计量经济学

第 二 章

简单线性回归模型

引子：中国旅游业总收入将超过3000亿美元吗？

从2004中国国际旅游交易会上获悉，到2020年，中国旅游业总收入将超过3000亿美元，相当于国内生产总值的8%至11%。（资料来源：国际金融报2004年11月25日第二版）

- 是什么决定性的因素能使中国旅游业总收入到2020年达到3000亿美元？
- 旅游业的发展与这种决定性因素的数量关系究竟是什么？
- 怎样具体测定旅游业发展与这种决定性因素的数量关系？

应当考虑的问题：

（1）确定作为研究对象的经济变量

（如中国旅游业总收入）

（2）分析影响研究对象变动的主要因素

（如中国居民收入的增长）

（3）分析各种影响因素与所研究经济现象的相互关系

（决定相互联系的数学关系式）

（4）确定所研究的经济问题与影响因素间具体的数量关系

（需要特定的方法）

（5）分析并检验所得数量结论的可靠性

（需要统计检验）

（6）运用数量研究结果作经济分析和预测

（对数量分析的实际应用）

第一节 回归分析与回归方程

一、回归与相关（对统计学的回顾）

1、经济变量间的相互关系

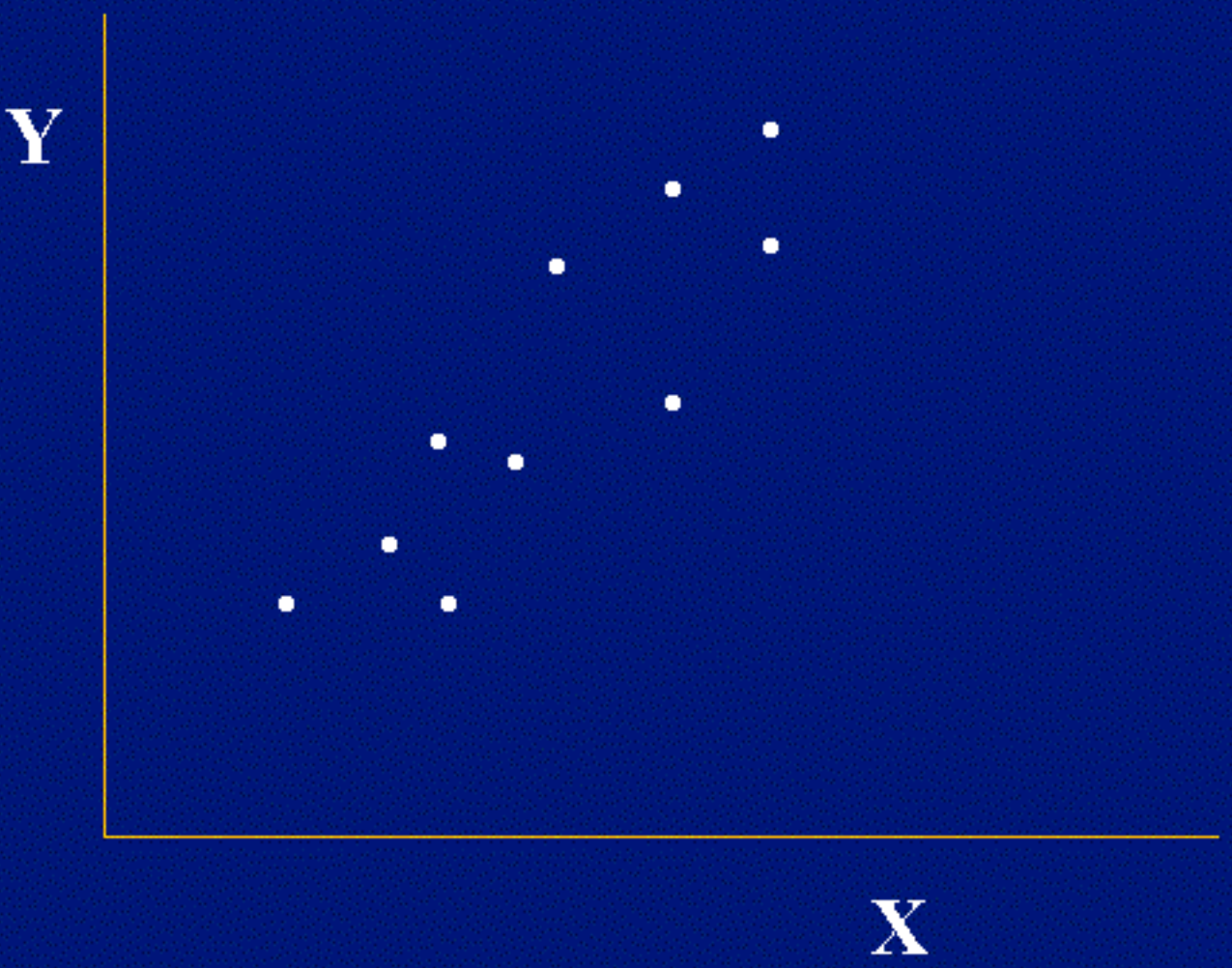
- ◆确定性的函数关系　　Y=f (X)
- ◆不确定性的统计关系—相关关系

  Y= f（X）+ ε　　（ε 为随机变量)
- ◆没有关系

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）



◆相关关系的类型

- 从涉及的变量数量看
  - 简单相关
  - 多重相关（复相关）
- 从变量相关关系的表现形式看
  - 线性相关——散布图接近一条直线
  - 非线性相关——散布图接近一条曲线
- 从变量相关关系变化的方向看
  - 正相关——变量同方向变化，同增同减
  - 负相关——变量反方向变化，一增一减
  - 不相关

3、相关程度的度量—相关系数

X和Y的**总体线性相关系数**：

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

其中： Var(X)-----X 的方差　Var(Y)-----Y的方差

Cov（X，Y）-----X和Y的协方差

X和Y的**样本线性相关系数**：

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

其中：X_i和 Y_i 分别是变量X和Y的样本观测值，

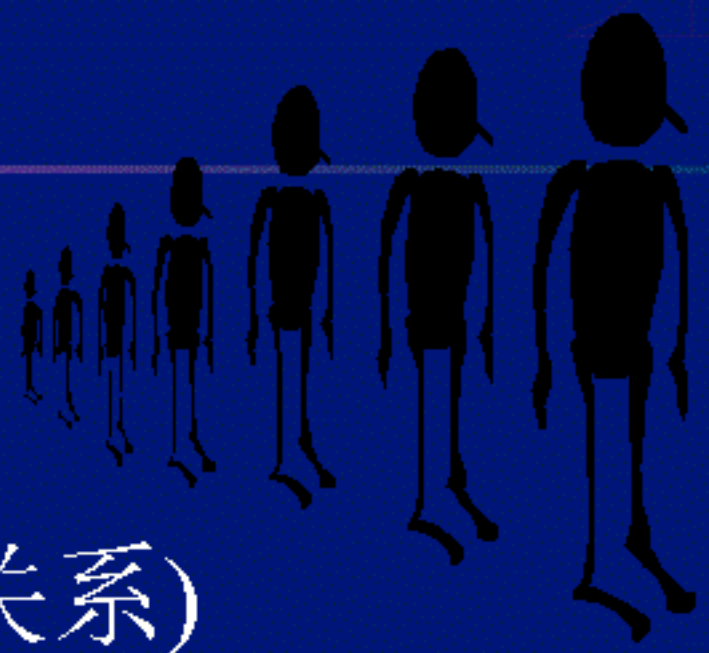
$\bar{X}$ 和 $\bar{Y}$ 分别是变量 X 和Y 样本值的平均值

使用相关系数时应注意：

- X和Y 都是相互对称的随机变量，
  - 线性相关系数只反映变量间的线性相关程度，不能说明非线性相关关系
  - 样本相关系数是总体相关系数的样本估计值，由于抽样波动，样本相关系数是个随机变量，其统计显著性有待检验
- 相关系数只能反映线性相关程度，不能确定因果关系，不能说明相关关系具体接近哪条直线

计量经济学关心：变量间的因果关系及隐藏在随机性后面的统计规律性，这有赖于回归分析方法

4、回归分析



回归的**古典意义**：

高尔顿遗传学的回归概念
(父母身高与子女身高的关系)

