is some variation in the X variable. X cannot be constant.
- **Assumption 3**. Disturbance term has zero expectation.
  $E(u_i) = 0$ for all i or $E(u_i | X_1, X_2, ..., X_n) = 0$ for all i
- **Assumption 4**. The disturbance term is homoscedastic. We assume that the error term has a constant variance, $E(u_i^2) = \sigma_u^2$ for all i
- **Assumption 5**. Values of disturbance term have independent distributions. We assume that the error terms are absolutely **independent** of each other.
  u_i is independently distributed from u_j for all $i \neq j$
- **Assumption 6**. The disturbance term has a normal distribution. Once we assume that the error term has a normal distribution, we can assume that $\hat{\beta}_1$ and $\hat{\beta}_2$ will also have a normal distribution, allowing us to carry out hypothesis tests on them.

Simple Regression Model: OLS Assumptions

- **Assumption 1**. Model is linear in parameters and correctly specified.

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{1}$$

- **Assumption 2**. There is some variation in the X variable. X cannot be constant.
- **Assumption 3**. Disturbance term has zero expectation.
  $E(u_i) = 0$ for all i or $E(u_i|X_1, X_2, ..., X_n) = 0$ for all i
- **Assumption 4**. The disturbance term is homoscedastic. We assume that the error term has a constant variance, $E(u_i^2) = \sigma_u^2$ for all i
- **Assumption 5**. Values of disturbance term have independent distributions. We assume that the error terms are absolutely **independent** of each other.
  u_i is independently distributed from u_j for all $i \neq j$
- **Assumption 6**. The disturbance term has a normal distribution. Once we assume that the error term has a normal distribution, we can assume that $\hat{\beta}_1$ and $\hat{\beta}_2$ will also have a normal distribution, allowing us to carry out hypothesis tests on them.

Multiple Regression Model: OLS Assumptions

- As with the simple regression model, a set of assumptions hold for multiple regression model
- **Assumption 1.** Model is linear in parameters and correctly specified:

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

- **Assumption 2.** There is no exact linear relationship amongst the X variables in the sample (more about this later)
- **Assumption 3.** Disturbance term has zero expectation.

$$E(u_i) = 0 \text{ for all } i$$

Multiple Regression Model: OLS Assumptions

- As with the simple regression model, a set of assumptions hold for multiple regression model
- **Assumption 1.** Model is linear in parameters and correctly specified:

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

- **Assumption 2.** There is no exact linear relationship amongst the X variables in the sample (more about this later)
- **Assumption 3.** Disturbance term has zero expectation.

$$E(u_i) = 0 \text{ for all } i$$

Multiple Regression Model: OLS Assumptions

- As with the simple regression model, a set of assumptions hold for multiple regression model
- **Assumption 1.** Model is linear in parameters and correctly specified:

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

- **Assumption 2.** There is no exact linear relationship amongst the X variables in the sample (more about this later)
- **Assumption 3.** Disturbance term has zero expectation.

$$E(u_i) = 0 \text{ for all } i$$

Multiple Regression Model: OLS Assumptions

- As with the simple regression model, a set of assumptions hold for multiple regression model
- **Assumption 1.** Model is linear in parameters and correctly specified:

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

- **Assumption 2.** There is no exact linear relationship amongst the X variables in the sample (more about this later)
- **Assumption 3.** Disturbance term has zero expectation.

$$E(u_i) = 0 \text{ for all } i$$

Multiple Regression Model: OLS Assumptions

- **Assumption 4.** The disturbance term is homoscedastic. We assume that the error term has a constant variance, $E(u_i^2) = \sigma_u^2$ for all i
- **Assumption 5.** Values of disturbance term have independent distributions. We assume that the error terms are absolutely **independent** of each other. u_i is independently distributed from u_j for all $i \neq j$
- **Assumption 6.** The disturbance term has a normal distribution.

Multiple Regression Model: OLS Assumptions

- **Assumption 4.** The disturbance term is homoscedastic. We assume that the error term has a constant variance, $E(u_i^2) = \sigma_u^2$ for all i
- **Assumption 5.** Values of disturbance term have independent distributions. We assume that the error terms are absolutely **independent** of each other. u_i is independently distributed from u_j for all $i \neq j$
- **Assumption 6.** The disturbance term has a normal distribution.

Multiple Regression Model: OLS Assumptions

- **Assumption 4.** The disturbance term is homoscedastic. We assume that the error term has a constant variance, $E(u_i^2) = \sigma_u^2$ for all i
- **Assumption 5.** Values of disturbance term have independent distributions. We assume that the error terms are absolutely **independent** of each other. u_i is independently distributed from u_j for all $i \neq j$
- **Assumption 6.** The disturbance term has a normal distribution.

