



University of  
Nottingham  
UK | CHINA | MALAYSIA



---

# APPLIED ECONOMETRIC I NOTES 3

---



NOVEMBER 11, 2024  
UNIVERSITY OF NOTTINGHAM  
HAOLING WANG

## Applied econometrics note

### Hypothesis testing

#### Table of Contents

<i>Formulation of hypothesis</i> .....	4
<i>Example of conducting a hypothesis testing (1)</i> .....	5
<i>Assumption MLR6: Normality of u</i> .....	8
Valuation .....	8
Explanation .....	9
<i>Classical Linear Model (CLM)</i> .....	9
<i>Decision rule</i> .....	11
Content .....	11
Explanation .....	11
Problem .....	12
<i>Hypothesis testing using the t statistic</i> .....	13
<i>p-values</i> .....	15
Definition .....	15
<i>Example of conducting a hypothesis testing (2)</i> .....	17
<i>Type I and Type II errors</i> .....	18
Type I error .....	19
Type II error .....	20
<i>Confidence intervals of <math>\beta_j</math></i> .....	22

<b><i>Assess regression model performance</i></b> .....	<b>25</b>
Decomposing the variation in the data .....	25
Coefficient of determination.....	25
<b><i>F test</i></b> .....	<b>28</b>
<b><i>Conducting F tests</i></b> .....	<b>33</b>

## Preface

This set of notes originates from my Applied Econometrics I module at UNNC. Rather than adding new material, I have reorganized the logical flow of the key points presented in the lecture slides. The module is closely linked to ECON1049 Mathematical Economics and Econometrics (SPR)—now renamed ECON1057 Introductory Econometrics—so many topics will already be familiar. If you find something related to statistics or econometrics, please go back to the lecture notes in ECON1049 (/ECON 1057). This will benefit to your memory and smooth your further study.

That said, the course should not be taken lightly. Its true value lies not in memorizing the slide-deck formulas, but in learning how to apply the statistical and econometric theory covered in Year 2 to real economic-data analysis. From a practical standpoint, I strongly recommend gaining proficiency in STATA, as intensive coding will return in Applied Econometrics II and even your final year dissertation.

Because I studied this module on exchange at UNUK, I also completed a piece of coursework. That assignment—requiring the direct application of PPT concepts, extensive STATA work, and a basic research framework—proved immensely beneficial. If you know classmates heading to UNUK, do not hesitate to ask for a look at their coursework. At UNNC, Applied Econometrics I is assessed solely by a final exam; therefore, if you are curious about econometrics or research methods, consider obtaining the UNUK coursework and experimenting with it yourself (perhaps over the summer, or even with GPT's guidance). Treat it as a scaffold for practicing real applications, but ultimately try to write the code on your own.

My guiding principle is to look beyond earning marks (though I know marks are important) for isolated knowledge points and instead ask whether each step you take builds the capabilities you will need next year or even in the future. Much like the Permanent Income Hypothesis in macroeconomics, cultivating a habit of extension and forward thinking will smooth your progression to the next stage, or if you learn repeated game in game theory, you will find out that the choices in each stage will be different depending on whether the game is one-stage or multi-stage (I think that sometimes, in real life, acting in player 1 or 2 in game theory is important, because they always have rational strategy).

This is the synthesized note 3 (4 in total). If you do not have the previous notes, please visit the website: <https://github.com/wang95483/notebook>. And since this is just a synthesized lecture notes, its function is limited. So if you have any other ideas of how to learn this course, feel free to drop me an email first [hmyhw8@nottingham.edu.cn](mailto:hmyhw8@nottingham.edu.cn) (or if this doesn't work, use [wang95483@gmail.com](mailto:wang95483@gmail.com), my personal email).

Good luck with your Applied Econometric I study!

### Formulation of hypothesis

Our aim is to infer somethings about the **population parameters**,  $\beta_0, \beta_1, \beta_2, \beta_3 \dots \beta_k$ , i.e., the  $\beta_j$

\*\* the hypotheses are set up with reference to the population parameters **not the estimates** ( $\hat{\beta}_j$ )

The Null hypothesis, $H_0$ :		The Alternative hypothesis, $H_1$ :	
the hypothesis we test		the <b>logical conclusion</b> if the null hypothesis is false	
The Null always refers to: (1) a specific value (e.g. $H_0: \beta_j = 0$ ) (2) a range with one margin defined by a specific value (e.g. $H_0: \beta_j \geq 0$ )		The theory is captured in the Alternative hypothesis The Alternative refers to: (1) all values other than the value (e.g. $H_1: \beta_j \neq 0$ ) (2) range (with a defined margin) referred to in the Null. (e.g. $H_0: \beta_j < 0$ )	
A Null-Alternative pair encompasses ALL the possibilities.			
two-tailed test		one-tailed test	
$H_0: \beta_1 = 0 \quad H_1: \beta_1 \neq 0$		$H_0: \beta_1 \geq 0 \quad H_1: \beta_1 < 0$	
Q: Should $x_j$ be in the model? - Null: $x_j$ has no effect on $y$ (unimportant, should not be in the model) - Alternative: $x_j$ has an effect on $y$ (important, should be in the model)			
rejecting the null implies that a relationship exists we are not confining the relationship to be either positive or negative		rejecting the null implies that the relationship is negative	
reject the null	Fail to reject the null		
the estimate $\hat{\beta}_j$ is statistically <b>significant</b> at the $\alpha\%$ level	The estimate $\hat{\beta}_j$ is statistically insignificant at the $\alpha\%$ level		

---

---

### Example of conducting a hypothesis testing (1)

Demand theory suggests that:

- (1) **demand is affected by price, specifically, it is inversely related to price**
- (2) demand may be positively or negatively related to income
- (3) individual characteristics affect demand in various ways

To test the theory, set up the model for the whole theory

$$Q = \beta_0 + \beta_1 P + \beta_2 Y + \beta_3 age + \dots + \beta_k x_k + u$$

( $Q$  = demand,  $P$  = price,  $Y$  = income,  $age \dots x_k$  = individual characteristics)

Turn the statement into parameter in the model:

- (1)  **$\beta_1 \neq 0$ , specifically,  $\beta_1 < 0$**
- (2)  $\beta_2 > 0$  or  $\beta_2 < 0$
- (3)  $\beta_3$  to  $\beta_k$  sign depends on the characteristic and the good

(1.1) Is demand affected by price? (theory predicts that it is)

$$H_0: \beta_1 = 0 \quad H_1: \beta_1 \neq 0$$

- **reject**  $H_0: \beta_1 = 0$ , i.e., reject the hypothesis of no relationship
- **fail to reject**  $H_0: \beta_1 = 0$ , cannot rule out the possibility of no relationship

## Decision Rule: When to reject the Null

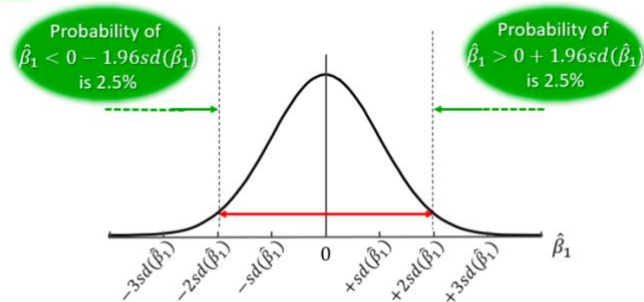
**E.g., Theory of demand:**  $Q = \beta_0 + \beta_1 P + \beta_2 Y + \beta_3 age + \dots + \beta_k x_k + u$

Is demand affected by price?

