

第二章 简单线性回归模型

引子：中国旅游业成为国民经济战略性支柱产业

改革开放以来，中国实现了从旅游短缺型国家到旅游大国的历史性跨越。

“十二五”期间，旅游业全面融入国家战略体系，走向国民经济建设的前沿，成为国民经济战略性支柱产业。“十三五”旅游业发展的主要目标是：到2020年，旅游市场总规模达到67亿人次，旅游投资总额2万亿元，旅游业总收入达到7万亿元。旅游业综合贡献度达12%。

（来源：《国务院关于印发“十三五”旅游业发展规划的通知》）

- 什么决定性因素能使中国旅游业成为战略性支柱产业？
- 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？

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

为什么先讨论古典假定下的模型呢？

这是一种研究方式：

不使问题复杂化，先比较单一的（理想的）情况下去讨论！在比较单一的情况下，某些复杂的理论问题才更容易被阐述，也才更容易被接受，所以我们从完全满足某些条件的理想状态入手去加以分析。

本章的思想和原理是理解整个计量经济学的基础。

比喻：

学习经济学时，总是先要熟悉“完全竞争理论”，然后再接触“垄断和寡头等非完全竞争理论”。但是，并不是说“完全竞争理论”就总是真实的或总是符合实际的。

为什么先讨论简单线性回归模型呢？

在计量经济模型中，只有两个变量且为线性的回归模型最简单，称为简单线性回归模型。简单线性回归的原理可以直接用代数式去表述，较为直观，更容易理解和接受。

先讨论简单线性回归模型，然后很容易拓展到多元的情况。

本章主要讨论的问题：

- ▶ 回归分析的基本概念
- ▶ 线性回归模型参数的估计
- ▶ 参数的区间估计和假设检验
- ▶ 回归方程的拟合优度
- ▶ 回归模型预测

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

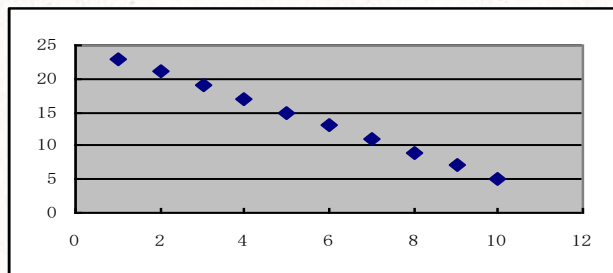
$Y=f(X) + \varepsilon$ (ε 为随机变量) 可用统计方法分析

◆没有关系 不用分析

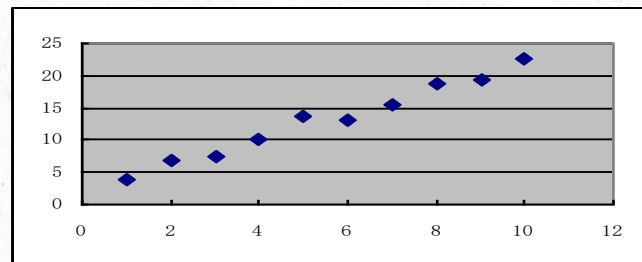
2、相关关系

► 相关关系的描述

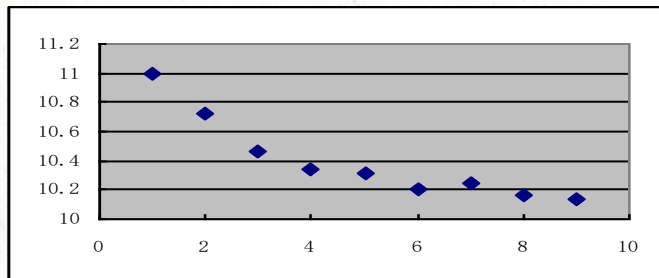
最直观的描述方式——坐标图（散布图、散点图）



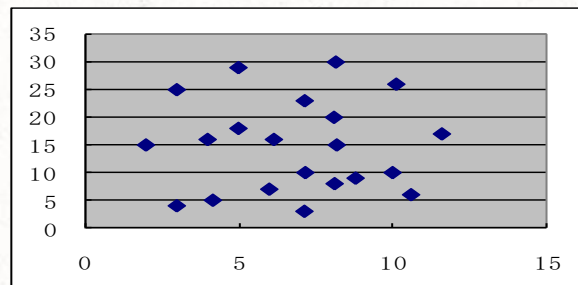
函数关系



相关关系(线性)



相关关系(非线性)



没有关系

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

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。

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。

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

只是相关分析还不能达到经济计量分析的目的

相关分析的局限:

相关系数只能反映变量间的线性相关程度, 不能确定变量间的因果关系; 相关系数只能说明两个变量线性相关的方向和程度, 不能说明相关关系具体接近哪条直线, 也就不能说明一个变量的变动会导致另一个变量变动的具体数量规律。

计量经济学关心的问题:

是经济变量间的因果关系以及隐藏在随机性后面的具体统计规律性
在这方面回归分析方法可以发挥更为重要的作用。

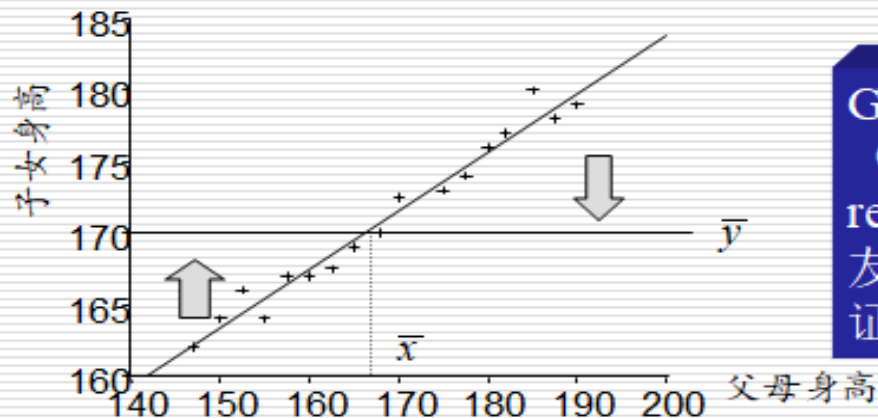
4. 回归分析

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”



“回归”(regression)一词最早由英国生物学家Francis Galton提出。

(1886年F. Galton的论文《Family Likeness in Stature》)

Galton的普遍回归定律 (law of universal regression) 被他的朋友Karl Pearson(1903)证实。

4、回归分析

回归的现代含义:

- 回归分析是关于研究一个叫做因变量的变量 (Y) 对另一个或多个叫做自变量的变量 (X) 的依赖关系;
- 其用意在于通过自变量在重复抽样中的已知或设定值, 去估计或预测因变量的总体均值。
- 回归 (Regression) 是计量经济学的主要工具

姚明9岁女儿身高“失控”，比成年人还高，郭敬明躺枪



猫眼电影
09月09日 17:46

+ 关注

有网友在近日晒出了一段关于姚明的视频，视频中是姚明与女儿姚沁蕾，以及老婆叶莉的出席活动的照片。

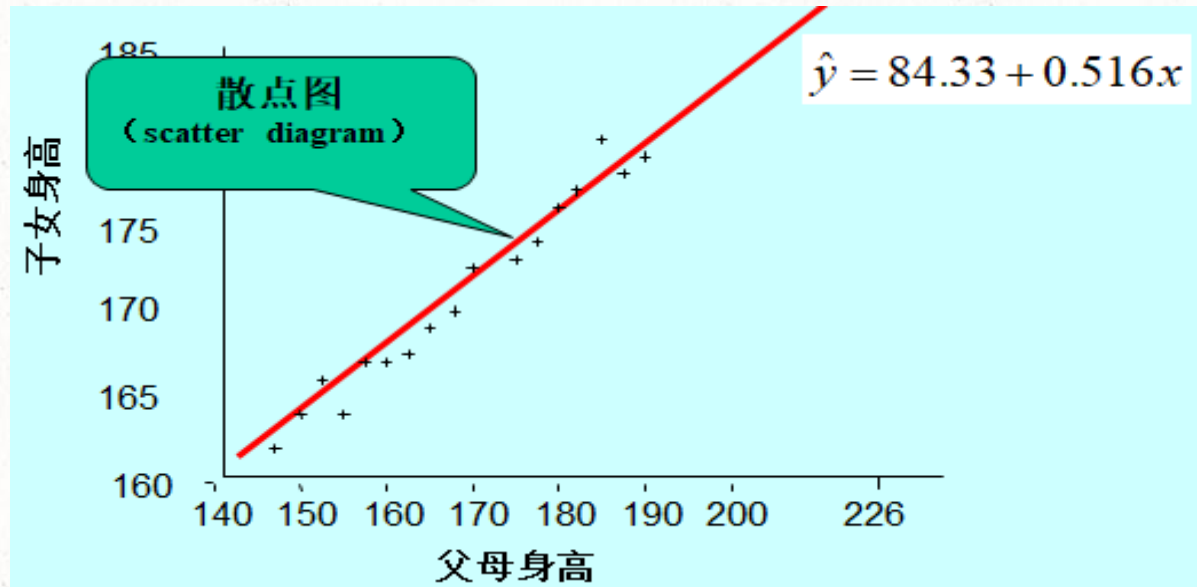


说说你的看法



例子：姚明身高2.26米，姚明的子女会有多高呢？

2.01米 可信吗？



因此，一旦知道了父母的身高，就可以按照上述关系式（回归线）来预测子女的平均身高（而不是具体身高）

(1) 注意明确几个概念（为深刻理解“回归”）

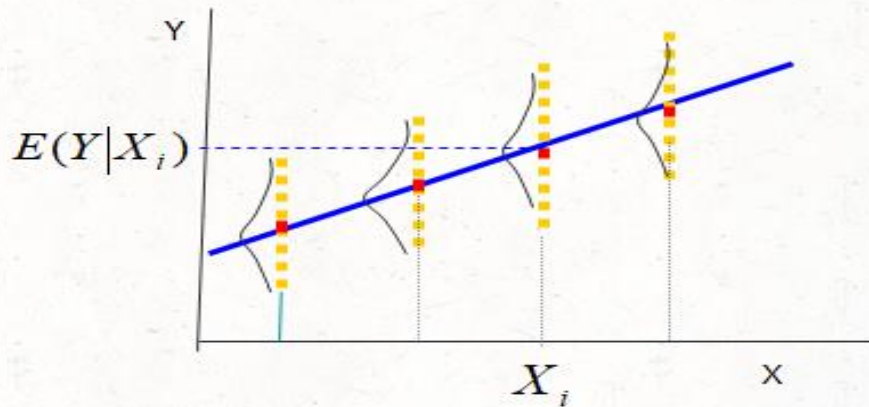
●被解释变量Y的条件分布和条件概率：

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，都有Y的条件期望

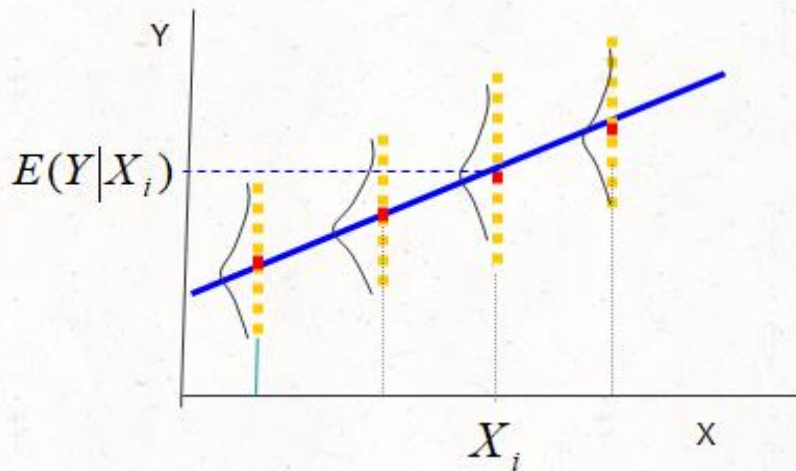
$E(Y|X_i)$ 与之对应，代表Y的条件期望的点的轨迹形成的直线或曲线称为回归线。

