ECONOMETRICS

第 二 章

简单线性回归模型

引子：中国旅游业将达到世界旅游强国水平

《中国旅游业“十二五”发展规划纲要》提出，到“十二五”期末，中国的旅游业初步建设成为国民经济的战略性支柱产业和人民群众更加满意的现代服务业。2015年中国旅游业总收入达到2.3万亿元，年均增长率为10%，旅游业增加值占全国GDP的比重提高到4.5%，占服务业增加值的比重达到12%，旅游消费相当于居民消费总量的比例达到10%。力争2020年我国旅游产业规模、质量、效益基本达到世界旅游强国水平。

是什么决定性的因素能使中国旅游业基本达到世界旅游强国水平？旅游业的发展与这种决定性因素的数量关系究竟是什么？

- 什么决定性因素能使中国旅游业总收入超过2.3万亿元?
- 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?

ECONOMETRICS 需要研究经济变量之间数量关系的方法

显然，对旅游起决定性影响作用的是“中国居民的收入水平”以及“入境旅游人数”等因素。

“旅游业总收入”（Y）与“居民平均收入”（X1）或者“入境旅游人数”（X2）有怎样的数量关系呢？

能否用某种线性或非线性关系式 Y= f(X) 去表现这种数量关系呢？具体该怎样去表现和计量呢？

为了不使问题复杂化，我们先在某些标准的(古典的)假定条件下，用最简单的模型，对最简单的变量间数量关系加以讨论

ECONOMETRICS

第一节 回归分析与回归函数

一、相关分析与回归分析

（对统计学的回顾）

1、经济变量之间的相互关系

性质上可能有三种情况：

◆确定性的函数关系　　Y=f (X)　　可用数学方法计算

◆不确定的统计关系—相关关系

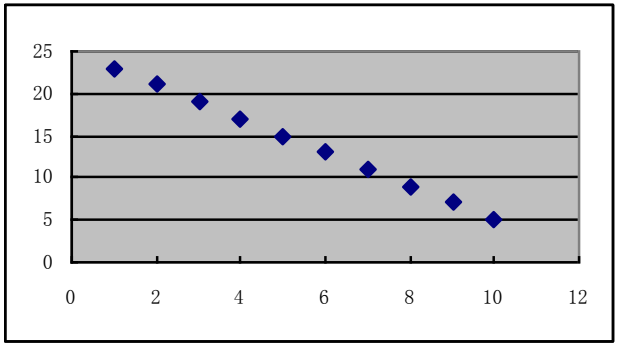
Y= f（X）+ε　　(ε为随机变量)　可用统计方法分析

◆没有关系　　不用分析

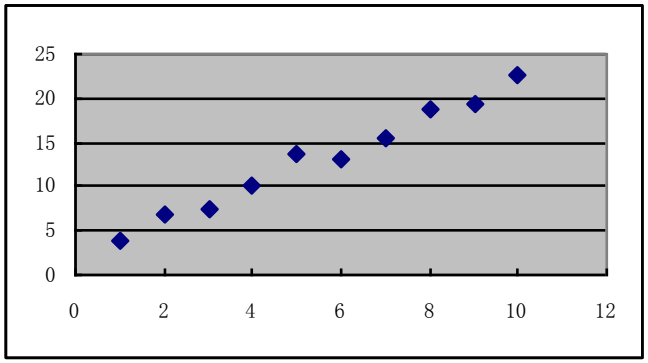
2、相关关系

◆ 相关关系的描述

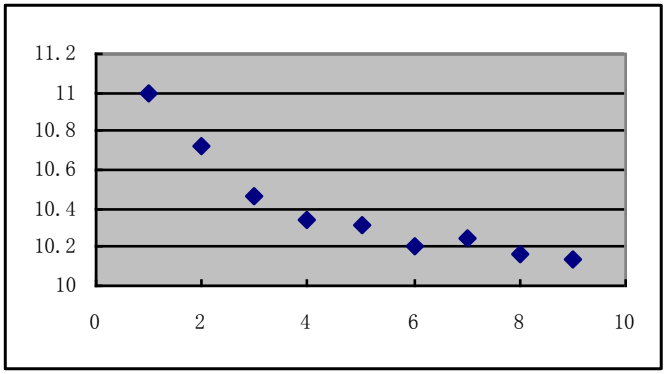
最直观的描述方式——坐标图（散布图、散点图））



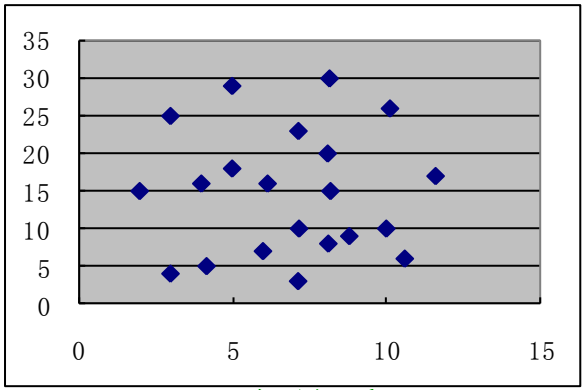
函数关系



相关关系(线性)



相关关系(非线性)



没有关系

ECONOMETRICS

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

6

3、相关程度的度量——相关系数

如果X和Y 总体的全部数据都已知，X和Y 的方差和协方差也已知，则

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

特点：

●总体相关系数只反映总体两个变量X和Y 的线性相关程度

●对于特定的总体来说，X 和 Y 的数值是既定的，总体相关系数 ρ 是客观存在的特定数值。

●总体的两个变量 X 和 Y的全部数值通常不可能直接观测，所以总体相关系数一般是未知的。

ECONOMETRICS

X和Y的样本线性相关系数：

如果只知道 X 和 Y 的样本观测值，则X和Y的样本线性相关系数为：

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

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

注意：r_{XY} 是随抽样而变动的随机变量。

相关系数较为简单，也可以在一定程度上测定变量间的数量关系，但是对于具体研究变量间的数量规律性还有局限性。

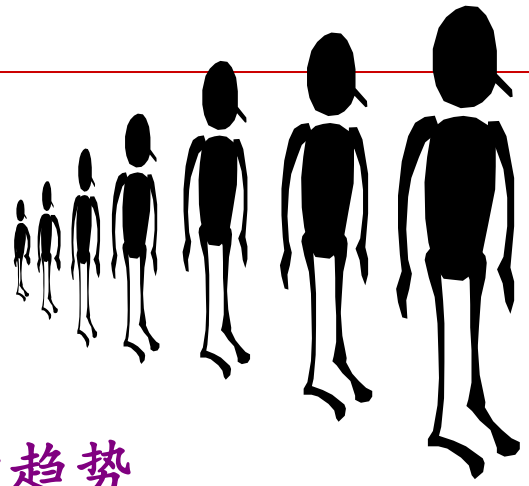
ECONOMETRICS

对相关系数的正确理解和使用

- X和Y 都是相互对称的随机变量，$r_{XY} = r_{YX}$
- 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，样本相关系数是随抽样而变动的随机变量，其统计显著性还有待检验

ECONOMETRICS

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）

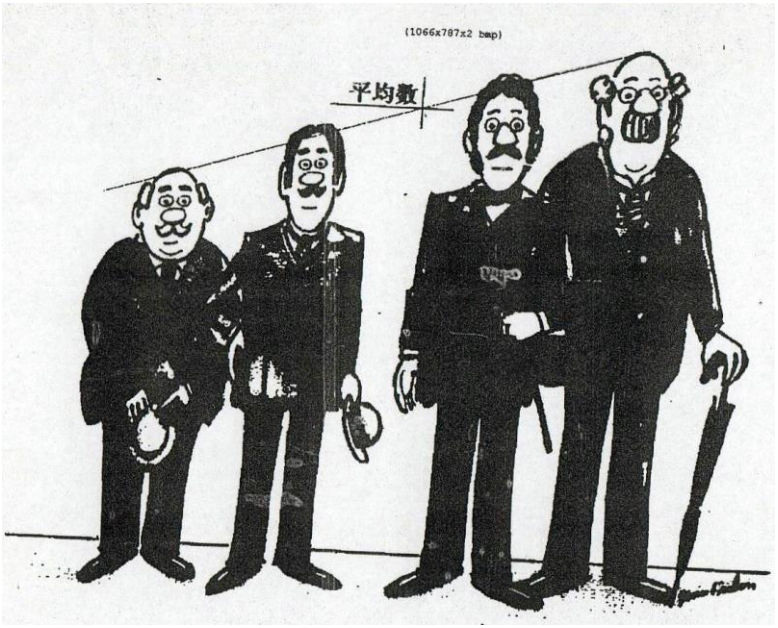
子女的身高有向人的平均身高"回归"的趋势

回归的现代意义：

一个被解释变量对若干个解释变量依存关系的研究

回归的目的（实质）：

由解释变量去估计被解释变量的平均值

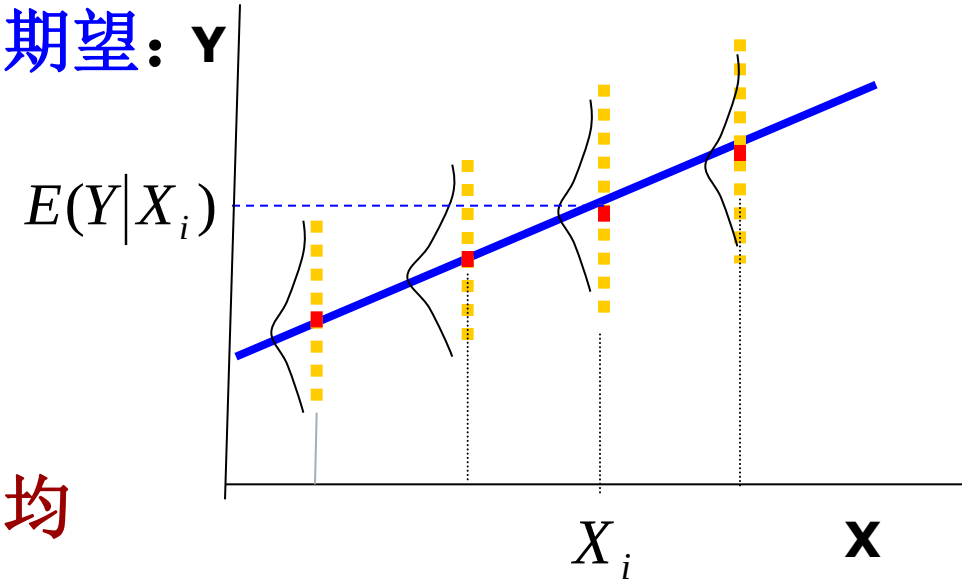


明确几个概念（为深刻理解“回归”）

●被解释变量Y的条件分布和条件概率：

当解释变量X取某固定值时（条件），Y 的值不确定，Y 的不同取值会形成一定的分布，这是 Y 的条件分布。 X 取某固定值时，Y 取不同值的概率称为条件概率。

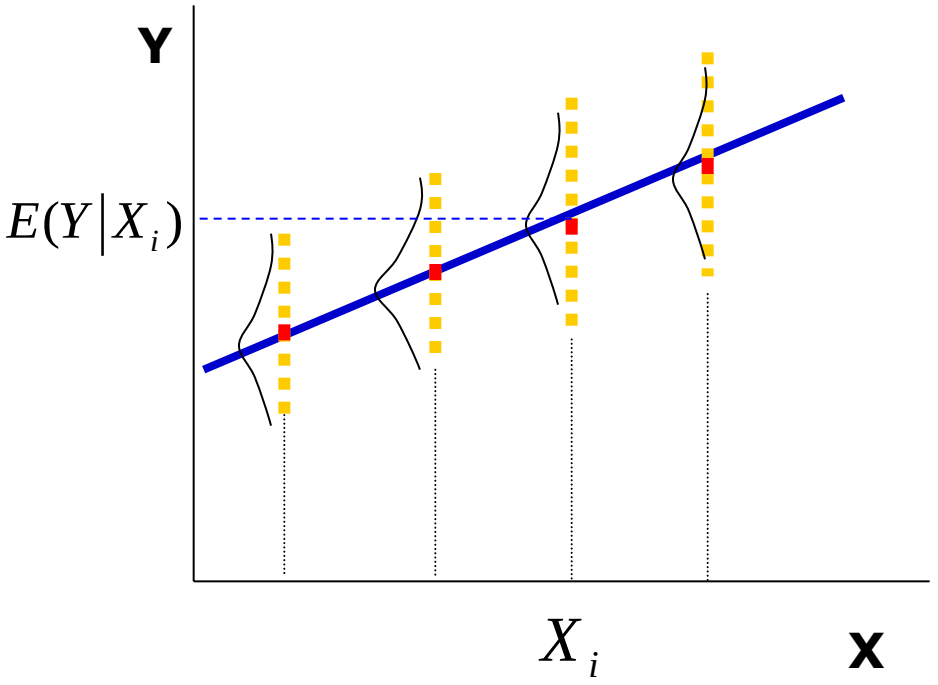
●被解释变量 Y 的条件期望：对于 X 的每一个取值，对 Y 所形成的分布确定其期望或均值，称为 Y 的条件期望或条件均值，用 $E(Y|X_i)$表示。