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

If  $H_0$  is true under the CLM assumptions, the sampling distribution of  $\hat{\beta}_1$  is...



and

- 68% of the distribution lies within 1 standard deviation of the mean ↔
- 95% of the distribution lies within 1.96 standard deviations of the mean ↔

(1.2) Does demand fall when price increases? (theory predicts that it does)

$$H_0: \beta_1 \geq 0 \quad H_1: \beta_1 < 0$$

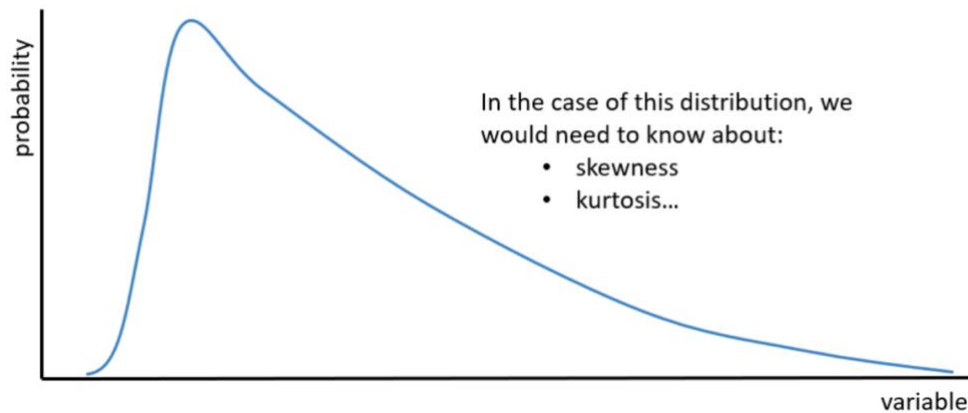
- **reject**  $H_0: \beta_1 \geq 0$ , i.e., reject the hypothesis of no or a positive relationship
  - **fail to reject**  $H_0: \beta_1 \geq 0$ , i.e., cannot rule out the possibility of no or a positive relationship
- 
-

but the **population parameters are unobservable**, we use our **estimates** and what we know about their dispersion (variance, standard error) **to draw our inferences**

Firstly, we would need to know everything about the **sampling distributions of our estimators**.

**\*\*when a distribution is not normal**, everything there is to know about it is not contained in its mean and variance

We need to know more than just its mean and variance in order to know the likelihood of the value of the variable falling within any given range.



(+) Any **normal distribution** is completely defined by its **mean** and **variance**

→ When the distribution is **normal**, knowing the **mean** and **variance** of the distribution equals to know everything about the distribution.

(+) **Sampling distribution**

The  $\hat{\beta}_j$  are random variables made up of a deterministic part ( $\beta_j$ , given certain assumptions) and a stochastic part which relates to  $u$

$$\hat{\beta}_1 = \beta_1 + \frac{Cov(x, u)}{Var(x)} \text{ (SLR)}$$

$$E(\hat{\beta}_1) = \beta_{1(\text{unbiased})}$$



$$Var(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_{ji} - \bar{x}_j)^2 (1 - R_j^2)}$$

(smallest variance)(MLR)

→ the shape of the distributions of the  $\hat{\beta}_j$  depends on the shape of the distribution of  $u$

→ If we assume that the **disturbances ( $u$ ) are normally distributed** with  $E(u) = 0$ ,  $Var(u) = \sigma^2$ , then the **sampling distributions** of our  $\hat{\beta}_j$  are also **normally distributed**

(+) Assumption MLR4: **Zero conditional mean:  $E(u|x) = 0$**

Assumption MLR5: **Homoscedasticity:  $Var(u|x) = \sigma^2$**

→

#### Assumption MLR6: Normality of $u$

The population disturbances ( $u$ ) are **independent** of the **explanatory variables** ( $x_1, x_2, x_3, \dots, x_k$ ) and are **normally distributed** with zero mean and variance  $\sigma^2$

$$u \sim \text{Normal}(0, \sigma^2)$$

#### Valuation

MLR6 is **stronger** than MLR4 and MLR5 combined because it is about the **overall shape** of the distribution of  $u$ , not just its mean and variance.

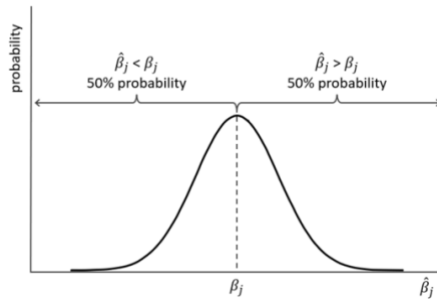
MLR6 is not required for the Gauss-Markov theorem, it is **not a Gauss- Markov assumption**.

### Explanation

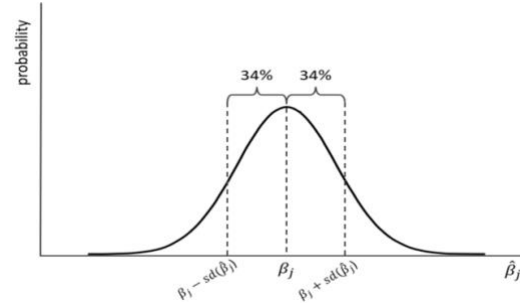
If we assume that the disturbances ( $u$ ) are normally distributed, because **we know the means and variances of the sampling distributions of the  $\hat{\beta}_j$** , we know everything about the sampling distributions of the  $\hat{\beta}_j$  and can, therefore, draw meaningful statistical inferences.

$$f(\hat{\beta}_j | \beta_j, sd) = \frac{1}{\sqrt{2\pi}sd} e^{-\frac{(\hat{\beta}_j - \beta_j)^2}{2sd^2}}$$

Assuming normality:



Assuming normality:



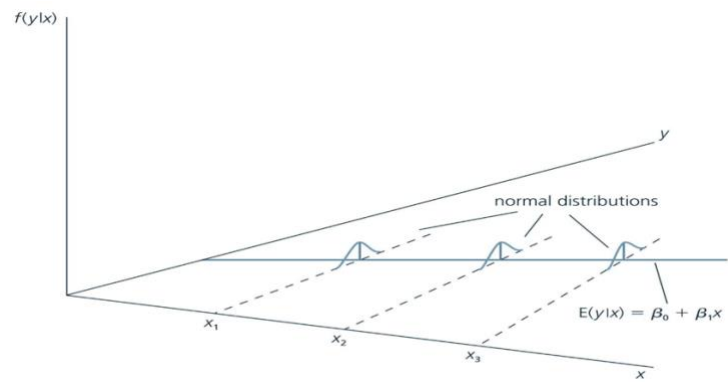
MLR1-MLR6 →

### Classical Linear Model (CLM)

conditional on  $x$ ,  $y$  has a normal distribution with a mean that is linear in  $x$  and a variance that is constant irrespective of  $x$

$$y|x \sim \text{Normal}(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k, \sigma^2)$$

Graph



### Decision rule

——about when to reject the Null hypothesis

### Content

Reject  $H_0$  if it implies that the probability of getting the estimate that OLS yields is less than a small preselected probability (e.g. 5%).

We call this probability the **significant level**. Using a significance level of 5% implies that we are willing to be wrong about rejecting the Null 5% of the time.

$$\frac{(\hat{\beta}_j - \beta_j)}{sd(\hat{\beta}_j)} \sim Normal[0,1]$$

Under the CLM,

When  $H_0: \beta_j = 0$ , to apply the decision rule, we use  $\frac{\hat{\beta}_j}{sd(\hat{\beta}_j)}$  to look at whether the resulting number falls in the standard normal distribution  $Z \sim Normal[0,1]$ .

Generalize  
 $\implies$  When  $H_0: \beta_j = h$ , to apply the decision rule, we use  $\frac{(\hat{\beta}_j - \beta_j)}{sd(\hat{\beta}_j)} = \frac{(\hat{\beta}_j - h)}{sd(\hat{\beta}_j)}$  and then look at where the resulting number falls in the standard normal distribution  $Z \sim Normal[0,1]$ .