回归的**现代意义**：

一个应变量对若干解释变量依存关系的研究

回归的**目的（实质）**：

由固定的解释变量去估计应变量的平均值

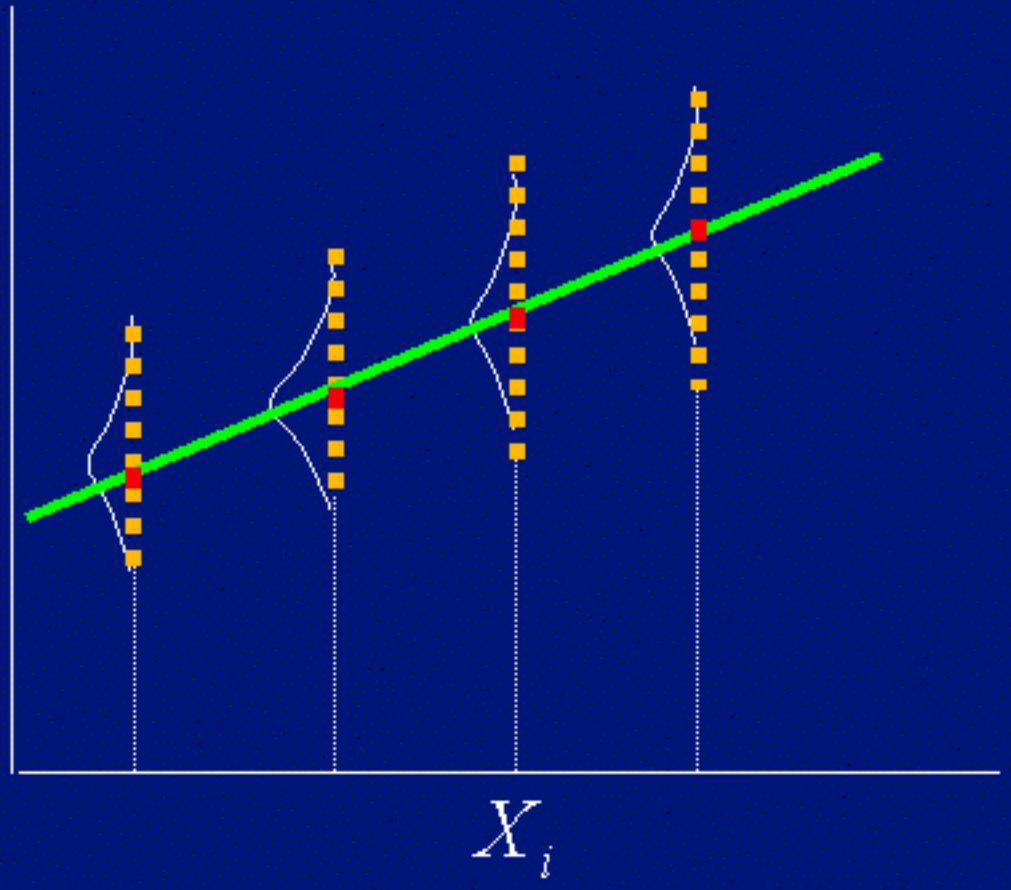
注意几个概念

- 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，这是Y的条件分布。

- Y的条件期望：

  对于X 的每一个取值，Y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。
- **回归函数：**应变量Y的条件期望E（Y | X_i）随解释变量X的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

$$E(Y \mid X_i) = f(X_i)$$

这个函数称为回归函数。

回归函数分为：总体回归函数

样本回归函数

三、总体回归函数（PRF）

举例：假如已知100个家庭构成的总体

	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				
$E(Y \mid X_i)$	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（随机设定形式）

对于一定的 X_i，Y的各个别值 Y_i 分布在 $E(Y|X_i)$ 的周围，若令各个别值 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

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

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

3、注意几点

- 实际的经济研究中总体回归函数通常是未知的，只能根据经济理论和实践经验去设定。“计量”的目的就是寻求PRF。
- 总体回归函数中Y与X的关系可是线性的，也可是非线性的。

对线性回归模型“线性”的两种解释：

就变量而言是线性的

——Y的条件均值是X的线性函数

就参数而言是线性的

——Y的条件均值是参数β的线性函数

判断：

$E(Y_i|X_i) = \beta_1 + \beta_2 X_i$ 变量、参数均”线性”

$E(Y_i|X_i) = \beta_1 + \beta_2 X^2{}_i$ 参数“线性”，变量”非线性”

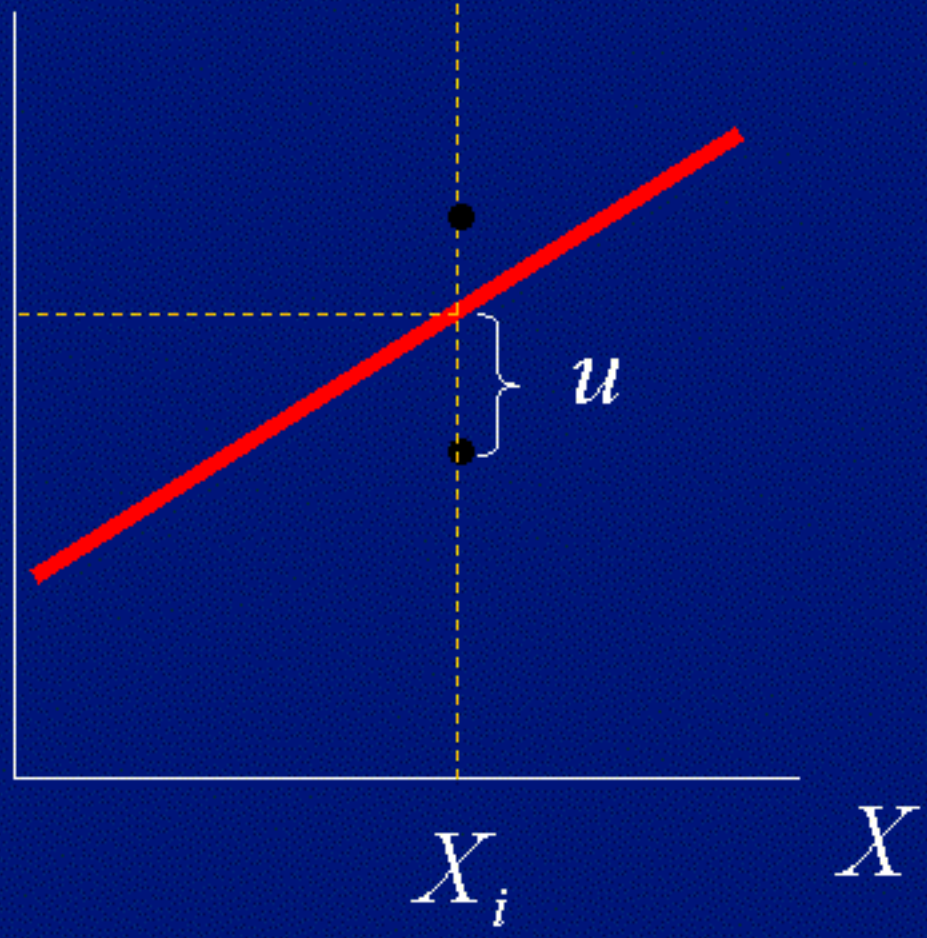
$E(Y_i|X_i) = \beta_1 + \sqrt{\beta_2} X_i$ 变量“线性”，参数”非线性”

计量经济学中线性回归模型主要指就参数是“线性”