● **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

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

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



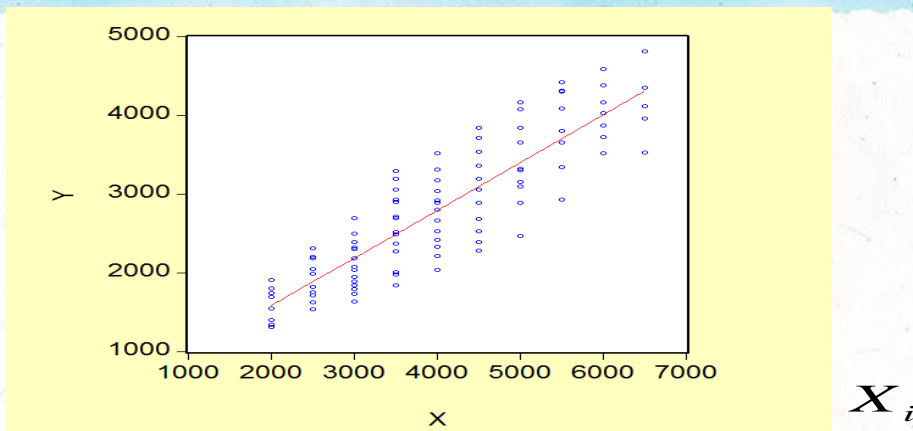
二、总体回归函数 (PRF)

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 | | | | | | | | | |
|------------|-------------|------|------|------|------|------|------|------|------|------|
| | 2000 | 2500 | 3000 | 3500 | 4000 | 4500 | 5000 | 5500 | 6000 | 6500 |
| 每月家庭消费支出 Y | 1312 | 1530 | 1631 | 1843 | 2037 | 2277 | 2469 | 2924 | 3515 | 3521 |
| | 1340 | 1619 | 1726 | 1974 | 2210 | 2388 | 2889 | 3338 | 3721 | 3954 |
| | 1400 | 1713 | 1786 | 2006 | 2325 | 2526 | 3090 | 3650 | 3865 | 4108 |
| | 1548 | 1750 | 1835 | 2265 | 2419 | 2681 | 3156 | 3802 | 4026 | 4345 |
| | 1688 | 1814 | 1885 | 2367 | 2522 | 2887 | 3300 | 4087 | 4165 | 4812 |
| | 1738 | 1985 | 1943 | 2485 | 2665 | 3050 | 3321 | 4298 | 4380 | |
| | 1800 | 2041 | 2037 | 2515 | 2799 | 3189 | 3654 | 4312 | 4580 | |
| | 1902 | 2186 | 2078 | 2689 | 2887 | 3353 | 3842 | 4413 | | |
| | | 2200 | 2179 | 2713 | 2913 | 3534 | 4074 | | | |
| | | 2312 | 2298 | 2898 | 3038 | 3710 | 4165 | | | |
| | | | 2316 | 2923 | 3167 | 3834 | | | | |
| | | | 2387 | 3053 | 3310 | | | | | |
| | | | 2498 | 3187 | 3510 | | | | | |
| | | | 2689 | 3286 | | | | | | |
| | 1591 | 1915 | 2092 | 2586 | 2754 | 3039 | 3396 | 3853 | 4036 | 4148 |

家庭消费支出的条件期望与家庭收入的关系的图形：

$$E(Y|X_i)$$



对于本例的全体，家庭消费支出的条件期望 $E(Y|X_i)$ 与家庭收入 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,也就是要去寻求总体回归函数。

► 条件期望表现形式

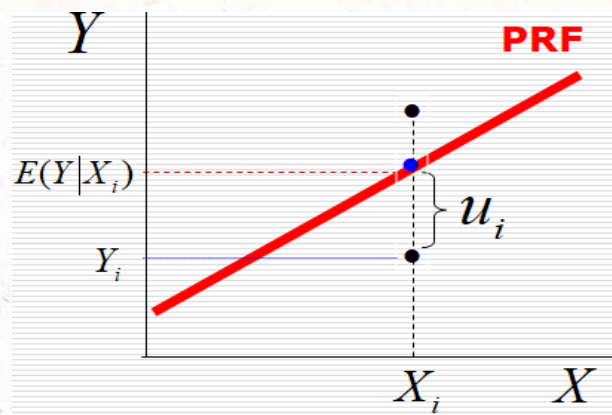
例如Y的条件期望 $E(Y|X_i)$ 是解释变量X的线性函数，可表示为：

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

► 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

$$\text{则有 } u_i = Y_i - E(Y|X_i) = Y_i - \beta_1 - \beta_2 X_i \quad Y_i = \beta_1 + \beta_2 X_i + u_i$$



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

◆概念

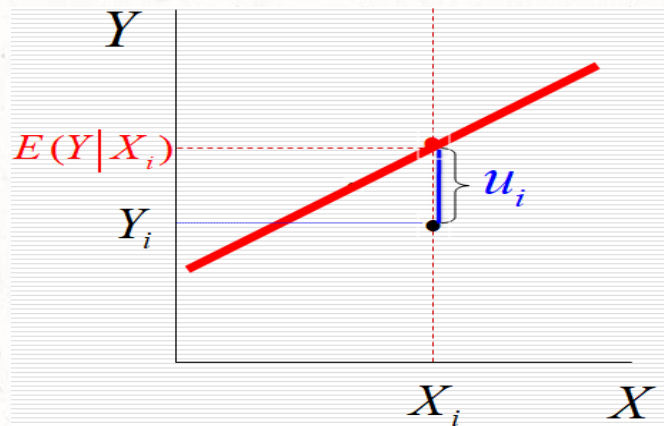
在总体回归函数中，各个 Y_i 的值与其条件期望 $E(Y_i|X_i)$ 的偏差 u_i 有很重要的意义。若只有 X 影响 Y ， Y_i 与 $E(Y_i|X_i)$ 不应有偏差。

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



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)

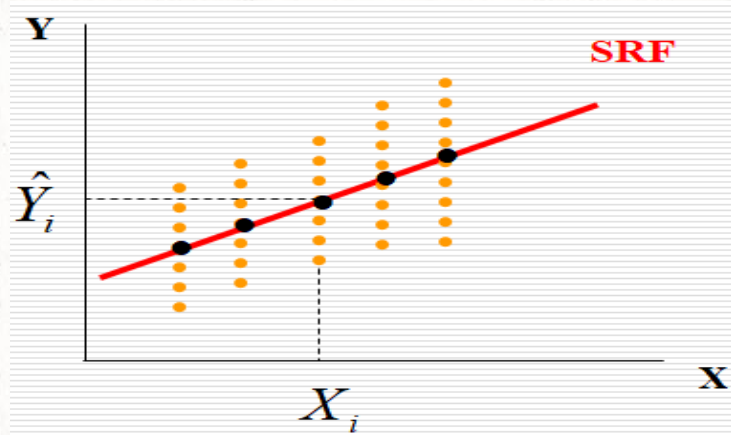
四、样本回归函数（SRF）

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 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**）。



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**：

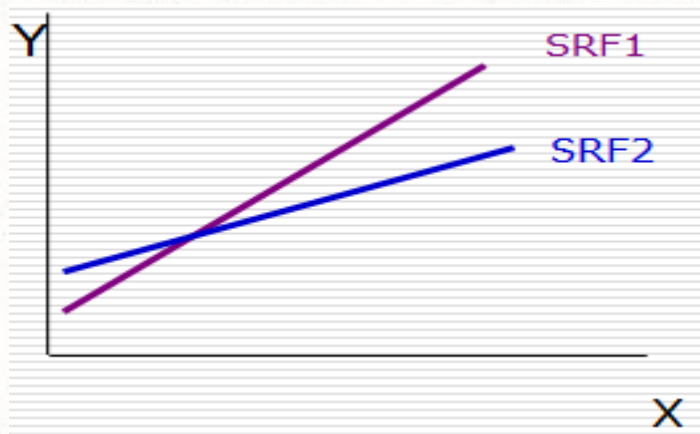
则 $e_i = Y_i - \hat{Y}_i$ 或 $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$

样本回归函数的特点

- 样本回归线随抽样波动而变化:

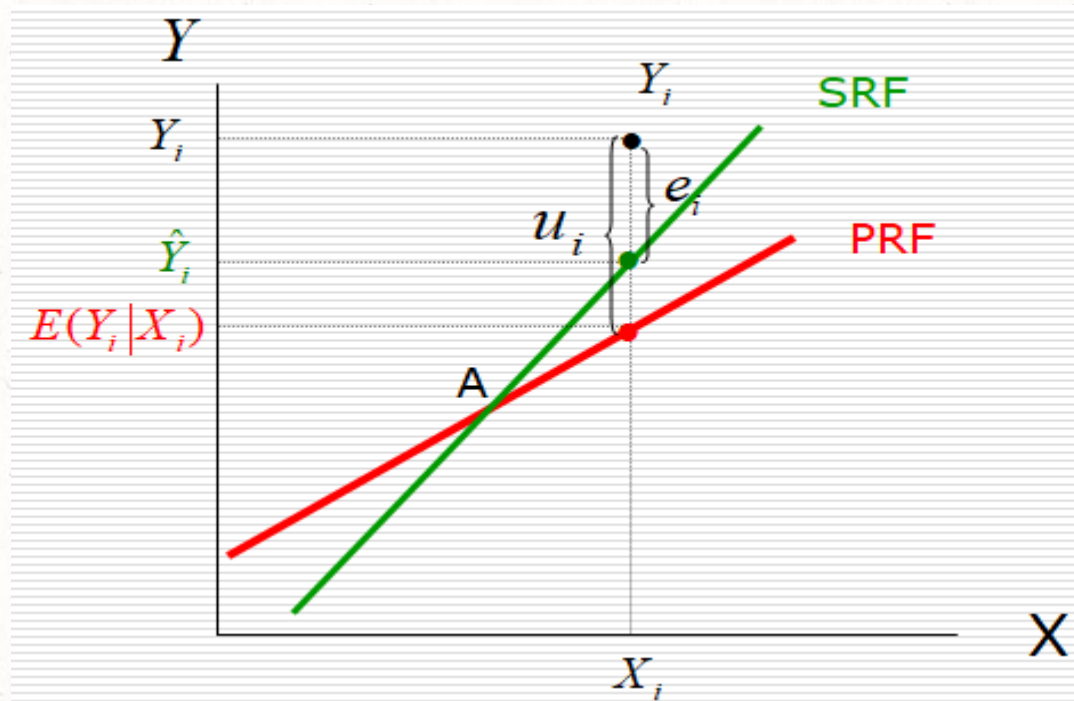
每次抽样都能获得一个样本，就可以拟合一条样本回归线，（**SRF不唯一**）

- 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

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

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

样本回归函数

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

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

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。

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。

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。
- 因为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

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)

注意：解释变量非随机在自然科学的实验研究中容易满足,经济领域变量的观测是被动不可控的， X 非随机的假定不容易满足。

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

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

假定3: 无自相关假定:

随机扰动项 u_i 的逐次值互不相关

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

假定4: 随机扰动项 u_i 与解释变量 X_i 不相关

(从随机扰动 u_i 角度看)

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

所有这些假定叫**古典假定**或者**高斯假定**，满足这些古典假定的线性回归模型叫**经典线性回归**（Classical Linear Regression Model, **CLRM**）。

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

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

(说明：正态性假定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)

注意：

并不是参数估计的每一具体步骤都要用到所有的假定，但对全部假定有完整的认识，对学习计量经济学的原理是有益的。

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，因此

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

$$\text{Var}(Y_i | X_i) = \sigma^2$$

假定3：无自相关假定

$$\text{Cov}(Y_i, Y_j) = 0$$

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。
- 理想的估计方法应使估计的 \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 $\hat{\beta}_2$ 。

2 正规方程和估计式

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

$$\begin{aligned}\sum Y_i &= n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i \\ \sum X_i Y_i &= \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2\end{aligned}$$