### Explanation

Our estimates  $\hat{\beta}_j$  are draws from these sampling distributions

We expect them to be equal to the  $\beta_j$ , but know that they may deviate from the  $\beta_j$  and that these **deviations** are governed by  $Var(\hat{\beta}_j)$

We must **take account of these possible deviations** when drawing our inferences, i.e., we must use our knowledge about the sampling distribution when we draw inference

### Problem

$$sd(\hat{\beta}_j) = \frac{\sigma^2}{\sqrt{\sum_{i=1}^n (x_{ji} - \bar{x}_j)^2 (1 - R_j^2)}}$$

$\sigma^2$  is unknown when calculating  $sd(\hat{\beta}_j)$ :

Our estimate of  $sd(\hat{\beta}_j)$  is the standard error of  $\hat{\beta}_j$ ,

$$se(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{\sqrt{\sum_{i=1}^n (x_{ji} - \bar{x}_j)^2 (1 - R_j^2)}}$$

$$\frac{(\hat{\beta}_j - \beta_j)}{se(\hat{\beta}_j)} = t, \text{ which has a t-distribution}$$

So now

It takes account of:

(1) how much information we used in the estimation, i.e., the sample size,  $n$

(2) how much of that information we used up, i.e., the number of model parameters we estimated,  $k$  slope parameters plus 1 constant

→  $n - k - 1$  = the degrees of freedom

→ the  $t$  statistic has a  $t$  distribution with  $n - k - 1$  degrees of freedom



## Hypothesis testing using the t statistic

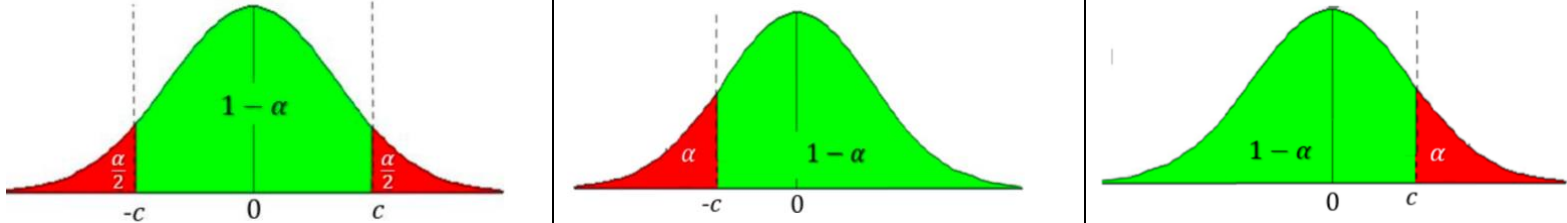
Under the CLM assumptions

$$t = \frac{(\hat{\beta}_j - \beta_j)}{se(\hat{\beta}_j)} \sim t_{n-k-1} \text{ if } H_0 \text{ is true}$$

When  $H_0: \beta_j = 0$ , to apply the decision rule, we use  $\frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$  and then look at where the resulting number falls in the  $t_{n-k-1}$  distribution

Generalize  $\Rightarrow$  When  $H_0: \beta_j = h$ , to apply the decision rule, we use  $\frac{(\hat{\beta}_j - h)}{se(\hat{\beta}_j)}$  and then look at where the resulting number falls in the  $t_{n-k-1}$  distribution.

Formulate the hypothesis (Null and Alternative)	(two-tailed) hypothesis (continued)	One tailed test (continued)	
	Alternative hypothesis is what you want to prove. i.e. your guess is written in alternative hypothesis $H_1$		
	$H_0: \beta_1 = 0 \quad H_1: \beta_1 \neq 0$ $H_0: \beta_j = h \quad H_1: \beta_j \neq h$	$H_0: \beta_1 \geq 0 \quad H_1: \beta_1 < 0$	$H_0: \beta_j \leq 0 \quad H_1: \beta_j > 0$
Select the significance level	Select a significance level, $\alpha$ , e.g. $\alpha = 5\%$ , which implies that we are willing to mistakenly reject $H_0$ 5% of the time.		
Find the critical value(s) and use it(them) to set up the Decision Rule	Find the critical values, $\pm c$ , such that the rejection regions are $\alpha = 5\%$ of the $t$ distribution with $n - k - 1$ d.f. (for a two-tailed test, $\alpha/2$ in each tail of the distribution)	Find the critical values, $-c$ , such that the rejection regions are $\alpha = 5\%$ of the $t$ distribution with $n - k - 1$ d.f. (for a one-tailed test, $\alpha$ in tail of the distribution)	Find the critical value, $c$ , such that the rejection region is $\alpha = 5\%$ of the $t$ distribution with $n - k - 1$ d.f. (for a one-tailed test, $\alpha$ in just one tail of the distribution)
	Given a significance level $\alpha$ (e.g. 5%), the critical value, $c$ , is: the $(1 - \frac{\alpha}{2})^{\text{th}}$ percentile of the $t$ distribution (with $n - k - 1$ df)	Given a significance level $\alpha$ (e.g. 5%), the critical value, $c$ , is: the $(1 - \alpha)^{\text{th}}$ percentile of the $t$ distribution (with $n - k - 1$ df)	
Calculate the test statistic, i.e., the $t$ statistic	Calculate the $t$ statistic, $t = \frac{(\hat{\beta}_j - \beta_j)}{se(\hat{\beta}_j)}$ which, under the CLM assumptions, is distributed as $t$ with $n - k - 1$ d.f. if $H_0$ is true		
Conduct the test, i.e., apply	Reject $H_0$ if $t < -c$ or $t > c$ Fail to reject $H_0$ if $-c < t < c$	Reject $H_0$ if $t < -c$ Fail to reject $H_0$ if $t > -c$	Reject $H_0$ if $t > c$ Fail to reject $H_0$ if $t < c$

the Decision Rule, i.e., compare the $t$ statistic and the critical value(s)			
Draw conclusion	<p>reject the Null <math>\rightarrow</math> there is a relationship: the estimate <math>\hat{\beta}_j</math> is statistically significant at the <math>\alpha\%</math> level.</p> <p>fail to reject the Null <math>\rightarrow</math> there may be no relationship: the estimate <math>\hat{\beta}_j</math> is statistically significant at the <math>\alpha\%</math> level.</p>		

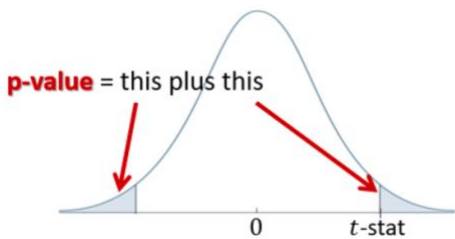
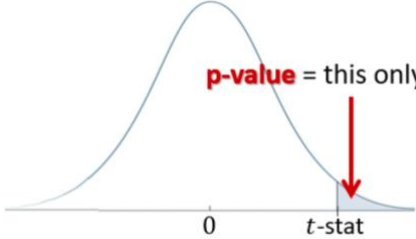
\*\*Note that there is a difference between **statistical significance** and **practical significance**; a  $\hat{\beta}_j$  can be statistically significant and, at the same time, very small  $se(\hat{\beta}_j)$ .

## p-values

### Definition

It is useful to know the specific significance level at which an estimate **ceases to be statistically significant** as this is a measure of the strength of the evidence against the null. This specific significance level is called the p-value.

The p-value is a probability.

