



经济学类各专业核心课程

# 计量经济学



# 计量经济学

## 第二章

### 简单线性回归模型

# 第二章 简单线性回归模型

## 本章主要讨论：

- 回归分析与回归函数
- 简单线性回归模型参数的估计
- 拟合优度的度量
- 回归系数的区间估计和假设检验
- 回归模型预测

# 第一节 回归分析与回归方程

## 本节基本内容:

- 回归与相关
- 总体回归函数
- 随机扰动项
- 样本回归函数

# 一、回归与相关（对统计学的回顾）

## 1. 经济变量间的相互关系

◆ 确定性的函数关系  $Y = f(X)$

◆ 不确定性的统计关系—相关关系

$$Y = f(X) + \varepsilon \quad (\varepsilon \text{ 为随机变量})$$

◆ 没有关系

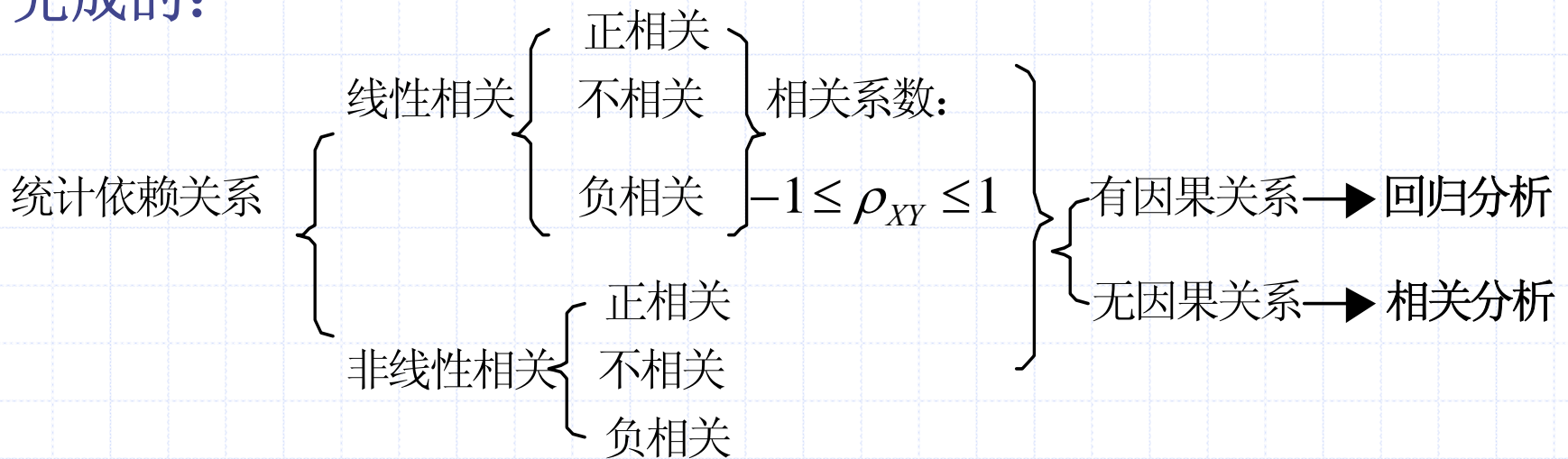
例如:

函数关系: 圆面积 =  $f(\pi, \text{半径}) = \pi \cdot \text{半径}^2$

统计依赖关系/统计相关关系:

农作物产量 =  $f(\text{气温}, \text{降雨量}, \text{阳光}, \text{施肥量})$

对变量间统计依赖关系的考察主要是通过**相关分析 (correlation analysis)**或**回归分析 (regression analysis)**来完成的:



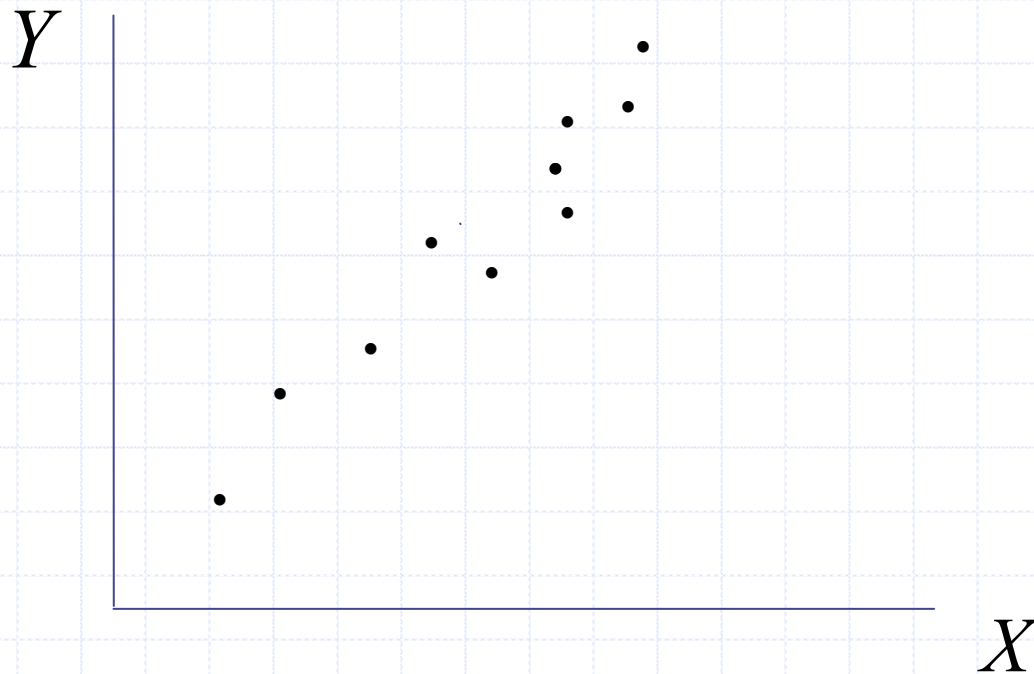
## ▲ 注意:

- ① 非线性相关并不意味着不相关;
- ② 有相关关系并不意味着一定有因果关系;
- ③ 相关分析对称地对待任何（两个）变量，两个变量都被看作是随机的。回归分析对变量的处理方法存在不对称性，即区分应变量（被解释变量）和自变量（解释变量）：前者是随机变量，后者不是。

## 2. 相关关系

### ◆ 相关关系的描述

相关关系最直观的描述方式——坐标图（散布图）





### 3. 相关程度的度量——相关系数

总体线性相关系数:

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

其中:  $\text{Var}(X)$ —— $X$ 的方差;  $\text{Var}(Y)$ —— $Y$ 的方差

$\text{Cov}(X, Y)$ —— $X$ 和 $Y$ 的协方差

样本线性相关系数:

$$r_{xy} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

其中:  $X_i$ 和  $Y_i$  分别是变量  $X$  和  $Y$  的样本观测值

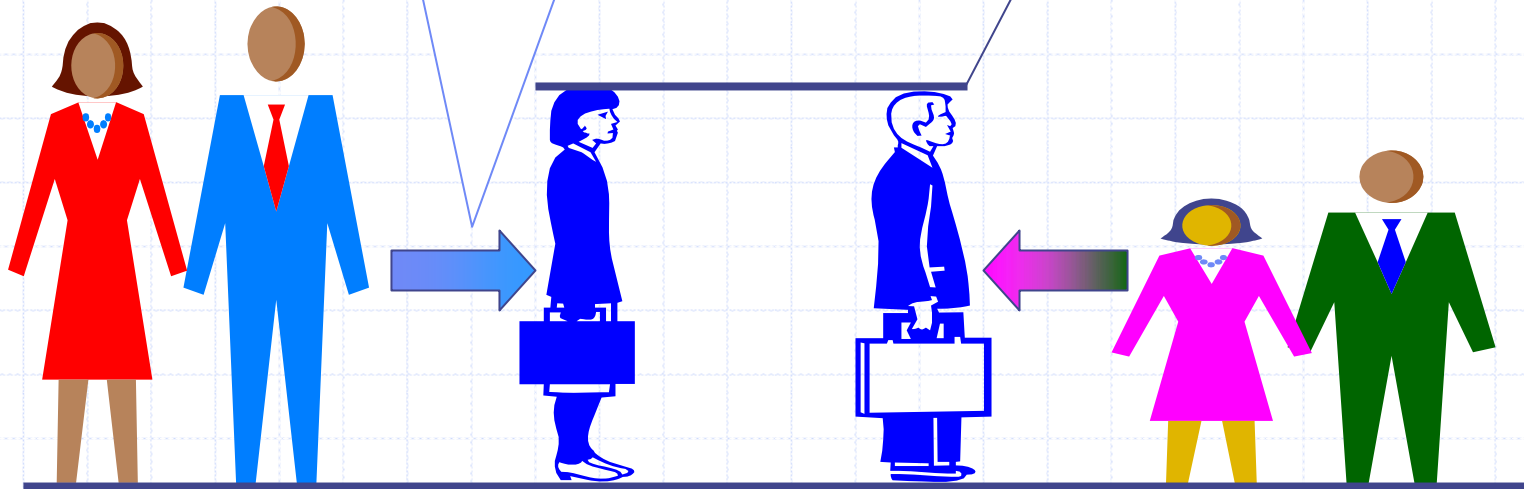
$\bar{X}$  和  $\bar{Y}$  分别是变量  $X$  和  $Y$  样本值的平均值

# 4. 回归分析

回归分析(regression):通过一个或几个变量的变化去解释另一变量的变化。包括找出自变量与因变量、设定数学模型、检验模型、估计预测等环节。

回归：退回  
regression

平均身高



1877年 弗朗西斯·高尔顿爵士 遗传学研究 回归线

回归分析(regression):通过一个或几个变量的变化去解释另一变量的变化。包括找出自变量与因变量、设定数学模型、检验模型、估计预测等环节。

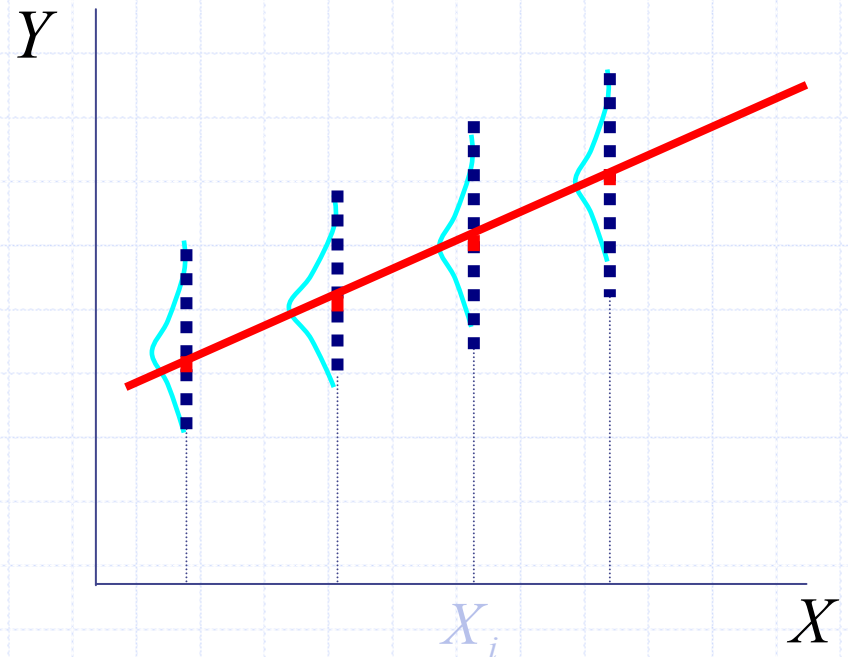
# 注意几个概念

- $Y$  的条件分布

当解释变量  $X$  取某固定值时（条件）， $Y$  的值不确定， $Y$  的不同取值形成一定的分布，即  $Y$  的条件分布。

- $Y$  的条件期望

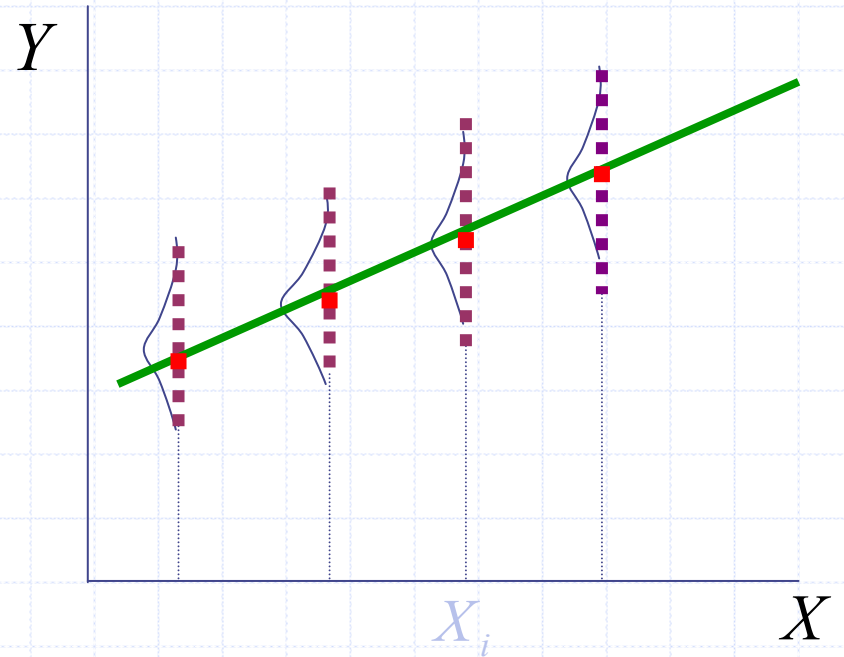
对于  $X$  的每一个取值，对  $Y$  所形成的分布确定其期望或均值，称为  $Y$  的条件期望或条件均值  $E(Y|X_i)$



# 回归线与回归函数

## ● 回归线:

对于每一个  $X$  的取值，  
都有  $Y$  的条件期望  
 $E(Y|X_i)$  与之对应，  
代表这些  $Y$  的条件期望的点的轨迹所形成的直线或曲线，称为回归线。