用克莱姆法则求解得以观测值表现的OLS估计式：

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

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：

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

可以证明：（见教材P29—P30证明）

（证明过程用到OLS正规方程的结论，但与基本假定无关）

- 回归线通过样本均值

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

（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）

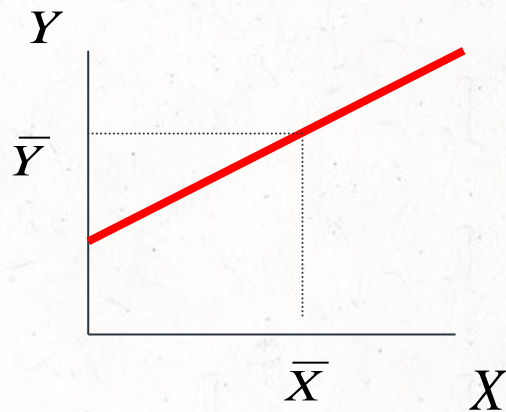
- 估计值 \hat{Y}_i 的均值等于实际观测值 Y_i 的均值

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

- 剩余项 e_i 的均值为零

（由OLS第一个正规方程直接得到）

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



被解释变量估计值 \hat{Y}_i 与剩余项 e_i 不相关

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

由OLS正规方程有: $\sum e_i = 0$ $\sum e_i X_i = 0$ (注意:红色的项为0)

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

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

- 解释变量 X_i 与剩余项 e_i 不相关

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

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

4. OLS估计式的统计性质

回顾第1章：参数估计式的优劣需要有评价的标准

- ◆参数无法通过观测直接确定，只能通过样本估计，但因存在抽样波动，参数估计值不一定等于总体参数的真实值。
- ◆参数估计方法及所确定的估计式不一定完备，不一定能得到总体参数的真实值，需要对估计方法作评价与选择。

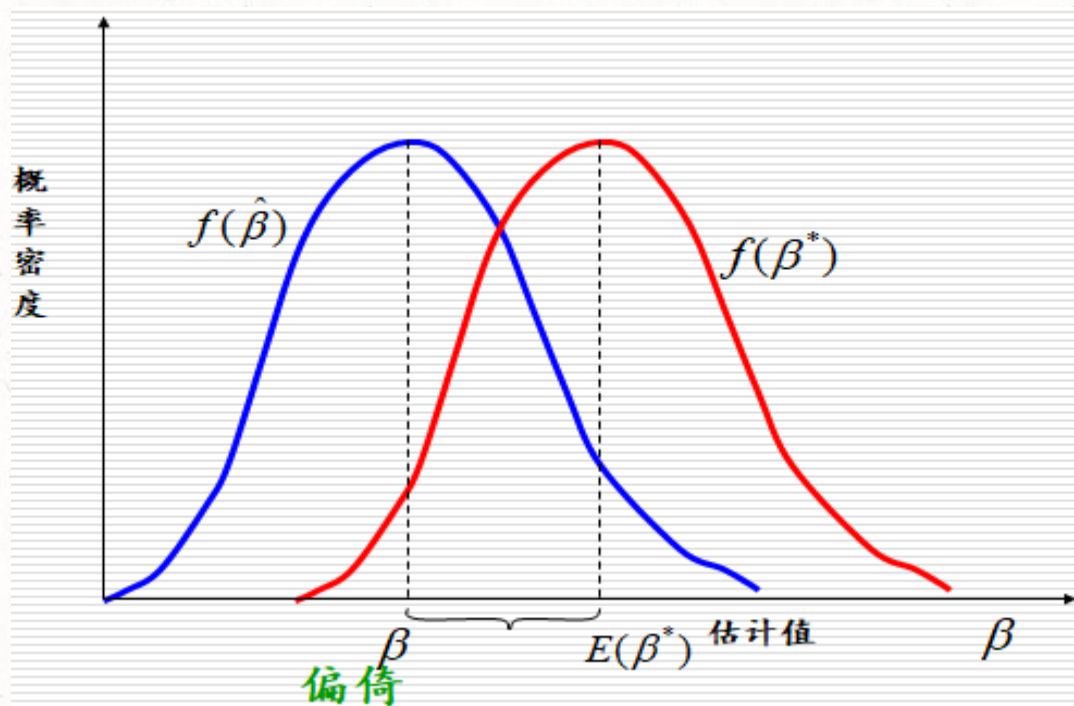
比较不同估计方法的估计结果时，需要有一定的评价标准

基本要求：参数估计值应尽可能地接近总体参数的真实值

估计准则：“尽可能地接近”原则

决定于参数估计式的统计性质：无偏性、有效性、一致性等。

(1) 无偏性



(1) 无偏性

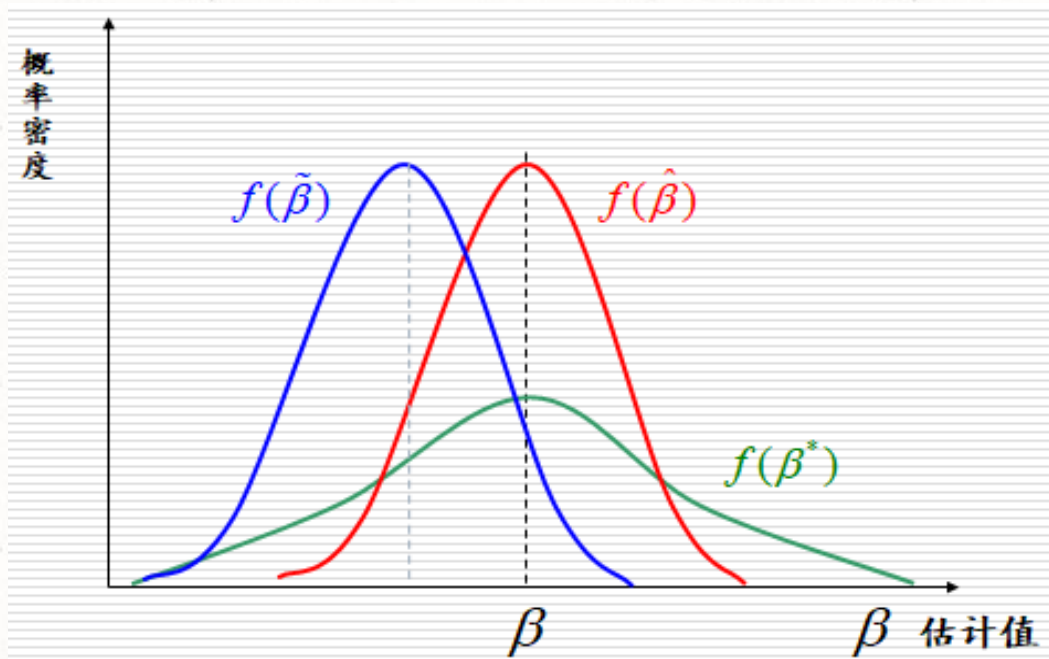
前提：重复抽样中估计方法固定、样本数不变、经 重复抽样的观测值, 可得一系列参数估计值 $\hat{\beta}$, $\hat{\beta}$ 的分布称为 $\hat{\beta}$ 的抽样分布, 其密度函数记为 $f(\hat{\beta})$

如果 $E(\hat{\beta}) = \beta$

称 $\hat{\beta}$ 是参数 β 的无偏估计式, 否则 $E(\hat{\beta}) \neq \beta$ 则称

$\hat{\beta}$ 是有偏的估计, 其偏倚为 $E(\hat{\beta}) - \beta$

(2) 有效性



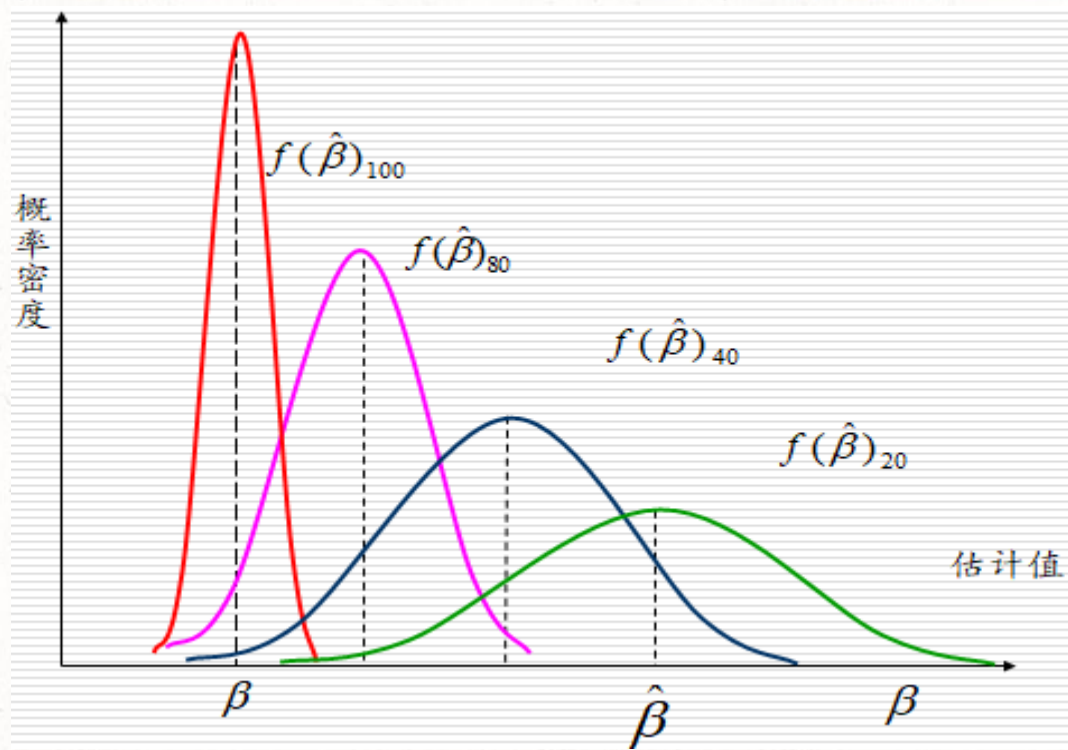
(2) 有效性

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式

既是无偏的同时又具有最小方差特性的估计式，称为**最佳（有效）估计式**。

(3) 渐近性质 (大样本性质)



(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

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

（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）

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。

4. 分析OLS估计式的统计性质

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

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

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

$\hat{\beta}_k$ 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估计式是点估计式

OLS估计式的统计性质——高斯定理

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

1、 线性特征 $\hat{\beta}_k$ 是Y的线性函数

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

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

2、 无偏特性

可以证明 $E(\hat{\beta}_k) = \beta_k$ (证明见教材P32-33)

3、最小方差特性 （证明见教材P59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:无偏性和最小方差性的证明中用到了基本假定1-假定4）

结论（高斯-马尔科夫定理）：

在古典假定条件下，OLS估计式是最佳线性无偏估计式（BLUE）。

OLS估计量的精度（标准误差）

- ▶ OLS估计量是样本数据的函数，若样本改变，则基于估计量计算出的数值也会改变，一个直观的想法是，若估计量随样本改变而变化的程度很低，即其标准误差（standard error）很小，则说明该估计量“很可靠”，或其精度高“重方法”。
- ▶ 根据CLRM的假设，可以计算OLS估计量的标准误：

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

$$se(\hat{\beta}_1) = \sqrt{\frac{\sum X_i^2}{n \sum x_i^2}} \sigma$$

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

自由度

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

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 $E(\sum e_i^2) = (n-2)\sigma^2$

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

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

(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

第三节 拟合优度的度量

概念:

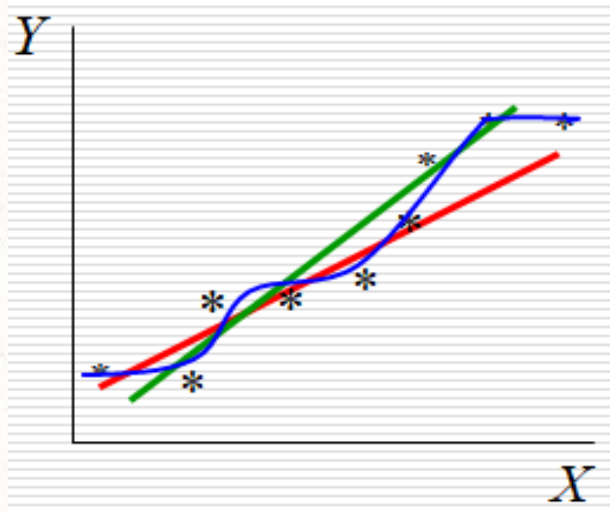
样本回归线是对样本数据的一种拟合。

● 不同的模型（不同函数形式）

可拟合出不同的回归线

● 相同的模型用不同方法估计

参数，可以拟合出不同的回归线



拟合优度的度量

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

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

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

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

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

$$\text{(TSS)} \quad \quad \text{(ESS)} \quad \quad \text{(RSS)}$$

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）

二、可决系数

以TSS同除总变差等式两边：

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

或

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

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 或 R^2 表示：

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

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数值关系

联系：数值上可决系数是相关系数的平方

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

可决系数与相关系数的区别

区别:

| 可决系数 | 相关系数 |
|--------------------------------|-------------------------------|
| 就模型而言 | 就两个变量而言 |
| 说明解释变量对被解释变量的解释程度 | 说明两变量线性依存程度 |
| 度量的不对称的因果关系 | 度量的对称的相关关系 |
| 取值 $0 \leq R^2 \leq 1$ 有非负性 | 取值 $-1 \leq r \leq 1$ 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**解释变量对被解释变量的**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）
- 如果回归的主要目的是经济结构分析，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。
- 如果研究的主要目的只是为了预测被解释变量的值，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。