三、随机扰动项u

◆概念

各个 Y_i 值与条件均值 $E(Y_i|X_i)$ 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

- 未知影响因素的代表
- 无法取得数据的已知影响因素的代表
- 众多细小影响因素的综合代表
- 模型的设定误差
- 变量的观测误差
- 变量内在随机性

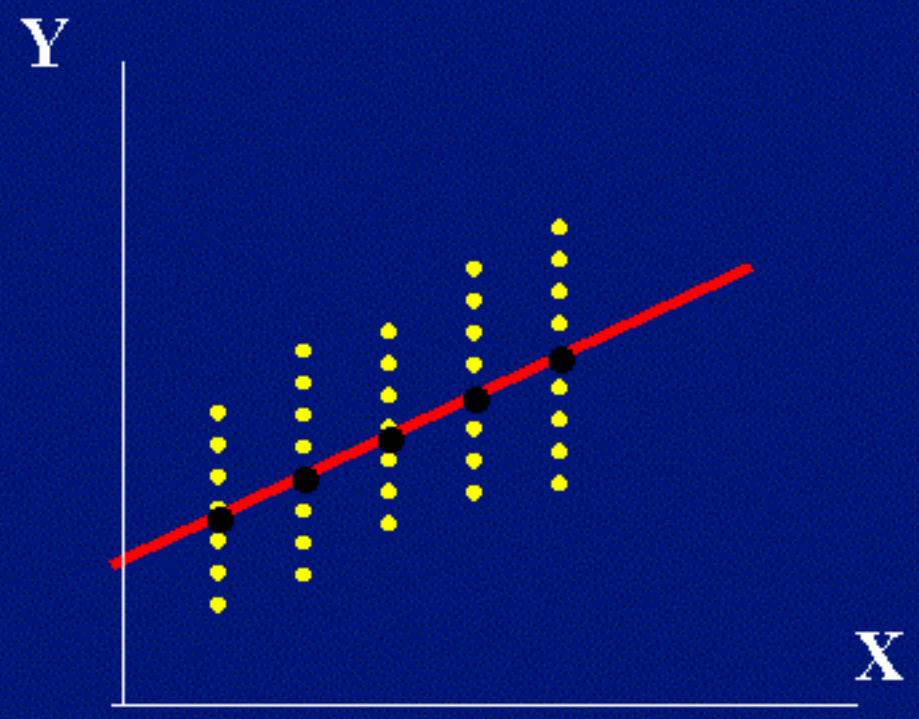
四、样本回归函数（SRF）

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。

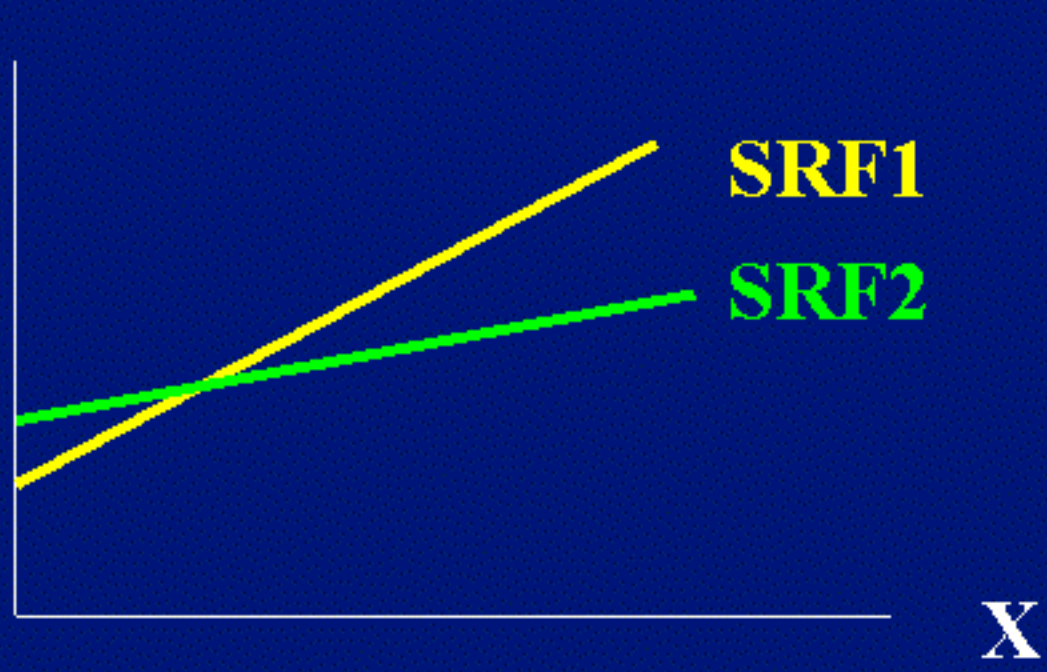
样本回归函数：

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，可以有许多条（**SRF不唯一**）。



- 样本回归函数的函数形式应与设定的总体回归函数的函数形式一致。
- 样本回归线还不是总体回归线，至多只是未知总体回归线的近似表现。

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：

或者

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

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

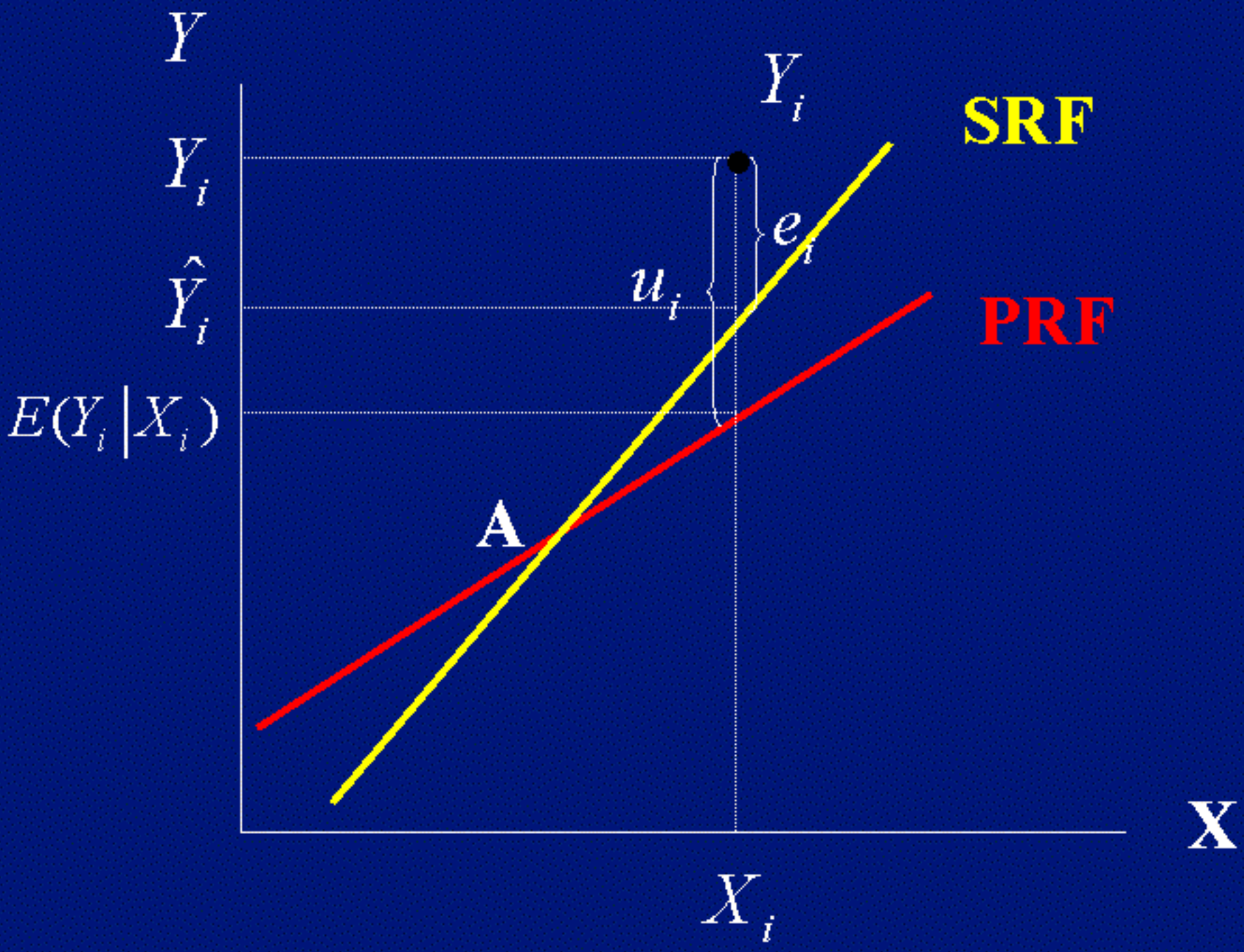
对样本回归的理解

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

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 β_1 和 β_2 的估计
- $\hat{Y}_i$ 是对总体条件期望 $\mathbf{E}(\mathbf{Y} \mid X_i)$ 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



回归分析的目的：

用样本回归函数SRF去估计总体回归函数PRF。

由于样本对总体总是存在代表性误差，SRF 总会过高或过低估计PRF。

要解决的问题：

寻求一种规则和方法，使得到的SRF的参数 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 尽可能“接近”总体回归函数中的参数 β_1 和 β_2。

这样的“规则和方法”有多种，最常用的是最小二乘法

第二节 简单线性回归模型的最小二乘估计

用样本去估计总体回归函数,除了样本以外,还需要一些前提条件——假定条件

一· 简单线性回归的基本假定

◆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。

基本假定的内容

- ◆对模型和变量的假定
- ◆对随机扰动项的假定

1、对模型和变量的假定

- ●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

$$\mathbf{E}(u_i \mid \mathbf{X}) = 0$$

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 σ^2

$$Var(u_i \mid X_i) = E[u_i - E(u_i \mid X_i)]^2 = \sigma^2$$

假定3：无自相关假定:

随机扰动项 u_i 的逐次值互不相关

$$\mathrm{Cov}(u_i, u_j)=E[u_i-E(u_i)][u_j-E(u_j)]$$

$$=E(u_i u_j)=0 \qquad (i \neq j)$$

假定4：随机扰动 u_i 与解释变量 X_i 不相关

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

（顺便提出）

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 σ^2 的正态分布

$$\mathbf{u} \sim \mathbf{N}(\mathbf{0}, \sigma^2)$$

（说明：正态性假定不影响对参数的点估计，所以可不列入基本假定，但这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）

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。 $\mathbf{Var}(Y \mid X_i) = \sigma^2$