注意:Y的条件期望是随X的变动而变动的

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

● 回归函数：被解释变量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的条件期望表现为 X 的某种函数 $\mathrm{E}(Y|X_i)=f(X_i)$ ，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

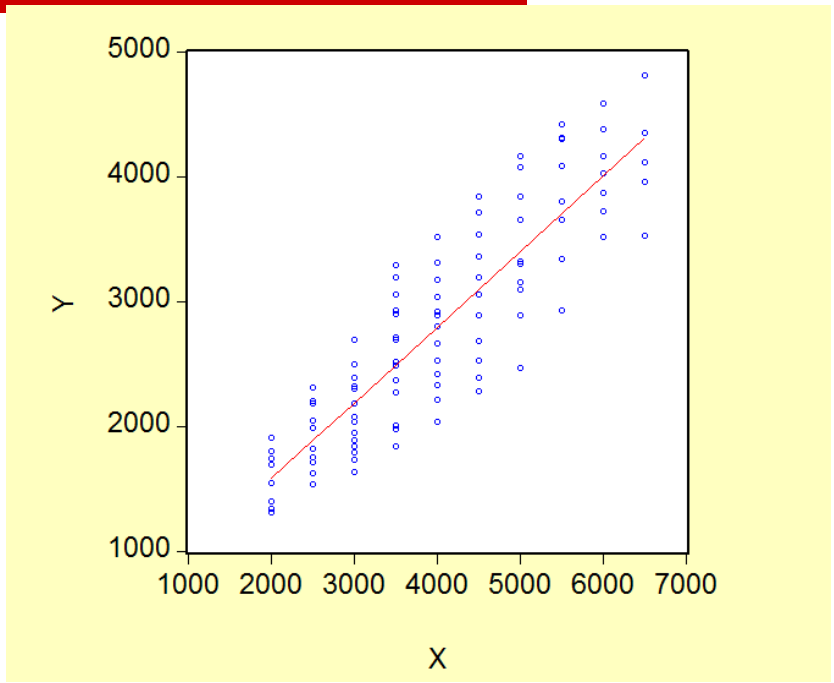
二、总体回归函数（PRF）

举例：假如已知由100个家庭构成的总体的数据　　(单位:元)

	每月家庭可支配收入X									
	3000	3500	4000	4500	5000	5500	6000	6500	7000	7500
每月家庭消费支出Y	1819	2027	2269	2304	2646	2917	3068	3383	4107	4267
	1847	2118	2364	2435	2819	3028	3488	3797	4313	4800
	1907	2212	2424	2467	2934	3166	3689	4109	4457	5004
	2055	2248	2473	2726	3028	3321	3755	4261	4618	5241
	2195	2313	2523	2828	3131	3527	3899	4546	4757	5408
	2245	2481	2581	2946	3244	3690	3920	4757	4972	
	2307	2541	2675	2976	3408	3829	4253	4771	5172	
	2409	2686	2716	3150	3496	3993	4441	4872		
		2702	2817	3174	3522	4174	4673			
		2812	2936	3349	3677	4350	4764			
			2954	3384	3776	4474				
			3025	3514	3919					
			3136	3658	4119					
			3327	3747						
$E(Y\|X_i)$	2098	2414	2730	3047	3363	3679	3995	4312	4628	4944

消费支出的条件期望与收入关系的图形

$E(Y|X_i)$



X_i

对于本例的总体，家庭消费支出的条件期望 $E(Y|X_i)$ 与家庭收入 X_i 基本是线性关系，可以把家庭消费支出的条件均值表示为家庭收入的线性函数：

$$E(Y|X_i)=\alpha+\beta X_i$$

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的被解释变量Y和解释变量X的每个观测值（通常这是不可能的！），那么，可以计算出总体被解释变量Y的条件期望 $E(Y|X_i)$，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）

本质：总体回归函数实际上表现的是特定总体中被解释变量随解释变量的变动而变动的某种规律性。

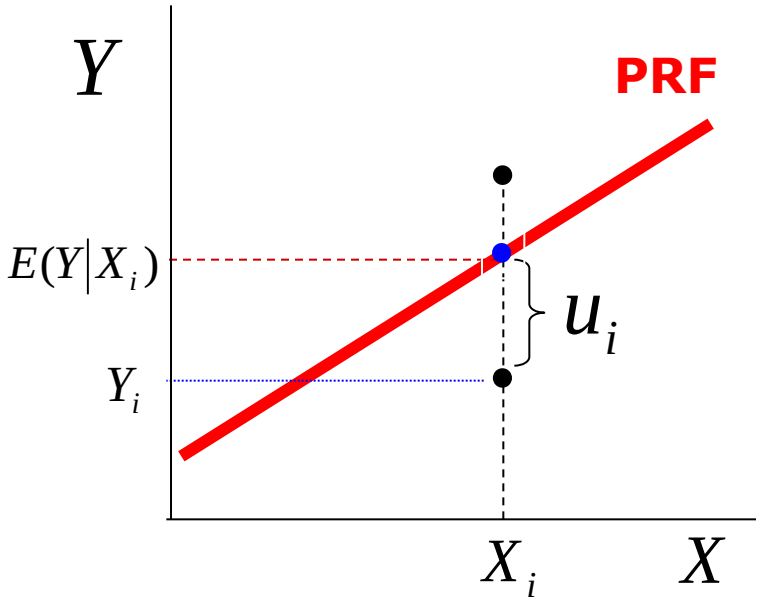
计量经济学的根本目的是要探寻变量间数量关系的规律,也就要努力去寻求总体回归函数。

2.总体回归函数的表现形式

●条件期望表现形式

例如Y的条件期望 $E(Y|X_i)$ 是解释变量X的线性函数，可表示为：

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



●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 并不一定等于条件期望，而是分布在 $E(Y|X_i)$ 的周围，若令各个 Y_i 与条件期望 $E(Y|X_i)$ 的偏差为 u_i，显然 u_i 是个随机变量

则有

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

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

ECONOMETRICS

3.如何理解总体回归函数

●作为总体运行的客观规律，总体回归函数是客观存在的，但在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。

计量经济学研究中“计量”的根本目的就是要寻求总体回归函数。

●我们所设定的计量模型实际就是在设定总体回归函数的具体形式。

●总体回归函数中 Y 与 X 的关系可以是线性的，也可以是非线性的。

“线性”的判断

计量经济学中，线性回归模型的**“线性”**有两种解释：

- **就变量而言**是线性的

  ——Y的条件期望（均值）是X的线性函数

- **就参数而言**是线性的

  ——Y的条件期望（均值）是参数β的线性函数

例如： $E(Y_i|X_i)=\beta_1+\beta_2X_i$ 对变量、参数均为“线性”

$E(Y_i|X_i)=\beta_1+\beta_2\ln X_i$ 对参数“线性”，对变量”非线性”

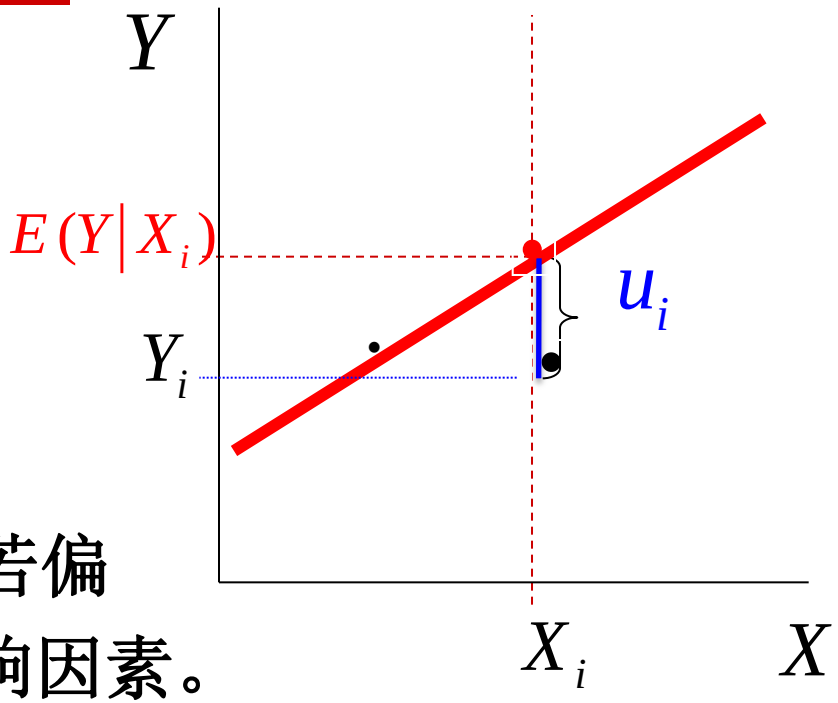
$E(Y_i|X_i)=\beta_1+\sqrt{\beta_2}X_i$ 对变量“线性”，对参数”非线性”

注意：在计量经济学中，线性回归模型主要指就参数而言是“线性”的，因为只要对参数而言是线性的，都可以用类似的方法去估计其参数，都可以归于线性回归。

三、随机扰动项

◆概念

在总体回归函数中，各个 Y_i 的值与其条件期望 $E(Y|X_i)$ 的偏差 u_i 有很重要的意义。若只有 X 的影响，Y_i 与 $E(Y|X_i)$ 不应有偏差。若偏差 u_i 存在，说明还有其他影响因素。



u_i 实际代表了排除在模型以外的所有因素对 **Y** 的影响。

◆性质

u_i 是其期望为 **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分析结果的性质和计量经济方法的选择

ECONOMETRICS

引入随机扰动项u_i的原因

- 是未知影响因素的代表(理论的模糊性)
- 是无法取得数据的已知影响因素的代表(数据欠缺)
- 是众多细小影响因素的综合代表(非系统性影响)
- 模型可能存在设定误差(变量、函数形式的设定)
- 模型中变量可能存在观测误差(变量数据不符合实际)
- 变量可能有内在随机性(人类经济行为的内在随机性)

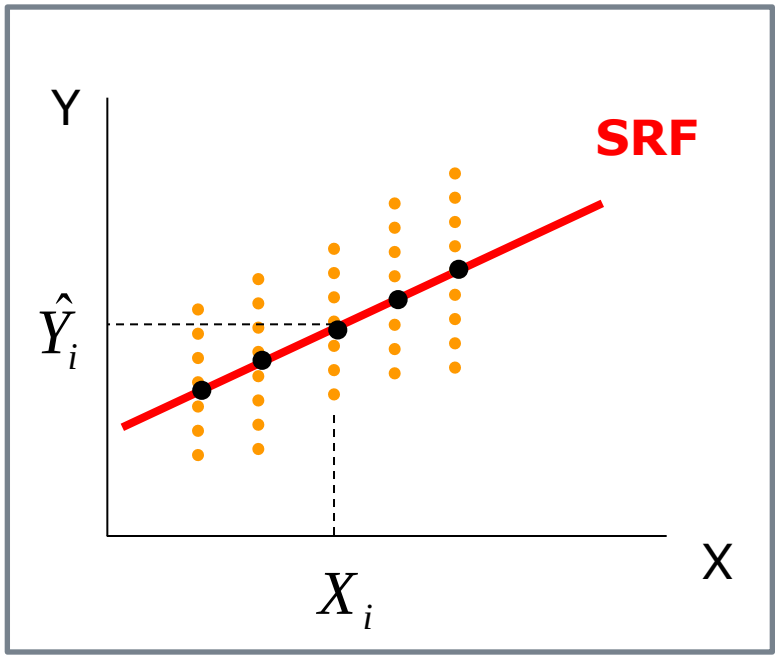
四、样本回归函数（SRF）

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：

如果把被解释变量**Y**的样本条件均值 $\hat{Y}_i$ 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**）