可决系数使用原则

- ▶ **切勿因为 R^2 的高或低轻易地肯定或否定一个模型：**
 - 视数据类型和样本容量
 - 视研究目的不同
 - 描述性判断而非显著性判断
- ▶ **可以比较不同模型的 R^2 但有前提：**
 - 样本相同
 - 被解释变量相同
- ▶ **R^2 具有两层含义， R^2 高意味着：**
 - 样本回归线对样本数据的拟合程度较高
 - 所有解释变量联合起来对被解释变量的影响程度较高

回顾：常用的统计分布

► 正态分布：

- 概率密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- 数字特征：

$$E(X) = \mu \quad D(X) = \sigma^2$$

$$3\text{阶矩: } E(X - \mu)^3 = 0$$

$$4\text{阶矩: } E(X - \mu)^4 = 3\sigma^4$$

► 正态分布标准化

若 $X \sim N(\mu, \sigma^2)$, 且 $\eta = \frac{X - \mu}{\sigma}$,

则 $\eta \sim N(0, 1)$

回顾：常用的统计分布

- ▶ 由正态分布可导出 χ^2 分布、 t 分布及 F 分布，它们与正态分布一起，是数理统计中常用的分布

- χ^2 分布

- 当 X_1, X_2, \dots, X_n 相互独立且都服从 $N(0,1)$ 时， $Z = \sum_i X_i^2$ 的分布称为自由度等于 n 的 χ^2 分布，记作 $Z \sim \chi^2(n)$
- χ^2 分布具有可加性，即当 Y 与 Z 相互独立，且 $Y \sim \chi^2(n)$ ， $Z \sim \chi^2(m)$ ，则 $Y+Z \sim \chi^2(n+m)$

回顾：常用的统计分布

○ t 分布

■ 若 X 与 Y 相互独立，且 $X \sim N(0,1)$ ， $Y \sim \chi^2(n)$ ，则

$$Z = \frac{X}{\sqrt{Y/n}}$$

的分布称为自由度等于 n 的 t 分布，记作 $Z \sim t(n)$ 。

回顾：常用的统计分布

○ F分布

■ 若 X 与 Y 相互独立，且 $X \sim \chi^2(n)$ ， $Y \sim \chi^2(m)$ ，则

$$Z = \frac{X}{n} \bigg/ \frac{Y}{m}$$

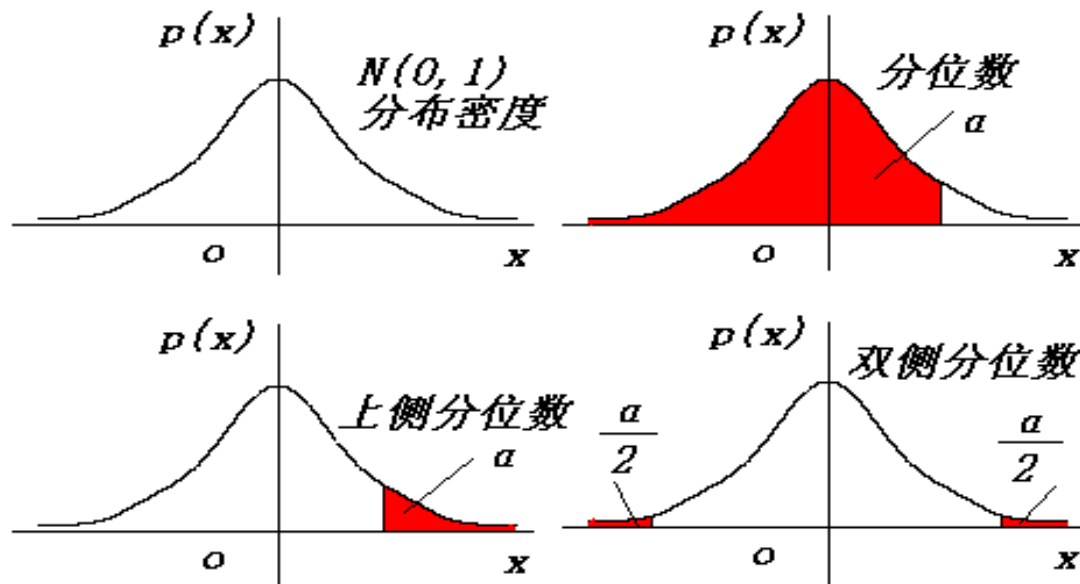
的分布称为第一自由度等于 n 、第二自由度等于 m 的F分布，记作 $Z \sim F(n, m)$ 。

回顾：常用的统计分布

- ▶ 各种分布的分位数：随机变量 X 的分布函数为 $F(x)$ ，实数 α 满足 $0 < \alpha < 1$ 时：
 - α 分位数是使 $P\{X < x_\alpha\} = F(x_\alpha) = \alpha$ 的数 x_α
 - 上侧 α 分位数是使 $P\{X > \lambda\} = 1 - F(\lambda) = \alpha$ 的数 λ
 - 双侧 α 分位数是使 $P\{X < \lambda_1\} = F(\lambda_1) = 0.5\alpha$ 的数 λ_1 、使 $P\{X > \lambda_2\} = 1 - F(\lambda_2) = 0.5\alpha$ 的数 λ_2

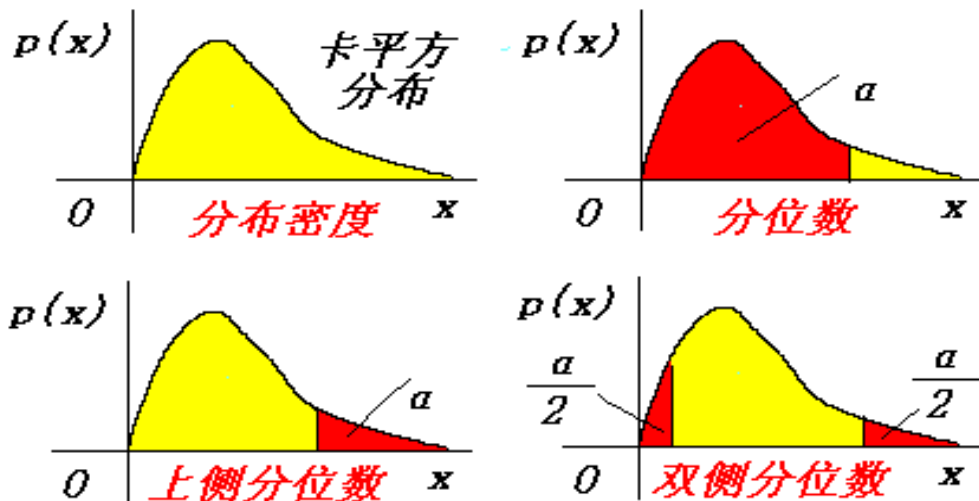
回顾：常用的统计分布

► 各种分布的图形和分位数

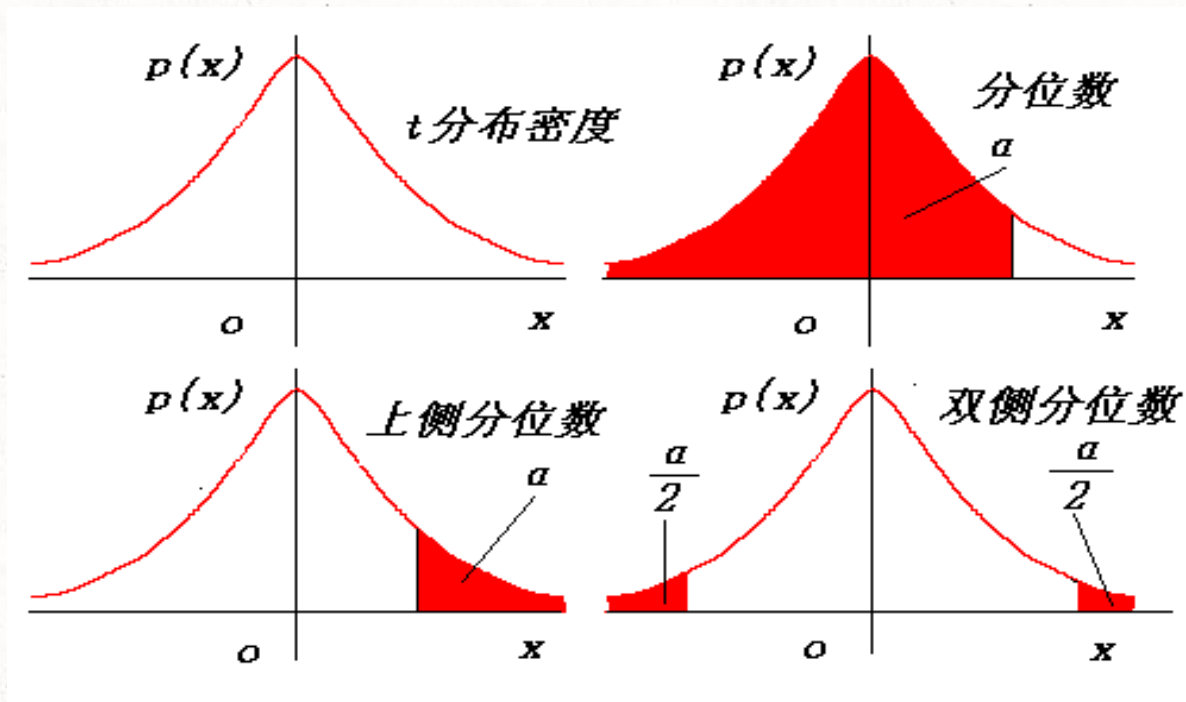


回顾：常用的统计分布

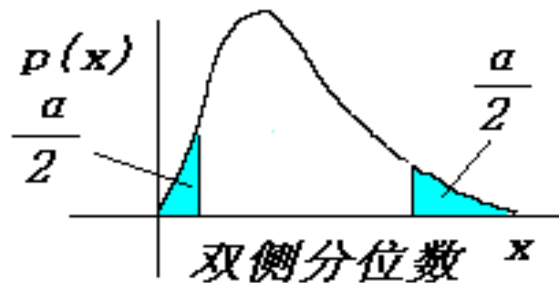
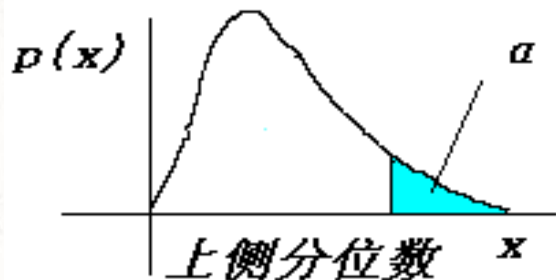
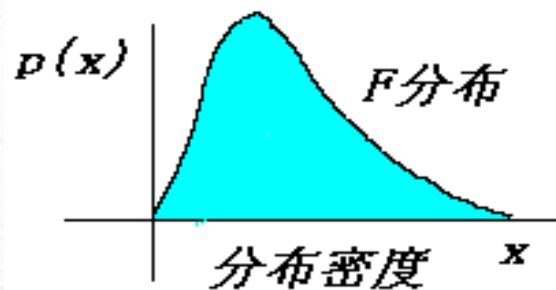
► 各种分布的图形和分位数



回顾：常用的统计分布



回顾：常用的统计分布



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。

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；

$\hat{\beta}_k$ 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 \longrightarrow Y_i 正态 \longrightarrow $\hat{\beta}_k$ 正态

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

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

线性特征 $\hat{\beta}_2 = \sum k_i Y_i$

$\hat{\beta}_k$ 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P33, 要求看懂!)
(标准误差是方差的平方根)

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

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

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

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

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 $\hat{\beta}_k$ 作标准化变换

分布函数

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

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

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

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:

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

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 α ($0 \leq \alpha \leq 1$)，能使得 $(\hat{\beta}_k - \delta, \hat{\beta}_k + \delta)$

这样的区间包含真实 β_k 的概率为 $1 - \alpha$ ，即

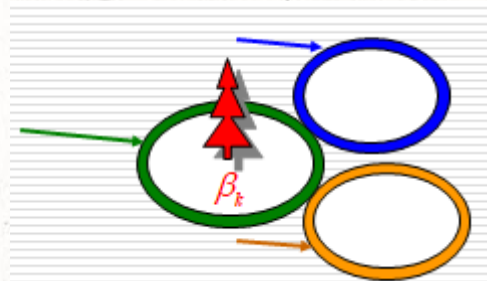
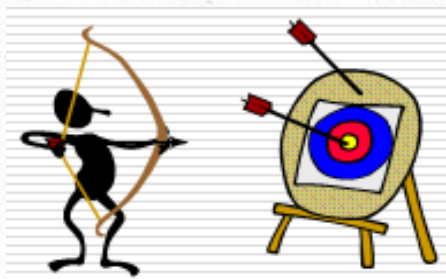
$$P(\hat{\beta}_k - \delta \leq \beta_k \leq \hat{\beta}_k + \delta) = 1 - \alpha$$

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 $1 - \alpha$ ”。这种说法对吗?