假定3：无自相关假定。 $\mathbf{Cov}(Y_i, Y_j) = 0$

假定5：正态性假定。 $Y_i \sim \mathbf{N}(\beta_1 + \beta_2 X_i, \sigma^2)$

二、普通最小二乘法（OLS）
（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

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

取偏导数为0，得正规方程

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

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

用克莱姆法则求解得观测值形式的OLS估

计式：

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

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

为表达得简洁，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：

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

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

三、OLS回归线的性质

可以证明：（见P26—P27证明）

- 回归线通过样本均值

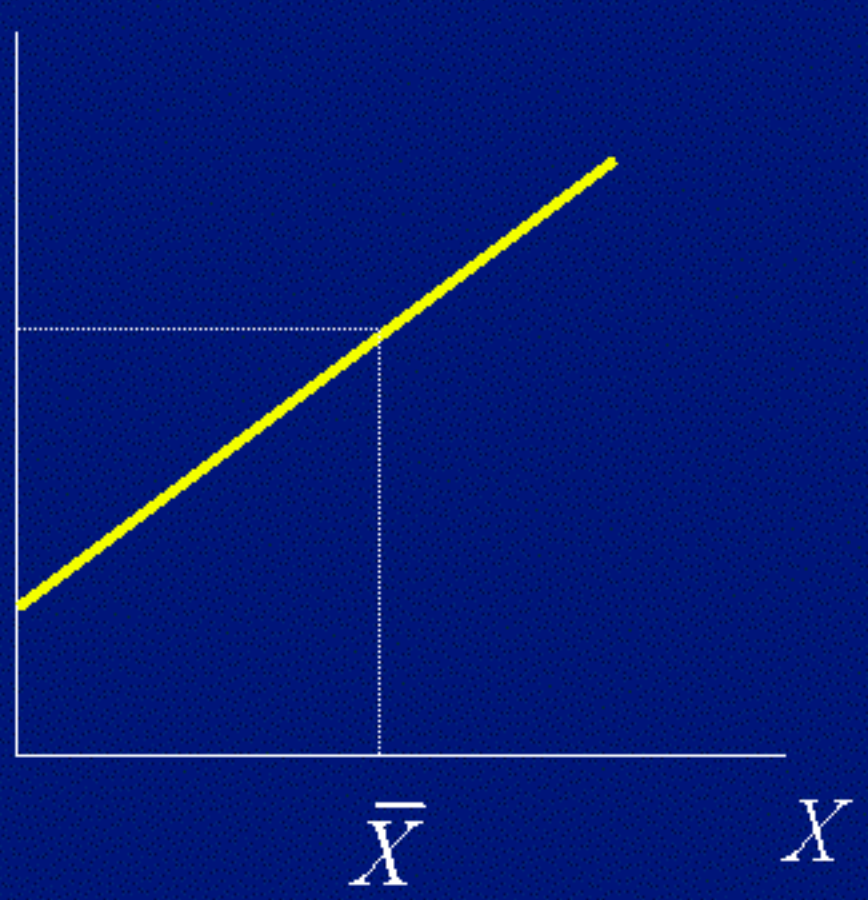
$$\overline{Y} = \hat{\beta}_1 + \hat{\beta}_2 \overline{X}$$

- 估计值 $\hat{Y}_i$ 的均值等于实际观测值 Y_i 的均值

$$\frac{\sum \hat{Y}_i}{N} = \overline{Y}$$

- 剩余项 e_i 的均值为零

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



- 应变量估计值 $\hat{Y}_i$ 与剩余项 e_i 不相关

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

- 解释变量 X_i 与剩余项 e_i 不相关

$$Cov(X_i, e_i) = 0$$

四、参数估计式的统计性质

（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

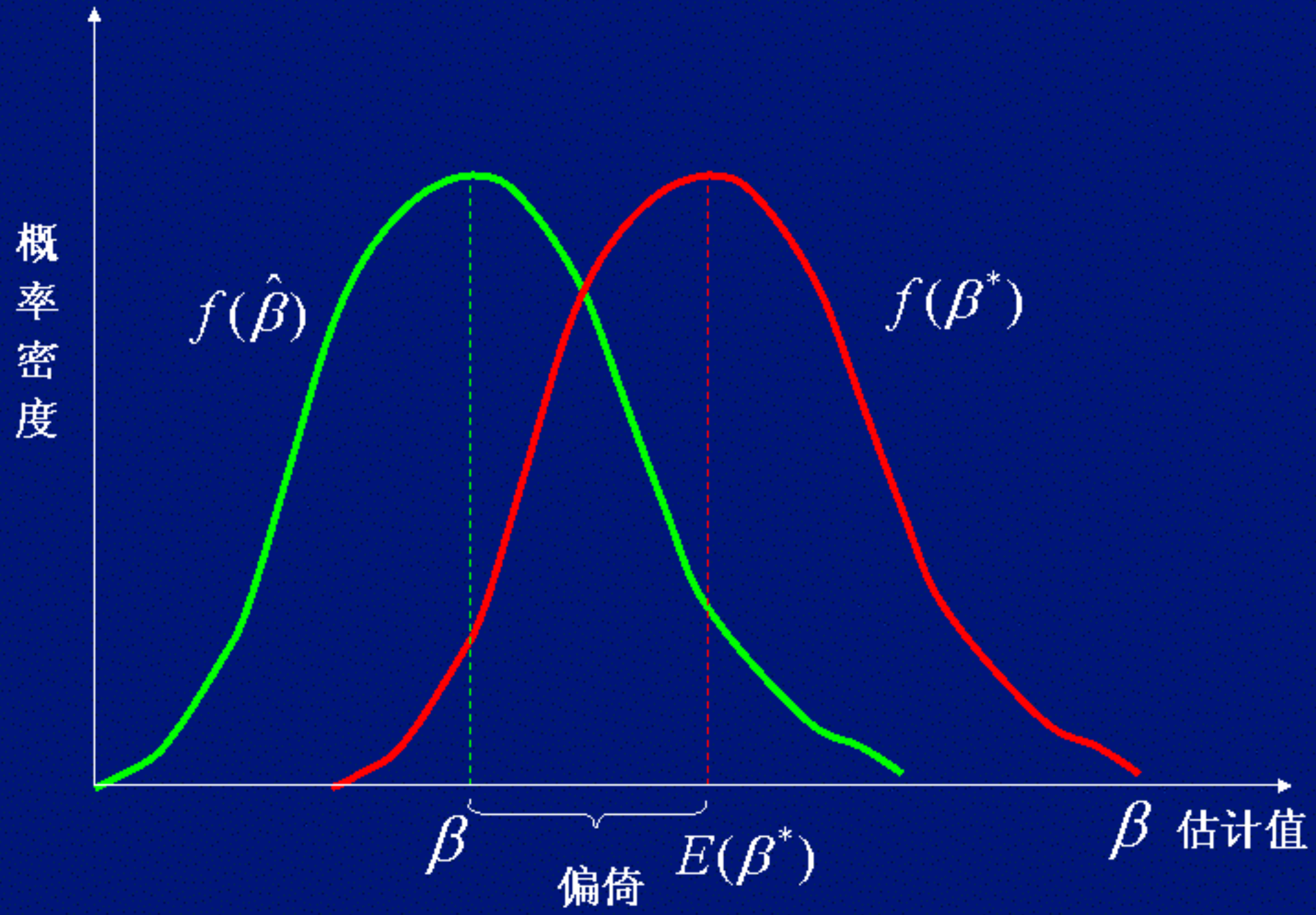
参数估计值 $\hat{\beta}$ 的分布称为 $\hat{\beta}$ 的抽样分布，其密度函数记为f($\hat{\beta}$)