ECONOMETRICS

样本回归函数的函数形式

条件均值形式：

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

个别值（实际值）形式：

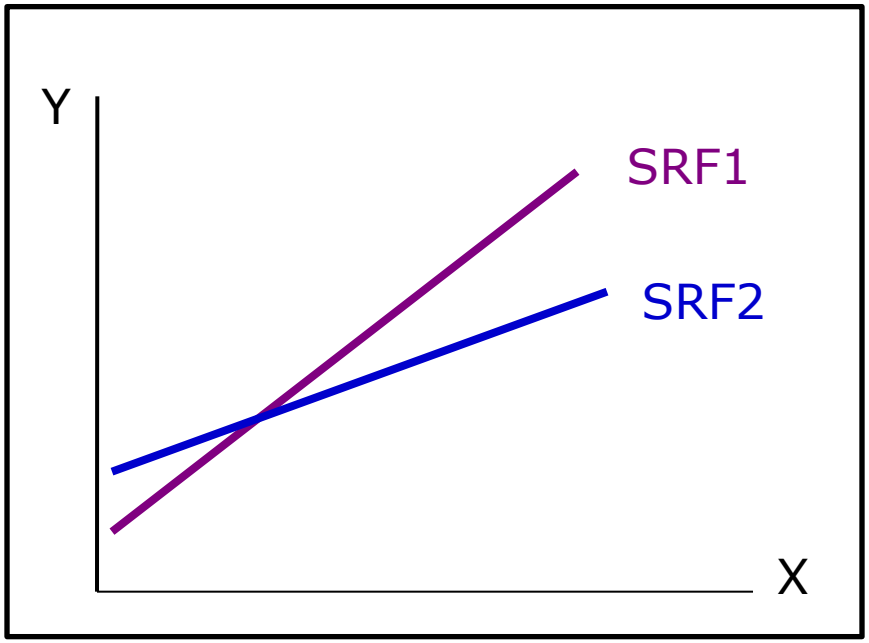
被解释变量**Y**的实际观测值 Y_i 不完全等于样本条件均值 $\hat{Y}_i$，二者之差用 e_i 表示，e_i 称为剩余项或残差项：

则 $e_i = Y_i - \hat{Y}_i$ 或 $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$

样本回归函数的特点

●样本回归线随抽样波动而变化：

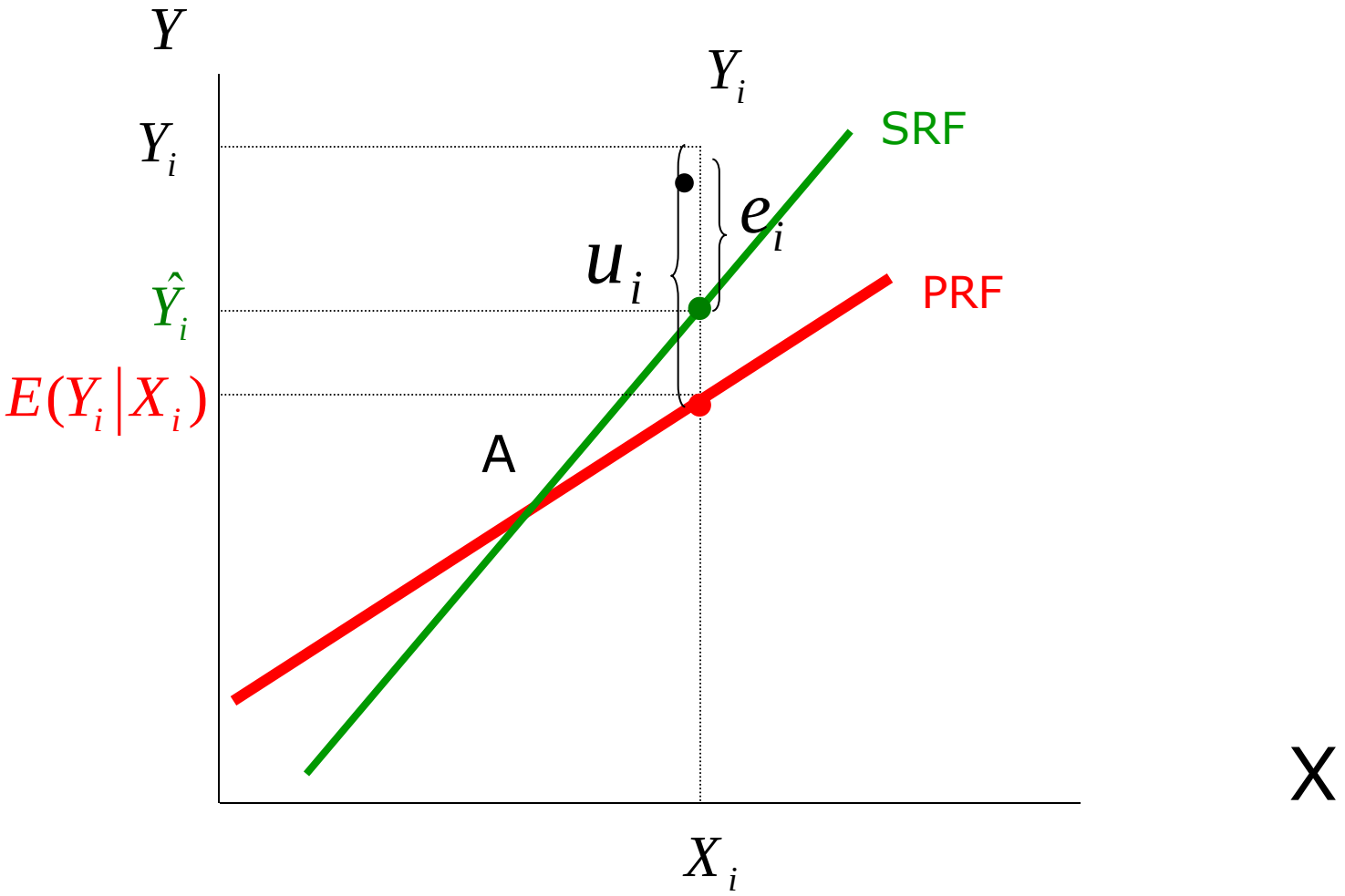
每次抽样都能获得一个样本，就可以拟合一条样本回归线，（SRF不唯一）



●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。

●样本回归线只是样本条件均值的轨迹，还不是总体回归线，它至多只是未知的总体回归线的近似表现。

样本回归函数与总体回归函数的关系



对样本回归的理解

对比：

总体回归函数

$$E(Y_i|X_i)=\beta_1+\beta_2X_i$$

$$Y_i=\beta_1+\beta_2X_i+u_i$$

样本回归函数

$$\hat{Y}_i=\hat{\beta}_1+\hat{\beta}_2X_i$$

$$Y_i=\hat{\beta}_1+\hat{\beta}_2X_i+e_i$$

如果能够通过某种方式获得 $\hat{\beta}_1$ 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 $E(Y_i|X_i)$ 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。

ECONOMETRICS

回归分析的目的

目的：

计量经济分析的目标是寻求总体回归函数。即用样本回归函数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。

1

要解决的问题：

寻求一种规则和方法，使其得到的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 的真实值。这样的“规则和方法”有多种，如矩估计、极大似然估计、最小二乘估计等。其中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，总要使用特定的方法，而任何估计参数的方法都需要有一定的前提条件——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？

- 只有具备一定的假定条件，所作出的估计才具有良好的统计性质。
- 模型中有随机扰动项，估计的参数是随机变量，显然参数估计值的分布与扰动项的分布有关，只有对随机扰动的分布作出假定，才能比较方便地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等统计推断。

假定分为：◆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

例如对于 $Y_i = \beta_1 + \beta_2 X_i + u_i$

- 假定模型设定是正确的（变量和模型无设定误差）
- 假定解释变量X在重复抽样中取固定值。
- 假定解释变量X是非随机的，或者虽然X是随机的，但与扰动项u是不相关的。(从变量X角度看是外生的)

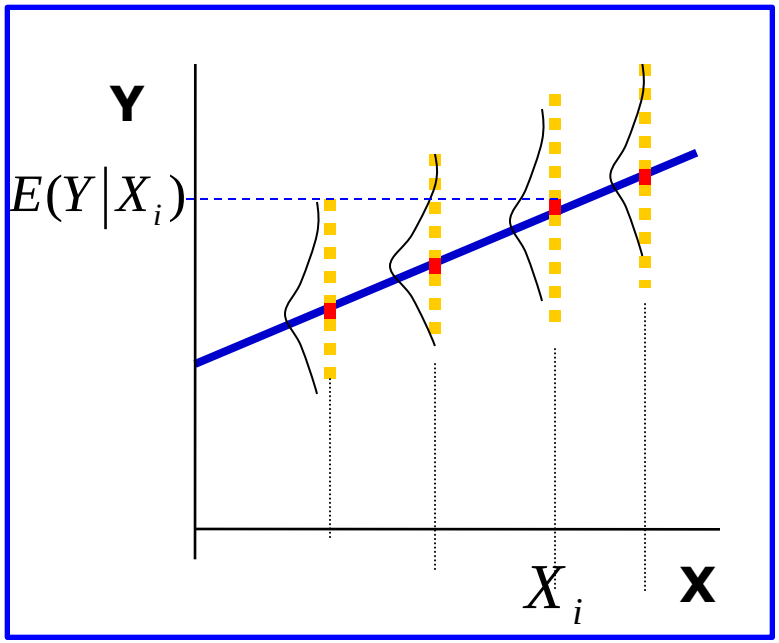
注意: 解释变量非随机在自然科学的实验研究中相对容易满足，经济领域中变量的观测是被动不可控的，X非随机的假定并不一定都满足。

2.对随机扰动项u的假定

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 σ^2

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

假定3：无自相关假定:

随机扰动项 u_i的逐次值互不相关

$$Cov(u_i, u_j) = E[u_i - E(u_i)][u_j - E(u_j)]$$
$$= E(u_i u_j) = 0 \qquad (i \neq j)$$

假定4：解释变量 X_i 是非随机的，或者虽然 X_i是随机的但与扰动项 u_i不相关(从随机扰动 u_i 角度看)

$$Cov(u_i, X_i) = E[u_i - E(u_i)][X_i - E(X_i)] = 0$$

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σ^2的正态分布

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

（说明：正态性假定并不影响对参数的点估计，所以有时不列入基本假定，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。所以正态性假定有合理性）

在对u_i的基本假定下Y的分布性质

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

其中的 β_1, β_2和 X_i是非随机的， u_i是随机变量，因此Y是随机变量，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 $$Var(Y_i|X_i) = \sigma^2$$

假定3：无自相关假定 $$Cov(Y_i, Y_j) = 0$$

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

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

1. OLS的基本思想

●对于 $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$，不同的估计方法可以得到不同的样本回归参数$\hat{\beta}_1$和 $\hat{\beta}_2$，所估计的 $\hat{Y}_i$ 也就不同。

●理想的估计结果应使估计的$\hat{Y}_i$与真实的Y_i的差(即剩余e_i)总的来说越小越好

●因 e_i可正可负，总有 $\sum e_i = 0$，所以可以取 $\sum e_i^2$ 最小，即

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

在观测值**Y**和**X**确定时，$\sum e_i^2$的大小决定于 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 。

要解决的问题**:**： 如何寻求能使 $\sum e_i^2$ 最小的 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 。

2. 正规方程和估计量

取偏导数并令其为**0**，可得正规方程

$$\frac{\partial(\sum e_i^2)}{\partial\hat{\beta}_1} = -2\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

$$\frac{\partial(\sum e_i^2)}{\partial\hat{\beta}_2} = -2\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)X_i = 0$$

即

$$\sum e_i = 0$$

$$\sum e_i X_i = 0$$

或整理得

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

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

用克莱姆法则求解得以观测值表现的**OLS**估计量：

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

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

用离差表现的OLS估计量

为表达得更简洁，或者用离差形式的OLS估计量：

容易证明

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

由正规方程： $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$

注意：其中： $x_i = X_i - \bar{X}$ $y_i = Y_i - \bar{Y}$

本课程中:大写的 X_i 和 Y_i 均表示观测值；

小写的 x_i 和 y_i 均表示观测值的离差

而且由 $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ $\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$

样本回归函数可用离差形式写为 $\hat{y}_i = \hat{\beta}_2 x_i$

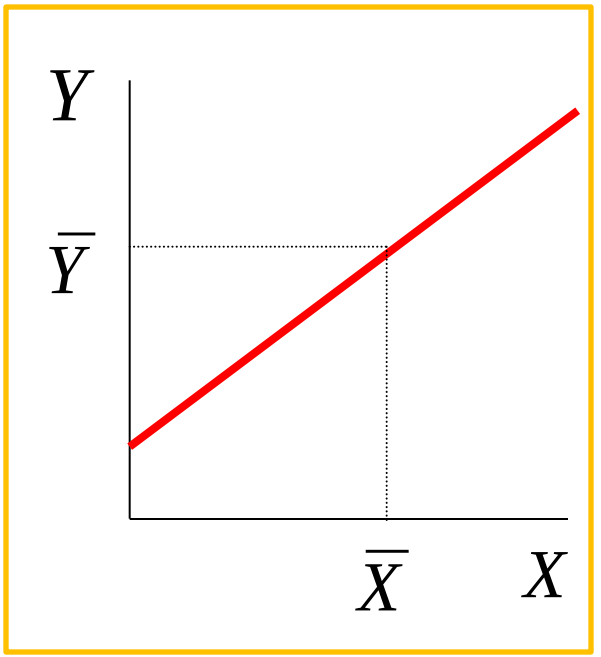
3. OLS回归线的数学性质

- 剩余项 e_i 的均值为零 $\bar{e} = \dfrac{\sum e_i}{n} = 0$

- **OLS回归线通过样本均值**

  (由OLS第一个正规方程直接得到)

- 估计值 $\hat{Y}_i$ 的均值等于实际观测值 Y_i 的均值 $\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$



(由OLS正规方程 $\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$ 两边同除n得到)

$$\frac{\sum \hat{Y}_i}{n} = \frac{1}{n}\sum(\hat{\beta}_1 + \hat{\beta}_2 X_i) = \hat{\beta}_1 + \hat{\beta}_2 \bar{X} = \bar{Y}$$

● 被解释变量估计值 $\hat{Y}_i$ 与剩余项 e_i 不相关

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

由OLS正规方程有： $\sum e_i = 0$ $\sum e_i X_i = 0$

$$\mathrm{Cov}(\hat{Y}_i, e_i) = \frac{1}{n}\sum(\hat{Y}_i - \bar{Y})(e_i - \bar{e}) = 0$$ 因为

$$\sum(\hat{Y}_i - \bar{Y})(e_i - \bar{e}) = \sum \hat{Y}_i e_i - \bar{Y}\sum e_i = \sum e_i(\hat{\beta}_1 + \hat{\beta}_2 X_i) = \hat{\beta}_1 \sum e_i + \hat{\beta}_2 \sum e_i X_i = 0$$

● 解释变量 X_i 与剩余项 e_i 不相关

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

$$\mathrm{Cov}(X_i, e_i) = \frac{1}{n}\sum(X_i - \bar{X})(e_i - \bar{e}) = \sum e_i X_i - \bar{X}\sum e_i = 0$$

4. OLS估计量的统计性质

面临的问题： 参数估计值 ≠ 参数真实值

对参数估计式的优劣需要有评价的标准 为什么呢?