怎样正确理解置信区间？

注意： β_k 是未知但**确定**的数， $(\hat{\beta}_k - \delta, \hat{\beta}_k + \delta)$ 是随抽样而变化的**随机区间**。

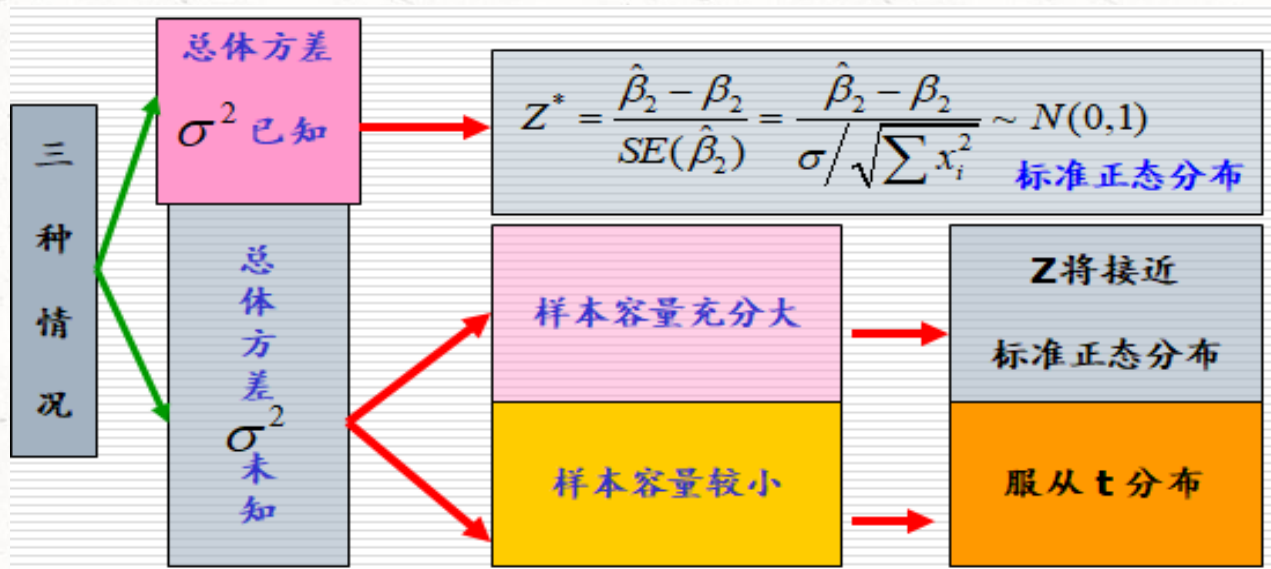


从重复抽样的观点看，每次抽样都可构造一个区间，象这样构造的区间，平均来说有 $(1-\alpha)$ 比例的次数包含 β_k 的真实值。但对**特定样本**，一旦估计出特定的 $\hat{\beta}_k$ ，区间 $(\hat{\beta}_k - \delta, \hat{\beta}_k + \delta)$ 就不再是随机的，而是特定的，这时它或者包含 β_k ，或者不包含 β_k 。

问题: α 是给定的, 如何去寻找合适的 δ 呢?

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

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 :

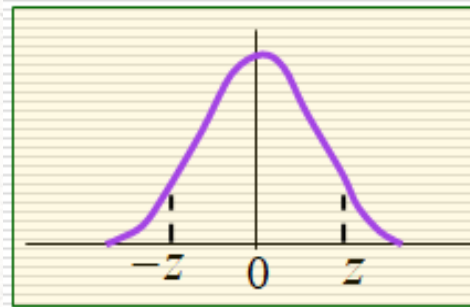


回归系数的区间估计 (分三种情况寻找合适的 δ)

(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 α 对应的临界值 $z_{\alpha/2}$ (1.96), 则标准化变量 **z*** (统计量)

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

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



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

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

(2) 当总体方差 σ^2 未知，而样本容量充分大时

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 \mathbf{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“ \wedge ”，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 $\delta = z_{\alpha/2} \hat{SE}(\hat{\beta}_2) = z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{\sum x_i^2}}$

(3) 当总体方差 σ^2 未知，且样本容量较小时

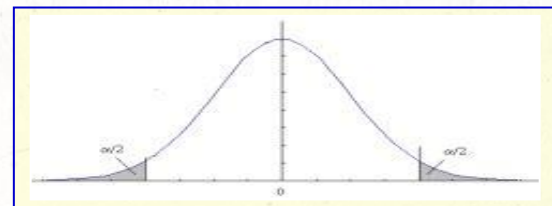
方法：用无偏估计 $\hat{\sigma}^2$ 去代替未知的 σ^2 ，由于样本容量较小，“标准化变量”**t**（统计量）不再服从正态分布，而服从**t**分布。

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

这时可用**t**分布去建立参数估计的置信区间。选定 α ，查**t**分布表得自由度为**n-2**的双侧分位数 $t_{\alpha/2}(n-2)$ 作为临界值，则有

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

即



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

例1:研究某市城镇居民人均鲜蛋需求量 Y (公斤)与人均可支配收入 X (元,1980年不变价计)的关系

设定模型:

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

1995-2005年**样本数据:**

| 年份 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 |
|-----|-------|-------|-------|-------|-------|--------|--------|--------|--------|--------|--------|
| Y | 14.4 | 14.4 | 14.4 | 14.7 | 17.0 | 16.3 | 18.0 | 18.5 | 18.2 | 19.3 | 17.1 |
| X | 847.3 | 821.0 | 884.2 | 903.7 | 984.1 | 1035.3 | 1200.9 | 1289.8 | 1432.9 | 1539.0 | 1633.6 |

估计参数:

$$\bar{Y} = 16.57, \quad \bar{X} = 1142.89, \quad n = 11$$

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{4489.94}{858661.8} = 0.005$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 16.57 - 0.005 \times 1142.89 = 10.60$$

估计结果： $\hat{Y}_t = 10.60 + 0.005 X_t$

计算可决系数

例1:由前面的估计结果可计算出 $\sum e_i^2 = 10.56$

由数据Y可计算出:

$$\sum (Y_i - \bar{Y})^2 = \sum y_i^2 = 34.0316$$

则

$$R^2 = 1 - \frac{\sum e_i^2}{\sum y_i^2} = 1 - \frac{10.56}{34.0316}$$

$$R^2 = 1 - 0.3103 = 0.6897$$

估计 σ^2 :

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{10.56}{11-2} = 1.1736$$

$$\hat{SE}(\hat{\beta}_2) = \frac{\hat{\sigma}}{\sqrt{\sum x_i^2}} = \frac{\sqrt{1.1736}}{\sqrt{858661.8}} = 0.001$$

给定 $\alpha=0.05$ 查df=n-2=9的t分布临界值 $t_{0.025}(9) = 2.262$

参数区间估计: $P[\hat{\beta}_2 - t_{\alpha/2} \hat{SE}(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2} \hat{SE}(\hat{\beta}_2)] = 1 - \alpha$

$$P[0.005 - 2.262 \times 0.001 \leq \beta_2 \leq 0.005 + 2.262 \times 0.001] = 1 - 0.05$$

$$P(0.0027 \leq \beta_2 \leq 0.0073) = 0.95$$

若给定 $\alpha=0.10$ 查df=9的t分布临界值 $t_{0.05}(9) = 1.833$

则

$$P(0.0032 \leq \beta_2 \leq 0.0068) = 0.90$$

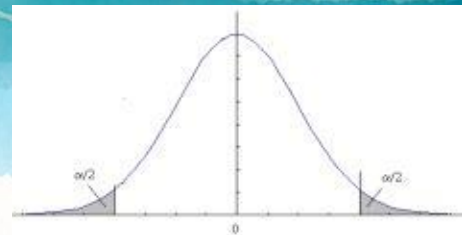
若给定 $\alpha=0.20$ 则

$$P(0.0036 \leq \beta_2 \leq 0.0064) = 0.80$$

若给定 $\alpha=0.50$ 则

$$P(0.0043 \leq \beta_2 \leq 0.0057) = 0.50$$

三、回归系数的假设检验

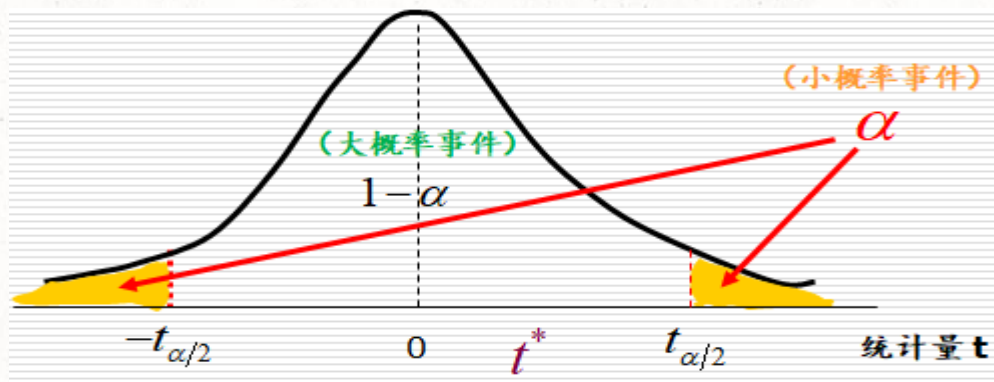


目的：简单线性回归中，检验X对Y是否真有显著影响

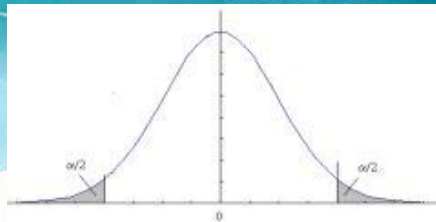
基本概念回顾：临界值与概率、大概率事件与小概率事件

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^*

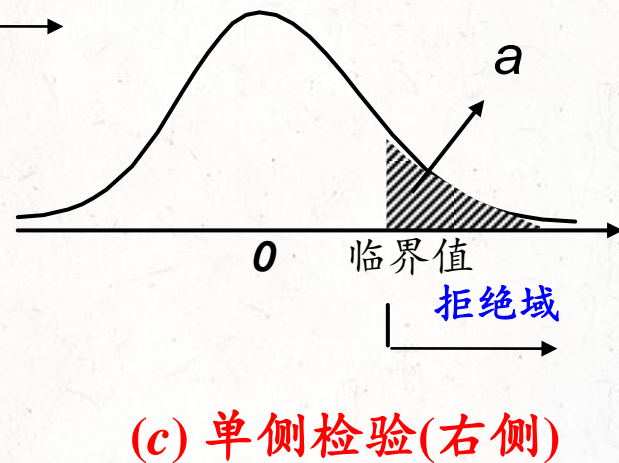
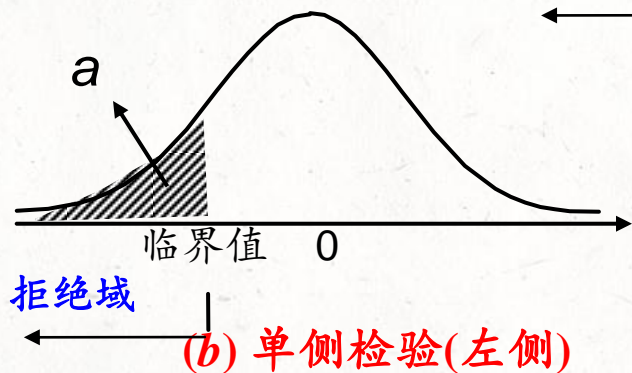
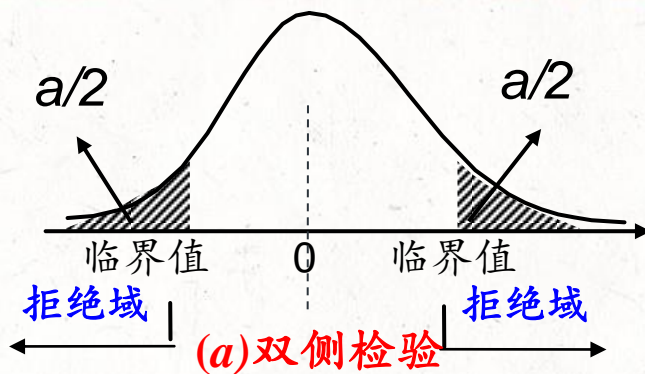