# 回归线与回归函数

回归函数：应变量  $Y$  的条件期望  $E(Y|X_i)$  随解释变量  $X$  的变化而有规律的变化，如果把  $Y$  的条件期望  $E(Y|X_i)$  表现为  $X$  的某种函数

$$E(Y | X_i) = f(X_i)$$

这个函数称为回归函数。

回归函数分为：总体回归函数和样本回归函数

举例：假如已知**100**个家庭构成的总体。



# 例:100个家庭构成的总体 (单位:元)

	每月家庭可支配收入 $X$									
	1000	1500	2000	2500	3000	3500	4000	4500	5000	5500
每月家庭消费支出 $Y$	820	962	1108	1329	1632	1842	2037	2275	2464	2824
	888	1024	1201	1365	1726	1874	2110	2388	2589	3038
	932	1121	1264	1410	1786	1906	2225	2426	2790	3150
	960	1210	1310	1432	1835	1068	2319	2488	2856	3201
		1259	1340	1520	1885	2066	2321	2587	2900	3288
		1324	1400	1615	1943	2185	2365	2650	3021	3399
			1448	1650	2037	2210	2398	2789	3064	
			1489	1712	2078	2289	2487	2853	3142	
			1538	1778	2179	2313	2513	2934	3274	
			1600	1841	2298	2398	2538	3110		
		1702	1886	2316	2423	2567				
			1900	2387	2453	2610				
			2012	2498	2487	2710				
				2589	2586					
$E(Y X_i)$	900	1150	1400	1650	1900	2150	2400	2650	2900	3150



## 二、总体回归函数 (PRF)

### 1. 总体回归函数的概念

前提：假如已知所研究的经济现象的总体应变  
量  $Y$  和解释变量  $X$  的每个观测值，可以计算出总体  
应变变量  $Y$  的条件均值  $E(Y|X_i)$ ，并将其表现为解释  
变量  $X$  的某种函数

$$E(Y|X_i) = f(X_i)$$

这个函数称为总体回归函数 (**PRF**)

## 2. 总体回归函数的表现形式

### (1) 条件均值表现形式

假如  $Y$  的条件均值  $E(Y|X_i)$  是解释变量  $X$  的线性函数，可表示为：

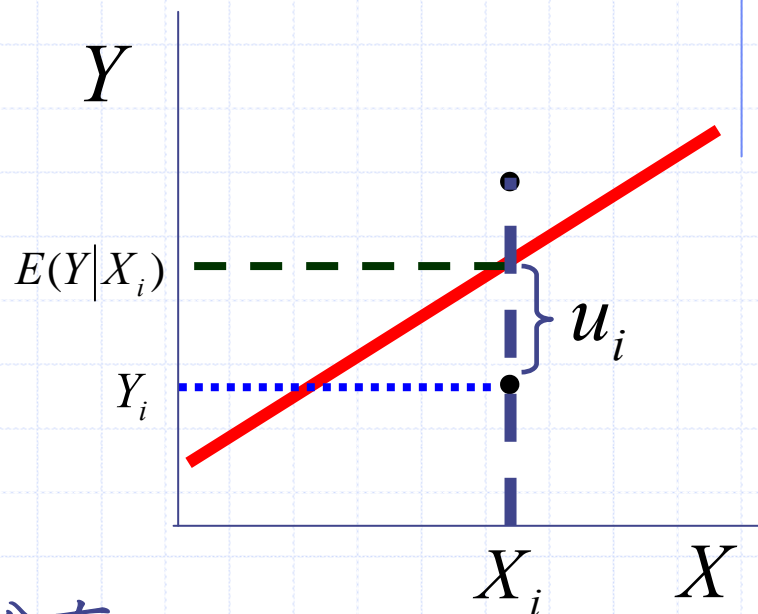
$$E(Y_i|X_i) = f(X_i) = \beta_1 + \beta_2 X_i$$

### (2) 个别值表现形式

对于一定的  $X_i$ ， $Y$  的各个个别值  $Y_i$  分布在  $E(Y|X_i)$  的周围，若令各个  $Y_i$  与条件均值  $E(Y|X_i)$  的偏差为  $u_i$ ，显然  $u_i$  是随机变量，则有

$$u_i = Y_i - E(Y_i|X_i) = Y_i - \beta_1 - \beta_2 X_i$$

或 
$$Y_i = \beta_1 + \beta_2 X_i + u_i$$



## ◆ 二、总体回归函数

在给定解释变量 $X_i$ 条件下被解释变量 $Y_i$ 的期望轨迹称为**总体回归线** (population regression line)，或更一般地称为**总体回归曲线** (population regression curve)。

相应的函数：

$$E(Y | X_i) = f(X_i)$$

称为（双变量）**总体回归函数** (population regression function, **PRF**)。

## ◆含义:

回归函数（PRF）说明被解释变量Y的平均状态（总体条件期望）随解释变量X变化的规律。

- 函数形式： 可以是线性或非线性的。

例2.1中，将居民消费支出看成是其可支配收入的线性函数时：

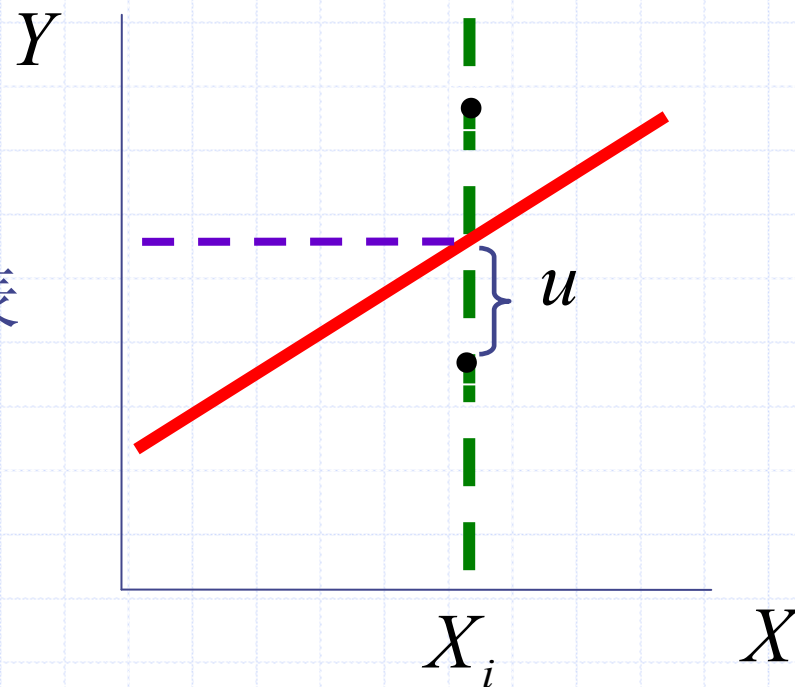
$$E(Y | X_i) = \beta_0 + \beta_1 X_i$$

为一线性函数。其中， $\beta_0$ ， $\beta_1$ 是未知参数，称为**回归系数**（regression coefficients）。

# 三、随机扰动项

## ◆ 概念:

各个  $Y_i$  值与条件均值  $E(Y|X_i)$  的偏差  $u_i$  代表排除在模型以外的所有因素对  $Y$  的影响。



◆ 性质:  $u_i$  是期望为  $0$  有一定分布的随机变量

重要性: 随机扰动项的性质决定着计量经济方法的选择

# 引入随机扰动项的原因

- 理论的模糊性
- 数据的缺失
- 核心变量与周边变量
- 糟糕的替代变量
- 人类行为的内在随机性
- 节省原则
- 错误的函数形式

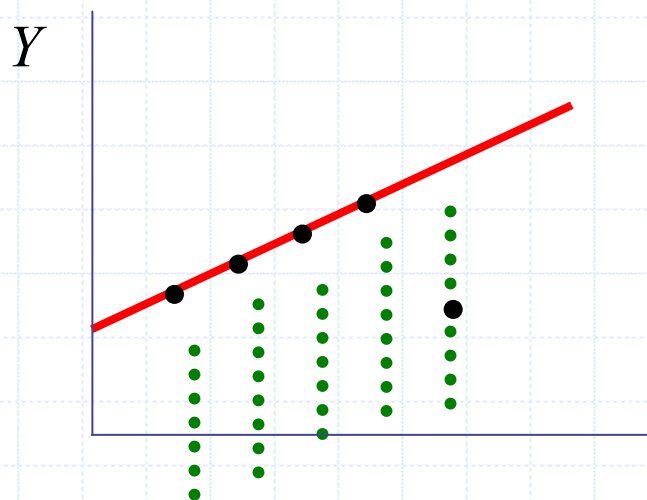
## 四、样本回归函数（SRF）

样本回归线：

对于  $X$  的一定值，取得  $Y$  的样本观测值，可计算其条件均值，样本观测值条件均值的轨迹称为样本回归线。

样本回归函数：

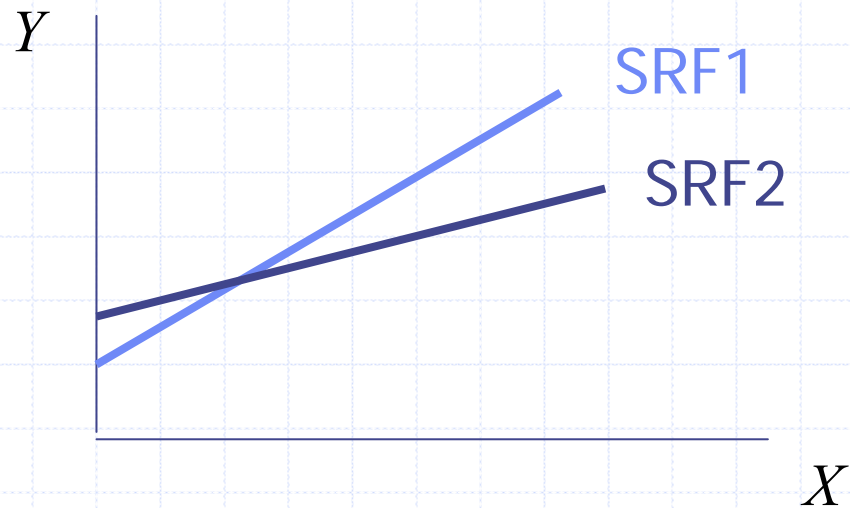
如果把应变量  $Y$  的样本条件均值表示为解释变量  $X$  的某种函数，这个函数称为样本回归函数（SRF）。





# SRF 的特点

- 每次抽样都能获得一个样本，就可以拟合一条样本回归线，所以样本回归线随抽样波动而变化，可以有許多条（SRF不唯一）。





# 样本回归函数的表现形式

样本回归函数如果为线性函数，可表示为

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

其中： $\hat{Y}_i$  是与  $X_i$  相对应的  $Y$  的样本条件均值  
 $\hat{\beta}_1$  和  $\hat{\beta}_2$  分别是样本回归函数的参数

应变量  $Y$  的实际观测值  $Y_i$  不完全等于样本条件均值，二者之差用  $e_i$  表示， $e_i$  称为剩余项或残差项：

$$e_i = Y_i - \hat{Y}_i$$

或者

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$$

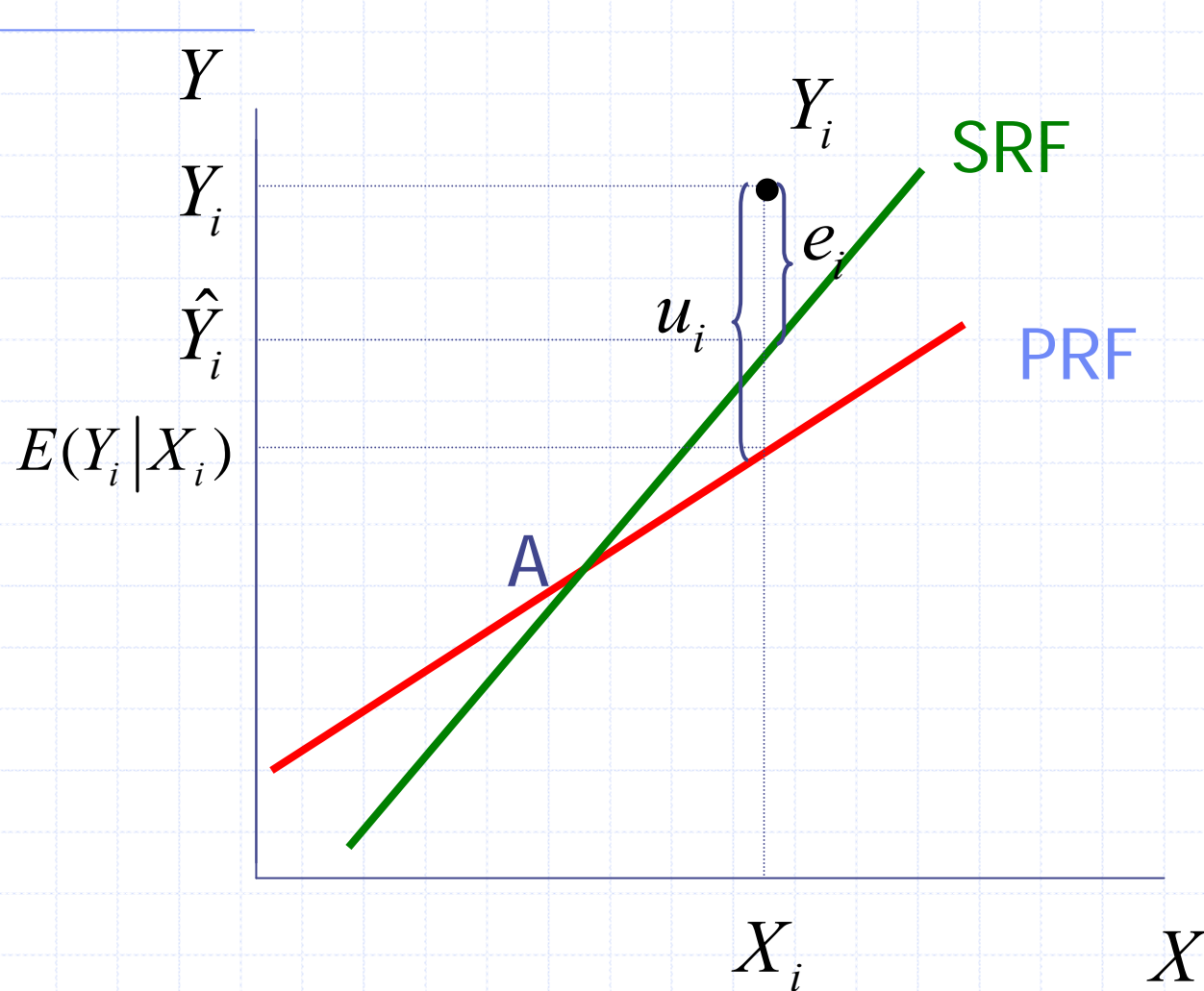
# 对样本回归的理解

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$$

如果能够获得  $\hat{\beta}_1$  和  $\hat{\beta}_2$  的数值，显然：

- $\hat{\beta}_1$  和  $\hat{\beta}_2$  是对总体回归函数参数  $\beta_1$  和  $\beta_2$  的估计
- $\hat{Y}_i$  是对总体条件期望  $E(Y|X_i)$  的估计
- $e_i$  在概念上类似总体回归函数中的  $u_i$ ，可视为对  $u_i$  的估计。