- 参数无法直接观测，只能通过样本去估计。样本的获得存在抽样波动，不同样本的估计结果不一致。
- 估计参数的方法有多种，不同方法的估计结果可能不相同，通过样本估计参数时，估计方法及所确定的估计量不一定完备，不一定能得到理想的总体参数估计值。

对各种估计方法优劣的比较与选择需要有评价标准。

估计准则的基本要求:

参数估计值应"尽可能地接近"总体参数真实值”。

什么是“尽可能地接近” 原则呢?

用统计语言表述就是: 无偏性、有效性、一致性等

(1) 无偏性

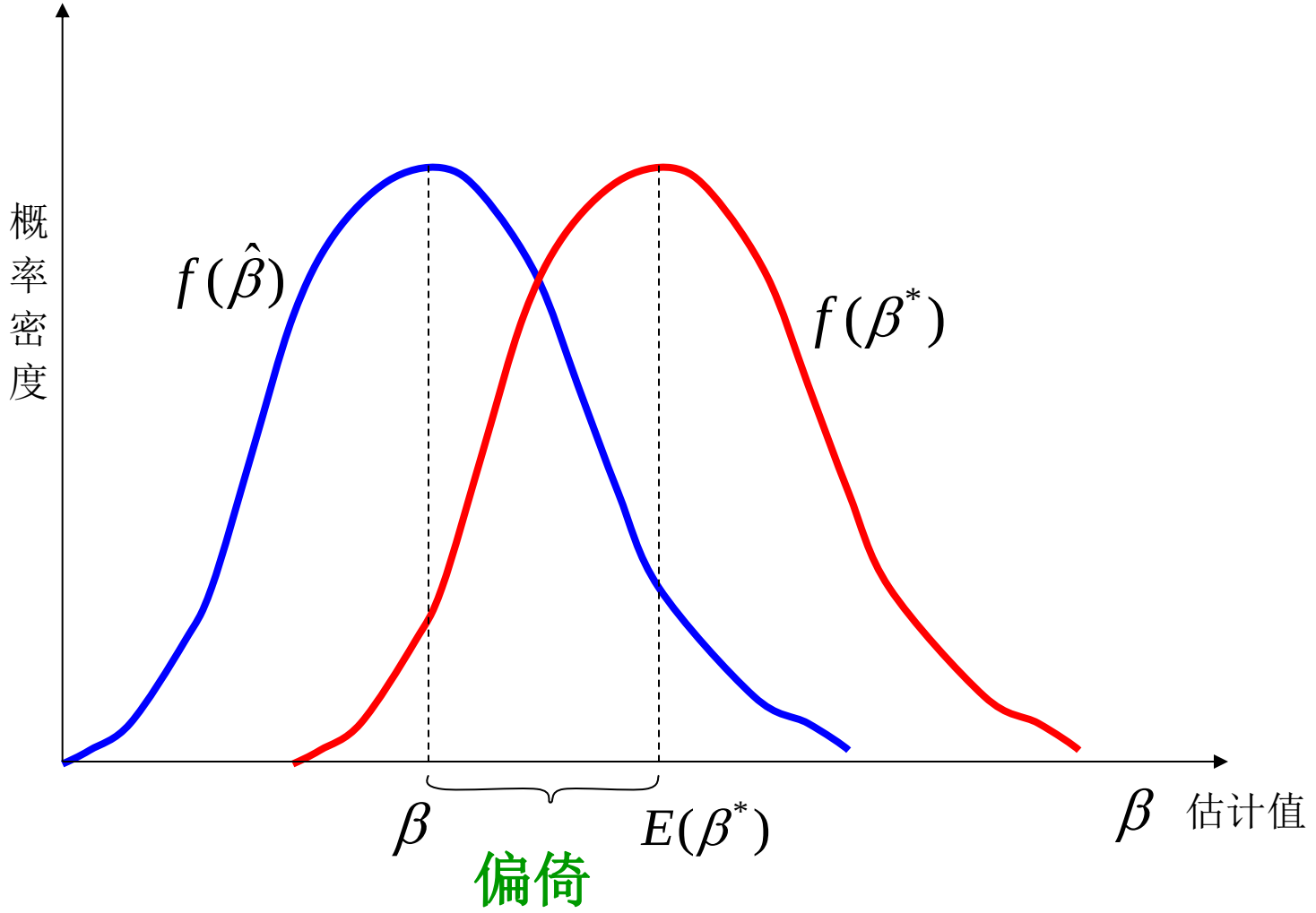
前提： 重复抽样中估计方法固定、样本数不变、

由重复抽样得到的观测值, 可得一系列参数估计值 $\hat{\beta}$，$\hat{\beta}$ 的分布称为 $\hat{\beta}$ 的抽样分布，其密度函数记为 $f(\hat{\beta})$

概念：

如果 $E(\hat{\beta})=\beta$ ，则称 $\hat{\beta}$ 是参数 β 的无偏估计量，

如果 $E(\hat{\beta})\neq\beta$ ，则称 $\hat{\beta}$ 是有偏的估计，其偏倚为 $E(\hat{\beta})-\beta$ （见下页图）


概率密度
$f(\hat{\beta})$
$f(\beta^*)$
β
$E(\beta^*)$
β 估计值
偏倚

ECONOMETRICS

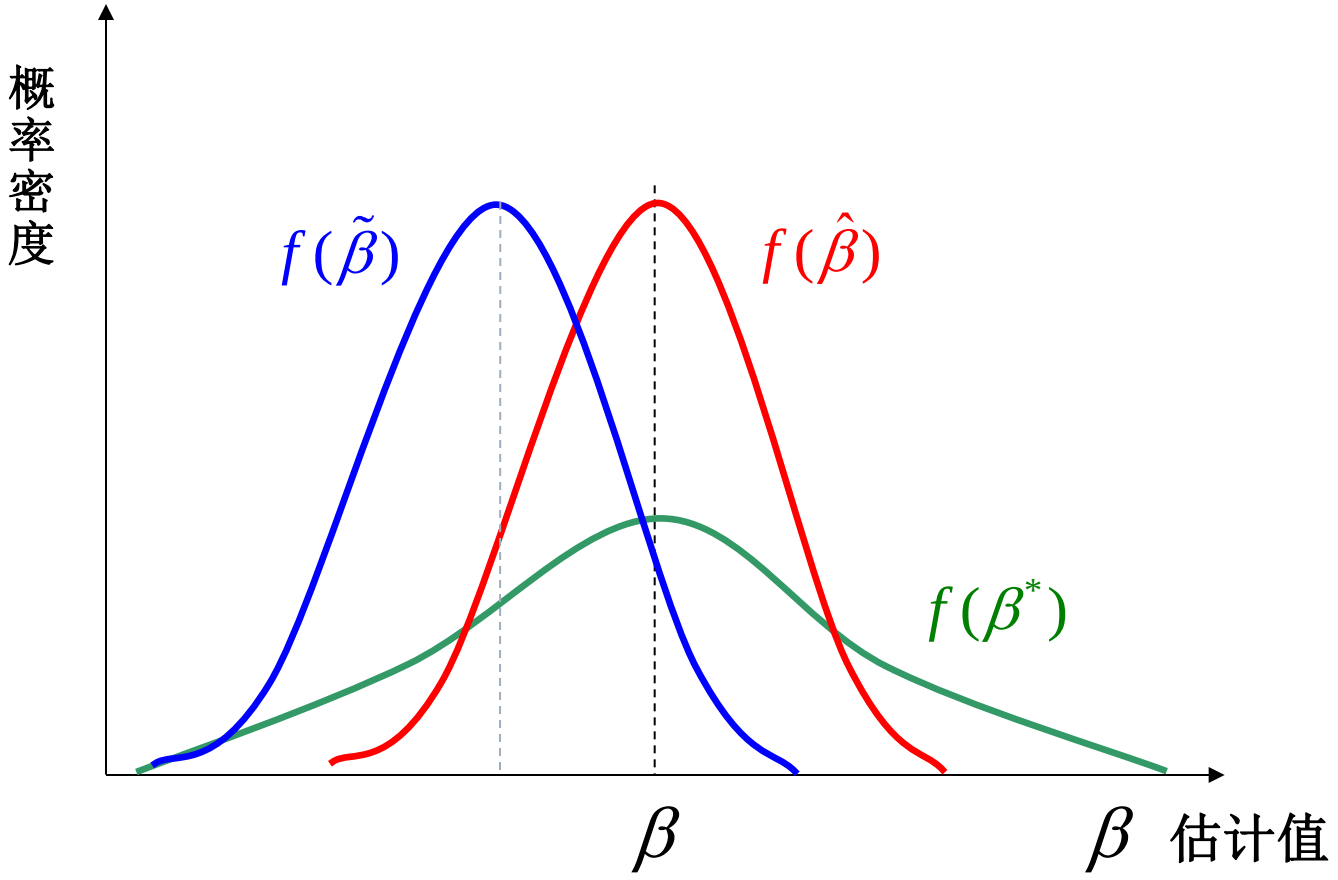
(2)有效性

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无偏估计式

目标：努力寻求其抽样分布具有最小方差的估计量

（见下页图）

既是无偏的同时又具有最小方差特性的估计量，称为最佳（有效）估计量。


概率密度
$f(\tilde{\beta})$
$f(\hat{\beta})$
$f(\beta^*)$
β
β 估计值

3、渐近性质（大样本性质）

思想:当样本容量较小时，有时很难找到方差最小的无偏估计，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）

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

或

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

$$P\lim_{n\to\infty}(\hat{\beta})=\beta$$

（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）(见下页图)