(two-tailed) hypothesis <a href="#">(continued)</a>	One tailed test <a href="#">(continued)</a>
the p-value is the area under the pdf for $t_{n-k-1}$ in the tails beyond $\pm t$ -statistic	the p-value is the area under the pdf for $t_{n-k-1}$ in the tails beyond $t$ -statistic
	
<p>Decision rule</p> <p>Compare the p-value to <math>\alpha</math>:</p> <ul style="list-style-type: none"><li>• If <math>p \leq \alpha</math>:<ul style="list-style-type: none"><li>• Reject the null hypothesis (<math>H_0</math>).</li><li>• Conclude that the results are statistically significant and provide sufficient evidence to support the alternative hypothesis (<math>H_a</math>).</li></ul></li><li>• If <math>p &gt; \alpha</math>:<ul style="list-style-type: none"><li>• Fail to reject the null hypothesis (<math>H_0</math>).</li><li>• Conclude that the results are not statistically significant, meaning there is insufficient evidence to support the alternative hypothesis.</li></ul></li></ul>	





### Example of conducting a hypothesis testing (2)

Assume *wages* (\$/hour) are determined by education and experience (both in years) and other unobservable factors, contained in  $u$ .

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u \quad (1)$$

Applying OLS to data for a sample of 28 people:

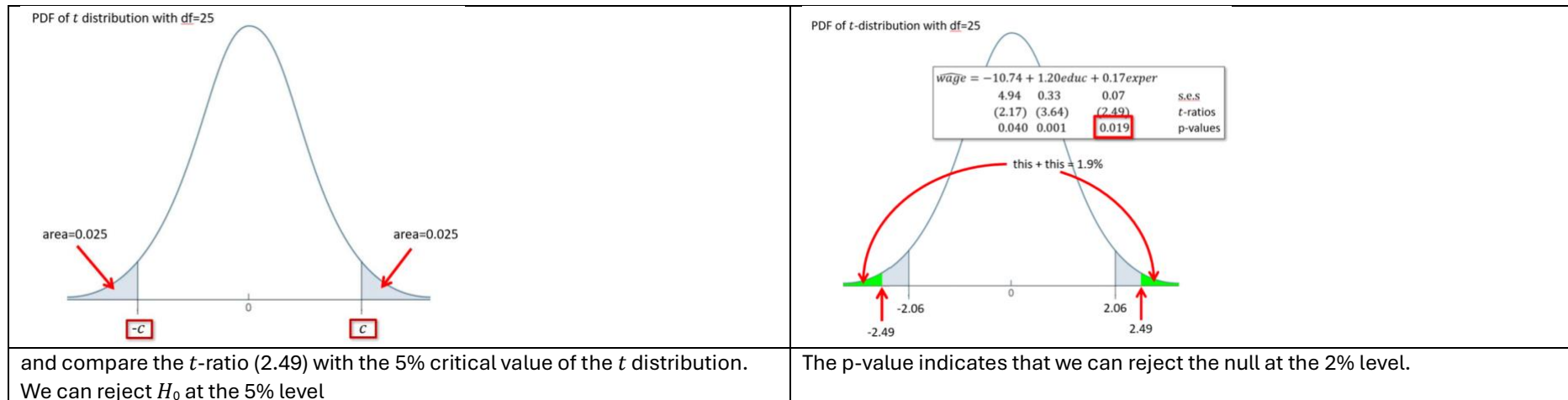
$$\widehat{wage} = -10.74 + 1.20educ + 0.17exper$$

	4.94	0.33	0.07	s.e.s
	(2.17)	(3.64)	(2.49)	t-ratios
	0.040	0.001	0.019	p-values

To evaluate whether there is a statistically significant relationship between wages and experience at the 5% significance level, we use the two-sided hypothesis:

$$H_0: \beta_2 = 0 \quad H_1: \beta_2 \neq 0$$

$$t\text{-ratio} = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)} = t\text{-statistic when } H_0: \beta_j = 0$$



### Type I and Type II errors

	$H_0$ is true	$H_0$ is false
Do not reject $H_0$	✓	✗ Type II error
Reject $H_0$	✗ Type I error	✓

### Relationship

The probabilities of making a Type I and a Type II error are inversely related

→ We tolerate a small Type I probability so that we don't drive the Type II probability to unacceptably high levels.

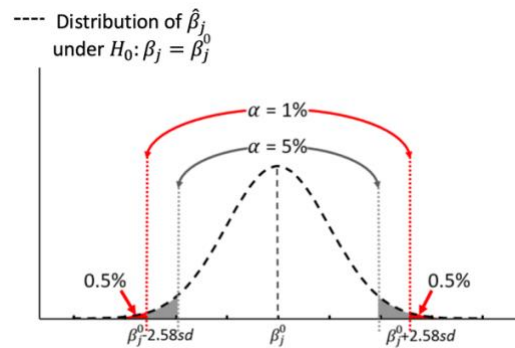
## Type I error

### Definition

reject a true null

### Formula

Significance level of a test =  $\alpha = \text{Prob}(\text{Type I})$



Suppose the Null is true: reducing  $\alpha$  to 1% reduces the risk of making a Type I error to 1% (expected to happen in 1 in 100 samples).

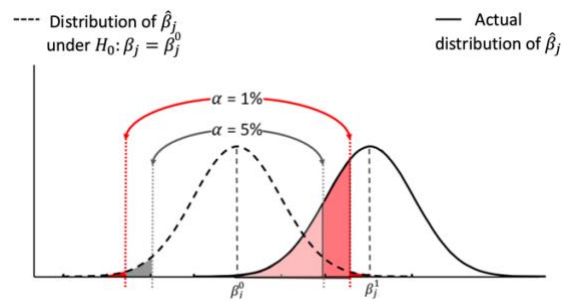
## Type II error

### Definition

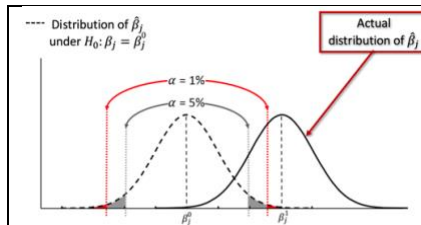
fail to reject a false null

### Formula

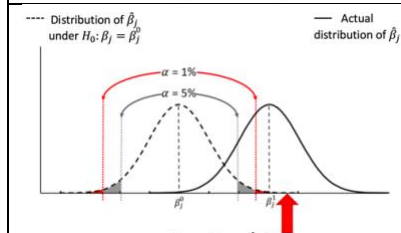
Power of a test =  $1 - \text{Prob}(\text{Type II})$



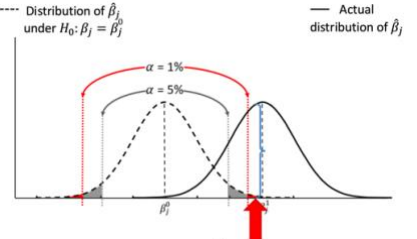
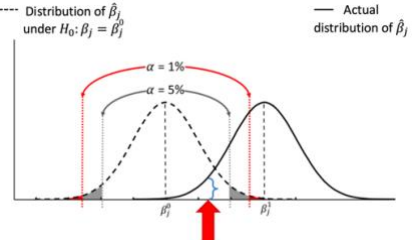
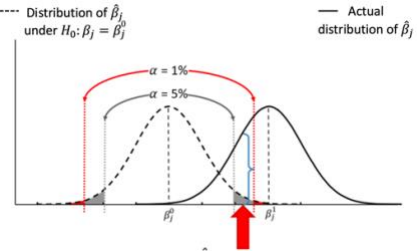
The probability of making a Type II error if  $\beta_j = \beta_j^1$  is given by the area under the actual distribution within the critical values of the null distribution



suppose the Null is untrue and that the true value of  $\beta_j$  is  $\beta_j^1$



If our estimate,  $\hat{\beta}_j$ , is here (quite likely), we will correctly reject  $H_0$  irrespective of whether  $\alpha = 1\%$  or  $5\%$