# 样本回归函数与总体回归函数的关系



## 四、样本回归函数 (SRF)

将样本回归线看成总体回归线的近似替代

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$



$$\begin{aligned} Y_i &= E(Y | X_i) + \mu_i \\ &= \beta_0 + \beta_1 X_i + \mu_i \end{aligned}$$

则  $\hat{Y}_i$  为  $E(Y | X_i)$  的估计量;  
 $\hat{\beta}_i$  为  $\beta_i$  的估计量,  $i = (0,1)$

## 样本回归函数的随机形式/样本回归模型：

同样地，样本回归函数也有如下的随机形式：

$$Y_i = \hat{Y}_i + \hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$$

式中， $e_i$  称为 **（样本）残差（或剩余）项** (residual)，代表了其他影响  $Y_i$  的随机因素的集合，可看成是  $\mu_i$  的估计量  $\hat{\mu}_i$ 。

由于方程中引入了随机项，成为计量经济模型，因此也称为 **样本回归模型** (sample regression model)。

▼ **回归分析的主要目的：** 根据样本回归函数SRF，估计总体回归函数PRF。

即，根据  $Y_i = \hat{Y}_i + e_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$

估计  $Y_i = E(Y | X_i) + \mu_i = \beta_0 + \beta_1 X_i + \mu_i$

这就要求：

设计一“方法”构造 SRF，以使 SRF 尽可能“接近” PRF，或者说使  $\hat{\beta}_i (i = 0,1)$  尽可能接近  $\beta_i (i = 0,1)$ 。

**注意：** 这里PRF可能永远无法知道。

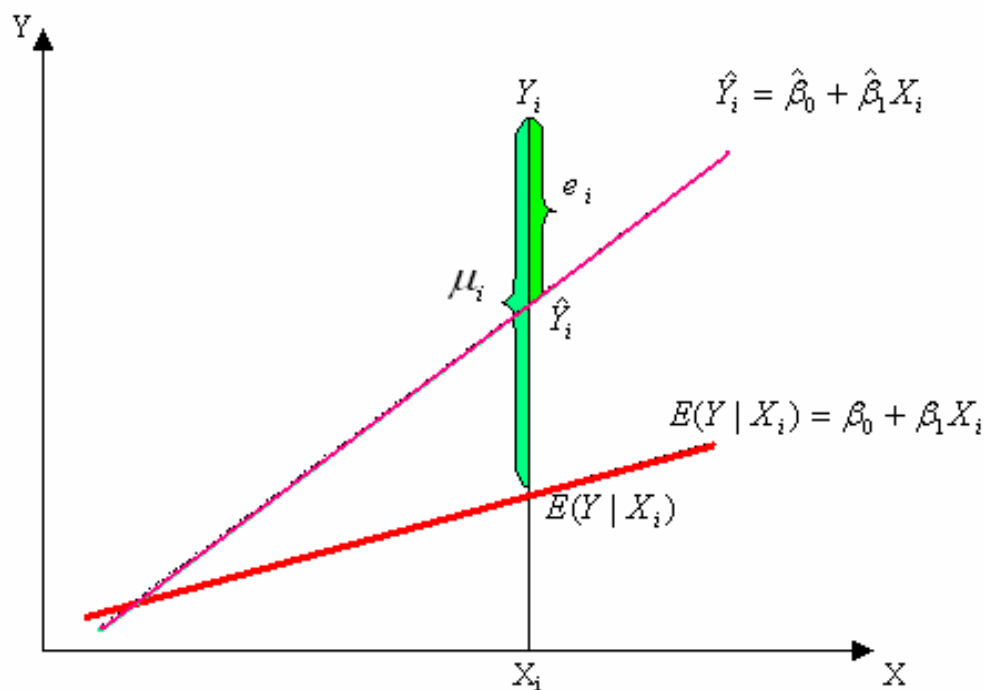


图 2.1.3 总体回归线与样本回归线的基本关系

## 第二节 简单线性回归模型的最小二乘估计

### 本节基本内容:

- 简单线性回归的基本假定
- 普通最小二乘法
- OLS回归线的性质
- 参数估计式的统计性质



# 一、简单线性回归的基本假定

## 1. 为什么要作基本假定？

- 模型中有随机扰动，估计的参数是随机变量，只有对随机扰动的分布作出假定，才能确定所估计参数的分布性质，也才可能进行假设检验和区间估计
- 只有具备一定的假定条件，所作出的估计才具有较好的统计性质。



## 2、基本假定的内容

### (1) 对模型和变量的假定

如 
$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

假定解释变量  $X$  是非随机的，或者虽然是随机的，但与扰动项  $u$  是不相关的

假定解释变量  $X$  在重复抽样中为固定值

假定变量和模型无设定误差

## (2) 对随机扰动项 $u$ 的假定

又称高斯假定、古典假定

假定1：零均值假定

在给定  $X$  的条件下， $u_i$  的条件期望为零

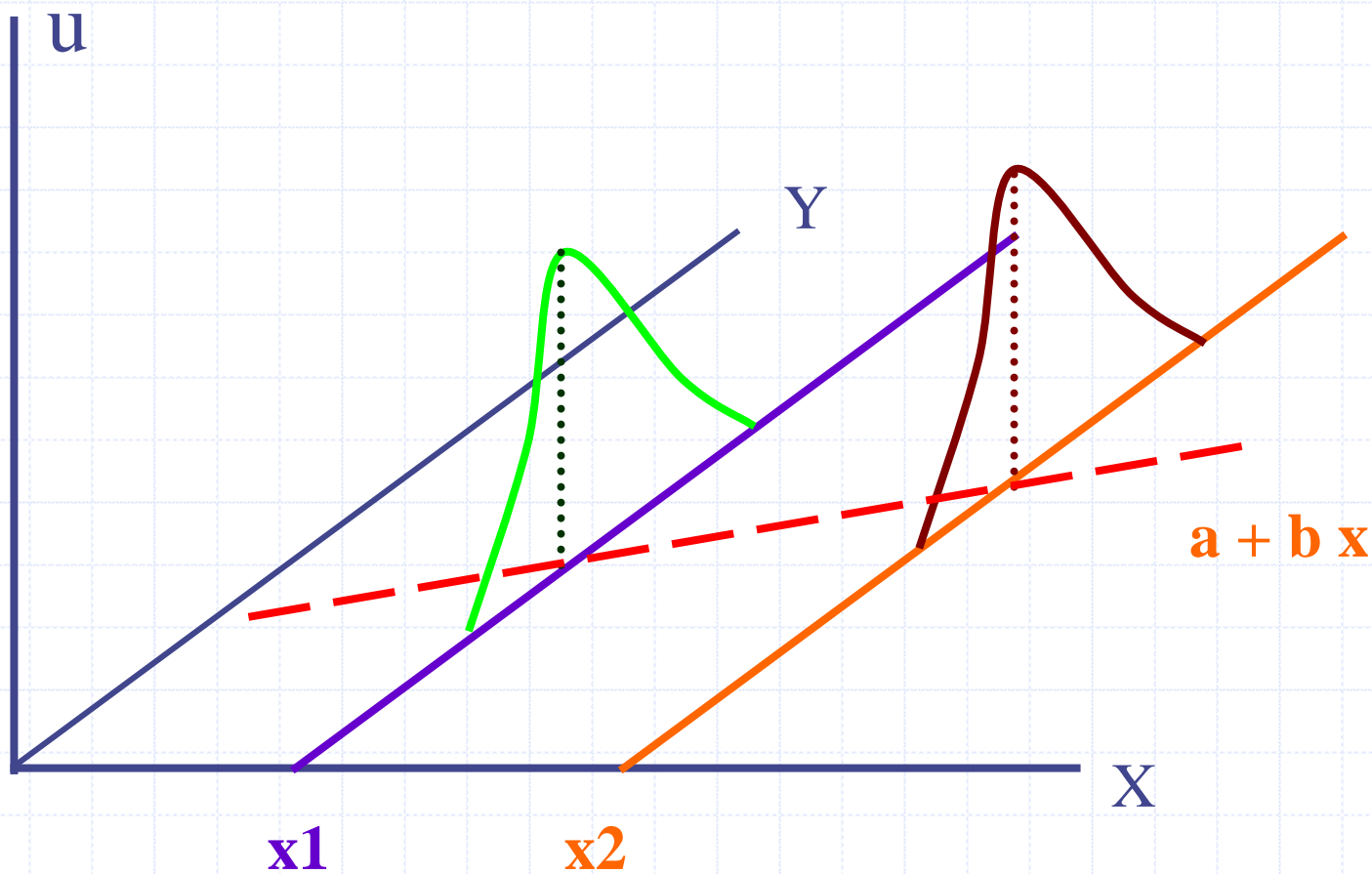
$$E(u_i | X_i) = 0$$

假定2：同方差假定

在给定  $X$  的条件下， $u_i$  的条件方差为某个常数  $\sigma^2$

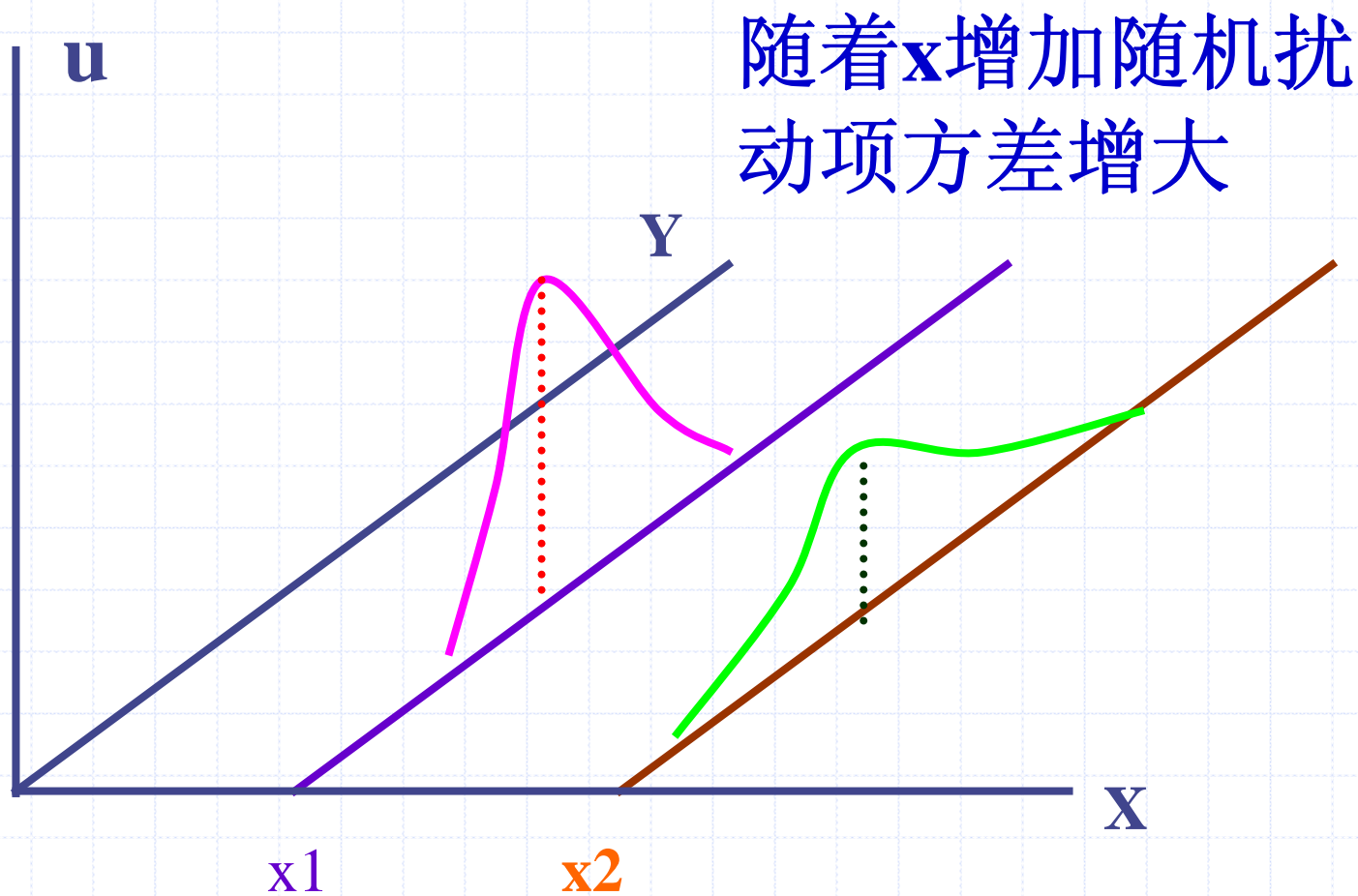
$$\text{Var}(u_i | X_i) = E[u_i - E(u_i | X_i)]^2 = \sigma^2$$

# 同方差



随着 $x$ 变化随机扰动项 $u$ 的方差不变<sup>25</sup>

# 异方差



假定3：无自相关假定

随机扰动项  $u_i$  的逐次值互不相关

$$\begin{aligned} \text{Cov}(u_i, u_j) &= E[u_i - E(u_i)][u_j - E(u_j)] \\ &= E(u_i u_j) = 0 \quad (i \neq j) \end{aligned}$$

假定4：随机扰动  $u_i$  与解释变量  $X$  不相关

$$\text{Cov}(u_i, X_i) = E[u_i - E(u_i)][X_i - E(X_i)] = 0$$

假定5：对随机扰动项分布的正态性假定

即假定  $u_i$  服从均值为零、方差为  $\sigma^2$  的正态分布

$$u_i \sim N(0, \sigma^2)$$

（说明：正态性假定不影响对参数的点估计，但对确定所估计参数的分布性质是需要的。且根据中心极限定理，当样本容量趋于无穷大时， $u_i$  的分布会趋近于正态分布。所以正态性假定是合理的）