如果 $E(\hat{\beta}) = \beta$

称 $\hat{\beta}$ 是参数 β 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



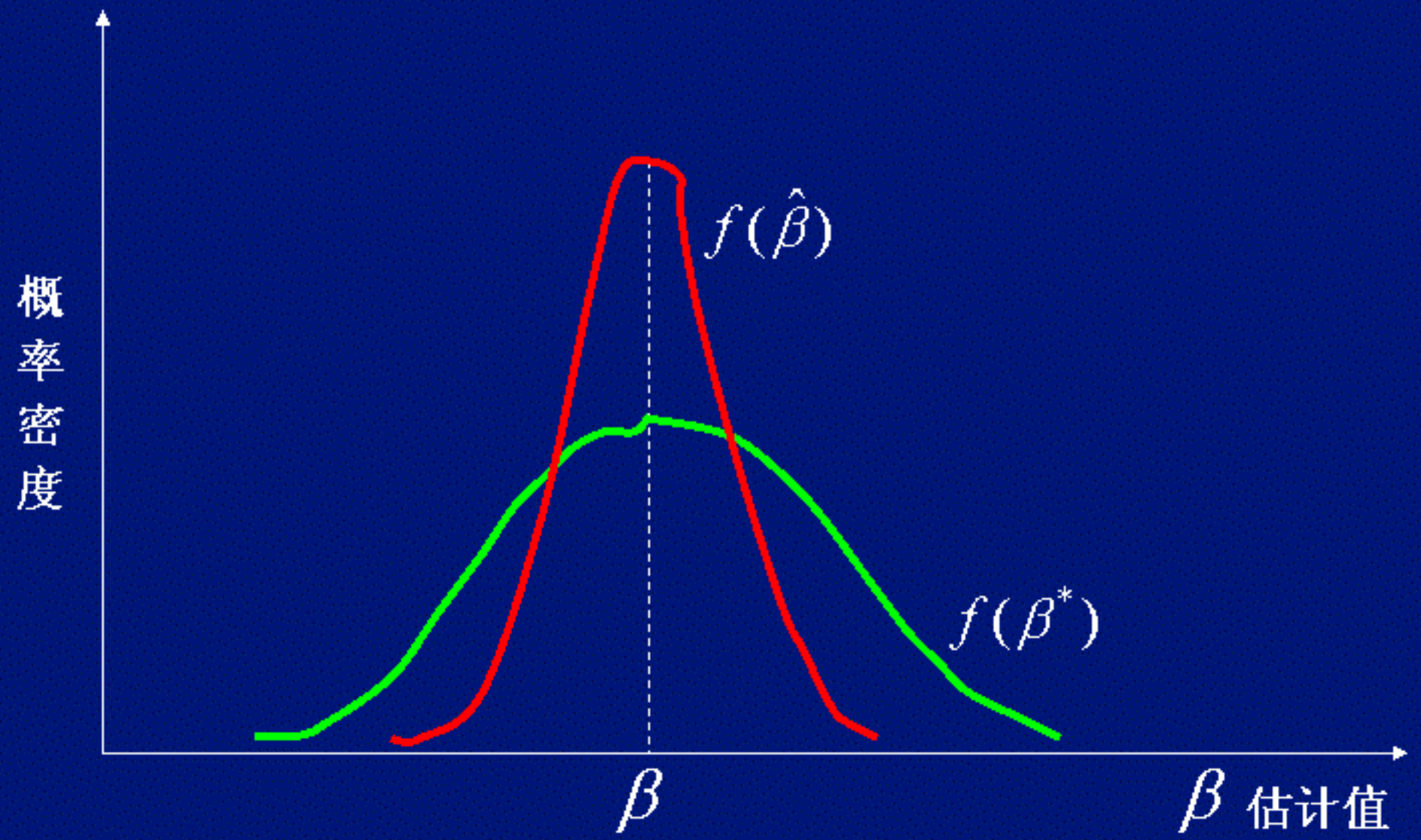
2、最小方差性

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



3、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

（估计方法不变，样本数逐步增大，分析其性质是否改善）

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，就称这个估计式 $\hat{\beta}$ 是 β 的一致估计式。即

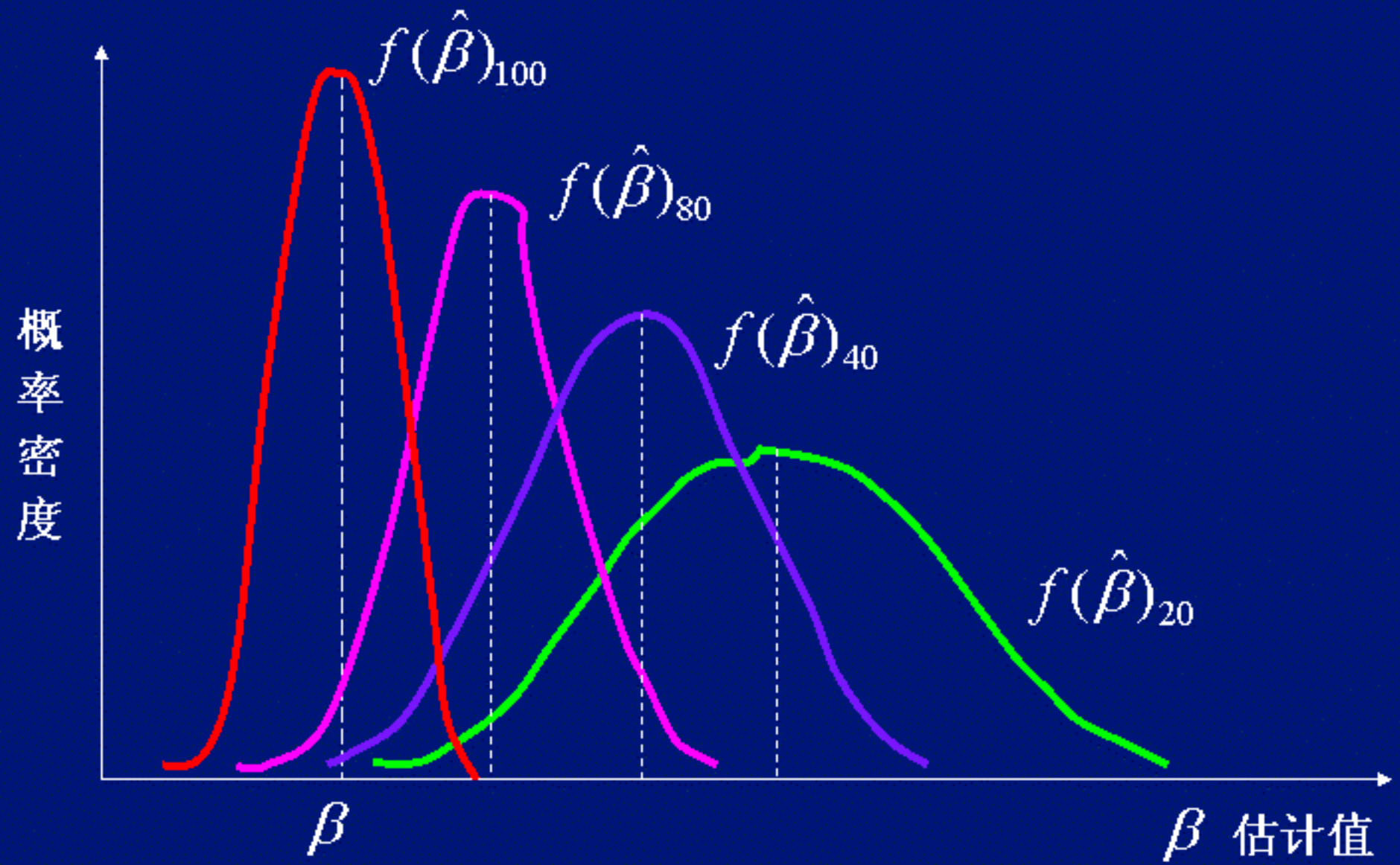
$$\lim P(|\hat{\beta}-\beta|\leq\varepsilon)=1$$

或 $$P\ \mathrm{Lim}_{n\to\infty}\hat{\beta}=\beta$$

（渐近无偏估计式是当样本容量变得足够大时其偏倚趋于零的估计式）

(见图1.4)

图 1.4


$f(\hat{\beta})_{100}$
$f(\hat{\beta})_{80}$
$f(\hat{\beta})_{40}$
$f(\hat{\beta})_{20}$
概率密度
β
β 估计值

（二） OLS估计式的统计性质

- 由**OLS**估计式可以看出

$$\hat{\beta}_2 = \frac{N\sum X_i Y_i - \sum X_i \sum Y_i}{N\sum X_i^2 - (\sum X_i)^2} \qquad \hat{\beta}_1 = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{N\sum X_i^2 - (\sum X_i)^2}$$

$\hat{\beta}$ 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 $\hat{\beta}$ 是随机变量
- **OLS**估计式是点估计式