 <p>----- Distribution of <math>\hat{\beta}_j</math> under <math>H_0: \beta_j = \beta_j^0</math></p> <p>— Actual distribution of <math>\hat{\beta}_j</math></p> <p><math>\alpha = 1\%</math></p> <p><math>\alpha = 5\%</math></p> <p><math>\beta_j^0</math></p>	
 <p>----- Distribution of <math>\hat{\beta}_j</math> under <math>H_0: \beta_j = \beta_j^0</math></p> <p>— Actual distribution of <math>\hat{\beta}_j</math></p> <p><math>\alpha = 1\%</math></p> <p><math>\alpha = 5\%</math></p> <p><math>\beta_j^0</math></p>	<p>If our estimate, <math>\hat{\beta}_j</math>, is here (unlikely), we will erroneously fail to reject <math>H_0</math>, i.e., make a Type II error irrespective of whether <math>\alpha = 1\%</math> or <math>5\%</math></p>
 <p>----- Distribution of <math>\hat{\beta}_j</math> under <math>H_0: \beta_j = \beta_j^0</math></p> <p>— Actual distribution of <math>\hat{\beta}_j</math></p> <p><math>\alpha = 1\%</math></p> <p><math>\alpha = 5\%</math></p> <p><math>\beta_j^0</math></p>	<p>If our estimate, <math>\hat{\beta}_j</math>, is here (quite likely), we will:</p> <ol style="list-style-type: none"> <li>(1) reject <math>H_0</math> if <math>\alpha = 5\%</math>;</li> <li>(2) erroneously fail to reject <math>H_0</math>, i.e. make a Type II error, if <math>\alpha = 1\%</math></li> </ol> <p>so, in selecting <math>\alpha = 5\%</math> (rather than <math>\alpha = 1\%</math>), we make it a little more likely to commit a Type I error but much less likely to commit a Type II error.</p> <p>→ as the selected significance level is increased, e.g. from 1% to 5%, the likelihood of making a:</p> <ol style="list-style-type: none"> <li>(1) Type I error (erroneously reject <math>H_0</math>) increases.</li> <li>(2) Type II error (erroneously failing to reject <math>H_0</math>) declines.</li> </ol>

## Confidence intervals of $\beta_j$

### Definition

Is known as the **interval estimates** with the range of hypothetical values of  $\beta_j$  that is consistent with the sample estimate obtained using OLS.

The (100- $\alpha$ )% confidence interval for  $\beta_j = \hat{\beta}_j \pm [c \times se(\hat{\beta}_j)]$

A 95% confidence interval means that if we were **to repeat the sampling process** many times, **95% of the calculated intervals** would **contain the true parameter** value.

### Formula

$$-c < t < c$$

$$\rightarrow -c < \frac{(\hat{\beta}_j - \beta_j)}{se(\hat{\beta}_j)} < c$$

$$\rightarrow \hat{\beta}_j - [c \times se(\hat{\beta}_j)] < \beta_j < \hat{\beta}_j + [c \times se(\hat{\beta}_j)]$$

### Properties

- (1)  $\hat{\beta}_j \pm [c \times se(\hat{\beta}_j)]$ : symmetric around the estimate  $\hat{\beta}_j$
- (2) depends on  $c$ , which is linked to  $\alpha$ , the selected significance level
  - The smaller the significance level, the wider the interval.
  - In the t distribution, the actual critical value will depend on the degrees of freedom.
  - For large population, the t distribution will be asymptotically approach normal distribution, so use the normal distribution table when the population is large, and the confidence interval become:
    - (I) 95% confidence interval; 5% significance level;  $c \sim \pm 1.96$  (if  $n$  large)
    - (II) 99% confidence interval; 1% significance level;  $c \sim \pm 2.58$  (if  $n$  large)

(3) depends on  $se(\hat{\beta}_j)$

The more accurate our estimate = the smaller its  $se$ ,  
the narrower the confidence interval.

### **Interpretation**

#### Example

Based on the estimated model above, as long as the Gauss-Markov assumptions are valid, we are 99% confident that, ceteris paribus, forwards earn between 4.2 percent less and 36.5 percent more than players in other positions.



### Example

Assume *wages* (\$/hour) are determined by education and experience (both in years) and other unobservable factors, contained in  $u$ .

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u \quad (1)$$

Applying OLS to data for a sample of 28 people:

$$\widehat{wage} = -10.74 + 1.20educ + 0.17exper$$

4.94	0.33	0.07	s.e.s
(2.17)	(3.64)	(2.49)	<i>t</i> -ratios
0.040	0.001	0.019	p-values

95% confidence interval for  $\beta_2$

$$\hat{\beta}_2 \pm [c \times se(\hat{\beta}_2)] = 0.17 \pm [2.06 \times 0.07]$$

i.e.,

$$0.026 < \beta_2 < 0.314$$

## Assess regression model performance

### Decomposing the variation in the data

Regression analysis decomposes <b>each yi</b> in the <b>sample</b> into	
	$y_i = \hat{y}_i + \hat{u}_i$
fitted value $\hat{y}_i$	explained by the model
	a linear function of the explanatory variables $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \dots + \hat{\beta}_k x_{ki}$
Residual $\hat{u}_i$	Unexplained $\hat{u}_i = y_i - \hat{y}_i$

The <b>variation in yi</b> can be decomposed in a similar way	
$SST = SSE + SSR$ $1 = \frac{SSE}{SST} + \frac{SSR}{SST}$	
$SST = \sum_{i=1}^n (y_i - \bar{y})^2$	measures the total variation in yi
$SSE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	measures <b>the variation in <math>\hat{y}_i</math></b> (the variation in yi that is explained by the model)
$SSR = \sum_{i=1}^n (\hat{u}_i - \bar{u})^2 = \sum_{i=1}^n \hat{u}_i^2$	measures <b>the variation in <math>\hat{u}_i</math></b> (the variation in yi that is unexplained)

### Coefficient of determination

#### Definition

A measure of model performance or 'fit'

The ratio of the explained variation to total variation.

The fraction of the sample variation in y that is explained by x.

A higher  $R^2$  indicates a better 'fit' of the model to the data.

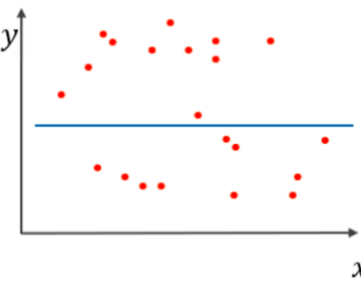
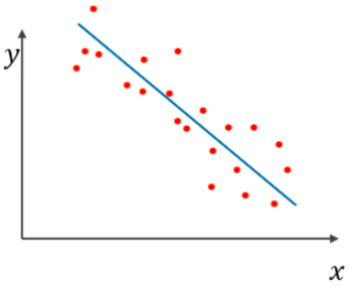
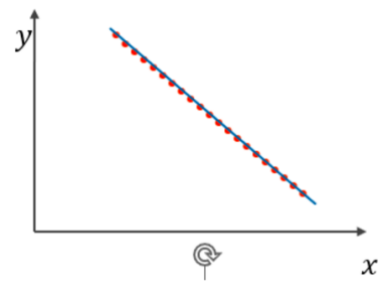
#### Formula

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (\hat{u}_i - \bar{u})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$$0 \leq R^2 \leq 1$$

### Example

If  $R^2 = 0.10$ , 10% of variation in  $y$  is explained by variation in the variables in  $x$ , the remainder (90%) by other factors ( $u$ ).

$R^2 = 0.00$ Model has no explanatory power	$R^2 = 0.85$	$R^2 = 1.00$ Model has 'perfect fit'
		

### Properties

$R^2$  encompasses the **co-linearity of the regressors**; we cannot attribute this part of the variation in  $y$  to either  $x_1$  alone or  $x_2$  alone, but we can attribute it to them both and, hence, to the model.

$R^2$  can be used to compare rival models as long as they have the **same dependent variable**.

### Problem

Including an **additional regressor** in a model never reduces and usually **increases  $R^2$** .