## 假设 6 数据产生过程是线性的 (Linearity of the Model)

$$\diamond y_i = a + bx_i + u_i \quad (i=1, 2, \dots, n)$$

◆ 因变量  $y_i$  = 自变量的线性组合再加上一个随机扰动项。自然，因变量  $y_i$  也是一个随机变量，于是必须对  $y_i$  的分布做一番讨论。

◆ 如果是非线性就不能采用最小二乘法。

# $Y$ 的分布性质

由于  $Y_i = \beta_1 + \beta_2 X_i + u_i$

$u_i$  的分布性质决定了  $Y_i$  的分布性质。

对  $u_i$  的一些假定可以等价地表示为对  $Y_i$  的假定：

假定1：零均值假定  $E(Y_i | X_i) = \beta_1 + \beta_2 X_i$

假定2：同方差假定  $\text{Var}(Y | X_i) = \sigma^2$

假定3：无自相关假定  $\text{Cov}(Y_i, Y_j) = 0 \quad (i \neq j)$

假定5：正态性假定  $Y_i \square N(\beta_1 + \beta_2 X_i, \sigma^2)$



## 二、普通最小二乘法 ( Ordinary Least Squares )

### ◆ OLS的基本思想

● 不同的估计方法可得到不同的样本回归参数  $\hat{\beta}_1$  和  $\hat{\beta}_2$ ，所估计的  $\hat{Y}_i$  也不同。

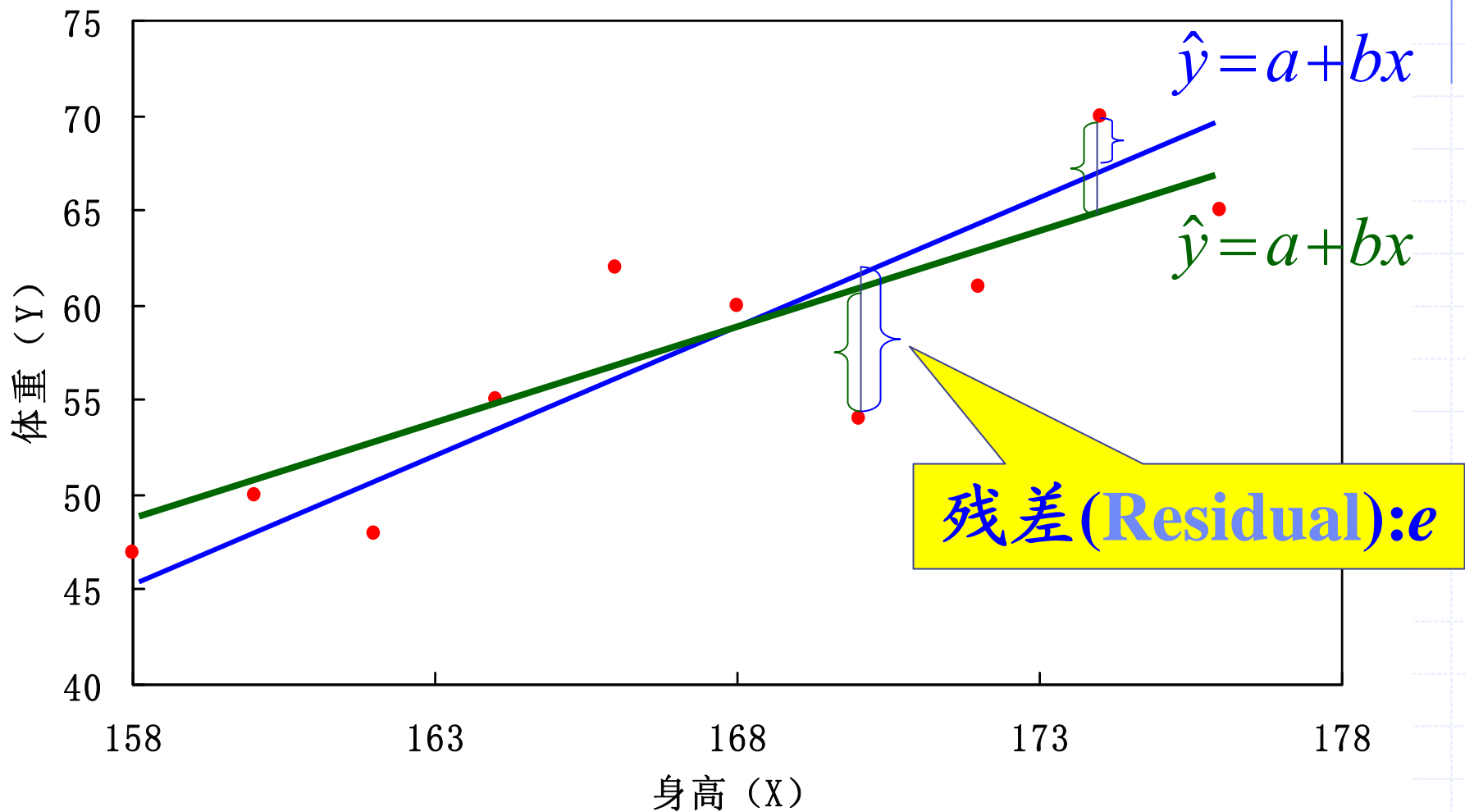
● 理想的估计方法应使  $Y_i$  与  $\hat{Y}_i$  的差即剩余  $e_i$  越小越好

● 因  $e_i$  可正可负，所以可以取  $\sum e_i^2$  最小

即 
$$\min(\sum e_i^2) = \min(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

## 二、普通最小二乘法 ( Ordinary Least Squares )

10名学生的身高与体重散点图



# 正规方程和估计式

取偏导数为**0**，得正规方程

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i$$

$$\sum X_i Y_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2$$

用克莱姆法则求解得观测值形式的OLS估计式：

$$\hat{\beta}_2 = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2}$$

$$\hat{\beta}_1 = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - (\sum X_i)^2}$$