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。

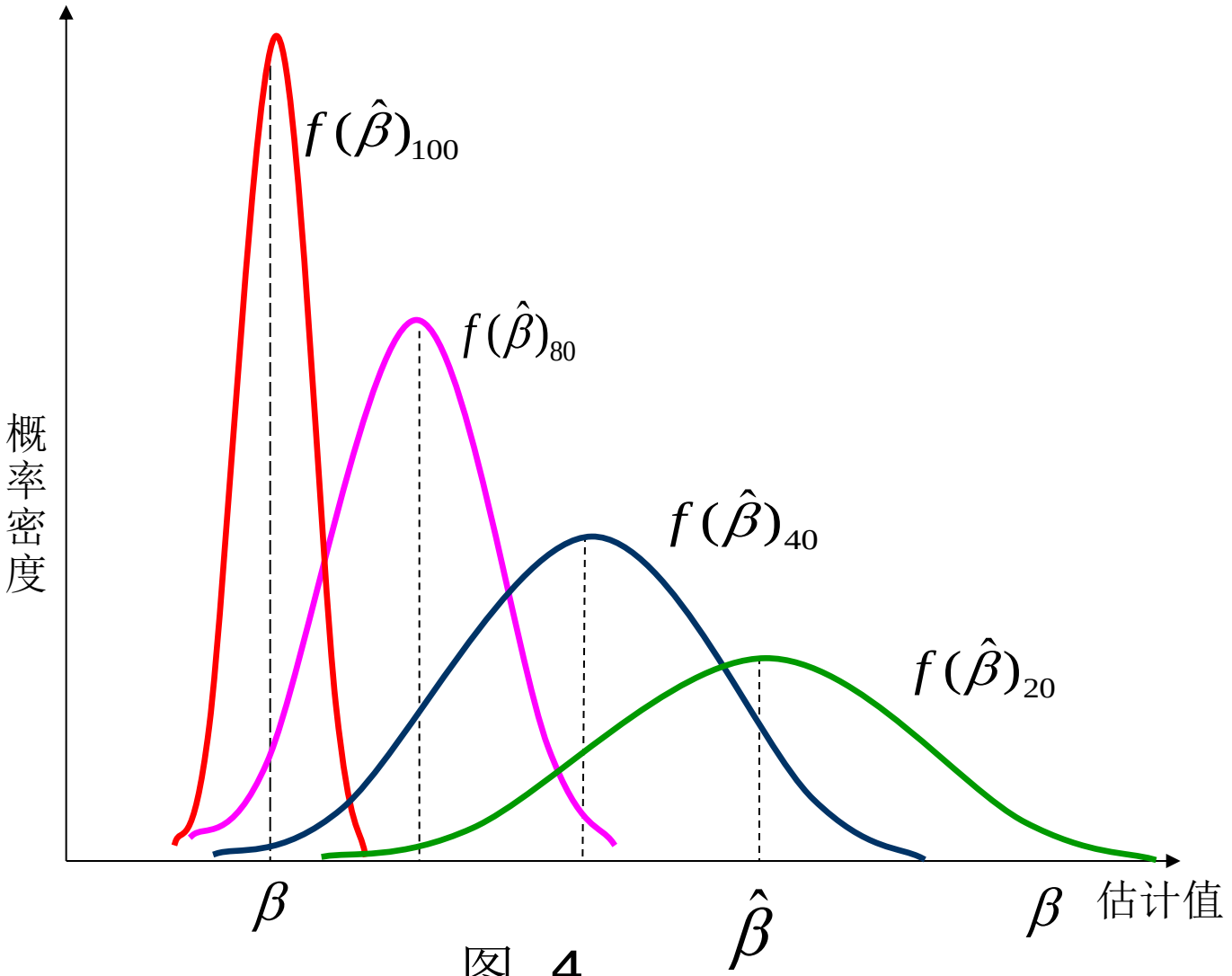

$f(\hat{\beta})_{100}$
$f(\hat{\beta})_{80}$
$f(\hat{\beta})_{40}$
$f(\hat{\beta})_{20}$
概率密度
β
$\hat{\beta}$
β
估计值

图 4

4. 分析OLS估计量的统计性质

OLS估计是否符合“尽可能地接近总体参数真实值”的要求呢?

先明确几点:

- 由OLS估计式可以看出

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

$\hat{\beta}_k$ 都由可观测的样本值 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估计式是点估计量

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

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

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

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = \bar{Y} - \bar{X}\sum k_i Y_i = \sum (\frac{1}{n} - \bar{X}k_i) Y_i$$

2、 无偏特性

可以证明 $E(\hat{\beta}_k) = \beta_k$ （证明见教材**P37**）

（注意：$\hat{\beta}_k$ 无偏性的证明中用到了基本假定中 u_i 零均值等假定）

3、 有效性 （证明见教材附录2·1）

可以证明：在所有的线性无偏估计中，**OLS**估计 $\hat{\beta}_k$ 具有最小方差

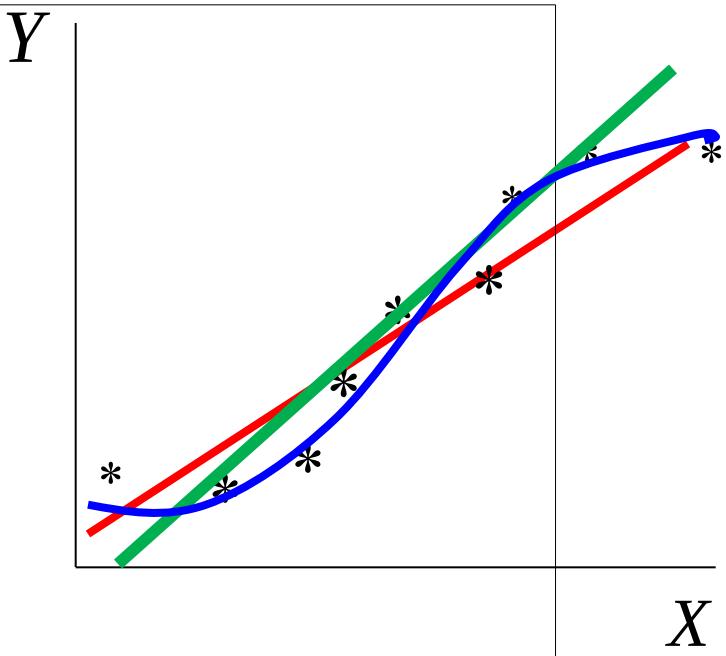
（注意:最小方差性的证明中用到了基本假定中的同方差、无自相关等假定）

结论（高斯定理）：

在古典假定条件下，OLS估计量是最佳线性无偏估计量（BLUE）

第三节　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。

●不同的模型（不同函数形式)可拟合出不同的样本回归线

●相同的模型用不同方法去估计参数，也可以拟合出不同的回归线

拟合的回归线与样本观测值总是有偏离。样本回归线对样本观测数据拟合的优劣程度，可称为拟合优度。

如何度量拟合优度呢？

拟合优度的度量建立在对 **Y** 的总变差分解的基础上

ECONOMETRICS

一、总变差的分解

分析Y的观测值 Y_i、估计值$\hat{Y}_i$ 与平均值$\bar{Y}$ 有以下关系

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

将上式两边平方加总，可证得（提示：交叉项 $\sum(\hat{Y}_i-\bar{Y})e_i=0$ ）

$$\underset{\text{（TSS）}}{\sum(Y_i-\bar{Y})^2}=\underset{\text{（ESS）}}{\sum(\hat{Y}_i-\bar{Y})^2}+\underset{\text{（RSS）}}{\sum(Y_i-\hat{Y}_i)^2}$$

或者表示为

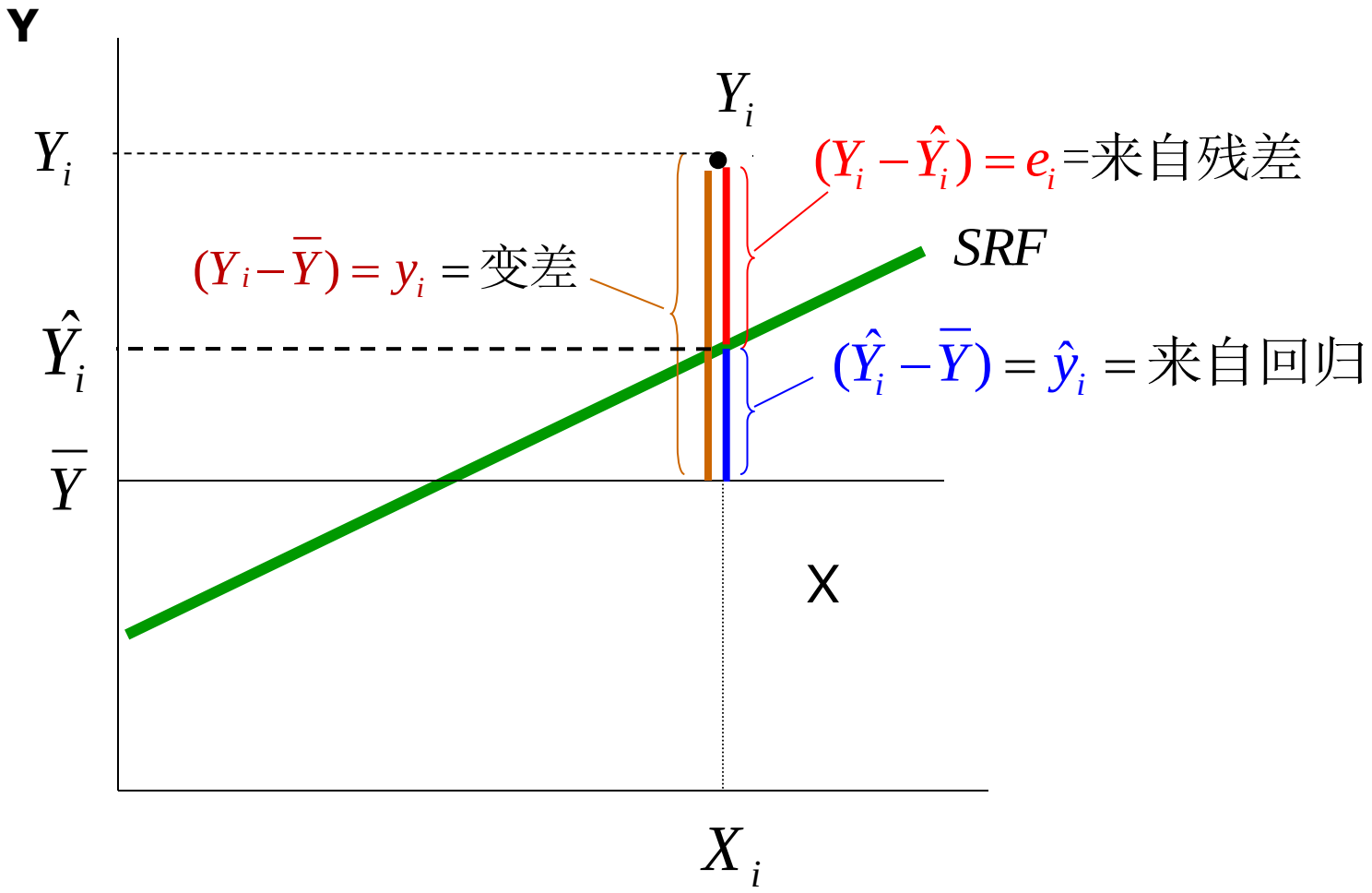
$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2$$

总变差 $\sum y_i^2$ （**TSS**）：被解释变量**Y**的观测值与其平均值的离差平方和（总平方和）(说明 Y 的总变动程度）

解释了的变差 $\sum \hat{y}_i^2$ （**ESS**）：被解释变量**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**）：被解释变量观测值与估计值之差的平方和（未解释的平方和）

变差分解的图示(以某一个观测值为例)



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

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

二、可决系数

以TSS同除总变差等式 $\sum(Y_i-\bar{Y})^2=\sum(\hat{Y}_i-\bar{Y})^2+\sum(Y_i-\hat{Y}_i)^2$ 两边：

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

或

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

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 或 R^2 表示：

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

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： $0 \leq R^2 \leq 1$
- 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

$$R^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{\sum (\hat{\beta}_2 x_i)^2}{\sum y_i^2}$$

$$\hat{y}_i = \hat{\beta}_2 x_i$$

$$= \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum y_i^2} = \frac{(\sum x_i y_i)^2}{(\sum x_i^2)^2} \cdot \frac{\sum x_i^2}{\sum y_i^2}$$

$$= \frac{(\sum x_i y_i)^2}{(\sum x_i^2)(\sum y_i^2)} = \left\{ \frac{\sum x_i y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}} \right\}^2$$

$$= r^2$$

区别：

可决系数	相关系数
是就模型而言	是就两个变量而言
说明解释变量对被解释变量的解释程度	说明两变量线性依存程度
度量不对称的因果关系	度量对称的相关关系
取值 $0 \leqq R^2 \leqq 1$ 有非负性	取值 $-1 \leqq r \leqq 1$ 可正可负

第四节 回归系数的区间估计和假设检验

为什么要作区间估计？ 运用**OLS**法可以估计出参数的一个估计值，但**OLS**估计只是通过样本得到的点估计，它不一定等于真实参数，还需要寻求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呢？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 $\hat{\beta}_k$ 概率分布性质的基础上。

ECONOMETRICS

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

怎样确定 $\hat{\beta}_k$ 的分布性质呢?