[Explanation] It can be shown that  $\overline{R^2}$  rises when a regressor is added to a model provided that **regressor's  $t$  ratio  $t\text{-ratio} = \frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)} > 1$** . So, adding statistically **insignificant** regressors may **increase  $\overline{R^2}$** .

### Formula

Adjustment  $\overline{R^2}$

$$\overline{R^2} = 1 - \left[ (1 - R^2) \times \frac{n - 1}{n - k - 1} \right]$$

### Properties

- (1) For any given model,  $\overline{R^2} < R^2$ .
- (2) As  $k$  **increases**, the **size of the adjustment increases**.
- (3)  $\overline{R^2}$  is often used to **compare** the fit of rival models with **different  $k$** .

## F test

Econometricians often also need to draw inferences about sets of coefficients as a group.

→ To do this we need a different test, the **F** test

### Definition

Let  $Q_1 \sim \chi^2_{(n_1)}$  and  $Q_2 \sim \chi^2_{(n_2)}$  be independent RVs then

$$F = \frac{Q_1/n_1}{Q_2/n_2}$$

is said to have *F distribution with  $n_1$  and  $n_2$  degrees of freedom*. We write

$$F \sim F_{(n_1, n_2)}.$$

T test	F test	
test a hypothesis about a single parameter or coefficient	test a hypothesis relating to 2 or more parameters ( $\beta_1, \beta_2, \beta_3 \dots$ ) at the same time	
	multiple hypothesis test <b>F</b> test for a set of exclusion restrictions, $q < k$	joint hypothesis test <b>F</b> test for the significance of the regression
Exclusion restriction	set of exclusion restrictions	
$H_0: \beta_1 = 0$	$H_0: \beta_{k-q+1} = 0, \beta_{k-q+2} = 0, \dots, \beta_k = 0$	$H_0: \beta_1 = 0, \beta_2 = 0, \beta_3 = 0 \dots$ $\leftrightarrow H_0: \beta_1, \beta_2, \beta_3, \dots = 0$ $\leftrightarrow H_0: \beta_1 = 0 \text{ and } \beta_2 = 0 \text{ and } \beta_3 = 0 \dots$ $\leftrightarrow$ there is no relationship between $x_1$ and $y$ , $x_2$ and $y$ , $x_3$ and $y \dots$ $\leftrightarrow H_0: x_1, x_2, \dots, x_k$ do not help explain $y$ $\leftrightarrow H_0: R^2 = 0$
$H_1: \beta_1 \neq 0$	$H_1: \beta_{k-q+1} \neq 0, \beta_{k-q+2} \neq 0, \dots, \beta_k \neq 0$ $\leftrightarrow H_1: \beta_{k-q+1} \neq 0 \text{ and/or } \beta_{k-q+2} \neq 0 \text{ and/or } \dots \text{ and/or } \beta_k \neq 0$ $\leftrightarrow$ there is a relationship between at least one of the $x$ variables ( $x_{k-q+1}, x_{k-q+2}, \dots$ and $x_k$ ) and $y$	$H_1: \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 \neq 0$ $\leftrightarrow H_1: \beta_1 \neq 0 \text{ and/or } \beta_2 \neq 0 \text{ and/or } \beta_3 \neq 0$ $\leftrightarrow H_1: \text{At least one of } \beta_1, \beta_2, \dots, \beta_k \neq 0$ $\leftrightarrow$ there is a relationship between at least one of the $x$ variables ( $x_1, x_2$ , and $x_3$ ) and $y$

	test whether the <b>last <math>q</math> regressors</b> in the model have no effect on $y$	When $R^2$ is small, it is useful to have a formal test of whether <b>the model</b> is explaining any of the variation in $y$ .
	$F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} = \frac{(R_u^2 - R_r^2)/q}{(1-R_u^2)/(n-k-1)}$	$F = \frac{(SSR_r - SSR_u)/k}{SSR_u/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$
<b>The relationship between t test and F test</b> (1) Apparent <b>disagreement</b> between $t$ tests and $F$ test is a sign of <b>multicollinearity</b> . (2) When groups of regressors <b>all measuring closely related dimensions of the same thing</b> and are <b>co-linear</b> , an <b><math>F</math> test</b> may be <b>superior</b> to a set of individual $t$ tests.		

$$F = \frac{(SSR_r - SSR_u)/k}{SSR_u/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

Proof.

In the model,  $0 \leq R^2 \leq 1$ , so when  $R^2=0$ , it means all the  $\beta_i$ 's cannot explain the  $y$ .

In  **$F$  test** for the significance of the regression, if  $H_0$  is true, then the restricted model has only  $\beta_0$ , all the other  $\beta_i = 0$ ; so now it equals to say:  $H_0: R_r^2 = 0$

For the unrestricted model, it is still  $R_u^2 = R^2$

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

$$F = \frac{(SSR_r - SSR_u)/k}{SSR_u/(n-k-1)} = \frac{(SSR_r/SST - SSR_u/SST)/k}{(SSR_u/SST)/(n-k-1)} = \frac{((1-R_r^2) - (1-R_u^2))/k}{(1-R_u^2)/(n-k-1)} = \frac{(R_u^2 - R_r^2)/k}{(1-R_u^2)/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

If  $H_0$  is true, i.e.  $H_0: R_r^2 = 0; R_u^2 = R^2$ .

Restricted model	Unrestricted model
one that embodies the exclusion restrictions	one that does not embody the exclusion restrictions
test whether the last $q$ regressors in the model have no effect on $y$	
$H_0: \beta_{k-q+1} = 0, \beta_{k-q+2} = 0, \dots, \beta_k = 0$	
$H_1: \beta_{k-q+1} \neq 0, \beta_{k-q+2} \neq 0, \dots, \beta_k \neq 0$ $\leftrightarrow H_1: \beta_{k-q+1} \neq 0 \text{ and/or } \beta_{k-q+2} \neq 0 \text{ and/or } \dots \text{ and/or } \beta_k \neq 0$ $\leftrightarrow$ there is a relationship between at least one of the $x$ variables ( $x_{k-q+1}, x_{k-q+2}, \dots$ and $x_k$ ) and $y$	
If $H_0$ is true: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_{k-q} x_{k-q} + u$ (which has $q$ fewer parameters)	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k + u$
$SSR_r$ = sum of the squared residuals for the restricted model $\beta_i$ 的数量变了, $\hat{\beta}_i$ 的数量变了, $u_i$ 和 $\hat{u}_i$ 也随之改变 但是 $x_i$ 的数量没变, sample size 没变, $n$ 没变	$SSR_u$ = sum of the squared residuals for the unrestricted model
Test $H_0: \beta_{k-q+1} = 0, \beta_{k-q+2} = 0, \dots, \beta_k = 0$ comparing the <b>fit</b> of the restricted model to the <b>fit</b> of the unrestricted model $\leftrightarrow$ compare <b>how much</b> of the variation in $y$ each model <b>leaves unexplained</b> , $SSR = \sum_{i=1}^n (\hat{u}_i - \bar{u})^2 = \sum_{i=1}^n (\hat{u}_i)^2$	
The $F$ statistic measures the proportional increase in the $SSR$ when the restrictions set out in $H_0$ are applied, with adjustments made for the number of restrictions and degrees of freedom.	
$F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n - k - 1)}$	
Numerator = $q$ = number of restrictions imposed under $H_0$ Denominator = $n - k - 1$ = df in the unrestricted model	
(1) Since $SSR_r > SSR_u$ , the $F$ statistic is always $> 0$	
<b>1. Model Flexibility:</b> <u>The unrestricted model has more parameters, allowing it to fit the data more flexibly.</u> This generally results in a smaller sum of squared residuals ( $SSR_u$ ) because it minimizes residuals without constraints.	
<b>2. Constraints in the Restricted Model:</b> <u>The restricted model imposes constraints on some parameters (e.g., setting them to zero). These constraints reduce the model's flexibility, which often leads to larger residuals and hence a larger <math>SSR_r</math>.</u>	
<b>3. Optimization Principle:</b> The unrestricted model minimizes the residuals as part of its optimization process, so $SSR_r$ is the smallest possible value achievable with the given data and model structure. <u>Adding restrictions (as in <math>SSR_r</math>) can only increase or maintain the sum of squared residuals, not reduce it.</u>	

Therefore, restricted model has less estimator, in general, it has more part that remains unexplained.