# 用离差表现的OLS估计式

为表达得更简洁，或者用离差形式OLS估计式：

$$\hat{\beta}_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

注意其中：

$$x_i = X_i - \bar{X}$$

$$y_i = Y_i - \bar{Y}$$

而且样本回归函数可写为  $\hat{y}_i = \hat{\beta}_i x_i$

由于参数的估计结果是通过最小二乘法得到的，故称为普通最小二乘估计量（ordinary least squares estimators）。

### 三、OLS回归线的性质

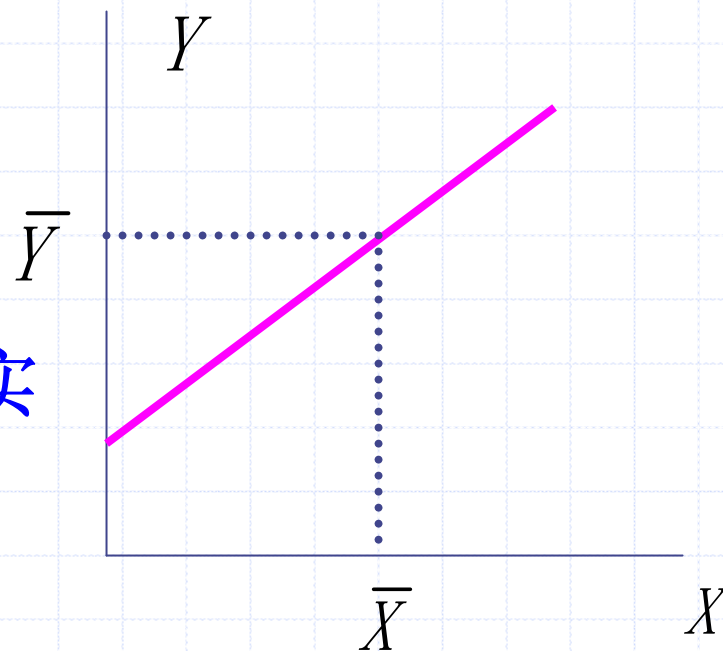
可以证明：

- 回归线通过样本均值

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$$

- 估计值  $\hat{Y}_i$  的均值等于实际观测值  $Y_i$  的均值

$$\frac{\sum \hat{Y}_i}{n} = \bar{Y}$$



- 剩余项  $e_i$  的均值为零

$$\bar{e} = \frac{\sum e_i}{n} = 0$$

- 应变量估计值  $\hat{Y}_i$  与剩余项  $e_i$  不相关

$$\text{Cov}(\hat{Y}_i, e_i) = 0$$

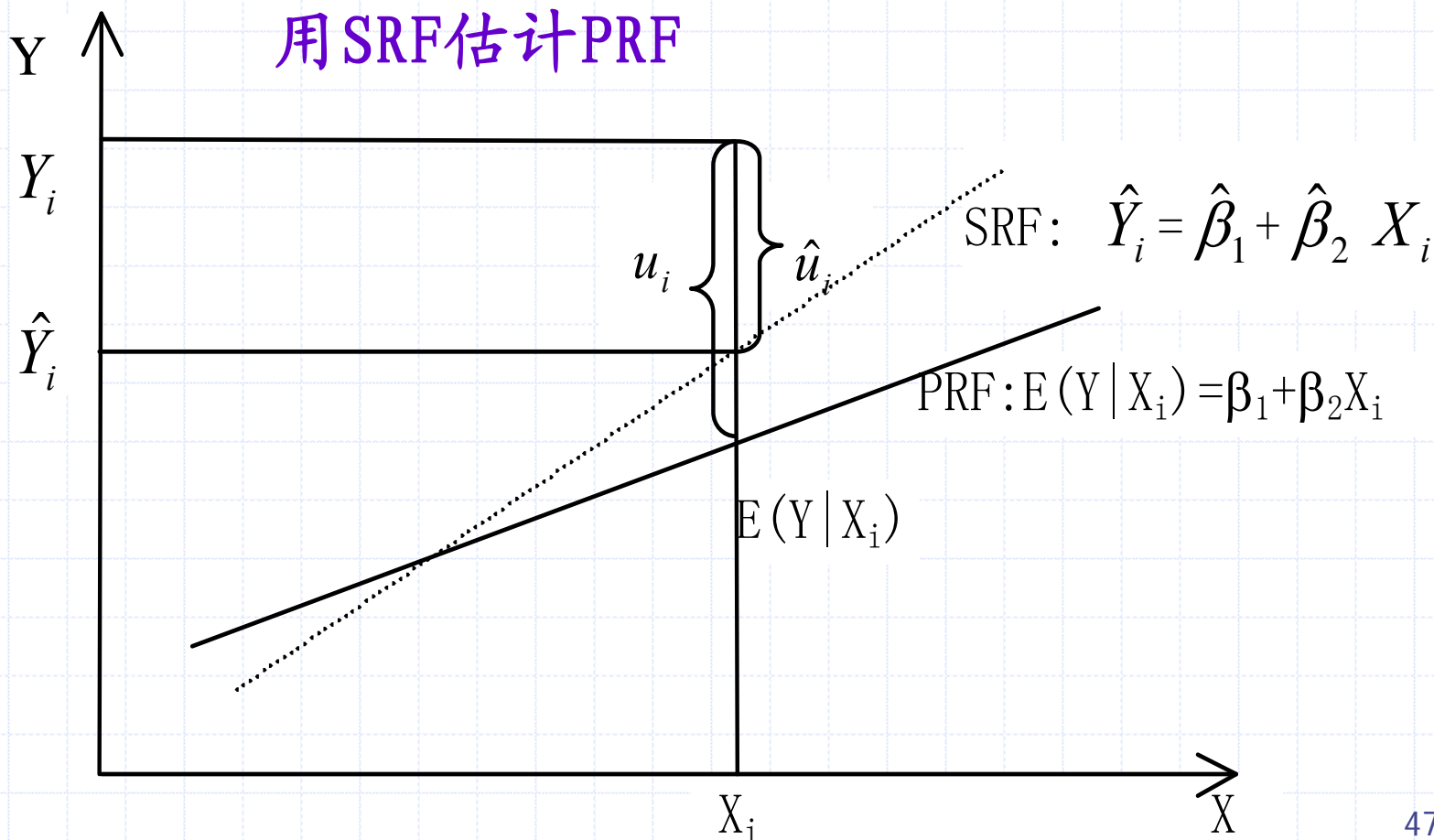
- 解释变量  $X_i$  与剩余项  $e_i$  不相关

$$\text{Cov}(X_i, e_i) = 0$$

# 样本回归线的几何意义

SRF来自总体，围绕着PRF变动

用SRF估计PRF





# 四、OLS估计式的统计性质

## (一)参数估计式的评价标准

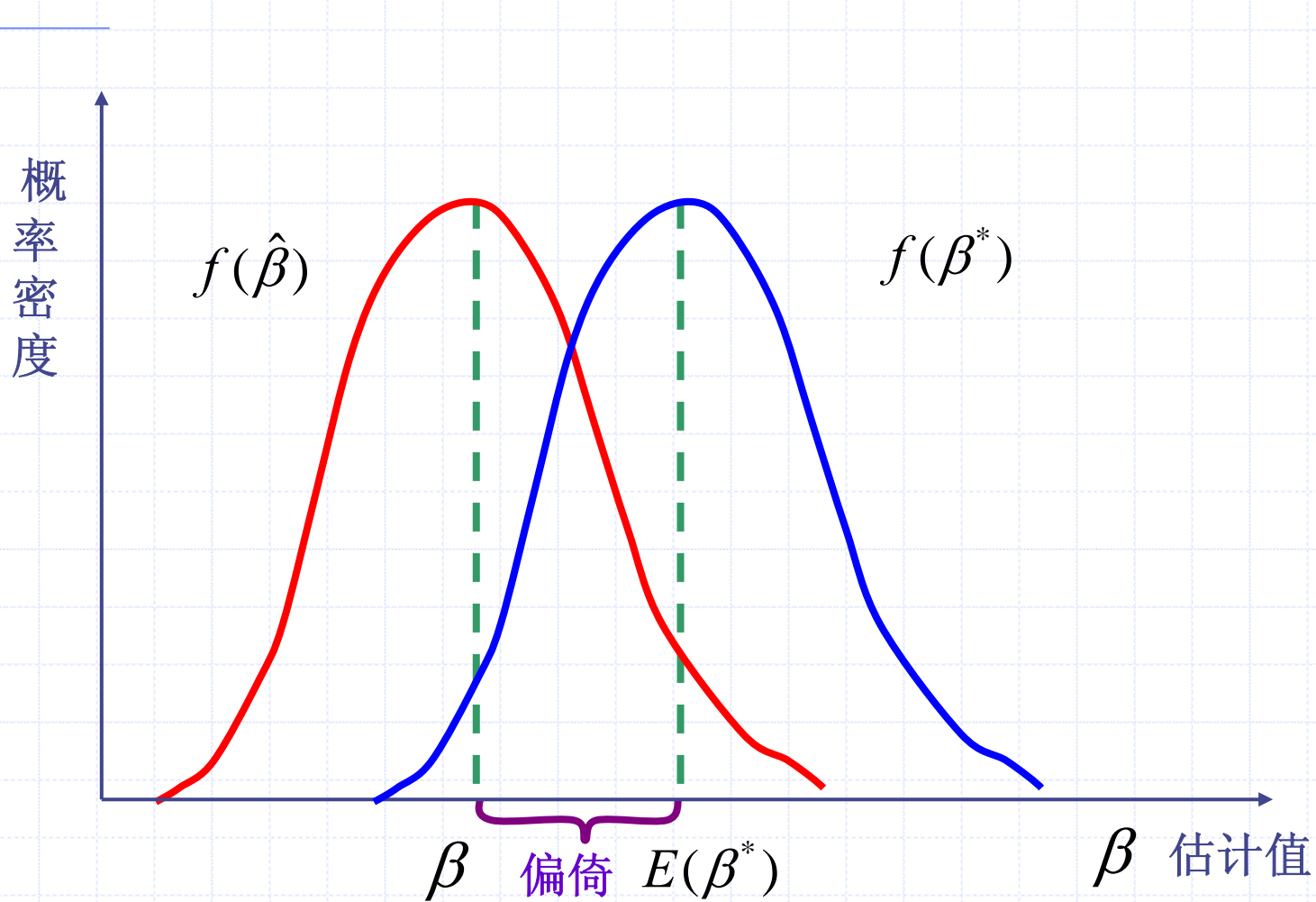
### 1. 无偏性

前提：重复抽样中估计方法固定、样本数不变、经重复抽样的观测值，可得一系列参数估计值  
参数估计值  $\hat{\beta}$  的分布称为  $\hat{\beta}$  的抽样分布，密度函数记为  $f(\hat{\beta})$

如果  $E(\hat{\beta}) = \beta$ ，称  $\hat{\beta}$  是参数  $\beta$  的无偏估计式，否则称  $\hat{\beta}$  是有偏的，其偏倚为  $E(\hat{\beta}) - \beta$

(见图1.2)

图 1.2



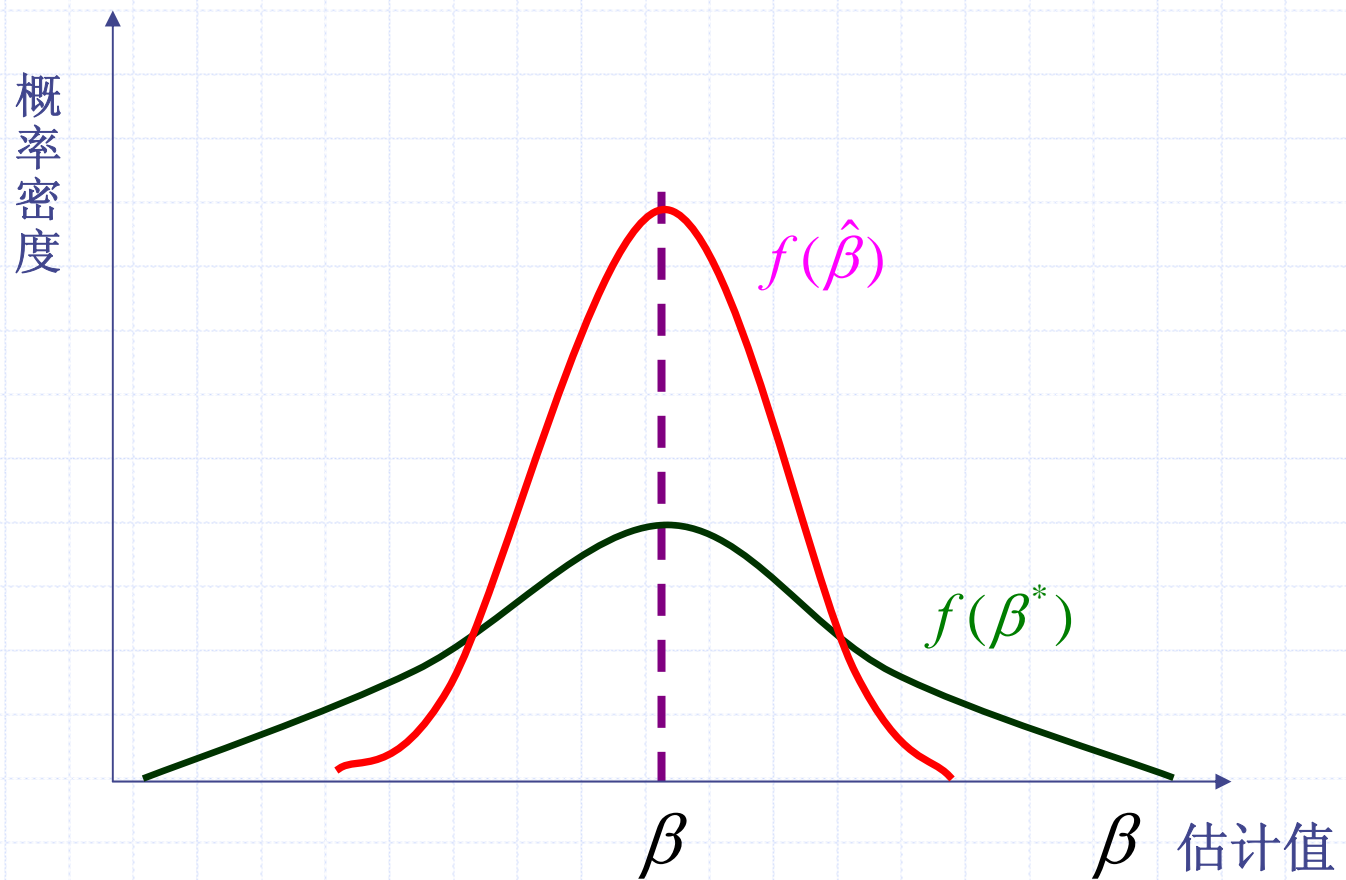
## 2. 最小方差性

前提：样本相同、用不同的方法估计参数，  
可以找到若干个不同的估计式

目标：努力寻求其抽样分布具有最小方差的  
估计式—— 最小方差准则，或称最佳  
性准则（见图1.3）

既是无偏的同时又具有最小方差的估计式，称为  
最佳无偏估计式。

图 1.3



## 4. 渐近性质（大样本性质）

思想:当样本容量较小时,有时很难找到最佳无偏估计,需要考虑样本扩大后的性质

一致性:

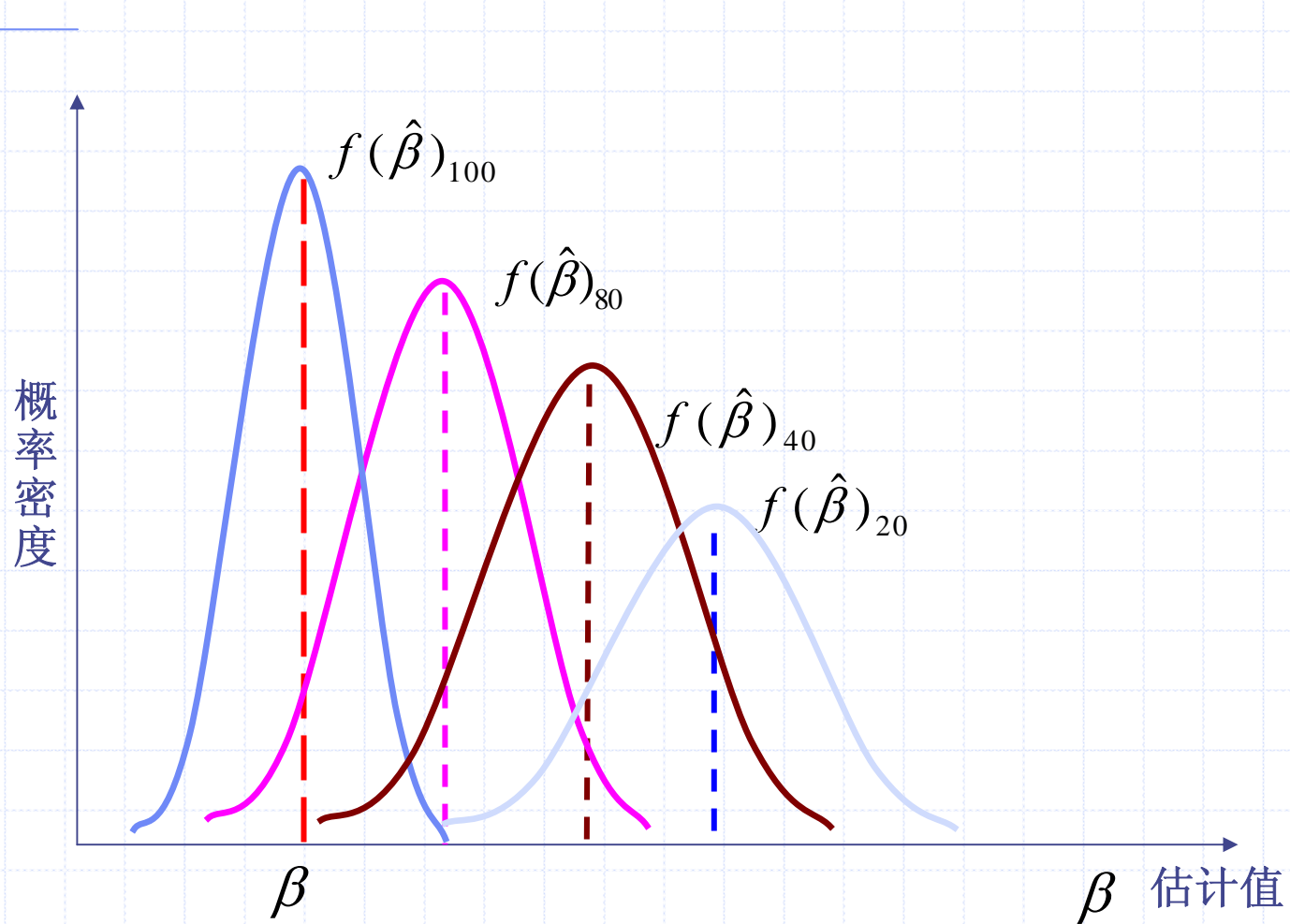
当样本容量  $n$  趋于无穷大时,如果估计式  $\hat{\beta}$  依概率收敛于总体参数的真实值,称这个估计式  $\hat{\beta}$  是  $\beta$  的一致估计式。即

$$\lim_{n \rightarrow \infty} P(|\hat{\beta} - \beta| \leq \varepsilon) = 1 \quad \text{或} \quad P \lim_{n \rightarrow \infty} (\hat{\beta}) = \beta$$

**渐近有效性:** 当样本容量  $n$  趋于无穷大时,在所有的一致估计式中,具有最小的渐近方差。

(见图1.4)

图 1.4



## (二) OLS估计式的统计性质

- 由OLS估计式可以看出

$$\hat{\beta}_2 = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2}$$

$$\hat{\beta}_1 = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - (\sum X_i)^2}$$

$\hat{\beta}_k$  由可观测的样本值  $X_i$  和  $Y_i$  唯一表示。

- 因存在抽样波动，OLS估计  $\hat{\beta}_k$  是随机变量
- OLS估计式是点估计式



# OLS估计式的统计性质——高斯定理

1. 线性特征  $\hat{\beta}_k$  是  $Y$  的线性函数

$$\hat{\beta}_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\sum x_i y_i}{\sum x_i^2} = \sum k_i y_i$$

$$k_i = \frac{x_i}{\sum x_i^2}$$

2. 无偏特性

$$E(\hat{\beta}_k) = \beta_k$$

3. 最小方差特性

在所有的线性无偏估计中，OLS估计  $\hat{\beta}_k$  具有最小方差  
结论：在古典假定条件下，OLS估计式是最佳线性无偏估计式（BLUE）

# 第三节 拟合优度的度量

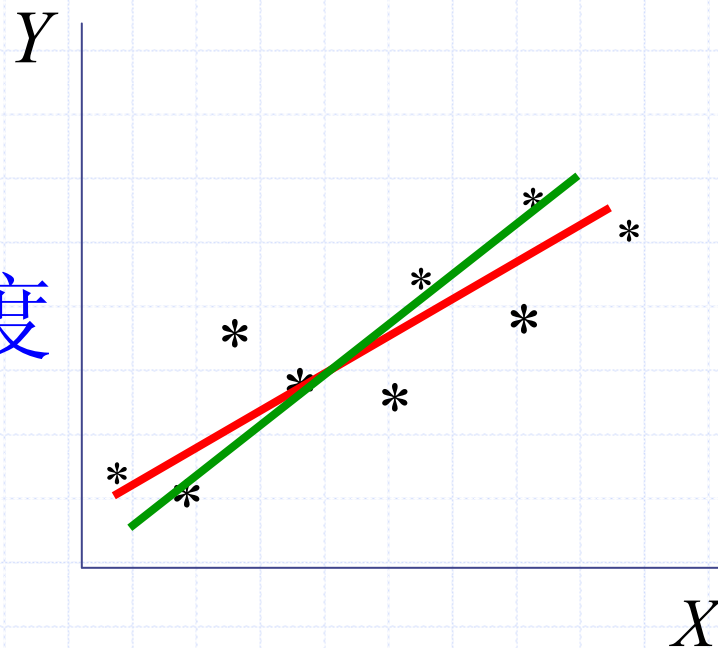
本节基本内容:

- 什么是拟合优度
- 总变差的分解
- 可决系数

# 一、什么是拟合优度？

概念：

样本回归线对样本观测数据拟合的优劣程度



其度量建立在对总变差分解的基础上

## 二、总变差的分解

分析  $Y$  的观测值、估计值与平均值的关系

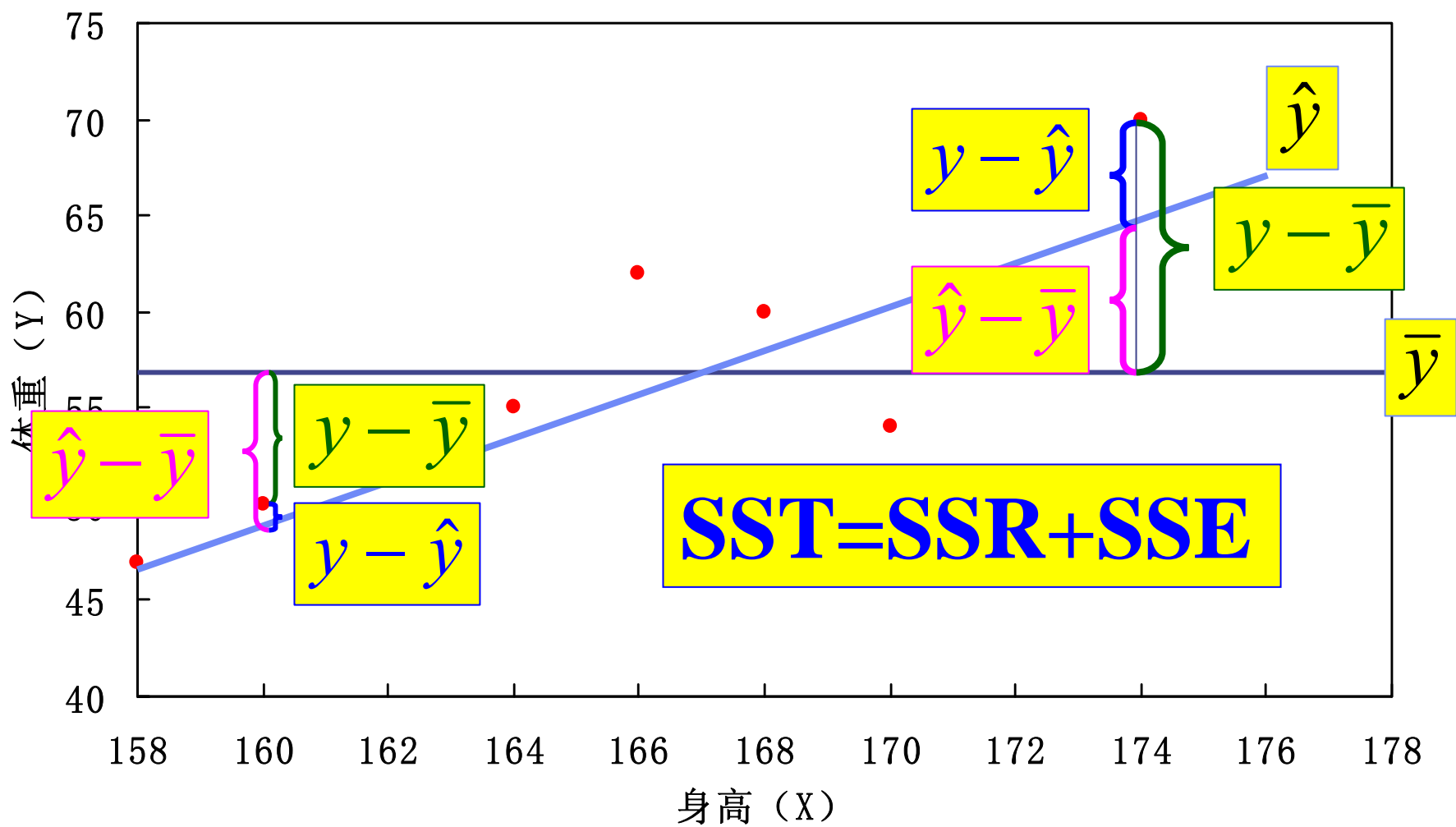
$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

将上式两边平方加总，可证得

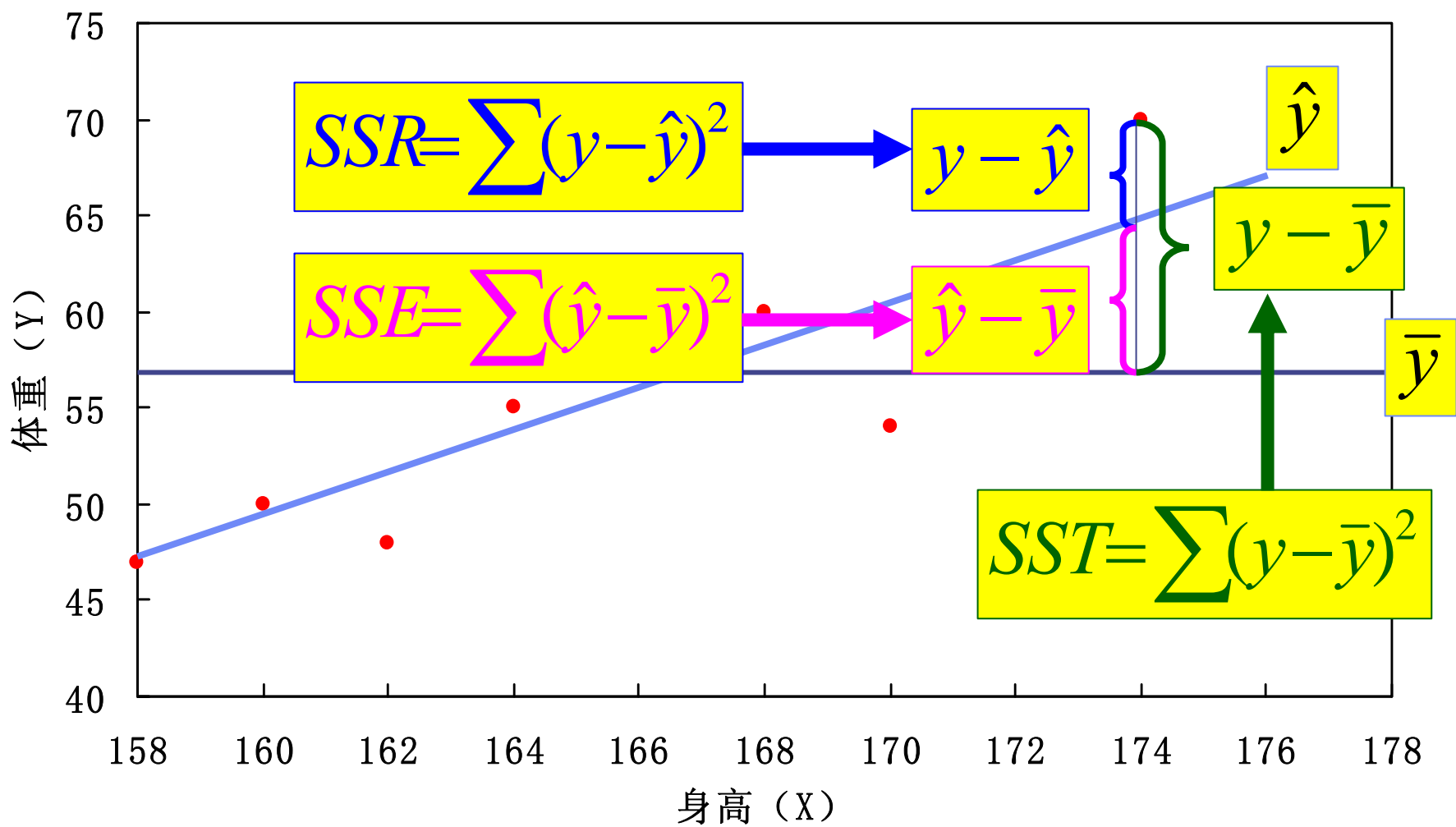
$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

(TSS)      (ESS)      (RSS)

10名学生的身高与体重散点图



# 10名学生的身高与体重散点图



**SST(Sum of squares of total):**

总的平方和

**SSR(Sum of squares of regression):**

回归平方和

**SSE(Sum of squares of errors):**

误差平方和



### 三、可决系数

以TSS同除总变差等式两边：

$$\frac{TSS}{TSS} = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

或

$$1 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} + \frac{\sum e_i^2}{\sum y_i^2}$$

**定义：**回归平方和（解释了的变差ESS） $\sum \hat{y}_i^2$  在总变差（TSS） $\sum y_i^2$  中所占的比重称为可决系数，用 $r^2$ 表示：

$$r^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2}$$

或

$$r^2 = 1 - \frac{\sum e_i^2}{\sum y_i^2}$$

# 可决系数的作用和特点

**作用：**可决系数越大，说明在总变差中由模型作出了解释的部分占的比重越大，模型拟合优度越好。反之可决系数小，说明模型对样本观测值的拟合程度越差。

- 特点：**
- 可决系数取值范围： $0 \leq r^2 \leq 1$
  - 随抽样波动，样本可决系数  $r^2$  是随抽样而变动的随机变量
  - 可决系数是非负的统计

# 可决系数与相关系数的关系

## (1) 联系

数值上，可决系数等于应变变量与解释变量之间简单相关系数的平方：

$$\begin{aligned} R^2 &= \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum y_i^2} = \frac{(\sum x_i y_i)^2}{(\sum x_i^2)^2} \cdot \frac{\sum x_i^2}{\sum y_i^2} \\ &= \frac{\sum x_i y_i}{(\sum x_i^2)(\sum y_i^2)} = r^2 \end{aligned}$$

# 可决系数与相关系数的关系

## (2) 区别

可决系数	相关系数
就模型而言	就两个变量而言
说明解释变量对应变量的解释程度	度量两个变量线性依存程度。
度量不对称的因果关系	度量不含因果关系的对称相关关系
取值: <b>[0,1]</b>	取值: <b>[-1,1]</b>

## 第四节 回归系数的区间估计和假设检验

本节基本内容：

- OLS估计的分布性质
- 回归系数的区间估计
- 回归系数的假设检验

# 一、OLS估计的分布性质

## 基本思想

$\hat{\beta}_k$  是随机变量，必须确定其分布性质才可能进行区间估计和假设检验

$u_i$  是服从正态分布的随机变量  $\rightarrow Y_i$  也是服从正态分布的随机变量， $\rightarrow \hat{\beta}_k$  也是服从正态分布的随机变量。

# $\hat{\beta}$ 的期望和方差

- $\hat{\beta}$  的期望:  $E(\hat{\beta}_k) = \beta_k$  (无偏估计)
- $\hat{\beta}$  的方差和标准误差  
(标准误差是方差的算术平方根)

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2}$$

$$\text{SE}(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum x_i^2}}$$

$$\text{Var}(\hat{\beta}_1) = \sigma^2 \frac{\sum X_i^2}{N \sum x_i^2}$$

$$\text{SE}(\hat{\beta}_1) = \sigma \sqrt{\frac{\sum X_i^2}{N \sum x_i^2}}$$

注意: 以上各式中  $\sigma^2$  未知, 其余均是样本观测值



## 对随机扰动项方差 $\sigma^2$ 的估计

可以证明（见教材P70附录2.2）

$\sigma^2$  的无偏估计为

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$$

( $n-2$ 为自由度,即可自由变化的样本观测值个数)

# 将 $\hat{\beta}$ 作标准化变换

● 在  $\sigma^2$  已知时

$$z_1 = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\sigma \sqrt{\frac{\sum X_i^2}{n \sum x_i^2}}} \sim N(0,1)$$

$$z_2 = \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} = \frac{\hat{\beta}_2 - \beta_2}{\frac{\sigma}{\sqrt{\sum x_i^2}}} \sim N(0,1)$$

## ● 当 $\sigma^2$ 未知时

(1) 当样本为大样本时，用估计的参数标准误差对  $\hat{\beta}$  作标准化变换，所得  $Z$  统计量仍可视为标准正态变量（根据中心极限定理）

(2) 当样本为小样本时，可用  $\hat{\sigma}^2$  代替  $\sigma^2$ ，去估计参数的标准误差，用估计的参数标准误差对  $\hat{\beta}$  作标准化变换，所得统计量不再服从正态分布（这时分母也是随机变量），而是服从  $t$  分布：

$$t = \frac{\hat{\beta}_k - \beta_k}{\hat{SE}(\hat{\beta}_k)} \sim t(n-2)$$

## 二、回归系数区间估计的方法

一般情况下, 总体方差  $\sigma^2$  未知, 用无偏估计  $\hat{\sigma}^2$  去代替, 由于样本容量较小, 统计量不再服从正态分布, 而服从 t 分布。可用 t 分布去建立参数估计的置信区间。

$$t^* = \frac{\hat{\beta}_2 - \beta_2}{\hat{SE}(\hat{\beta}_2)} \sim t(n-2)$$

选定  $\alpha$ ，查  $t$  分布表得显著性水平为  $\alpha/2$ ，自由度为  $n-2$  的临界值，则有

$$P\left[-t_{\alpha/2} \leq \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} \leq t_{\alpha/2}\right] = 1 - \alpha$$

即

$$P\left[\hat{\beta}_2 - t_{\alpha/2} SE(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2} SE(\hat{\beta}_2)\right] = 1 - \alpha$$