1、 线性特征—— $\hat{\beta}$ 是Y的线性函数

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

2、 无偏特性 （证明见**P28**）

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

3、 最小方差特性 （证明见**P48**附录**2·1**）

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：OLS估计式是最佳线性无偏估计式（BLUE）

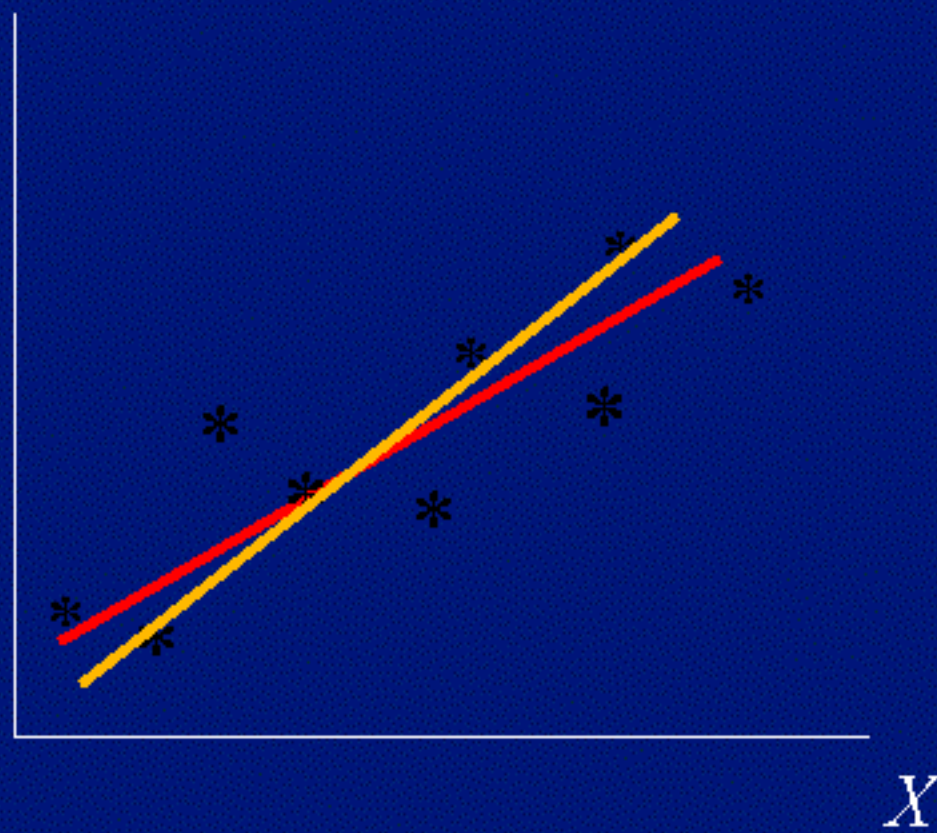
（高斯定理）

第四节　拟合优度的度量

概念：

样本回归线是对样本数据的一种拟合，不同估计方法可拟合出不同的回归线，拟合的回归线与样本观测值总有偏离。样本回归线对样本观测数据拟合的优劣程度

——拟合优度



拟合优度的度量建立在对总变差分解的基础上

一、总变差的分解

分析Y的观测值、估计值与平均值的关系

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

因为 $\sum (Y_i - \bar{Y}) = 0$,将上式两边平方加总，可证得

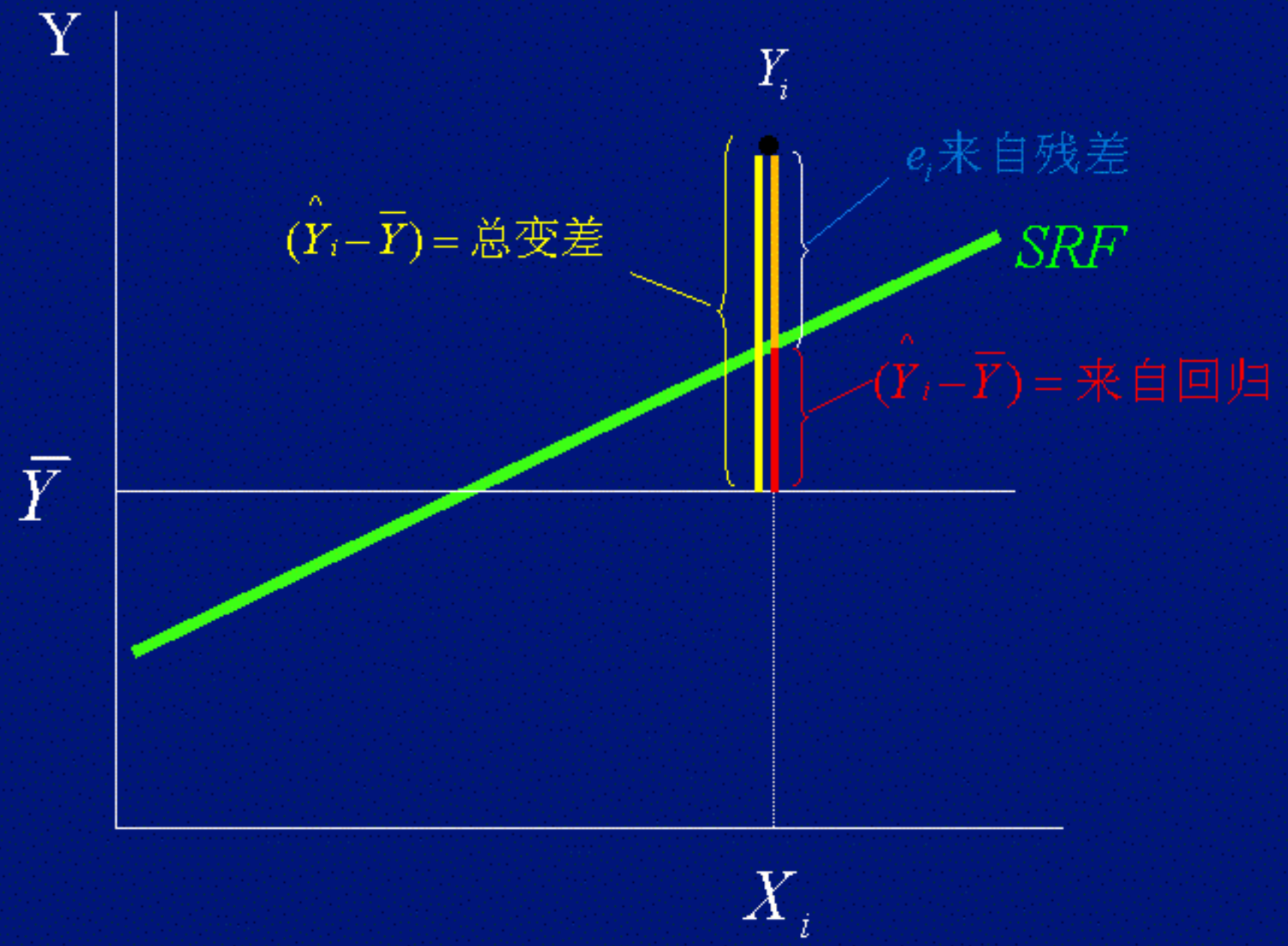
$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

（TSS） （ESS） （RSS）

总变差 $\sum y_i^2$（TSS）：应变量Y的观测值与其平均值的离差平方和（总平方和）

解释了的变差 $\sum \hat{y}_i^2$（ESS）：应变量Y的估计值与其平均值的离差平方和（回归平方和）

剩余平方和 $\sum e_i^2$（RSS）：应变量观测值与估计值之差的平方和（未解释的平方和）


Y
Y_i
e_i来自残差
$(\hat{Y}_i-\bar{Y})$=总变差
SRF
$(\hat{Y}_i-\bar{Y})$=来自回归
$\bar{Y}$
X_i

三、可决系数

以TSS同除总变差等式两边：

$$\frac{TSS}{TSS}=\frac{ESS}{TSS}+\frac{RSS}{TSS} \qquad \text{或} \qquad 1=\frac{\sum \hat{y}^2}{\sum y_i^2}+\frac{\sum e_i^2}{\sum y_i^2}$$

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：

$$r^2=\frac{\sum \hat{y}^2}{\sum y_i^2} \qquad \text{或} \qquad r^2=1-\frac{\sum e_i^2}{\sum y_i^2}$$

可决系数的作用：

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。反之可决系数越小，说明模型对样本观测值的拟合程度越差。

可决系数的特点：

- 可决系数取值范围： $0 \le r^2 \le 1$
- 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量

可决系数与相关系数的关系：

联系：数值上可决系数是相关系数的平方

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	说明两变量线性依存程度
度量的不对称的因果关系	度量的不含因果关系的对称相关关系
取值 $0 \le r^2 \le 1$ 有非负性	取值 $-1 \le r \le 1$ 可正可负

运用可决系数时应注意：

- 可决系数只是说明列入模型的所有解释变量对应变量的联合的影响程度，不说明模型中每个解释变量的影响程度（在多元中）
- 回归的主要目的如果是经济结构分析，不能只追求高的可决系数，而是要得到总体回归系数可信的估计量。可决系数高并不一定每个回归系数都可信任。
- 如果建模目的只是为了预测应变量值，不是为了正确估计回归系数，一般可考虑有较高的可决系数。

第三节 回归系数的区间估计和假设检验

为什么要作区间估计？

OLS估计只是通过样本得到的点估计，不一定等于真实参数，还需要找到真实参数的可能范围，并说明其可靠性

为什么要作假设检验？

OLS 估计只是用样本估计的结果，是否可靠？

是否抽样的偶然结果？还有待统计检验。

区间估计和假设检验都是建立在确定参数估计值 $\hat{\beta}$ 概率分布性质的基础上。

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区间估计和假设检验

u_i 是服从正态分布的随机变量，决定了 Y_i 也是服从正态分布的随机变量，$\hat{\beta}_k$ 是 Y_i 的线性函数，决定了 $\hat{\beta}_k$ 也是服从正态分布的随机变量

只要确定 $\hat{\beta}_k$ 的期望和方差，即可确定 $\hat{\beta}_k$ 的分布性质

$\hat{\beta}$ 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

  (标准误差是方差的平方根)

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

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

注意： 以上各式中 σ^2 未知，其余均是样本观测值

对随机扰动项方差 σ^2 的估计：

可以证明（见附录2.2)其无偏估计为

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

(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 σ^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 σ^2 未知时，

（1）当样本为小样本时，可用 $\hat{\sigma}^2$ 代替 σ^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，所得的 t 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：

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

（2）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）

二、回归系数的区间估计

概念：

对参数作出的点估计是随机变量，虽然是无偏估计，但还不能说明估计的可靠性和精确性，需要找到包含真实参数的一个范围，并确定这个范围包含参数真实值的可靠程度。

在确定参数估计式概率分布性质的基础上,可找到两个正数 δ 和 α （**0**≤ α ≤**1**），使得区间 $(\hat{\beta}_j-\delta,\hat{\beta}_j+\delta)$ 包含真实 β_j 的概率为**1**— α，

即

$$P(\hat{\beta}_j-\delta\le\beta_j\le\hat{\beta}_j+\delta)=1-\alpha$$

这样的区间称为所估计参数的置信区间。

讨论：怎样正确理解置信区间？

回归系数区间估计的方法

一般情况下，总体方差 σ^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 t 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。