(2) If  $H_0$  is true,  $SSR_r \cong SSR_u$  and  $F \cong 0$

$$F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} \sim F_{q, n-k-1}$$

有没有  $\beta_{k-q+1}, \beta_{k-q+2}, \dots, \beta_k$ , 无法解释的部分都变化不大, 说明被解释的部分变化也不大, 说明  $\beta_{k-q+1}, \beta_{k-q+2}, \dots, \beta_k$  是很小的, 不足以过多影响  $y$ , 说明因素  $x_{k-q+1}, x_{k-q+2}, \dots$  and  $x_k$  不足以解释  $y$ , 说明 fail to reject  $H_0, H_0$  is true.

(3) The larger the difference between  $SSR_r$  and  $SSR_u$ , the larger is  $F$

A large  $F$  indicates that the explanatory power of the restricted model is poor relative to that of the unrestricted model, implying  $H_0$  is false and should be rejected.

#### [Logic]

$SSR_r - SSR_u$  is large

→  $SSR$  increase a lot after the estimators  $\beta_{k-q+1}, \beta_{k-q+2}, \dots, \beta_k$  has been removed

→ the unexplained proportion increases a lot after the estimators  $\beta_{k-q+1}, \beta_{k-q+2}, \dots, \beta_k$  has been removed

→  $\beta_{k-q+1}, \beta_{k-q+2}, \dots, \beta_k$  as a group is significant to explain  $y$

→ reject  $H_0$

Decision rule

Select the **significnat level,  $\alpha$**

reject  $H_0$  when  $F$  is larger than the **critical value** implied by our pre-selected significance level,  $\alpha$ .

↔ reject  $H_0$  if  $F > c$

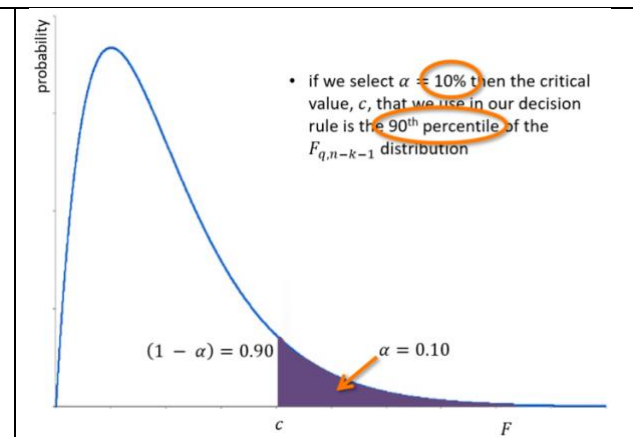
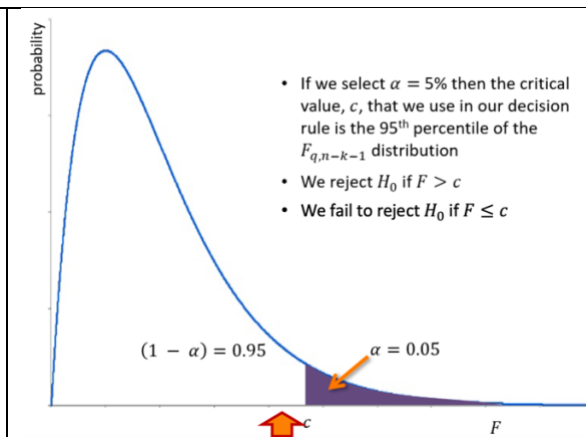
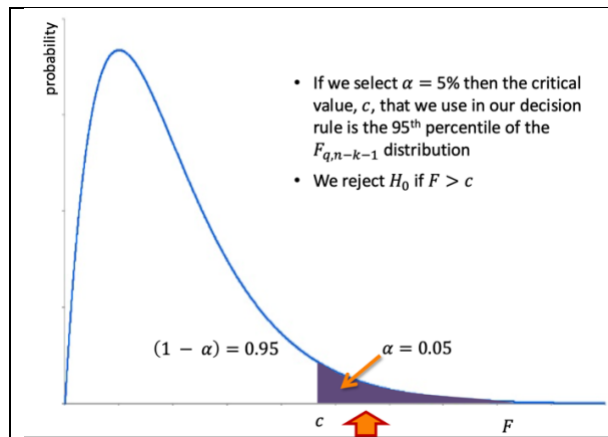
C, critical value depends on:

significance level (5, 10 or 1% are usual)
number of restrictions ( $q$ )
degrees of freedom when estimating the unrestricted model ( $n - k - 1$ )

Example

If we pre-selected  $\alpha = 5\%$  then the critical value,  $c$ , that we use in our decision rule is the 95th percentile of the  $F_{q, n-k-1}$  distribution.





### Conducting F tests

Formulate the **hypothesis** (Null and Alternative)

Select the **significance level**

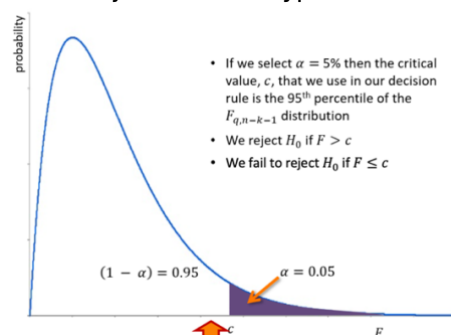
Find the **critical value** and use it to set up the **Decision Rule**

Calculate the **test statistic**, i.e., the  $F$  statistic

**Conduct the test**, i.e., apply the Decision Rule, i.e., compare the  $F$  statistic and the critical value

Draw **conclusion**

Formulate the <b>hypothesis</b> (Null and Alternative)	<b>Hypothesis (hypothesis for the overall non-linearity)</b>	
	<b>Null Hypothesis (<math>H_0</math>):</b> $\gamma_1 = \gamma_2 = 0$ (No quadratic relationship exists; the relationship is purely linear.)  <b>Alternative Hypothesis (<math>H_1</math>):</b> $\gamma_1 \neq 0$ and/or $\gamma_2 \neq 0$ (The relationship between profitability and female share in the directorate is non-linear.)	
The restricted and unrestricted models identified	<b>Restricted model</b>	<b>Unrestricted model</b>
	(specific model illustrated)	(specific model illustrated)
	$R_R^2 =$ _____	$R_U^2 =$ _____
	n = _____	n = _____
Degrees of freedom calculated	k = _____	k = _____
	<b>Df1=</b> _____	
	<b>Df2=</b> _____	
Select the <b>significance level</b>	Select a significance level, $\alpha = 5\%$ , which implies that we are willing to mistakenly reject $H_0$ 5% of the time.	