### 三、回归系数的假设检验

一般情况下，总体方差  $\sigma^2$  未知，只能用  $\hat{\sigma}^2$  去代替，可利用 t 分布作 t 检验

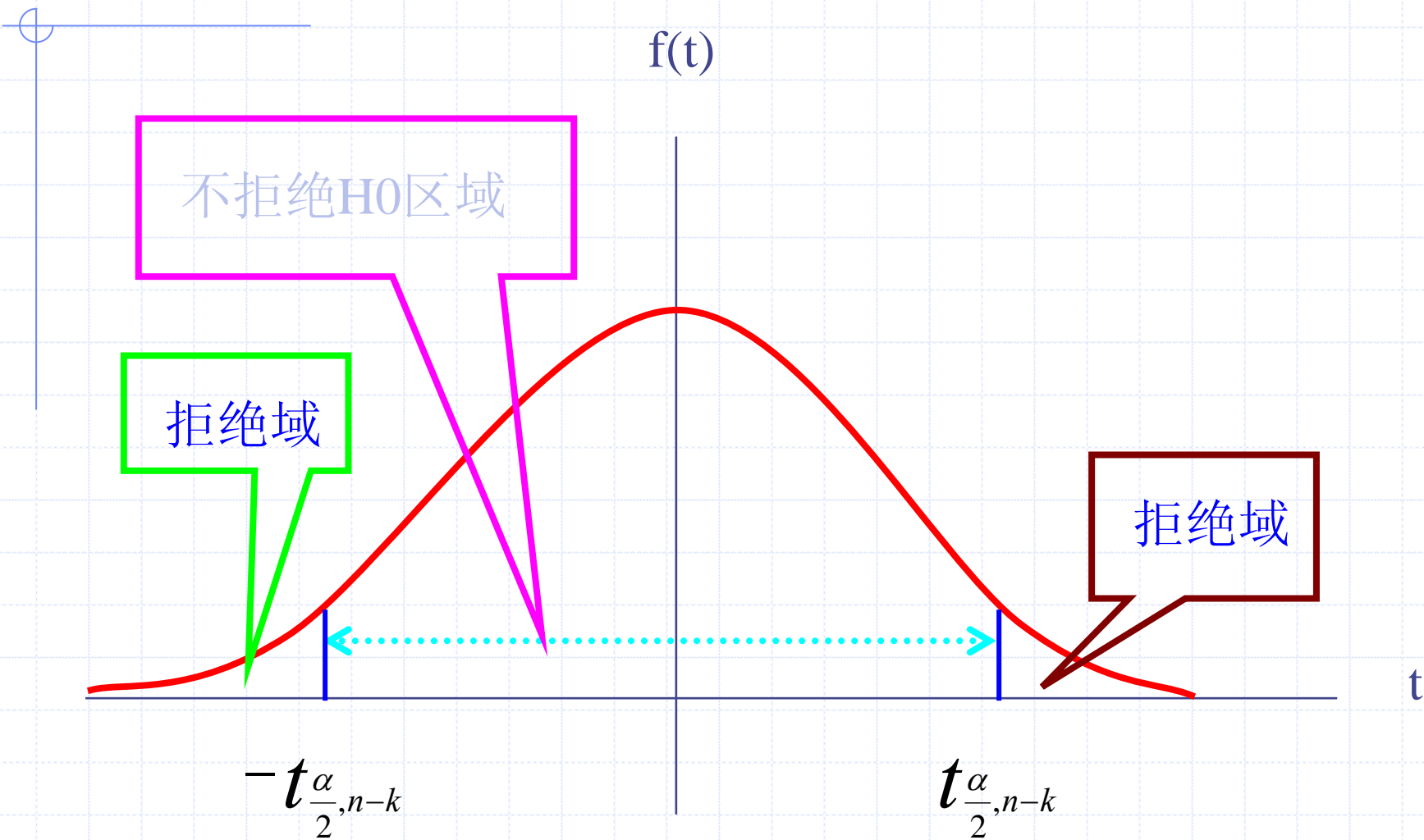
$$t^* = \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} \sim t(n-2)$$

给定  $\alpha$ ，查 t 分布表得  $t_{\alpha/2}(n-2)$

▼如果  $t^* \leq -t_{\alpha/2}(n-2)$  或者  $t^* \geq t_{\alpha/2}(n-2)$  则拒绝原假设  $H_0: \beta_2 = 0$ ，而接受备择假设  $H_1: \beta_2 \neq 0$

▼如果  $-t_{\alpha/2}(n-2) \leq t^* \leq t_{\alpha/2}(n-2)$  则接受原假设  $H_0: \beta_2 = 0$

# t检验的步骤







# 用 P 值判断参数的显著性

## 假设检验的 p 值:

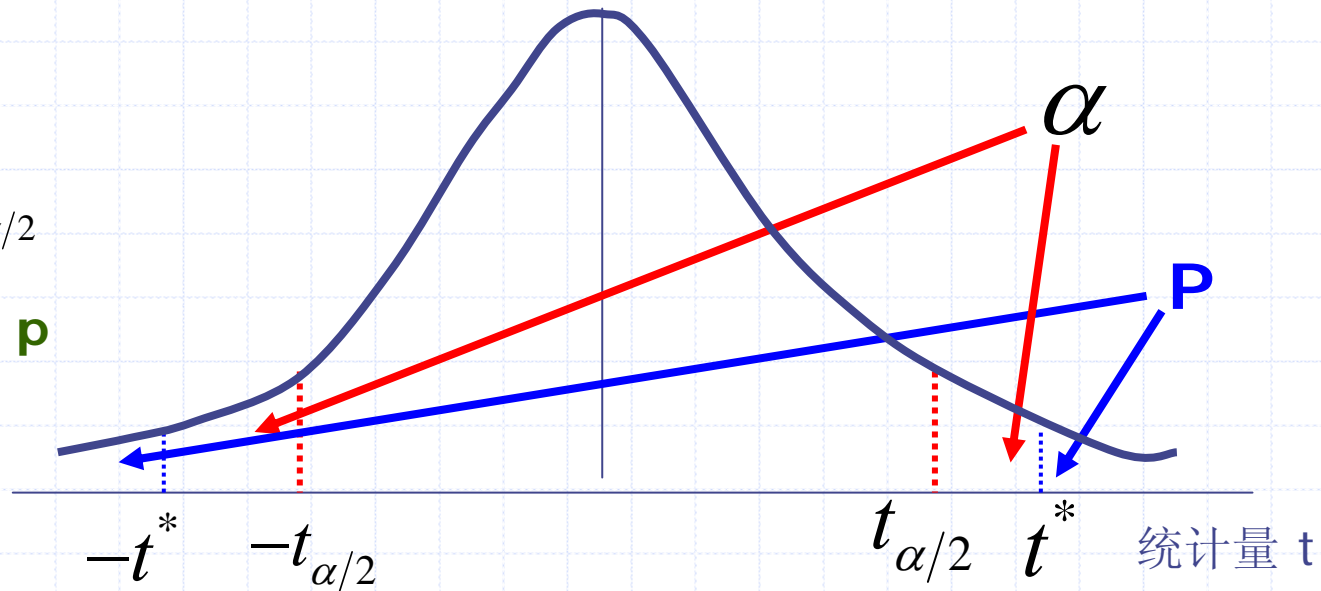
p 值是基于既定的样本数据所计算的统计量，是拒绝原假设的最低显著性水平。

统计分析软件中通常都给出了检验的 p 值

注意:

t 检验是比较  $t^*$  和  $t_{\alpha/2}$

P 值检验是比较  $\alpha$  和 p



# 第五节 回归模型预测

本节主要内容：

- 回归分析结果的报告
- 被解释变量平均值预测
- 被解释变量个别值预测

# 一、回归分析结果的报告

经过模型的估计、检验，得到一系列重要的数据，为了简明、清晰、规范地表述这些数据，计量经济学通常采用了以下规范化的方式：

例如：回归结果为

$$\hat{Y}_i = 352.00 + 0.5300X_i$$

(76.5826)    (0.0216)

标准误差 **SE**

$$t = (4.5963) \quad (24.5902)$$

**t** 统计量

$$r^2 = 0.9869 \quad df = 8$$

可决系数和自由度

## 二、被解释变量平均值预测

### 1. 基本思想

● 计量经济预测是一种条件预测：

条件：◆ 模型设定的关系式不变

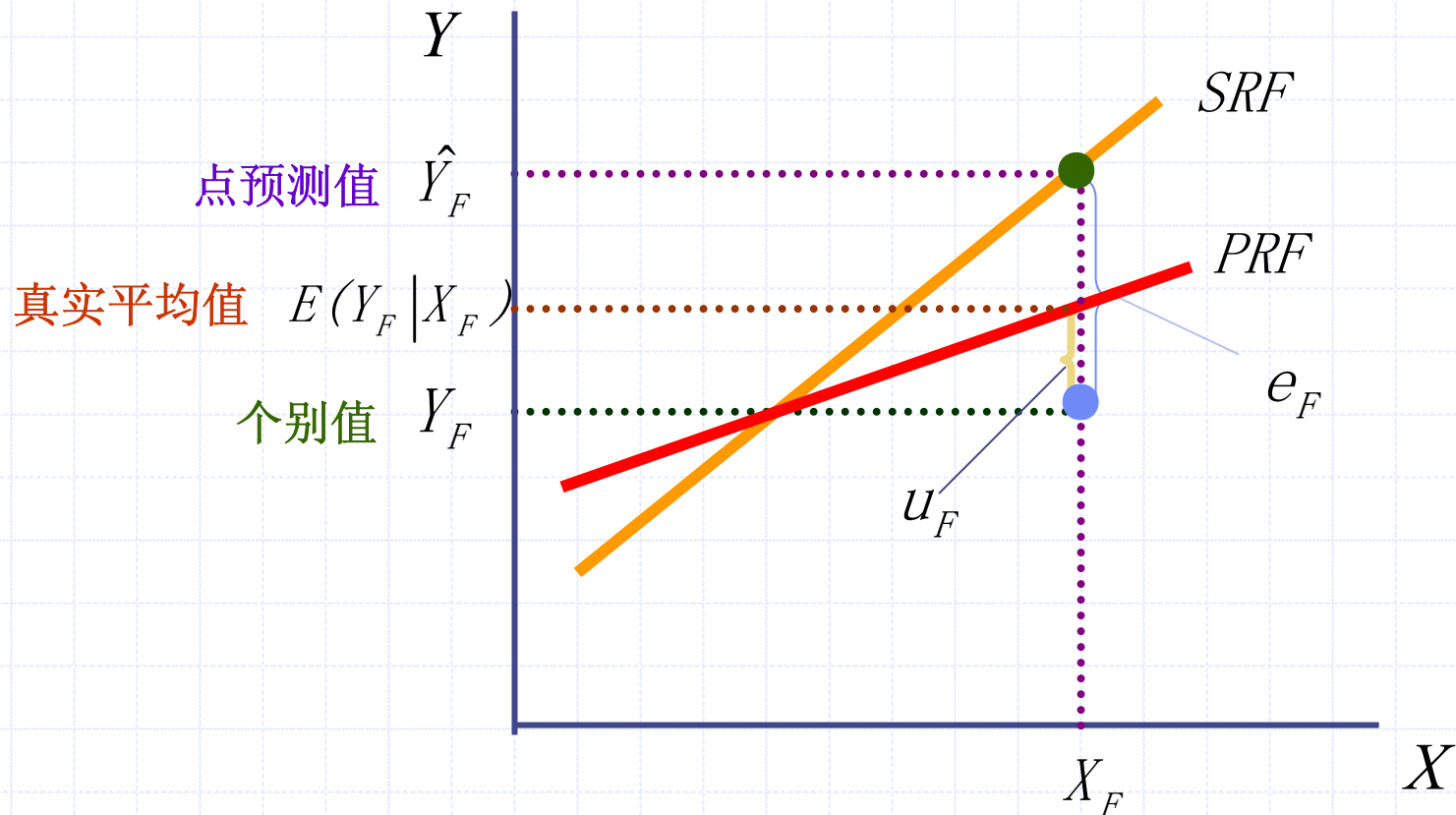
◆ 所估计的参数不变

◆ 解释变量在预测期的取值已作出预测

对应变量的预测分为平均值预测和个别值预测

对应变量的预测又分为点预测和区间预测

# 预测值、平均值、个别值的相互关系



$\hat{Y}_F$  是真实平均值的点估计,也是对个别值的点估计

## 2 . $Y$ 平均值的点预测

将解释变量预测值直接代入估计的方程

$$\hat{Y}_F = \hat{\beta}_1 + \hat{\beta}_2 X_F$$

这样计算的  $\hat{Y}_F$  是一个点估计值

### 3. $Y$ 平均值的区间预测

#### 基本思想:

- ◆ 由于存在抽样波动, 预测的平均值  $\hat{Y}_F$  不一定等于真实平均值  $E(Y_F|X_F)$ , 还需要对  $E(Y_F|X_F)$  作区间估计。
- ◆ 为对  $Y$  作区间预测, 必须确定平均值预测值的抽样分布, 必须找出与  $\hat{Y}_F$  和  $E(Y_F|X_F)$  都有关的统计量

# 具体作法 (从 $\hat{Y}_F$ 的分布分析)

已知

$$E(\hat{Y}_F) = E(Y_F | X_F) = \beta_1 + \beta_2 X_F$$

可以证明

$$\text{Var}(\hat{Y}_F) = \sigma^2 \left[ \frac{1}{n} + \frac{(X_F - \bar{X})^2}{\sum x_i^2} \right]$$

$\hat{Y}_F$  服从正态分布, 将其标准化,

当  $\sigma^2$  未知时, 只得用  $\hat{\sigma}^2 = \sum e_i^2 / (n-2)$  代替, 这时有

$$t = \frac{\hat{Y}_F - E(Y_F | X_F)}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_F - \bar{X})^2}{\sum x_i^2}}} \sim t(n-2)$$



有

$$P(-t_{\alpha/2} \leq t = \frac{\hat{Y}_F - E(Y_F | X_F)}{\hat{SE}(\hat{Y}_F)} \leq t_{\alpha/2}) = 1 - \alpha$$

Y平均值的置信度为  $1 - \alpha$  的预测区间为

$$[\hat{Y}_F - t_{\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_F - \bar{X})^2}{\sum x_i^2}}, \hat{Y}_F + t_{\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_F - \bar{X})^2}{\sum x_i^2}}]$$

### 三、应变变量个别值预测具体作法:

已知剩余项  $e_F = Y_F - \hat{Y}_F$  是与预测值  $\hat{Y}_F$  及个别值  $Y_F$  都有关的变量, 并且已知  $e_F$  服从正态分布, 且可证明

$$E(e_F) = 0$$

$$\text{Var}(e_F) = E(Y_F - \hat{Y}_F)^2 = \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(X_F - \bar{X})^2}{\sum x_i^2} \right]$$

当用  $\hat{\sigma}^2 = \sum e_i^2 / (n-2)$  代替  $\sigma^2$  时, 对  $e_F$  标准化的变量  $t$  为

$$t = \frac{e_F - E(e_F)}{\hat{SE}(e_F)} = \frac{Y_F - \hat{Y}_F}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(X_F - \bar{X})^2}{\sum x_i^2}}} \sim t(n-2)$$