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

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

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

即

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

三、回归系数的假设检验

1、假设检验的基本思想

◆为什么要作假设检验？

所估计的回归系数 $\hat{\beta}_1$ 、$\hat{\beta}_2$ 和方差 $\hat{\sigma}^2$ 都是通过样本估计的，都是随抽样而变动的随机变量，它们是否可靠？是否抽样的偶然结果呢？还需要加以检验。

从哪些方面检验？（回顾“导论”）

- 经济意义检验：用先验理论检验，看是否与经济理论一致，是否合乎情理
- 统计推断检验：

  各个回归系数的显著性检验

  回归总显著性检验

  模型拟合程度检验
- 计量经济学检验：

  是否符合估计方法的基本假定
- 预测检验：将估计的模型用于实际经济过程的预测，检验其预测效果

对回归系数假设检验的方式

目的：对简单线性回归，判断解释变量X是否是被解释变量Y一个显著的影响因素。在一元线性模型中，就是要判断X是否对Y具有显著的线性性影响。这就需要进行变量的显著性检验。

思想：变量的显著性检验的方法是假设检验。假设检验采用的逻辑推理方法是反证法。先假定原假设正确，然后根据样本信息，观察由此假设而导致的结果是否合理，从而判断是否接受原假设。判断结果合理与否，是基于“小概率事件不易生”的原理。

计量经济学中，主要是针对变量的参数真值是否为零来进行显著性检验的。

2、回归系数的检验方法

一般情况下，总体方差 σ^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 α ，查 t 分布表得 $t_{\alpha/2}(n-2)$

▼如果 $t^* \le -t_{\alpha/2}(n-2)$ 或者 $t^* \ge -t_{\alpha/2}(n-2)$ （小概率事件发生）

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 $-t_{\alpha/2}(n-2) \le t^* \le t_{\alpha/2}(n-2)$ （大概率事件发生）

则接受原假设 $H_0: \beta_2 = 0$

用 P 值判断参数的显著性

假设检验的 p 值：

p 值是根据既定的样本数据所计算的统计量，拒绝原假设的最小显著性水平

统计分析软件中通常都给出了检验的 p 值

方法：将给定的显著性水平 α 与 p 值比较：

- 若 $\alpha > p$ 值，则在显著性水平 α下拒绝原假设

  $H_0: \beta_k = 0$ ，即认为 X 对 Y 有显著影响

- 若 $\alpha \leq p$ 值，则在显著性水平 α下接受原假设

  $H_0: \beta_k = 0$ ，即认为 X 对 Y 没有显著影响

规则：当 $p < \alpha$ 时，P值越小，越能拒绝原假设 H_0

第五节　回归模型预测

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以下规范化的方式：

例如：回归结果为

$$\hat{Y}_i = 24{\cdot}4545 + 0{\cdot}5091 X_i$$

（6·4138）　（0·0357）　　标准误差SE

t = (3·8128)　(14·2605)　　t 统计量

$R^2 = 0{\cdot}9621$　　df = 8　　可决系数和自由度

F = 202·87　　DW = 2.3　　F 统计量　DW统计量

二、应变量平均值预测

1、基本思想

- 运用计量经济模型作预测是利用所估计的样本回归函数，用解释变量的已知值或预测值，对预测期或样本以外的应变量数值作出定量的估计。
- 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

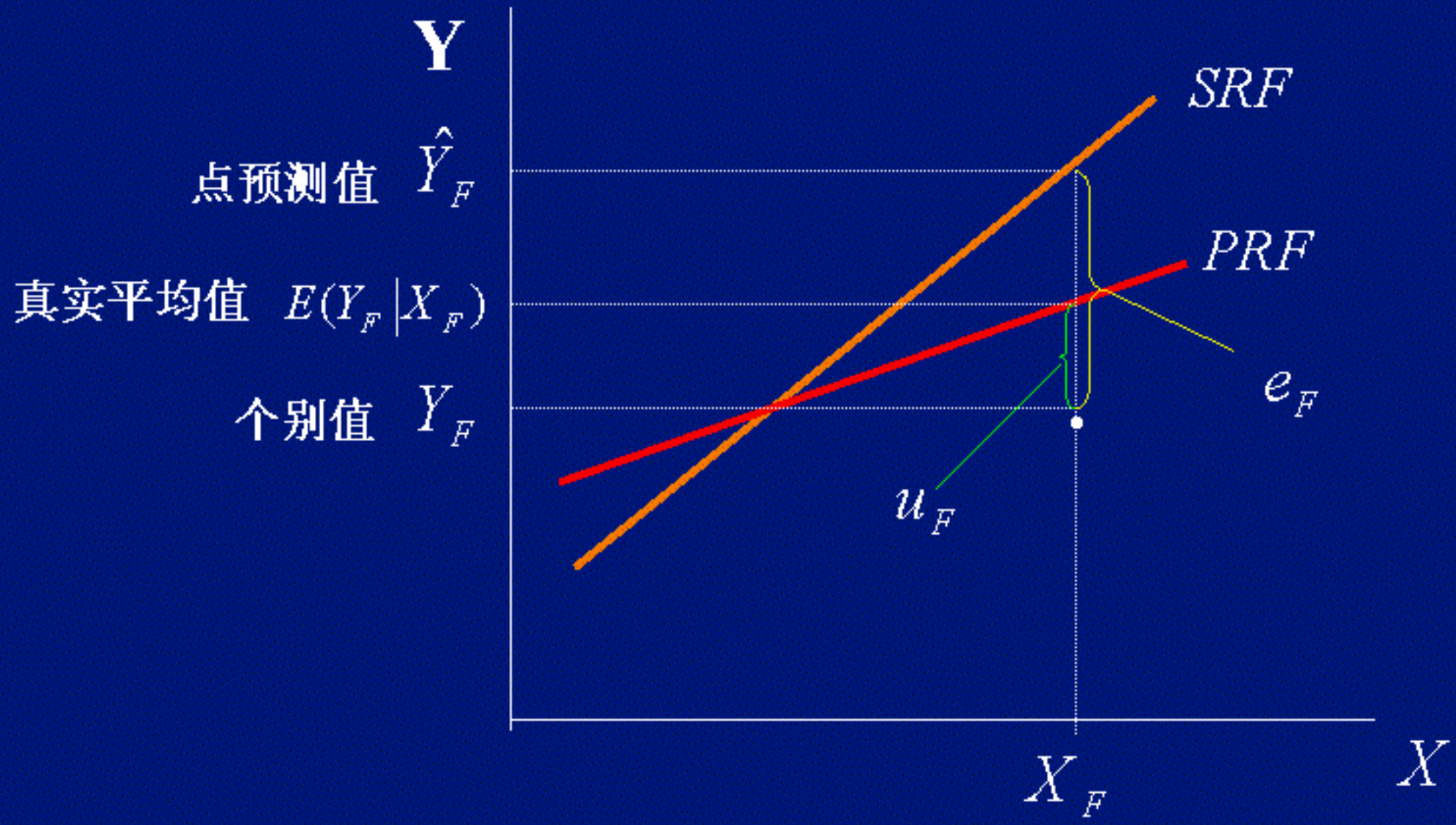
所估计的参数不变

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

可以证明 $$Var(\hat{Y}_F) = \sigma^2 [\frac{1}{N} + \frac{(X_F - \overline{X})^2}{\sum x_i^2}]$$

$$SE(\hat{Y}_F) = \sigma \sqrt{\frac{1}{N} + \frac{(X_F - \overline{X})^2}{\sum x_i^2}}$$

$\hat{Y}_F$ 服从正态分布(为什么?)，将其标准化,当 σ^2 未知时，只得用 $\hat{\sigma}^2 = \sum e_i^2 / (n-2)$ 代替，这时有

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

给定显著性水平α，查t分布表，得自由度n－2的临界值 $t_{\alpha/2}(n-2)$ 则有

$$p\{[\hat{Y}_F - t_{\alpha/2}\hat{SE}(\hat{Y}_F)] \le E(Y_F \mid X_F) \le [\hat{Y}_F + t_{\alpha/2}\hat{SE}(\hat{Y}_F)]\} = 1-\alpha$$

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、应变量个别值预测

基本思想：

- $\hat{Y}_F$ 既是对Y平均值的点预测，也是对Y个别值的点预测。

- 由于存在随机扰动 u_i 的影响，Y的平均值并不等于Y的个别值

- 为了对Y的个别值 Y_F 作区间预测，需要寻找与预测值 $\hat{Y}_F$ 和个别值 Y_F 有关的统计量，并要明确其概率分布

具体作法：

已知剩余项 $e_F = Y_F - \hat{Y}_F$ 是与预测值 $\hat{Y}_F$ 和个别值 Y_F 都有关的变量,并且已知 e_F 服从正态分布，且可证明 $E(e_F) = 0$

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

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

$$t = \frac{e_F - E(e_F)}{\hat{SE}(e_F)} = \frac{Y_F - \hat{Y}_F}{\hat{SE}(e_F)} \sim t(n-2)$$