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；

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

线性特征 $\hat{\beta}_2 = \sum k_i y_i$

$\hat{\beta}_k$ 是 Y_i 的线性函数，决定了 $\hat{\beta}_k$ 也服从正态分布

u_i 正态 → Y_i 正态 → $\hat{\beta}_k$ 正态 （线性估计的重要性）

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

$\hat{\beta}_k$ 的期望和方差

- $\hat{\beta}_k$ 的期望：$E(\hat{\beta}_k)=\beta_k$ （已证明是无偏估计）
- $\hat{\beta}_k$ 的方差和标准误差 （证明见**P38**）

（标准误差是方差的平方根） $\sigma^2=Var(u_i)$

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

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

注意：以上各式中 σ^2 均未知，但是个常数，其余均是已知的样本观测值，这时 $Var(\hat{\beta}_k)$ 和 $SE(\hat{\beta}_k)$ 都不是随机变量。

对随机扰动项方差σ^2的估计

基本思想：

σ^2 是 u_i 的方差，而 u_i 不能直接观测，只能从由样本得到的 e_i 去获得有关 u_i 的某些信息，去对 σ^2 作出估计。

可以证明（见附录2.2)其无偏估计为 $E(\sum e_i^2) = (n-2)\sigma^2$

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} \qquad E(\hat{\sigma}^2) = \sigma^2$$

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

(这里的n-2为自由度, 即可自由变化的样本观测值个数)

注意区别： σ^2 是未知的确定的常数；

$\hat{\sigma}^2$是由样本信息估计的，是个随机变量

ECONOMETRICS

对 $\hat{\beta}_k$ 作标准化变换

为什么要对 $\hat{\beta}_k$ 作标准化变换?

在 u_i 正态性假定下，由前面的分析已知

$$\hat{\beta}_k \sim N[\beta_k, Var(\hat{\beta}_k)]$$

但在对一般正态变量 $\hat{\beta}_k$ 作实际分析时，要具体确定 $\hat{\beta}_k$ 的取值及对应的概率，要通过正态分布密度函数或分布函数去计算是很麻烦的，为了便于直接利用“标准化正态分布的临界值”，需要对 $\hat{\beta}_k$ 作标准化变换。

标准化的方式：

$$z_k = \frac{\hat{\beta}_k - E(\beta_k)}{SE(\hat{\beta}_k)}$$

标准正态分布函数

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{x^2}{2}} dx$$

1. σ^2已知时，对$\hat{\beta}_k$作标准化变换

●**在σ^2已知时**对$\hat{\beta}_k$作标准化变换，所得**Z**统计量为标准正态变量。

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

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

注意:这时 $SE(\hat{\beta}_1)$ 和 $SE(\hat{\beta}_2)$ 都不是随机变量(**X**、σ、n 都是非随机的）

2. σ^2未知时，对$\hat{\beta}_k$作标准化变换

条件： 当σ^2未知时，可用 $\hat{\sigma}^2$（随机变量）代替 σ^2去估计参数的标准误差。这时参数估计的标准误差是个随机变量。

- 样本为大样本时,作标准化变换所得的统计量Z_k, 也可以视为标准正态变量（根据中心极限定理）。
- 样本为小样本时，

用估计的参数标准误差对 $\hat{\beta}_k$ 作标准化变换，所得的统计量用**t**表示，这时**t**将不再服从正态分布，而是服从 **t** 分布（注意这时分母是随机变量）：

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

二、回归系数的区间估计

基本思想：

对参数作出的点估计是随机变量，虽然是无偏估计，但还不能说明这种估计的可靠性和精确性。如果能找到包含真实参数的一个范围，并确定这样的范围包含参数真实值的可靠程度，将是对真实参数更深刻的认识。

方法： 如果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 $\alpha(0 \le \alpha \le 1)$，能使得这样的区间 $(\hat{\beta}_k - \delta, \hat{\beta}_k + \delta)$ 包含真实 β_k 的概率为 $1-\alpha$，即

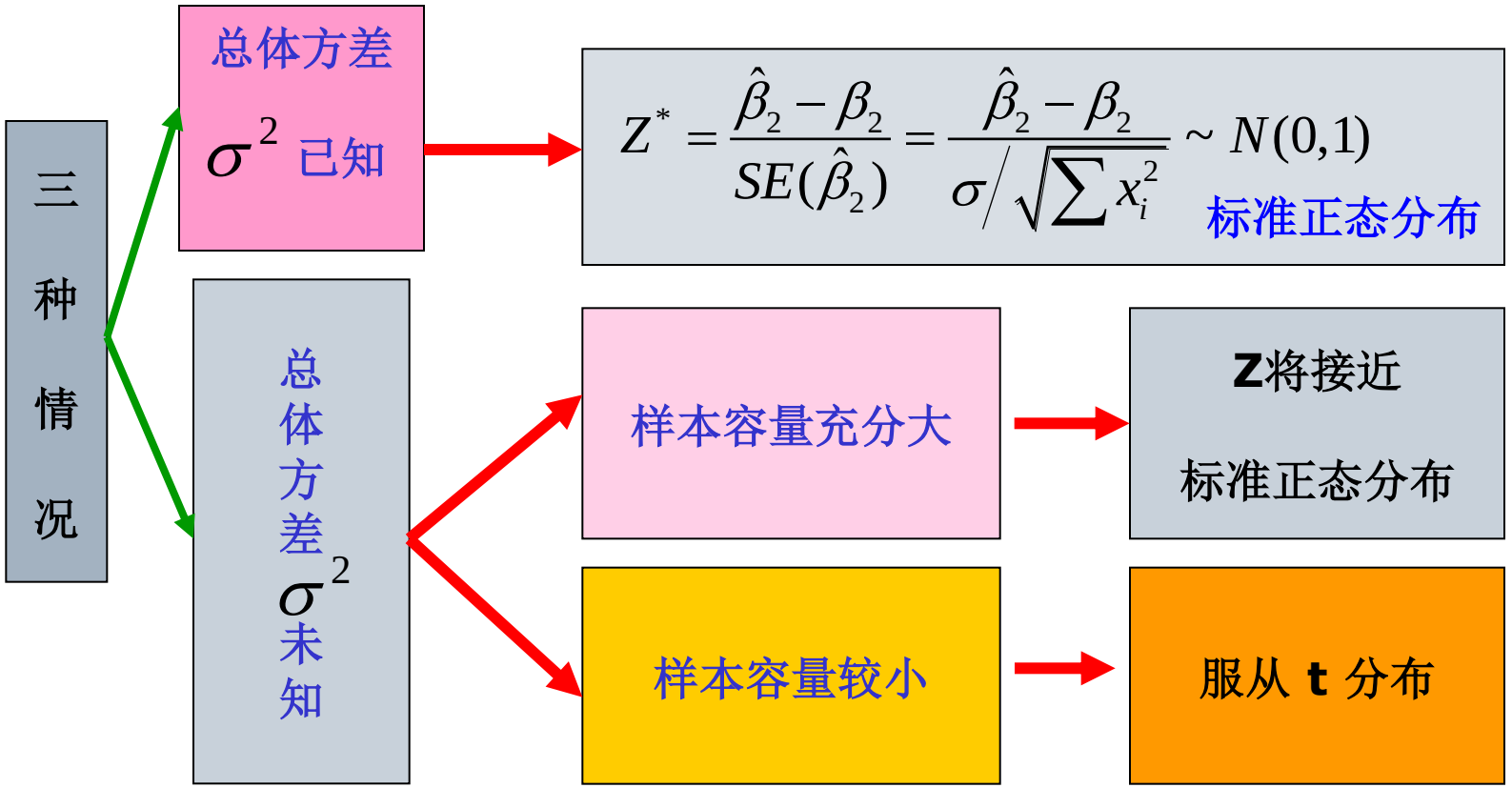
$$P(\hat{\beta}_k - \delta \le \beta_k \le \hat{\beta}_k + \delta) = 1 - \alpha$$

这样的区间称为所估计参数的置信区间。

讨论：“如果已经得出了 $\hat{\beta}_k$ 的特定估计值,并确定了某个置信区间，这说明真实参数落入这个区间的概率为1-α”。这种说法对吗？

置信区间： $P(\hat{\beta}_k - \delta \le \beta_k \le \hat{\beta}_k + \delta) = 1 - \alpha$

基本思想：利用 $\hat{\beta}_k$ 标准化后统计量的分布性质去寻求 δ ：

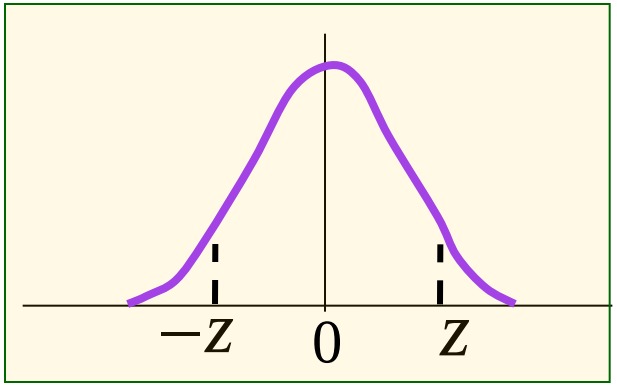


回归系数的区间估计 (分三种情况寻找合适的 δ)

（1） 当总体方差 σ^2 已知时(Z 服从正态分布)

取定 α（例如 $\alpha=0.05$），查标准正态分布表得与 对 α 应的临界值z (例如z为1.96)，则标准化变量Z*（统计量）

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



因为

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

或

即 $$P[\hat{\beta}_2 - zSE(\hat{\beta}_2) \le \beta_2 \le \hat{\beta}_2 + zSE(\hat{\beta}_2)] = 1-\alpha$$

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

2. 当总体方差 σ^2 未知，且样本容量充分大时

方法： 可用无偏估计 $\hat{\sigma}^2$ 去代替未知的 σ^2，

由于样本容量充分大，标准化变量**Z***（统计量）将接近标准正态分布

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

注意:这里的“ ^ ”，表示“估计的”，

这时区间估计的方式也可利用标准正态分布

只是这时

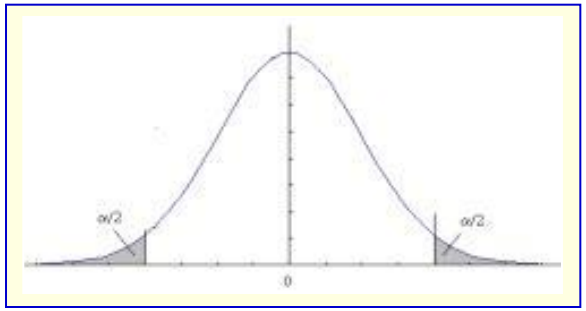
$$\delta = z\hat{SE}(\hat{\beta}_2) = z\frac{\hat{\sigma}}{\sqrt{\sum x_i^2}}$$

3、当总体方差 σ^2 未知，且样本容量较小时

方法：用无偏估计 $\hat{\sigma}^2$去代替未知的 σ^2，由于样本容量较小，“标准化变量” **t** （统计量）不再服从正态分布，而服从 **t** 分布。

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

这时可用 **t** 分布去建立参数估计的置信区间。选定**α**，查 **t** 分布表得显著性水平为 $\alpha/2$，自由度为**n-2**的临界值 $t_{\alpha/2}$ (n-2) ，则有



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

即

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

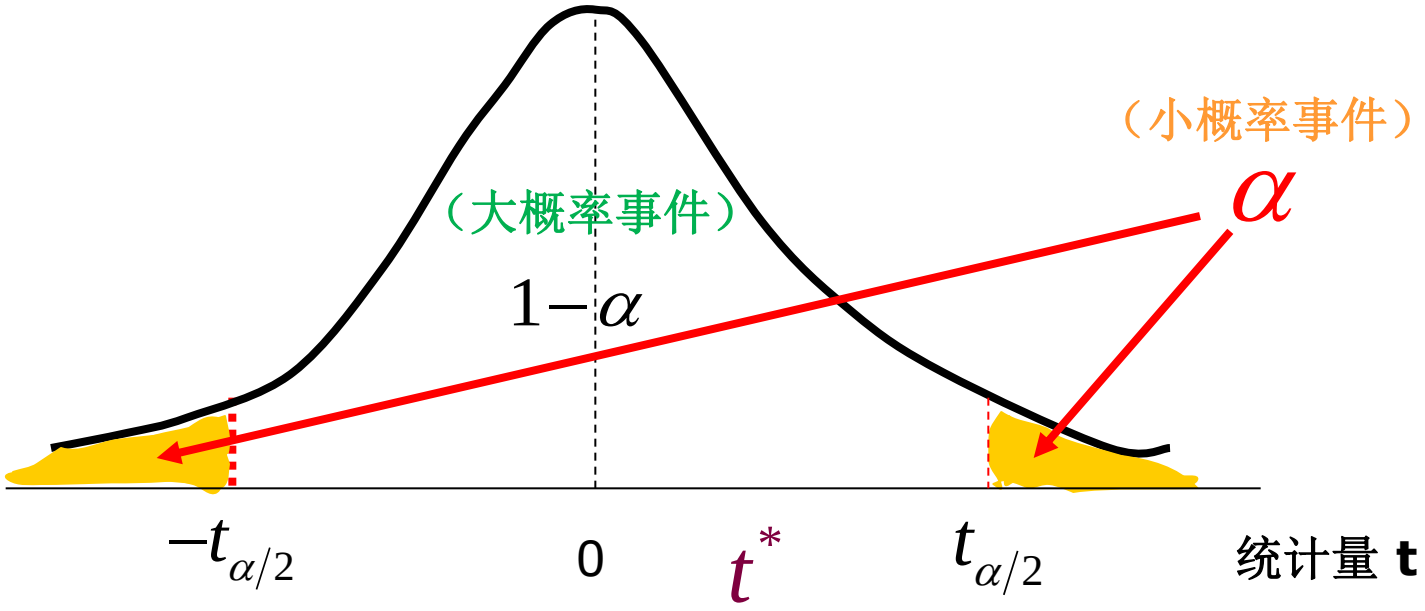
三、回归系数的假设检验

目的： 简单线性回归中，检验**X**对**Y**是否真有显著影响

基本概念回顾： 临界值与概率、大概率事件与小概率事件

相对于显著性水平 α 的临界值为：t_α（单侧）或 $t_{\alpha/2}$（双侧）

计算的统计量为： t^*



回归系数的检验方法

确立假设： 原假设为 $H_0: \beta_2 = 0$

备择假设为 $H_1: \beta_2 \neq 0$

(**本质**：检验 β_2 是否为**0**，即检验 X_i 是否对**Y**有显著影响)