1. 假设检验的基本思想

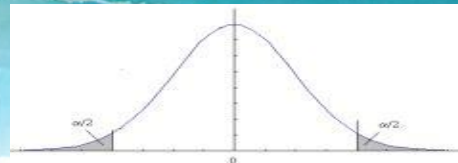


- ▶ 在某种条件下，在一次抽样中，大概率事件出现被认为是合理的，而小概率事件被认为基本不会发生，如果小概率事件竟然发生了，认为是不合理的。
- ▶ 在事先作出的某种原假设成立的条件下，利用样本构造适当统计量（一次抽样的结果），并确定统计量的抽样分布。给定显著性水平，构造一个小概率事件。如果在一次抽样中该小概率事件竟然发生，就认为原假设不真实，从而拒绝原假设，不拒绝备择假设。反之，如果大概率事件发生，则不拒绝原假设。

双侧检验与单侧检验



2. 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

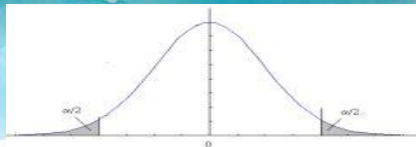
$$Z^* = \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} \sim N(0,1)$$

给定 α ，查正态分布表得临界值 Z

▼ 如果 $-z < Z^* < z$ (大概率事件发生) 则不拒绝原假设 H_0

▼ 如果 $Z^* \leq -z$ 或 $Z^* \geq z$ (小概率事件发生) 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:

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

给定 α , 查 t 分布表得 $t_{\alpha/2}(n-2)$

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

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

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

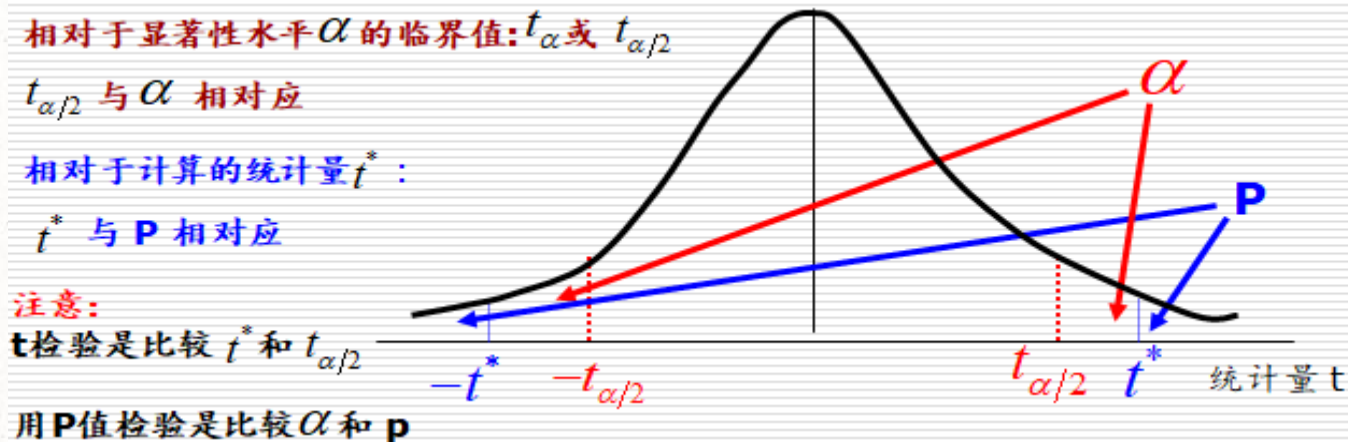
则不拒绝原假设 $H_0: \beta_2 = 0$

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，原假设可以被拒绝的**最高**显著性水平。

统计分析软件中通常都给出了检验的 p 值



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 $\alpha \geq p$ 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 $\alpha < p$ 值，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 $p \leq \alpha$ 时，P值越小，越能拒绝原

假设 H_0

举例：对例1参数的显著性检验

给定 $\alpha = 0.05$ 查 $df=9$ 的 t 分布临界值 $t_{0.025}(9) = 2.262$

计算统计量
$$t^* = \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = \frac{0.005}{0.001} = 5.00$$

判断：因 $t^* = 5.00 > t_{0.025}(9) = 2.262$ 拒绝 $H_0 : \beta_2 = 0$

说明 β_2 显著不为0，X对Y 确有显著影响

用P值检验：（需要确定与 $t^* = 5.00$ 对应的P值）

由 $t^* = 5.00$ ， $df=9$ ，查 t 分布表知道 $P < 0.001$ ($t = 4.781$ 时)

因 $t = 5.00$ 时的P值 $< 0.001 < \alpha = 0.05$

则在显著性水平 $\alpha = 0.05$ 下更应拒绝原假设 $H_0 : \beta_2 = 0$

即认为 X 对 Y 有显著影响

例如，给定 $\hat{\beta}$ 服从 t 分布， $\hat{\beta} - \beta_0$ 是否显著异于零，关键是看这个差值的绝对值等于估计值 $\hat{\beta}$ 的多少倍标准差。

$$t_{\alpha/2} = \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})}$$

知道了 $t_{\alpha/2}$ ，查表可得 α 的值，即置信水平（EViews输出为p值）。若这个置信水平满足研究要求，则认为这个“差异”显著，否则不显著。

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|----------|-------------|------------|-------------|--------|
| C | 24.45455 | 6.413817 | 3.812791 | 0.0051 |
| X | 0.509091 | 0.035743 | 14.24317 | 0.0000 |

假设检验

假设检验中的一些实际操作问题：

- ▶ 由于原假设通常存在一个接受域，因此在不能拒绝某个原假设时宁可说“不拒绝”而不是“接受”原假设（接受是一种排他性的、确定性的论断）。
- ▶ “2倍t”经验法则（简化了的t检验法）：如果自由度 >20 且显著性水平在0.05，若t值绝对值 >2 ，则可拒绝原假设，因t分布在自由度为20或更大，5%的显著水平上其临界值在2左右。
- ▶ 经典假设检验的不足在于选择 α 时的武断性，而P值是一个原假设被拒绝的最高显著性水平。对于给定的 α ，若p值小于 α ，则拒绝原假设，它与检验统计量值落在 α 水平下的拒绝域内是等价的

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

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

(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 X_i

F = 202.87 DW = 2.3 F 统计量 DW统计量

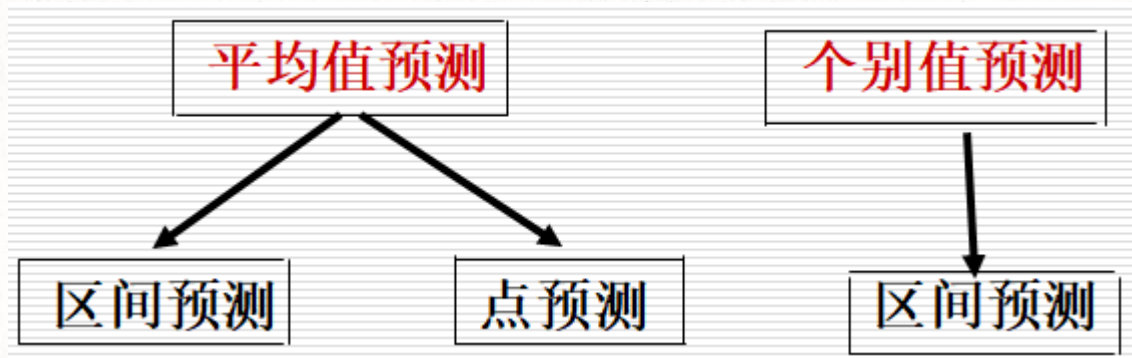
1. 基本思想

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

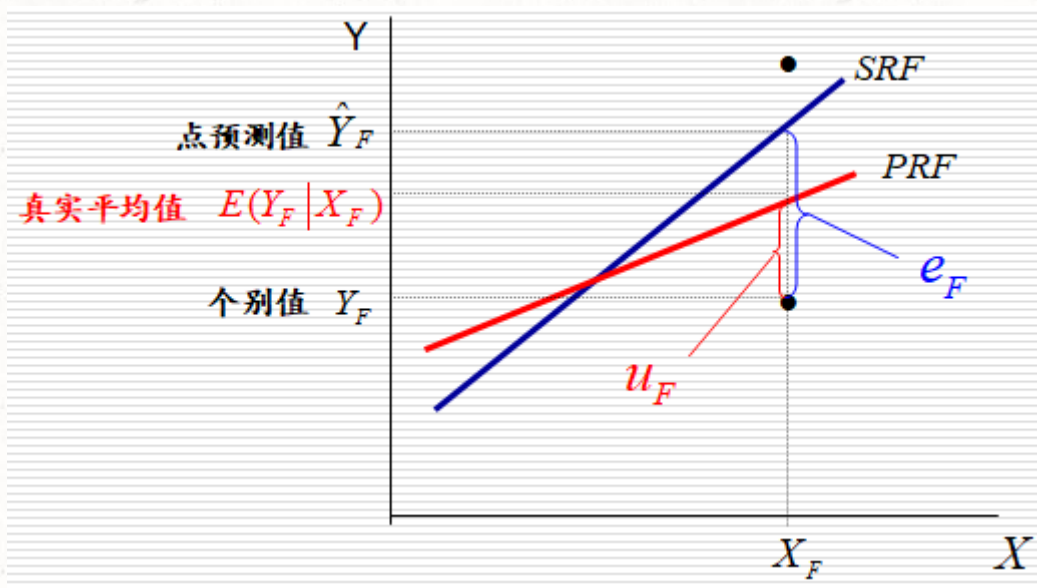
- 运用计量经济模型作预测：指利用所估计的样本回归函数作预测工具，用解释变量的已知值或预测值，对预测期或样本以外的被解释变量的数值作出定量的估计。
- 计量经济预测是一种条件预测：
 - 条件：◆模型设定的关系式不变
 - ◆所估计的参数不变
 - ◆解释变量在预测期的取值已作出预测

预测的类型

- 对被解释变量Y的预测分为：
 平均值预测和个别值预测
- 对被解释变量Y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 是对真实平均值的点估计

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

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 是随机变量，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的平均值作区间预测，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)

由 $\hat{Y}_F = \hat{\beta}_1 + \hat{\beta}_2 X_F$ \hat{Y}_F 服从正态分布(为什么?)