构建个别值置信区间

给定显著性水平 α，查 t 分布表得自由度为N—2的临界值 $t_{\alpha/2}(n-2)$，则有

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

因此，一元回归时Y的个别值的置信度为1－α 的预测区间上下限为

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

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

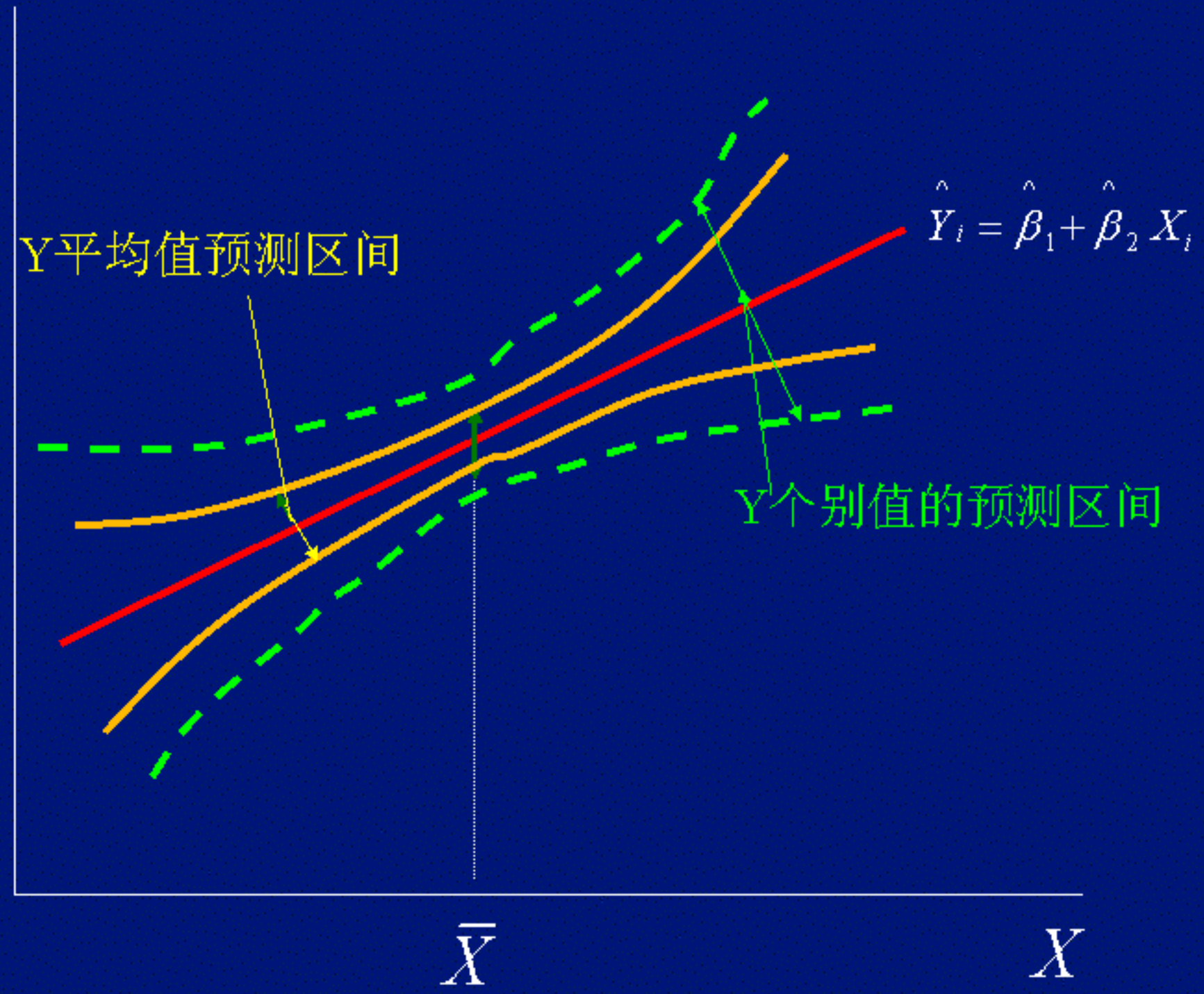
Y个别值的预测值与真实个别值的差异,不仅受抽样波动影响，而且还受随机扰动项的影响

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

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→∞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



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。但全国各地区经济发展速度不同，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)

数据收集：从2002年《中国统计年鉴》中得到数据：

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 $\hat{Y}_i = 282.2434 + 0.758511X_i$

（287.2649） (0.036928)

t=(0.982520) (20.54026)

$r^2 = 0.935685$ F=421.9023 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**：$\alpha = 0.05 >> p=0.0000$

4、经济意义检验： 估计的X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。

5、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

$$\hat{Y}_{F1} = 282.2434 + 0.758511 \times 12405 = 9691.577$$

区间预测：

平均值区间预测上下限：$$Y_F = \hat{Y}_F \mp t_{\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(X_F-\overline{X})^2}{\sum x_i^2}}$$

$X_{f1}=8270$时

$$Y_{f1}=6555.13\mp 2.045\times 413.1593\times\sqrt{\frac{1}{31}+\frac{569985.74}{125176492.59}}$$

$$=6555.13\mp 162.10$$

$X_{f2}=12405$时

$$Y_{f2}=9691.58\mp 2.045\times 413.1593\times\sqrt{\frac{1}{31}+\frac{23911845.72}{125176492.59}}$$

$$=9691.58\mp 499.25$$

即是说：

$X_{f1}=8270$时，平均值置信度95%的预测区间为（6393.03，6717.23）元。

$X_{f2}=12405$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系的研究

回归的实质：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。

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

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。

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

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：

对随机扰动项u的假定：

零均值假定：$E(u_i)=0$　　$E(Y_i)=\beta_1+\beta_2X_i$

同方差假定：$Var(u_i)=Var(Y_i)=\sigma^2$

无自相关假定：$Cov(u_i,u_j)=E(u_iu_j)=0$

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；

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

$$\hat{\beta}_1 = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{N\sum X_i^2 - (\sum X_i)^2} \qquad \hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}$$

OLS 估计式的分布性质

期望 $E(\hat{\beta}_k) = \beta_k$

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

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

OLS估计式是最佳线性无偏估计式。

7、σ^2 的无偏估计 $\hat{\sigma}^2 = \dfrac{\sum e_i^2}{n-2}$

8、对回归系数区间估计的思想和方法。

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

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。

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

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，

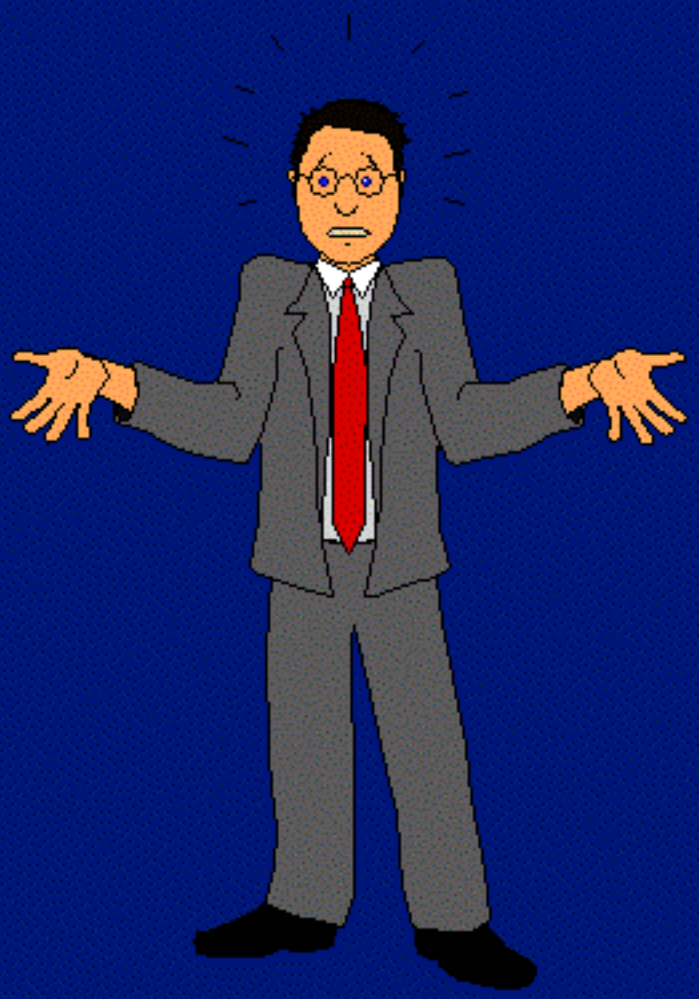
$$[\hat{Y}_F - t_{\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(X_F-\overline{X})^2}{\sum x_i^2}}, \hat{Y}_F + t_{\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(X_F-\overline{X})^2}{\sum x_i^2}}]$$

被解释变量个别值区间预测的方法。

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

12、运用EViews软件实现对简单线性回归模型的估计和检验。

第二章结束了！



THANKS