(1)当已知 σ^2 或样本容量足够大时

可利用正态分布作**Z**检验

$$Z^* = \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} \sim N(0.1)$$

给定 α，查正态分布表得临界值 **Z**

- 如果 $-z < Z^* < z$ 则不拒绝原假设 H_0
- 如果 $Z^* < -z$ 或 $Z^* > z$ 则 拒绝原假设 H_0

(2) 当σ^2未知，且样本容量较小时

只能用 $\hat{\sigma}^2$ 去代替 σ^2，可利用 t分布作 t 检验：

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

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 $t^* \ge t_{\alpha/2}(n-2)$

则拒绝原假设 $H_0: \beta_2 = 0$ 而不拒绝备择假设 $H_1: \beta_2 \ne 0$

▼如果 $-t_{\alpha/2}(n-2) \le t^* \le t_{\alpha/2}(n-2)$

则不拒绝原假设 $H_0: \beta_2 = 0$

用 P 值判断参数的显著性

假设检验的 p 值：

p 值是基于既定的样本数据所计算的统计量，拒绝原假设的最低显著性水平。

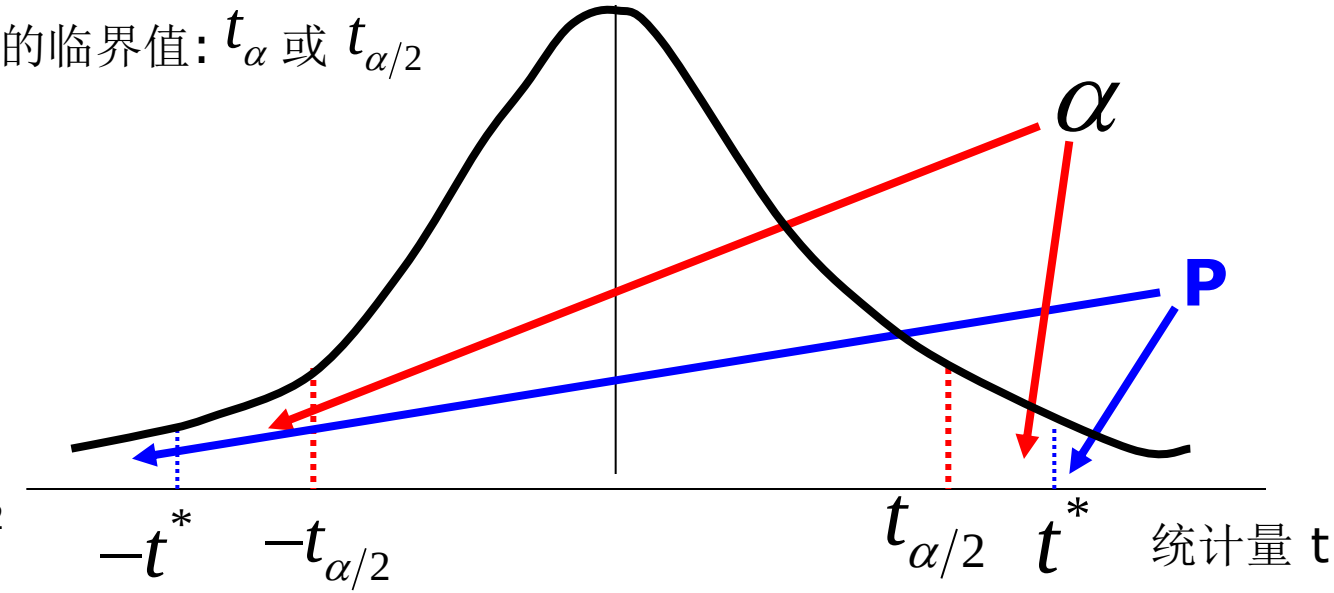
统计分析软件中通常都给出了检验的 p 值

相对于显著性水平 α 的临界值：t_α 或 $t_{\alpha/2}$

计算的统计量：t^*

$t_{\alpha/2}$ 与 α 相对应

t^* 与 P 相对应



注意：

t检验是比较 t^*和 $t_{\alpha/2}$

P值检验是比较α 和 p

用 P 值判断参数显著性的方法

方法： 将给定的显著性水平 α 与 p 值比较：

- 若 $\alpha > p$ 值，必有 $|t^*| > |t_{\alpha/2}|$ ，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，必有 $|t^*| \leq |t_{\alpha/2}|$ ，则在显著性水平 α 下不拒绝原假设 $H_0: \beta_k = 0$，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

ECONOMETRICS

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

$\hat{Y}_i$ = 24.4545 + 0.5091 X_i

（6.4138）　（0.0357）　　标准误差SE

t = (3.8128)　(14.2605)　　t 统计量

R^2 = 0.9621　　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、被解释变量平均值预测

1. 基本思想

经估计的计量经济模型可用于：经济结构分析　　经济预测

政策评价　　验证理论

●运用计量经济模型作预测：指利用所估计的样本回归函数作预测工具，用解释变量的已知值或预测值，对预测期或样本以外的被解释变量的数值作出定量的估计。

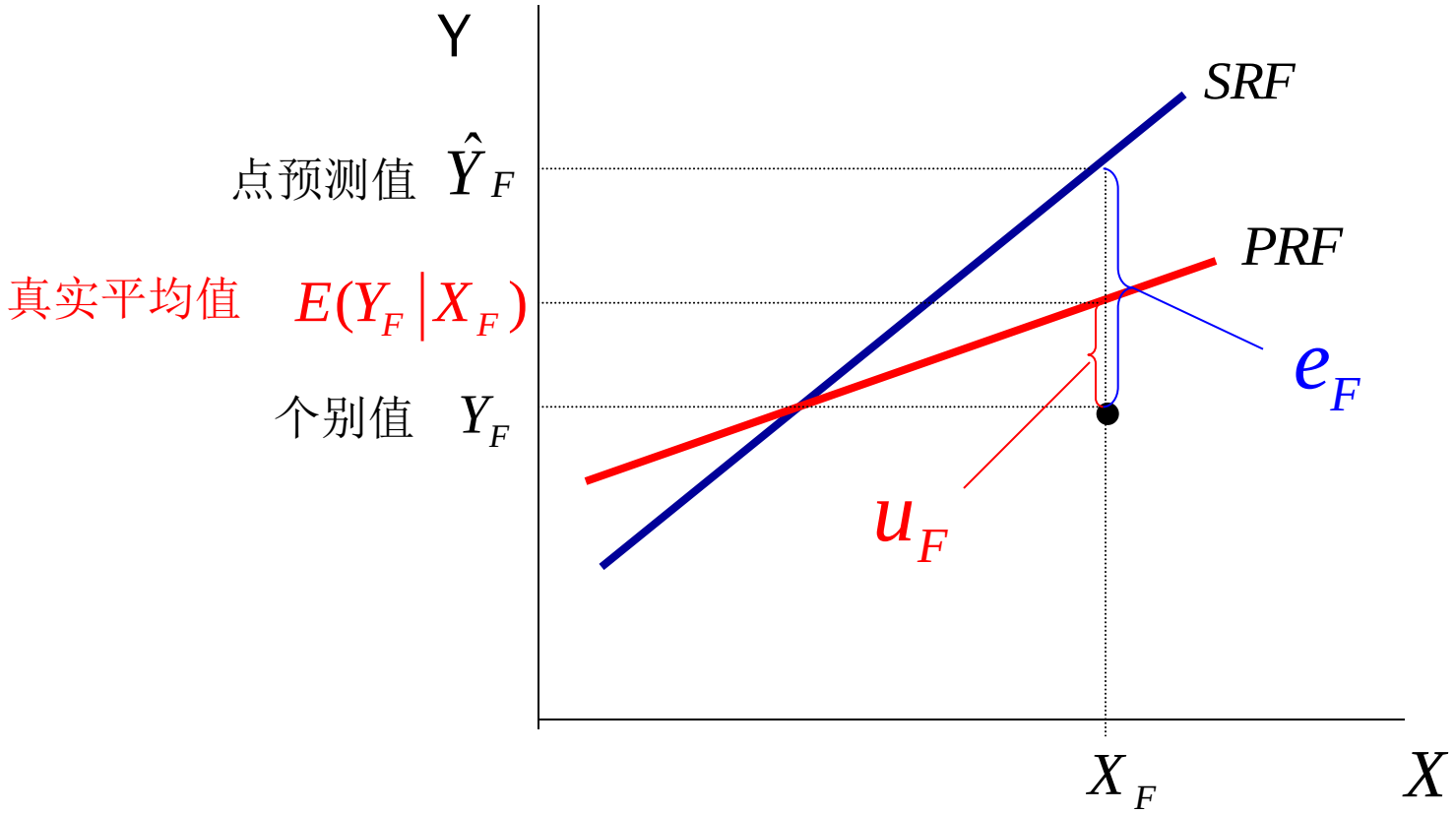
●计量经济预测是一种条件预测：

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

预测值、平均值、个别值的相互关系



$\hat{Y}_F$ 是对真实平均值的点估计，也是对个别值的点估计

ECONOMETRICS

2、Y平均值的点预测

点预测：

用样本估计的总体参数值所计算的Y的估计值直接作为Y的预测值

方法：

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

ECONOMETRICS

3、Y平均值的区间预测

基本思想：

- 预测的目标值是真实平均值，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，必须确定平均值点预测值 $\hat{Y}_F$ 的抽样分布
- 必须找出点预测值$\hat{Y}_F$与预测目标值$E(Y_F|X_F)$的关系，即找出与二者都有关的统计量

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

（较复杂不具体证明）

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

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

$\hat{Y}_F$ 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

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

注意：

构建平均值的预测区间

显然这样的 t 统计量与 $\hat{Y}_F$ 和 $E(Y_F|X_F)$ 都有关。

给定显著性水平**a**，查 t 分布表，得自由度n－2的临界值$t_{\alpha/2}(n-2)$，则有 $P(-t_{\alpha/2} \le t \le t_{\alpha/2}) = 1-\alpha$

即

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

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

Y平均值的置信度为 $1-\alpha$ 的预测区间为

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

ECONOMETRICS

三、被解释变量个别值预测

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，需要寻找与点预测值$\hat{Y}_F$ 和预测目标个别值 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 $\hat{Y}_F$ 及个别值 F 都有关的变量，并且已知 e_F 服从正态分布，且可证明

$$E(e_F) = 0$$

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

（较复杂不具体证明）

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

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

ECONOMETRICS

构建个别值的预测区间

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 $1-\alpha$的预测区间上下限为

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

被解释变量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

预测区间

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

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

预测区间

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

（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，当 $X_F = \overline{X}$ 时，预测区间最小。