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

可以证明 (较复杂不具体证明)

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

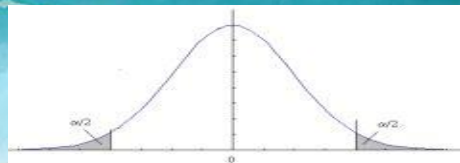
当 σ^2 未知时, 只得用 $\hat{\sigma}^2 = \sum e_i^2 / (n-2)$ 代替, 这时将 \hat{Y}_F 标准化, 有

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

注意:

$\hat{\sigma}$

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

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

$P\{[\hat{Y}_F - t_{\alpha/2} \hat{SE}(\hat{Y}_F)] \leq E(Y_F|X_F) \leq [\hat{Y}_F + t_{\alpha/2} \hat{SE}(\hat{Y}_F)]\} = 1 - \alpha$
 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、被解释变量个别值预测

基本思想：

- \hat{Y}_F 是对 Y 平均值的点预测。
- 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 Y_F 都有关的变量, 并且已知 e_F 服从正态分布, 且可证明 $E(e_F) = 0$

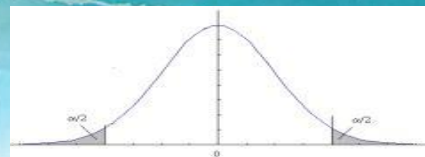
$$Var(e_F) = E(Y_F - \hat{Y}_F)^2 = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_F - \bar{X})^2}{\sum x_i^2} \right]$$

(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

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

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

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

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

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

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

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

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

预测区间

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

被解释变量Y区间预测的特点(续)

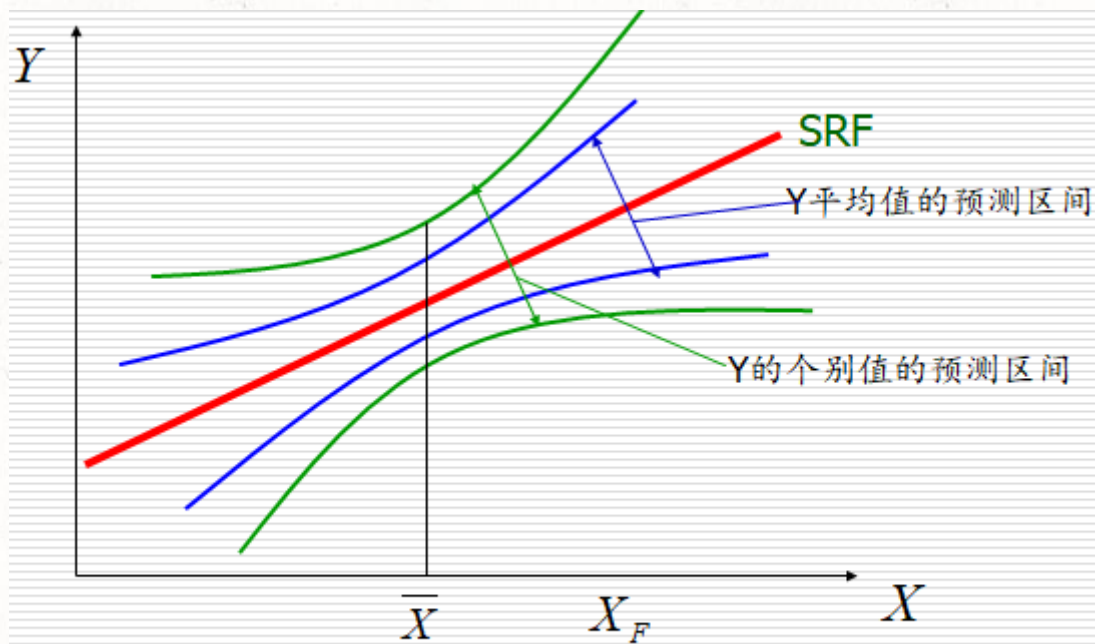
(2) 平均值和个别值预测区间都不是常数，是随 X_F 的变化而变化的，当 $X_F = \bar{X}$ 时，预测区间最小。

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

预测区间

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

各种预测值的关系



当 $X_F = \bar{X}$ 时，预测区间最小

第六节 案例分析

案例1:中国各地区居民交通通信消费支出和人均地区生产总值关系的分析

提出问题: 深入研究中国各地区居民人均交通通信消费与经济发展水平的数量关系, 对于探求居民交通通信消费增长的规律性, 分析各地区居民交通通信消费的差异, 认识地区发展不平衡不充分的影响程度, 预测其发展趋势, 通过发展经济减小地区差异, 预测各地区居民交通通信消费增长, 合理规划交通和信息化产业的发展, 都有重要的意义。

模型设定

为了分析各地区居民交通通信消费与经济发展水平的关系，选择“居民交通通信消费支出”（单位：元）为被解释变量（用 Y 表示）；选择“人均地区生产总值”（单位：元）为解释变量（用 X 表示）。表2.5为由《中国统计年鉴2017》得到的“分地区居民人均消费支出（2016年）”中的31个省市的“交通通信消费支出”和“人均地区生产总值”的数据。

（数据见教材P48-49）

模型设定

1. 建立工作文件

首先，双击EViews图标，进入EViews主页。直接点击下面对话框中的“Create a new EViews workfile”，

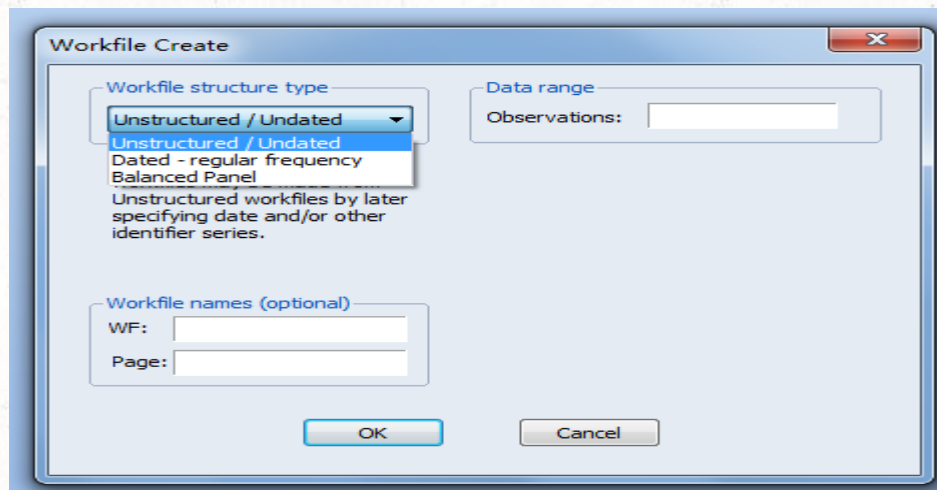
或者，关闭该对话框，
直接在最上面的菜单栏
中依此点击File\
New\Workfile



模型设定

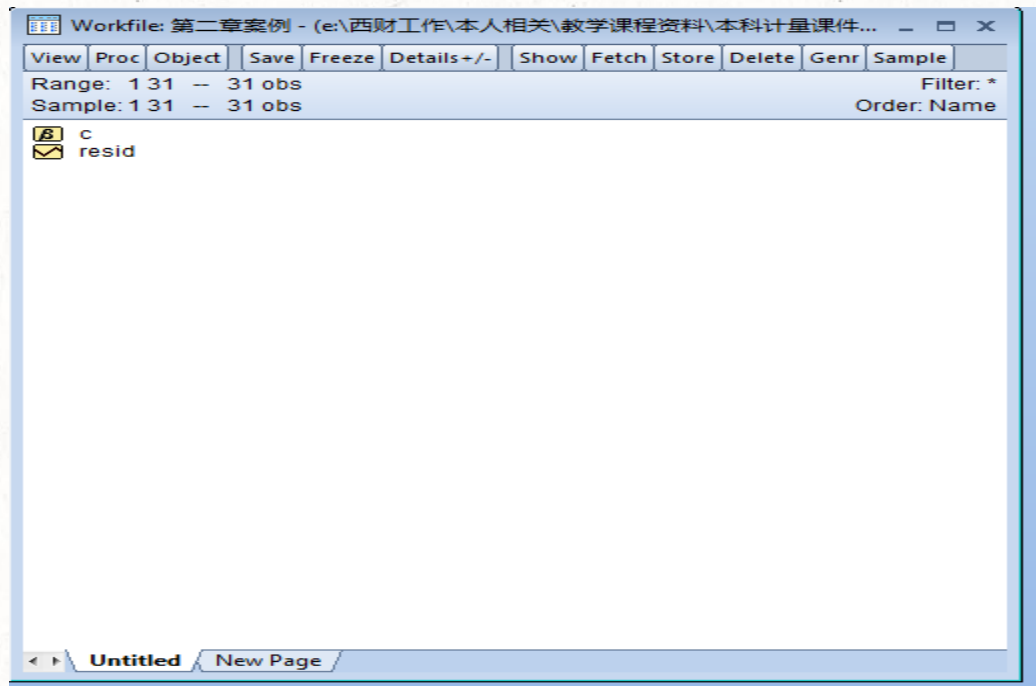
本案例数据为横截面数据，在出现的“Workfile Create”对话框里选择“Unstructured/Undated”类型，然后在右边的“Observations”一栏填入样本容量31。

注意：此处与书上做法不同，“Dated-regular frequency”的结构类型常用于时间序列数据。



模型设定

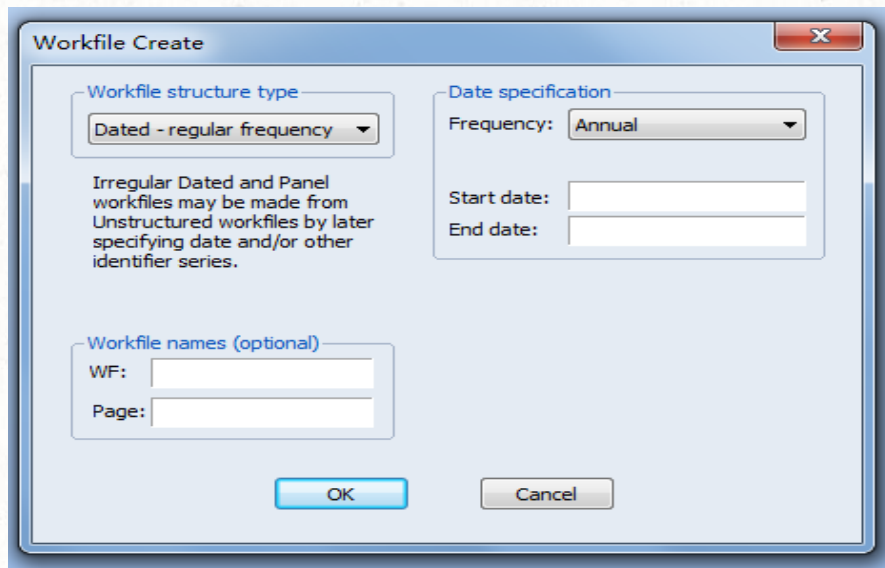
在“Workfile name”中输入文件名称，点击“ok”，出现如有的工作框。其中已有对象：“c”为截距项，“resid”为剩余项。若要将工作文件存盘，点击窗口上方的“Save”。



模型设定

对于时序数据，选择“Dated-regular frequency”后还可以在在“date specification”中选择数据频率：例如，multi-year（多年度）、Annual（年度）、Semi annual（半年）、Quartrly（季度）、Monthly（月度）、Weekly（周数据）、.....

注：第三种“Balanced Panel”类型可用于面板数据。



模型设定

2. 输入数据

在“Quick”菜单中点击“Empty Group”，出现数据编辑窗口。将第一列数据命名为Y：方法是按上行键“↑”，对应“obs”格自动上跳，在对应的第二行有边框的“obs”空格中输入变量名“Y”，再按下行键“↓”，变量名以下各格出现“NA”，依顺序输入Y的对应数据。按同样方法，可对“X”等其他变量命名，并输入对应数据。

模型设定

2. 输入数据

也可以在EViews命令框中直接输入“data Y X”(一元时，多元时类似)，回车出现“Group”窗口数据编辑框，在对应的Y、X下输入数据。还可以从Excel、Word等文档的数据表中直接将对应数据粘贴到EViews的数据表中。若要对数据存盘，点击“File/Save”。

还可以依次单击“Object/New Object”，然后选择“Group”类型，并命名。

模型设定

EViews

File Edit Object View Proc Quick Options Add-ins Window Help

DATA Y X

Workfile: 第二章案例 - (e:\西财工作\本人相关\教学课程资料\本科计量课件...