# 构建个别值的预测区间

有

$$P\{[\hat{Y}_F - t_{\alpha/2} \hat{SE}(e_F)] \leq Y_F \leq [\hat{Y}_F + t_{\alpha/2} \hat{SE}(e_F)]\} = 1 - \alpha$$

因此，一元回归对 的个别值的置信度为  $1 - \alpha$  的预测区间上下限为

$$Y_F = \hat{Y}_F \mp t_{\alpha/2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(X_F - \bar{X})^2}{\sum x_i^2}}$$

# 应变量Y 区间预测的特点

- 1、Y平均值的预测值与真实平均值有误差，主要是受抽样波动影响

$$Y_F = \hat{Y}_F \mp t_{\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_F - \bar{X})^2}{\sum x_i^2}}$$

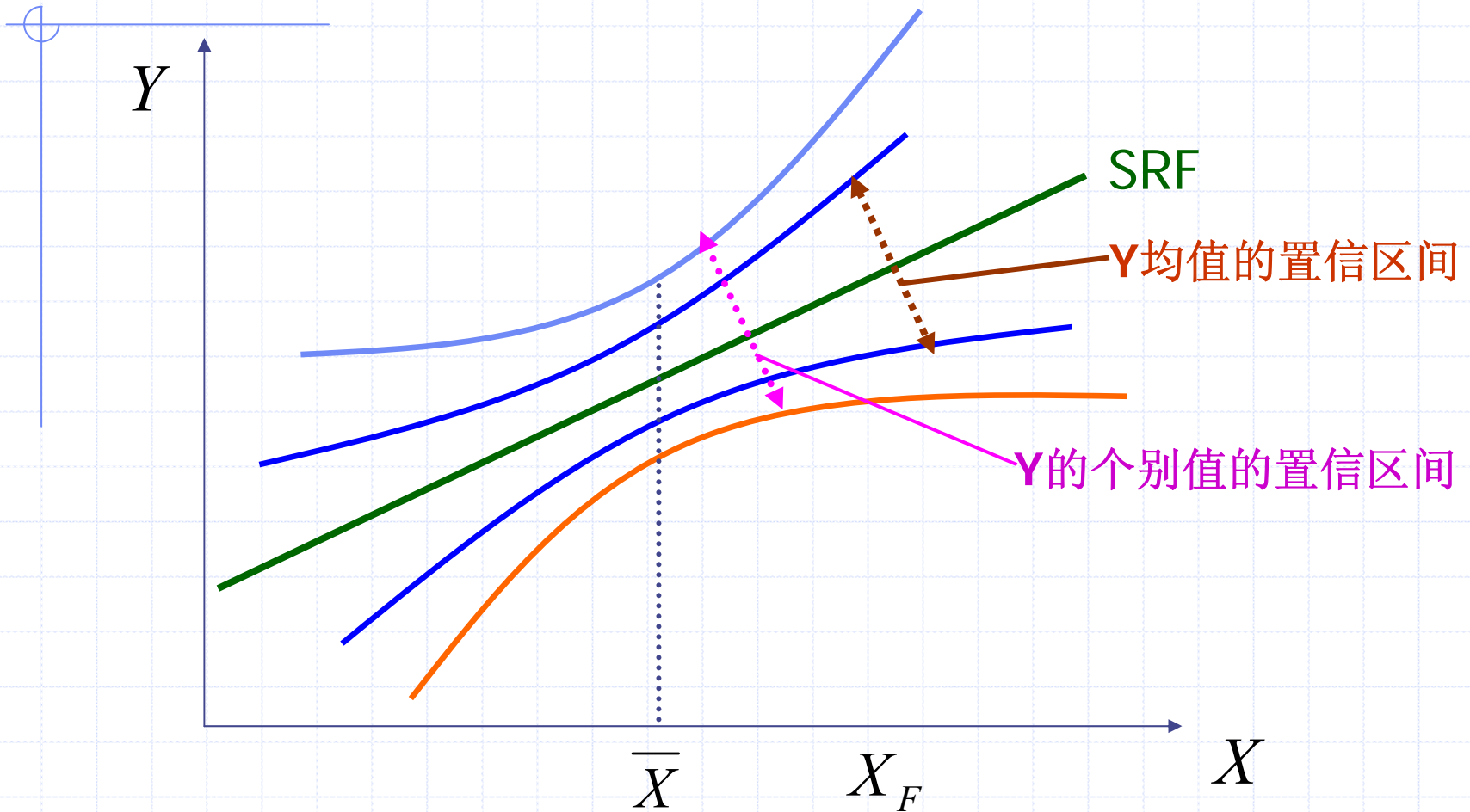
Y个别值的预测值与真实个别值的差异,不仅受抽样波动影响,而且还受随机扰动项的影响

$$Y_F = \hat{Y}_F \mp t_{\alpha/2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(X_F - \bar{X})^2}{\sum x_i^2}}$$

2、平均值和个别值预测区间都不是常数，是随  $X_F$  的变化而变化的

3、预测区间上下限与样本容量有关，当样本容量  $n \rightarrow \infty$  时个别值的预测误差只决定于随机扰动的方差

# 各种预测值的关系



当 $X_F = \bar{X}$ 时，置信区间最小

## 第六节 案例分析

**提出问题：** 改革开放以来随着中国经济的快速发展，居民的消费水平也不断增长。但全国各地经济发展速度不同，居民消费水平也有明显差异。

为了分析什么是影响各地区居民消费支出有明显差异的最主要因素，并分析影响因素与消费水平的数量关系，可以建立相应的计量经济模型去研究。

**研究范围：** 全国各省市2002年城市居民家庭平均每人每年消费截面数据模型。

**理论分析：**影响各地区城市居民人均消费支出的因素有多种，但从理论和经验分析，最主要的影响因素应是居民收入。从理论上说可支配收入越高，居民消费越多，但边际消费倾向大于0，小于1。

**建立模型：**  $Y_i = \beta_1 + \beta_2 X_i + u$

其中：**Y**—城市居民家庭平均每人每年消费支出(元)

**X**—城市居民人均年可支配收入(元)



# 数据：从2002年《中国统计年鉴》中得到

地 区	城市居民家庭平均每人每年消费支出(元) $Y$	城市居民人均年可支配收入(元) $X$
北京	10284.60	12463.92
天津	7191.96	9337.56
河北	5069.28	6679.68
山西	4710.96	5234.35
内蒙古	4859.88	6051.06
辽宁	5342.64	6524.52
吉林	4973.88	6260.16
黑龙江	4462.08	6100.56
上海	10464.00	13249.80
江苏	6042.60	8177.64
浙江	8713.08	11715.60
安徽	4736.52	6032.40
福建	6631.68	9189.36
江西	4549.32	6334.64
山东	5596.32	7614.36
河南	4504.68	6245.40
湖北	5608.92	6788.52

(接上页数据表)

地 区	城市居民家庭平均每人每年 消费支出(元) Y	城市居民人均年可支配收入 (元) X
湖南	5574.72	6958.56
广东	8988.48	11137.20
广西	5413.44	7315.32
海南	5459.64	6822.72
重庆	6360.24	7238.04
四川	5413.08	6610.80
贵州	4598.28	5944.08
云南	5827.92	7240.56
西藏	6952.44	8079.12
陕西	5278.04	6330.84
甘肃	5064.24	6151.44
青海	5042.52	6170.52
宁夏	6104.92	6067.44
新疆	5636.40	6899.64

# 估计参数

假定模型中随机扰动满足基本假定，可用**OLS**法。

具体操作：使用**EViews** 软件包。估计结果：

Method: Least Squares

Date: 02/25/05 Time: 03:15

Sample: 1 31

Included observations: 31

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	282.2434	287.2649	0.982520	0.3340
X	0.758511	0.036928	20.54026	0.0000
R-squared	0.935685	Mean dependent var	5982.476	
Adjusted R-squared	0.933467	S.D. dependent var	1601.762	
S.E. of regression	413.1593	Akaike info criterion	14.94788	
Sum squared resid	4950317.	Schwarz criterion	15.04040	
Log likelihood	-229.6922	F-statistic	421.9023	
Durbin-Watson stat	1.481439	Prob(F-statistic)	0.000000	

表示为  $\hat{Y}_i$

$$Y_i = 282.2434 + 0.758511X_i$$

$$(287.2649) \quad (0.036928)$$

$$t = (0.982520) \quad (20.54026)$$

$$r^2 = 0.935685 \quad F = 421.9023 \quad df = 29$$

# 模型检验

1. 可决系数:  $r^2 = 0.935685$  模型整体上拟合好。

2. 系数显著性检验: 给定  $\alpha = 0.05$ , 查 t 分布表,

在自由度为  $n-2=29$  时临界值为  $t_{0.025}(29) = 2.045$

因为  $t = 20.44023 > t_{0.025}(29) = 2.045$

说明“城镇人均可支配收入”对“城镇人均消费支出”有显著影响。

3. 用 P 值检验  $\alpha = 0.05$   $\gg p=0.0000$

#### 4. 经济意义检验:

估计的解释变量的系数为0.758511，说明城镇居民人均可支配收入每增加1元，人均年消费支出平均将增加0.758511元。这符合经济理论对边际消费倾向的界定。

# 经济预测

## 点预测:

西部地区的城市居民人均年可支配收入第一步争取达到**1000**美元(按现有汇率即人民币**8270**元), 代入估计的模型得

$$\hat{Y}_{f1} = 282.2434 + 0.758511 \times 8270 = 6555.132$$

第二步再争取达到**1500**美元(即人民币**12405**元), 利用所估计的模型可预测这时城市居民可能达到的人均年消费支出水平

$$\hat{Y}_{f2} = 282.2434 + 0.758511 \times 12405 = 9691.577$$

# 区间预测

平均值区间预测上下限:

$$Y_F = \hat{Y}_F \mp t_{\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_F - \bar{X})^2}{\sum x_i^2}}$$

$X_{f1} = 8270$ 时

$$\begin{aligned} Y_{f1} &= 6555.13 \mp 2.045 \times 413.1593 \times \sqrt{\frac{1}{31} + \frac{569985.74}{125176492.59}} \\ &= 6555.13 \mp 162.10 \end{aligned}$$

$X_{f2} = 12405$ 时

$$\begin{aligned} Y_{f2} &= 9691.58 \mp 2.045 \times 413.1593 \times \sqrt{\frac{1}{31} + \frac{23911845.72}{125176492.59}} \\ &= 9691.58 \mp 499.25 \end{aligned}$$



即是说：

$$X_{f1} = 8270 \text{时,}$$

平均值置信度**95%**的预测区间为 **(6393.03, 6717.23)** 元。

$$X_{f2} = 12405 \text{时,}$$

平均值置信度**95%**的预测区间为 **(9292.33, 10090.83)** 元。

个别值区间预测 (略)

# 第二章 小 结

## 1、变量间的关系：

函数关系——相关关系

相关系数——对变量间线性相关程度的度量

## 2、现代意义的回归：

一个被解释变量对若干个解释变量依存关系的研究

实质：由固定的解释变量去估计被解释变量的平均值

3、总体回归函数（PRF）：将总体被解释变量 $Y$ 的条件均值表现为解释变量 $X$ 的某种函数

样本回归函数（SRF）：将被解释变量 $Y$ 的样本条件均值表示为解释变量 $X$ 的某种函数。

总体回归函数与样本回归函数的区别与联系

4、随机扰动项：被解释变量实际值与条件均值的偏差，代表排除在模型以外的所有因素对 $Y$ 的影响。

## 5、简单线性回归的基本假定：

对模型和变量的假定

$$E(Y_i) = \beta_1 + \beta_2 X_i$$

对随机扰动项u的假定

零均值假定： $E(u_i) = 0$

同方差假定： $Var(u_i) = Var(Y_i) = \sigma^2$

无自相关假定： $Cov(u_i, u_j) = E(u_i u_j) = 0$

随机扰动与解释变量不相关假定： $Cov(u_i, X_i) = 0$

正态性假定： $u_i \sim N(0, \sigma^2)$

## 6、普通最小二乘法（OLS）估计参数的基本思想及估计式；

$$\hat{\beta}_1 = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{N \sum X_i^2 - (\sum X_i)^2}$$

$$\hat{\beta}_2 = \frac{N \sum X_i Y_i - \sum X_i \sum Y_i}{N \sum X_i^2 - (\sum X_i)^2} = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

## OLS 估计式的分布性质

期望:  $E(\hat{\beta}_k) = \beta_k$

方差:  $Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2}$        $Var(\hat{\beta}_1) = \sigma^2 \frac{\sum X_i^2}{N \sum x_i^2}$

标准差:  $SE(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum x_i^2}}$        $SE(\hat{\beta}_1) = \sigma \sqrt{\frac{\sum X_i^2}{N \sum x_i^2}}$

OLS估计式是最佳线性无偏估计式。

## 7、 $\sigma^2$ 的无偏估计

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$$

## 8、对回归系数区间估计的思想和方法

$$P[\hat{\beta}_2 - t_{\alpha/2} \hat{SE}(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2} \hat{SE}(\hat{\beta}_2)] = 1 - \alpha$$

9、拟合优度：样本回归线对样本观测数据拟合的优劣程度，

可决系数：在总变差分解基础上确定的，模型解释了的变差在总变差中的比重

可决系数的计算方法、特点与作用。

$$1 = \frac{\sum \hat{y}^2}{\sum y_i^2} + \frac{\sum e_i^2}{\sum y_i^2} \quad r^2 = \frac{\sum \hat{y}^2}{\sum y_i^2} \quad r^2 = 1 - \frac{\sum e_i^2}{\sum y_i^2}$$



## 10、对回归系数的假设检验

假设检验的基本思想

对回归系数 t 检验的思想与方法

$$t^* = \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} \sim t(n-2)$$

用 P 值判断参数的显著性

## 11、对被解释变量的预测

被解释变量平均值预测与个别值预测的关系

被解释变量平均值的点预测和区间预测的方法

$$\left[ \hat{Y}_F - t_{\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_F - \bar{X})^2}{\sum x_i^2}}, \hat{Y}_F + t_{\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_F - \bar{X})^2}{\sum x_i^2}} \right]$$

## 被解释变量个别值区间预测的方法

$$Y_F = \hat{Y}_F \mp t_{\alpha/2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(X_F - \bar{X})^2}{\sum x_i^2}}$$

## 12、运用EViews软件对简单的线性回归模型进行估计和检验

本章内容结束！