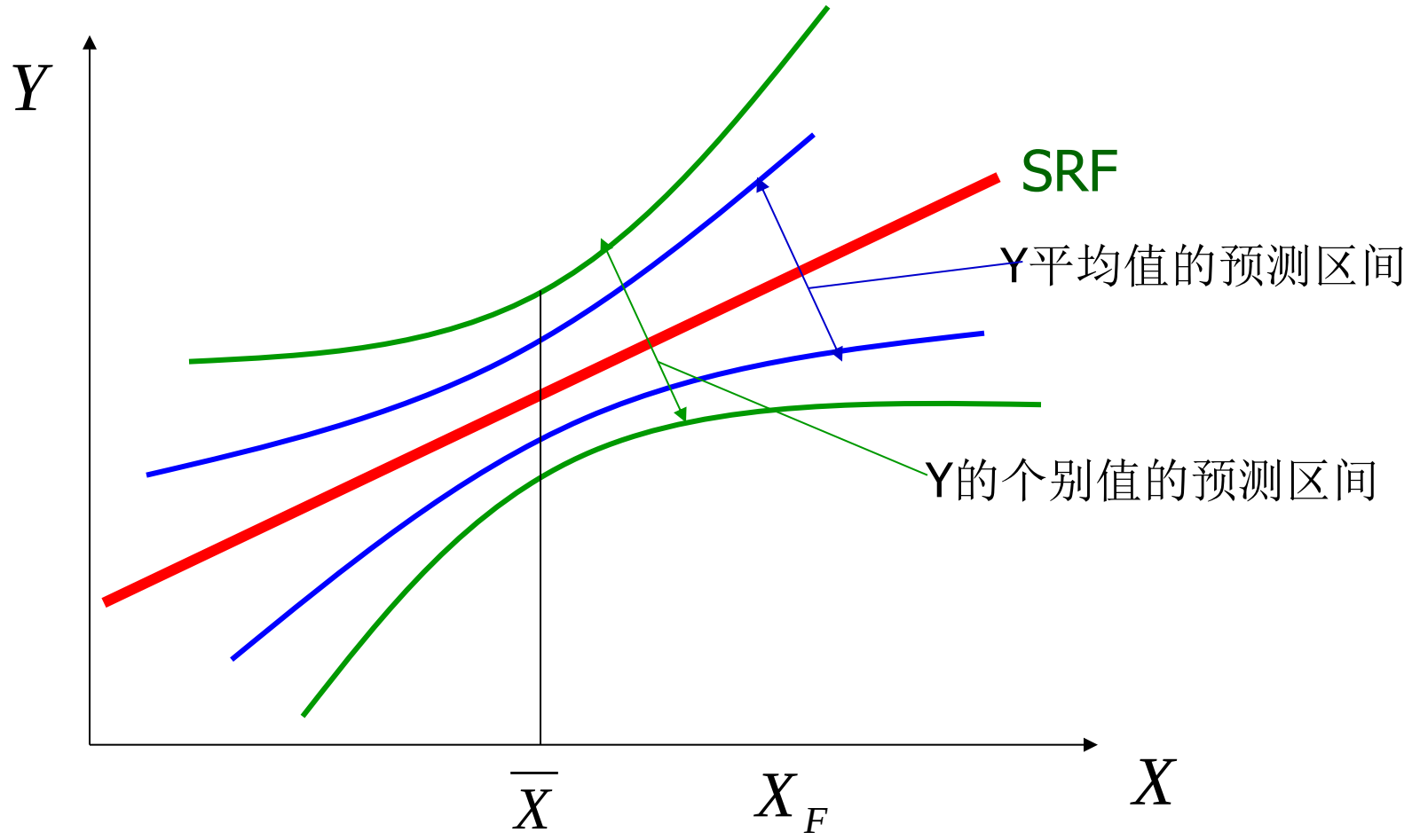
（3）预测区间上下限与样本容量有关，当样本容量n→∞时,个别值的预测区间只决定于随机扰 动的方差。

预测区间

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

ECONOMETRICS

各种预测值的关系



当$X_F = \overline{X}$时，预测区间最小

ECONOMETRICS

第八节　案例分析

案例：分析各地区城镇居民计算机拥有量与城镇居民收入水平的关系

提出问题：随着信息化程度和居民收入水平的提高，作为居民耐用消费品重要代表的计算机已为众多城镇居民家庭所拥有。研究中国各地区城镇居民计算机拥有量与居民收入水平的数量关系，对于探寻居民消费增长的规律性，分析各地区居民消费的差异，预测地区全体居民消费水平和结构的发展趋势，合理规划信息产业的发展，都有重要的意义。

理论分析：影响居民计算机拥有量的因素有多种，但从理论和经验分析，最主要的影响因素应是居民收入水平。从理论上说居民收入水平越高，居民计算机拥有量越多。

变量选择：被解释变量选择能代表城乡所有居民消费的“城镇居民家庭平均每百户计算机拥有量”(单位:台)；解释变量选择表现城镇居民收入水平的“城镇居民平均每人全年家庭总收入”（单位:元）

研究范围：全国各省市**2011**年底的城镇居民家庭平均每百户计算机拥有量和城镇居民平均每人全年家庭总收入数据。

ECONOMETRICS

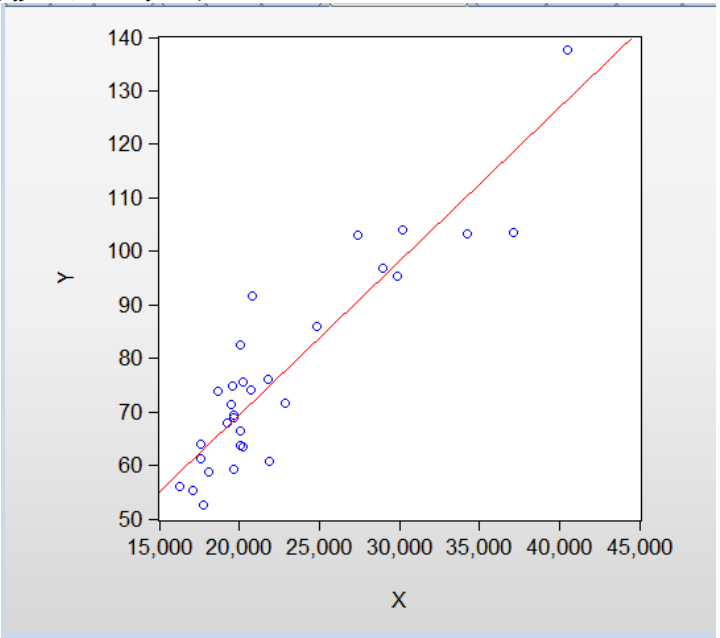
2011年中国各地区城镇居民每百户计算机拥有量和人均总收入

地区	2011年底城镇居民家庭平均每百户计算机拥有量(台)Y	城镇居民平均每人全年家庭总收入（元）X
北 京	103.51	37124.39
天 津	95.4	29916.04
河 北	74.74	19591.91
山 西	69.45	19666.1
内蒙古	60.83	21890.19
辽 宁	71.66	22879.77
吉 林	68.04	19211.71
黑龙江	55.36	17118.49
上 海	137.7	40532.29
江 苏	96.94	28971.98
浙 江	103.17	34264.38
安 徽	74.04	20751.11
福 建	103	27378.11
江 西	73.87	18656.52
山 东	85.88	24889.8

地区	2011年底城镇居民家庭平均每百户计算机拥有量 (台)Y	城镇居民平均每人全年家庭总收入（元）X
河 南	71.41	19526.92
湖 北	75.49	20193.27
湖 南	66.36	20083.87
广 东	104.13	30218.76
广 西	91.72	20846.11
海 南	63.82	20094.18
重 庆	76.07	21794.27
四 川	68.86	19688.09
贵 州	63.89	17598.87
云 南	63.55	20255.13
西 藏	58.83	18115.76
陕 西	82.43	20069.87
甘 肃	56.14	16267.37
青 海	52.65	17794.98
宁 夏	59.39	19654.59
新 疆	61.2	17631.15

模型设定：

为了初步分析城镇居民家庭平均每百户计算机拥有量**(Y)**与城镇居民平均每人全年家庭总收入**(X)**的关系，作以**X**为横坐标，以**Y**为纵坐标的散点图。



从散点图可以看出城镇居民家庭平均每百户计算机拥有量**(Y)**与城镇居民平均每人全年家庭总收入**(X)** 大体呈现线性关系。

可以建立如下简单线性回归模型：

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

ECONOMETRICS

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，估计结果是：

Dependent Variable: Y
Method: Least Squares
Date: 09/08/13 Time: 19:37
Sample: 1 31
Included observations: 31

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	11.95802	5.622841	2.126686	0.0421
X	0.002873	0.000240	11.98264	0.0000

R-squared	0.831966	Mean dependent var	77.08161
Adjusted R-squared	0.826171	S.D. dependent var	19.25503
S.E. of regression	8.027957	Akaike info criterion	7.066078
Sum squared resid	1868.995	Schwarz criterion	7.158593
Log likelihood	-107.5242	Hannan-Quinn criter.	7.096236
F-statistic	143.5836	Durbin-Watson stat	1.656123
Prob(F-statistic)	0.000000		

用规范的形式将参数估计和检验的结果写为：

$$\hat{Y}_t = 11.9580 + 0.002873 X_t$$

（5.6228） (0.00024)

t= (2.1267) (11.9826)

$R^2 = 0.8320$ F=143.5836 n=31

91

模型检验

1. 可决系数： $R^2 = 0.8320$ 模型整体上拟合较好。

2. 系数显著性检验： 取$\alpha = 0.05$，查t分布表得自由度为 $n-2=31-2=29$ 的临界值为 $t_{0.025}(29)=2.045$。

因为 $t(\hat{\beta}_1)=2.1267 > t_{0.025}(29)=2.045$ 应拒绝 $H_0:\beta_1=0$

$t(\hat{\beta}_2)=11.9826 > t_{0.025}(29)=2.045$ 应拒绝 $H_0:\beta_2=0$

3. 用P值检验 $\alpha = 0.05$ >> p=0.0000

表明，城镇居民人均总收入对城镇居民每百户计算机拥有量确有显著影响。

4. 经济意义检验：

所估计的参数 ，，说明城镇居民家庭人均总收入每增加1元，平均说来城镇居民每百户计算机拥有量将增加0.002873台，这与预期的经济意义相符。

ECONOMETRICS

经济预测

点预测：

如果西部地区某省城镇居民家庭人均总收入能达到25000元/人，利用所估计的模型可预测城镇居民每百户计算机拥有量，点预测值为

$$\hat{Y}_f = 11.9580 + 0.002873 \times 25000 = 83.7846 \quad (台)$$

区间预测： 平均值区间预测上下限：

$$Y_f = \hat{Y}_f \mp t_{\alpha/2}\,\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_f - \overline{X})^2}{\sum x_i^2}}$$

已知：

$Y_f = 83.7846$　　$t_{0.025}(29) = 2.045$　　$\hat{\sigma} = 8.027957$　　$n = 31$

平均值区间预测区间预测

由X和Y的描述统计结果

	X	Y
Mean	22666.97	77.08161
Median	20094.18	71.66000
Maximum	40532.29	137.7000
Minimum	16267.37	52.65000
Std. Dev.	6112.965	19.25503
Skewness	1.515854	1.185095
Kurtosis	4.384257	4.259649
Jarque-Bera	14.34708	9.305832
Probability	0.000767	0.009534
Sum	702676.0	2389.530
Sum Sq. Dev.	1.12E+09	11122.69
Observations	31	31

$$\bar{X} = 22666.97$$

$$(X_f - \bar{X})^2 = (25000 - 22666.97)^2$$

$$= 5443028.981$$

$$\sum x_i^2 = \sum (X_i - \bar{X})^2 = \sigma_X^2 (n-1)$$

$$= 6112.965^2 \times (31-1) = 1121050233$$

$X_f = 25000$时

$$83.7846\,\mathrm{m}\,2.045 \times 8.027957 \times \sqrt{\frac{1}{31} + \frac{5443028.981}{1121050233}} = 83.7846\,\mathrm{m}\,3.1627$$

即是说：当地区城镇居民人均总收入达到25000元时，城镇居民每百户计算机拥有量 平均值置信度95%的预测区间为（80.6219，86.9473）台。

个别值区间预测：

$$Y_f = \hat{Y}_f \mp t_{\alpha/2}\,\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(X_f-\overline{X})^2}{\sum x_i^2}}$$

$X_F = 25000$时：

$$83.7846\,\mathrm{m}\,2.045\times 8.027957\times\sqrt{1+\frac{1}{31}+\frac{5443028.981}{1121050233}} = 83.7846\,\mathrm{m}16.7190$$

即是说：当地区城镇居民人均总收入达到25000元时，城镇居民每百户计算机拥有量 个别值置信度95%的预测区间为（67.0656，100.5036）台。

96

ECONOMETRICS

本章小结

1、变量间的关系分为函数关系与相关关系。

相关系数是对变量间线性相关程度的度量。

2、现代意义的回归是一个被解释变量对若干个解释变量依存关系的研究，回归的实质是由解释变量去估计被解释变量的平均值。

3、总体回归函数（**PRF**）是将总体被解释变量**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**）是将被解释变量**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、随机扰动项是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、同方差假定、无自相关假定、随机扰动与解释变量不相关假定、正态性假定）

6、普通最小二乘法（**OLS**）估计参数的基本思想及估计量；**OLS** 估计量的分布性质及期望、方差和标准误差；**OLS**估计式是最佳线性无偏估计量。

7、简单线性回归模型极大似然估计的思想和方法。

8、对回归系数区间估计的思想和方法。

9、拟合优度是样本回归线对样本观测数据拟合的优劣程度，可决系数是在总变差分解基础上确定的。可决系数的计算方法、特点与作用。

10、对回归系数假设检验的基本思想。对回归系数t检验的思想与方法；用P值判断参数的显著性。

11、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，被解释变量个别值区间预测的方法。

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