Find the <b>critical value</b> and use it to set up the <b>Decision Rule</b>	Based on the degrees of freedom (2 and 831), the critical F-value at $\alpha = 0.05$ is <b>3.007</b> . If F-statistic > critical value or p-value < $\alpha$ , reject $H_0$ . Otherwise, fail to reject $H_0$ .
Calculate the <b>test statistic</b> , i.e., the $F$ statistic	$F_{(2,831)} = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)}$ or $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$ (from the STATA F test result) q = number of restrictions n = number of observations k = number of predictors in the unrestricted model $R^2$ = the coefficient of determination for model under consideration (e.g., the unrestricted model).
<b>Conduct the test</b> , i.e., apply the Decision Rule, i.e., compare the $F$ statistic and the critical value	$F_{(2,831)} = 1.51 < 3.007$ Fail to reject the null hypothesis.  <ul style="list-style-type: none"> <li>• If we select <math>\alpha = 5\%</math> then the critical value, <math>c</math>, that we use in our decision rule is the 95<sup>th</sup> percentile of the <math>F_{q,n-k-1}</math> distribution</li> <li>• We reject <math>H_0</math> if <math>F &gt; c</math></li> <li>• We fail to reject <math>H_0</math> if <math>F \leq c</math></li> </ul>
Draw <b>conclusion</b>	The results suggest that the coefficients of pfdir and pfdirsquare are not jointly significant in explaining the dependent variable (e.g., profitability). This implies that there is no strong evidence to suggest a relationship (linear or quadratic) between the female share in the directorate and the dependent variable in this model.

### Example

Are US baseball players' salaries affected by their performance?

$$\ln(\text{sal}) = \beta_0 + \beta_1 \text{yrs} + \beta_2 \text{gmsy} + \beta_3 \text{bavg} + \beta_4 \text{hrunsy} + \beta_5 \text{rbisy} + u$$

*sal* = annual salary in USD

*yrs* = years in league

*gmsy* = average number of games played per year

*bavg* = career batting average

*hrunsy* = number of home runs per year

*rbisy* = number of runs batted per year

**Formulate the hypothesis** that corresponds to the question and the model

$H_0: \beta_3, \beta_4, \beta_5 = 0$ , i.e., performance has no effect on salary

$H_1: \beta_3 \neq 0$  and/or  $\beta_4 \neq 0$  and/or  $\beta_5 \neq 0$ , i.e., at least one aspect of performance has an effect on salary

**Test required:** a joint test of three exclusion restrictions, i.e., an  $F$  test

**Selected significance level:**  $\alpha = 5\%$

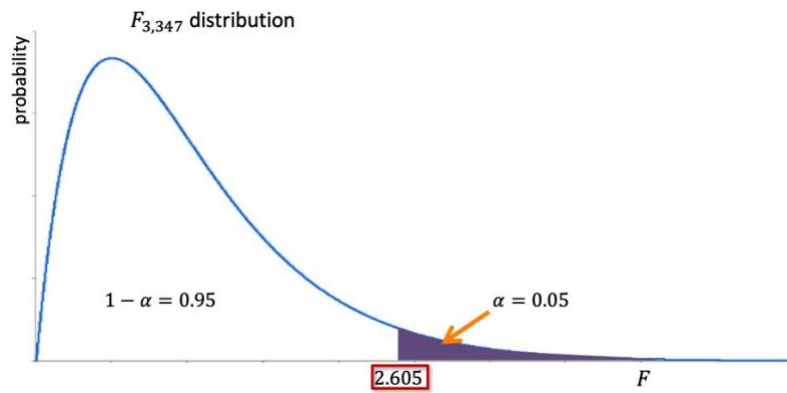
Applying OLS yields:

$$\ln(\widehat{\text{sal}}) = 11.19 + 0.689\text{yrs} + 0.126\text{gmsy} + 0.0098\text{bavg} + 0.0144\text{hrunsy} + 0.0108\text{rbisy}$$

$$\begin{array}{cccccc} (0.29) & (0.121) & (0.0026) & (0.0110) & (0.0161) & (0.0072) \\ [38.5] & [5.69] & [48.46] & [0.89] & [0.89] & [1.50] \end{array}$$

$$n = 353, \quad SSR = 183.186 \quad R^2 = 0.6278$$

$$df: 3, 347 (= 353 - 5 - 1)$$



Applying OLS to this yields:

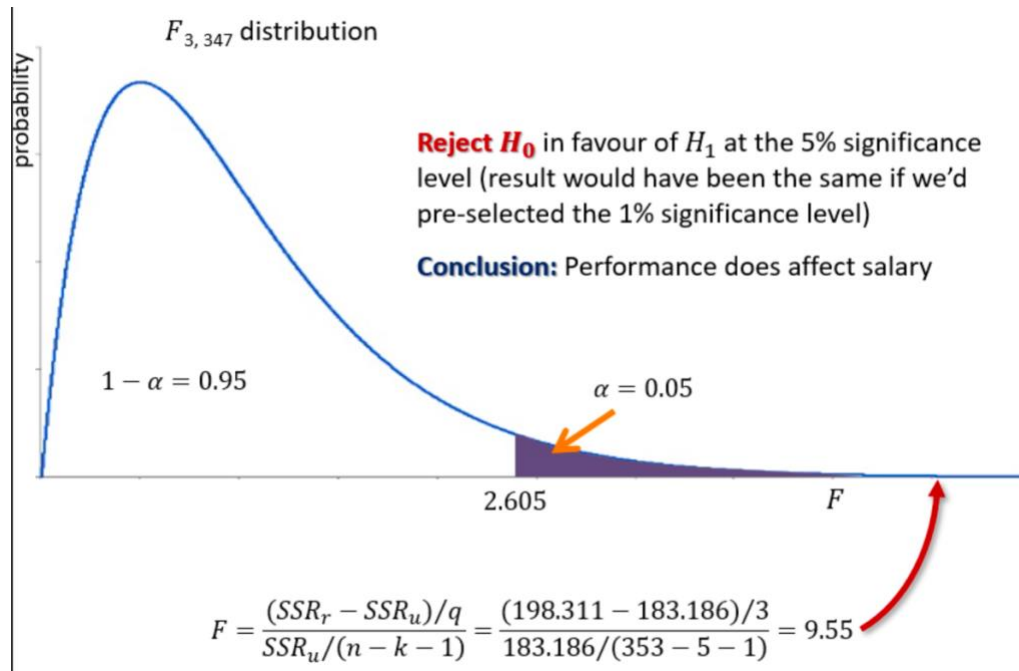
$$\ln(\widehat{sal}) = 11.22 + 0.713\text{yrs} + 0.202\text{gmsy}$$

$$(0.11) \quad (0.012) \quad (0.0013)$$

$$n = 353, \quad SSR = 198.311 \quad R^2 = 0.5971$$

So,

$$F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n - k - 1)} = \frac{(198.311 - 183.186)/3}{183.186/(353 - 5 - 1)} = 9.55$$



$H_0: \beta_3, \beta_4, \beta_5 = 0$ , i.e., performance has no effect on salary, is **rejected**

**Conclusion:** Performance does affect salary...

Even though:

- the coefficient on *bavg* is insignificant;
- the coefficient on *hrunsy* is insignificant; and
- the coefficient on *rbisy* is insignificant

**[logic]**

In t test, H0 for each individual estimator is insignificant because according to the critical value, H0 does not be rejected, this means estimators individually has almost no effect on x explaining to y.

However, F test tells us that H0 is rejected, means three estimators together can be used to make x explaining y.

This is the conflict between the results from t test and F test.

The reason is these three estimators are highly multicollinearity.

Recall:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^n (x_{ji} - \bar{x}_j)^2 (1 - R_j^2)} \quad \text{for } j = 1, 2, \dots, k$$

$R_j^2$  is a measure of the correlation between  $x_j$  and all of the other explanatory variables, i.e., all the other  $x$ 's,  $0 \leq R_j^2 \leq 1$ .

When estimators  $\hat{\beta}_i$  has high multicollinearity, then  $R_j^2 \rightarrow 1$ .

So  $Var(\hat{\beta}_i)$  is very high, so  $se(\hat{\beta}_i)$  is very large, so  $t = \frac{(\hat{\beta}_j - \beta_j)}{se(\hat{\beta}_j)}$  will be very small, so  $\hat{\beta}_i$  is insignificant, which may be different from the results of F test.

Graphically explanation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