Multiple Regression Model: BLUE

If all our OLS assumptions hold, we get Best (most efficient) Linear Unbiased Estimators (BLUE) $\hat{\beta}_i$

Unbiasedness

An estimator is unbiased when expected value equals population value.

$$E(\hat{\beta}_i) = \beta_i$$

Consistency

The larger the sample the closer our estimators $\hat{\beta}_i$ should be to the population value β_i

Efficiency / Precision

For the OLS estimator to be best (most efficient) it needs to have a lower variance than all the other estimators within the class

Multiple Regression Model: BLUE

Efficiency / Precision

Simple regression model:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{2}$$

$$variance(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X)}$$

where $MSD(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Multiple regression model:

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

$$variance(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X_2)} \times \frac{1}{1 - r^2_{X_2 X_3}}$$

where $MSD(X_2) = \frac{1}{n}\sum_{i=1}^{n}(X_2 - \bar{X}_2)^2$

$r^2_{X_2 X_3}$ is the squared sample correlation coefficient between X_2 and X_3

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. $H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0; H_1 : \textit{at least one } \beta \neq 0$
  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 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. $H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0; H_1 : \textit{at least one } \beta \neq 0$
    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 Overall Significance or Overall Fit of a model

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

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

- We will use *F*-Tests again to check the overall significance of the model
- There are two ways to write down the hypotheses:
  1. $H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0; H_1 : \textit{at least one } \beta \neq 0$
  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 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. $H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0; H_1 : \textit{at least one}\, \beta \neq 0$
  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 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. $H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0$; $H_1 : \textit{at least one}\ \beta \neq 0$
  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 Overall Significance or Overall Fit of a model

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

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

- We will use *F*-Tests again to check the overall significance of the model
- There are two ways to write down the hypotheses:
  1. $H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0$; 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 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. $H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0; H_1 : \textit{at least one}\ \beta \neq 0$
  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 Overall Significance or Overall Fit of a model

Example: educational attainment model

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SM_i + \beta_4 SF_i + u_i$$

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	$H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0$ or $H_0 : R^2 = 0$
Alternative Hypothesis	H_1 : *at least one* $\beta \neq 0$ or $H_1 : R^2 \neq 0$

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

Testing the Overall Significance or Overall Fit of a model

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SM_i + \beta_4 SF_i + u_i$$

Step 3. Select and calculate the test statistics

Test the entire regression model, so use F statistic. Two ways:

Option 1. $F(k-1, n-k) = \dfrac{ESS/(k-1)}{RSS/(n-k)} = \dfrac{\frac{ESS}{TSS}/(k-1)}{\frac{RSS}{TSS}/(n-k)}$

k: number of regression coefficients in the model, including the intercept

n: number of observations in the model, sample size

Using ESS and RSS we get by using anova() command in R

```
> anova(educfit)
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
```

```
> nobs(educfit)
[1] 500
```

$$F(k-1, n-k) = F(4-1, 500-4) = \frac{ESS/(k-1)}{RSS/(n-k)} = \frac{(1007.00 + 112.38 + 115.68)/(4-1)}{2518.97/(500-4)} = 81.06$$

Testing the Overall Significance or Overall Fit of a model

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SM_i + \beta_4 SF_i + u_i$$

Step 3. Select and calculate the test statistics

Option 2. $F(k-1, n-k) = \frac{ESS/(k-1)}{RSS/(n-k)} = \frac{\frac{ESS}{TSS}/(k-1)}{\frac{RSS}{TSS}/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$

k: number of regression coefficients in the model, including the intercept
n: number of observations in the model, sample size
Using R^2 from when using summary() command in R

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

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 10.59674    0.61428  17.251   < 2e-16 ***
ASVABC       1.24253    0.12359  10.054   < 2e-16 ***
SM           0.09135    0.04593   1.989    0.0473 *
SF           0.20289    0.04251   4.773 0.0000024 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.254 on 496 degrees of freedom
Multiple R-squared:  0.329,       Adjusted R-squared:  0.3249
F-statistic: 81.06 on 3 and 496 DF,  p-value: < 2.2e-16
```

$$F(k-1, n-k) = F(4-1, 500-4) = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} = \frac{0.329/(4-1)}{(1-0.329)/(500-4)} = 81.06$$

Testing the Overall Significance or Overall Fit of a model

Step 4. Set the decision rule.

$k = 4, n = 500.$ $F_{crit,5\%}(4-1, 500-4) = F_{crit,5\%}(3, 496) = 2.62$

Step 5. Make statistical decisions.

$F = 81.06 > F_{crit,5\%}(3, 496) = 2.62$

We can reject the null $H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0$ or $H_0 : R^2 = 0$.

Our model is statistically significant at 5% level.

Student Task

- We set up a model to identify the impact of attendance (*attend*), submitted homework (*hwrte*) and ability (*ACT*) on the final exam mark (*final*) at an American University. We get to following results:

$$final_i = \beta_1 + \beta_2 attend_i + \beta_3 hwrte_i + \beta_4 ACT_i + u_i$$

- Use an F-test to determine the overall statistical significance of the estimated model.
- Number of observations in the model, sample size $n = 674$
- The critical F-value at the 5% significance level is 2.6.

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

Response: final
           Df  Sum Sq Mean Sq  F value      Pr(>F)
attend      1   329.4  329.39  17.8108 0.00002776 ***
hwrte       1    71.7   71.72   3.8783    0.04933 *
ACT         1  2232.8 2232.81 120.7341  < 2.2e-16 ***
Residuals 670 12390.7   18.49
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```