View Proc Object

Range: 1 31
Sample: 1 31

☒ c
☒ resid
☒ x
☒ y

Group: UNTITLED Workfile: 第二章案例::Untitled\

View Proc Object Print Name Freeze Default Sort Edit+/- Smpl+/- Compare+/-

| | Y | X |
|----|----|----|
| 1 | NA | NA |
| 2 | NA | NA |
| 3 | NA | NA |
| 4 | NA | NA |
| 5 | NA | NA |
| 6 | NA | NA |
| 7 | NA | NA |
| 8 | NA | NA |
| 9 | NA | NA |
| 10 | NA | NA |
| 11 | NA | NA |
| 12 | NA | NA |
| 13 | NA | NA |
| 14 | NA | NA |
| 15 | NA | NA |
| 16 | NA | NA |
| 17 | NA | NA |
| 18 | NA | NA |
| 19 | NA | NA |
| 20 | NA | NA |
| 21 | NA | NA |
| 22 | NA | NA |
| 23 | | |

Untitled New Page

模型设定

File Edit Object View Proc Quick Options Add-ins Window Help

DATA Y X

Workfile: 第二章案例 - (e:\西财工作\本人相关\教学课程资料\本科计量课件... - □ X

View Proc Object **G** Group: UNTITLED Workfile: 第二章案例::Untitled\ - □ X

Range: 1 31
Sample: 1 31

☒ c
☒ resid
☒ x
☒ y

| | Y | X |
|----|----------|----------|
| 1 | 4701.700 | 118198.0 |
| 2 | 3752.200 | 115053.0 |
| 3 | 2062.200 | 43062.00 |
| 4 | 1709.000 | 35532.00 |
| 5 | 2525.500 | 72064.00 |
| 6 | 2837.300 | 50791.00 |
| 7 | 2073.000 | 53868.00 |
| 8 | 2040.900 | 40432.00 |
| 9 | 4228.500 | 116562.0 |
| 10 | 3372.200 | 96887.00 |
| 11 | 4377.300 | 84916.00 |
| 12 | 1975.200 | 39561.00 |
| 13 | 2504.200 | 74707.00 |
| 14 | 1576.900 | 40400.00 |
| 15 | 2324.800 | 68733.00 |
| 16 | 1550.800 | 42575.00 |
| 17 | 1931.000 | 55665.00 |
| 18 | 1915.500 | 46382.00 |
| 19 | 3296.500 | 74016.00 |
| 20 | 1541.600 | 38027.00 |
| 21 | 1762.200 | 44347.00 |
| 22 | 1941.600 | 58502.00 |
| 23 | | |

Untitled New Page

关闭“group”时可以选择命名并保存。

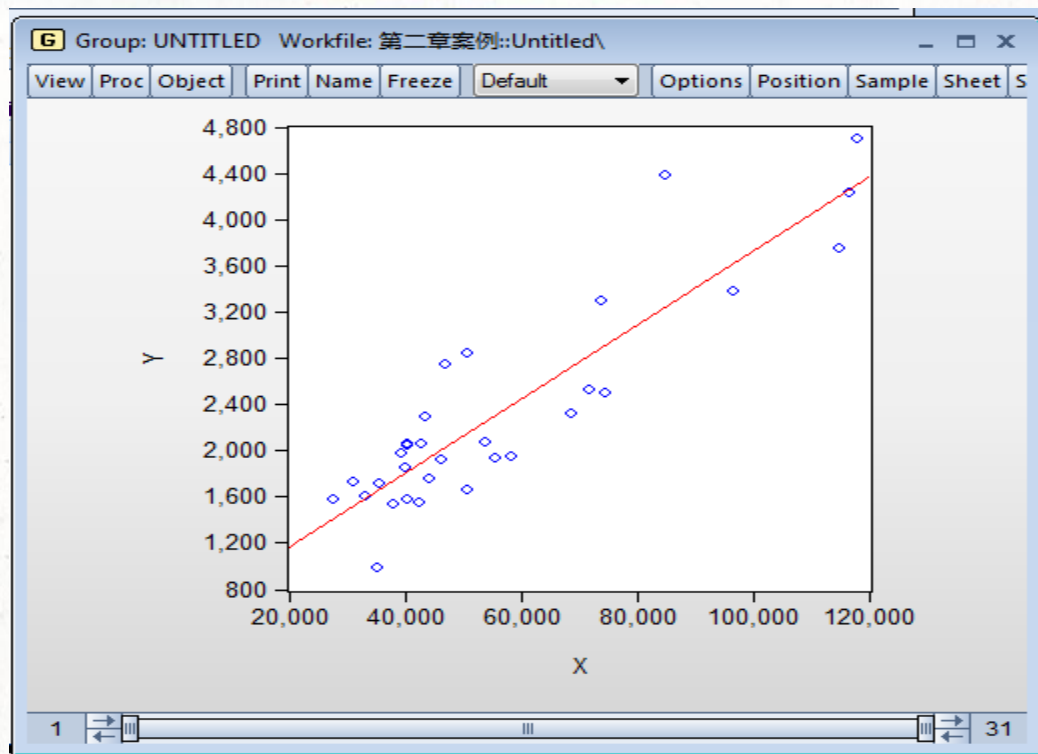
注意工作文件又出现了“x”和“y”两个序列。

模型设定

3.作Y与X的相关图形

为了初步分析居民交通通信消费支出 (Y)与人均地区生产总值(X)的关系, 可以作以X为横坐标, 以Y为纵坐标的散点图。方法是按住“ctrl”键依次选择工作文件中的对象X和Y（注：先选出现在横轴的变量），双击“open group”，点“View/Graph”在“Graph type”中选“scatter”，可在“Details”的“Fit lines”中选择“Regression line”，点“ok”，得到如下的带回归线的散点图：

模型设定



注：也可以直接在命令窗输入“scat X Y”（一样地，出现在横轴的变量先写），只是这种方式下得到的散点图没有回归线。

模型设定

从散点图可以看出各地区居民交通通信消费支出随着人均地区生产总值水平的提高而增加，近似于线性关系，为分析中国各地区居民居民交通通信消费支出随人均地区生产总值变动的数量规律性，可以考虑建立如下简单线性回归模型：

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

注：横截面数据下标惯用*i*而时间序列数据惯用*t*。

估计参数

假定所建模型及其中的随机扰动项 u 满足各项古典假定，可以用OLS法估计其参数。

方法一：在EViews工作框主页点击“Quick”菜单，点击“Estimate Equation”，出现“Equation specification”对话框，在“Method”中选“Least Squares”，在“Equation specification”对话框中键入“Y C X”，点“ok”或按回车，即出现回归结果。

方法二：在EViews命令框中直接键入“LS Y C X”，按回车，即出现回归结果。

估计参数

| Equation: UNTITLED Workfile: 第二章案例::Untitled\ | | | | |
|------------------------------------------------------------------------------------------------------------------------------|-------------|-----------------------|-------------|--------|
| View | Proc | Object | Print | Name |
| Freeze | Estimate | Forecast | Stats | Resids |
| Dependent Variable: Y Method: Least Squares Date: 09/07/19 Time: 16:43 Sample: 1 31 Included observations: 31 | | | | |
| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
| C | 521.5179 | 182.5074 | 2.857517 | 0.0078 |
| X | 0.032012 | 0.002937 | 10.90108 | 0.0000 |
| R-squared | 0.803834 | Mean dependent var | 2338.697 | |
| Adjusted R-squared | 0.797069 | S.D. dependent var | 918.3677 | |
| S.E. of regression | 413.7048 | Akaike info criterion | 14.95052 | |
| Sum squared resid | 4963399. | Schwarz criterion | 15.04304 | |
| Log likelihood | -229.7331 | Hannan-Quinn criter. | 14.98068 | |
| F-statistic | 118.8336 | Durbin-Watson stat | 1.840245 | |
| Prob(F-statistic) | 0.000000 | | | |

可用规范的形式将参数估计和检验的结果写为：

$$\hat{Y}_i = 521.5179 + 0.032012X_i$$

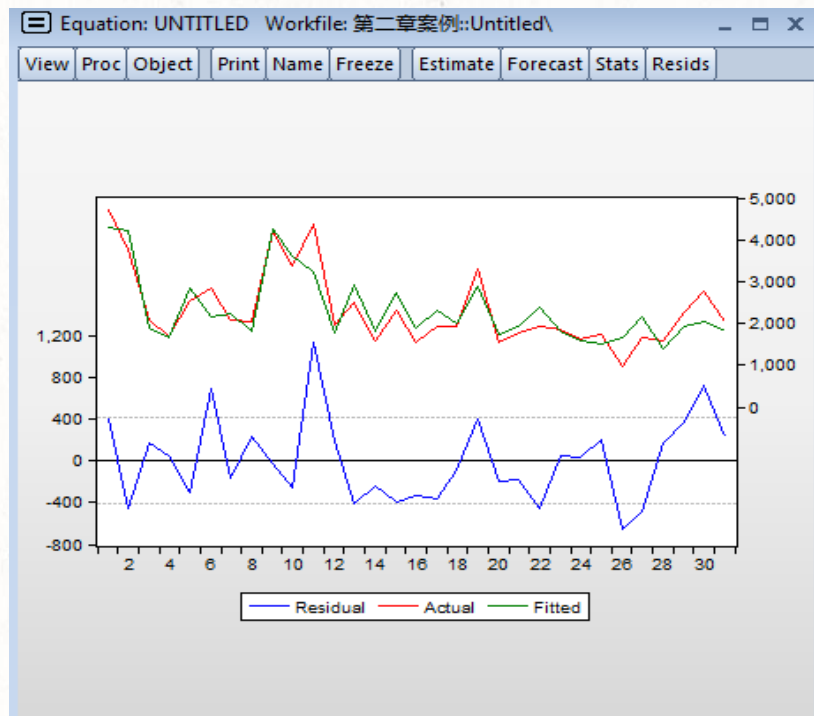
$$(182.5074) \quad (0.002937)$$

$$t = (2.857517) \quad (10.90108)$$

$$R^2 = 0.803834 \quad F = 118.8336 \quad n = 31$$

估计参数

若要显示回归结果的图形，在“Equation”框中，点击“Resids”，即出现剩余项（Residual）、实际值（Actual）、拟合值（Fitted）的图形：



模型检验

1、经济意义检验

所估计的参数 $\hat{\beta}_1 = 521.5179$, $\hat{\beta}_2 = 0.032012$, 说明人均地区生产总值每增加 1 元, 平均说来居民交通通信消费支出将增加 0.032012 元, 这与预期的经济意义相符。

2、拟合优度和统计检验

用 EViews 得出回归模型参数估计结果的同时, 已经给出了用于模型检验的相关数据。

拟合优度的度量: 本例中可决系数为 0.803834, 说明所建模型整体上对样本数据拟合较好, 即解释变量 “人均地区生产总值” 对被解释变量 “居民交通通信消费支出” 的绝大部分差异作出了解释。

模型检验

系数显著性检验：

a. 传统的t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 $t = 10.90108 > t_{0.025}(29) = 2.045$

说明“城镇人均可支配收入”对“城镇人均消费支出”确有显著影响。

b. 用P值检验 $\alpha = 0.05 >> p = 0.0000$

回归预测

点预测:

中国西部地区除了重庆市以外，2016年各省的人均地区生产总值都还低于全国人均国内生产总值水平，如果西部地区某省的人均地区生产总值能达到2016年全国人均国内生产总值54000元/人的水平，利用所估计的模型可预测该省居民交通通信消费支出可能达到的水平，点预测值的计算方法为：

$$\hat{Y}_f = 521.5179 + 0.032012 \times 54000 = 2250.145$$

回归预测

点预测软件操作方法：

利用 EViews 作回归预测，首先在“Workfile”窗口点击“Range”，将“End data”由“31”改为“32”，点“OK”，将“Workfile”中的“Range”扩展为 1—32。在“Workfile”窗口点击“sample”，将“sample”窗口中的“1 31”改为“1 32”，点“OK”，从而将样本区间改为 1—32。

为了输入 $X_f = 54000$ ，在 EViews 命令框键入 data x /回车，在 X 数据表中的“32”位置输入“54000”，将数据表最小化。然后在“Equation”框中，点击“Forecast”，打开对话框。在“Forecast name”(预测值序列名)中键入“ Y_f ”，回车即得到模型估计值及标准误差的图形。双击“Workfile”窗口中出现的“ Y_f ”，在“ Y_f ”数据表中的“32”位置出现 $Y_f = 2250.145$ ，这是当 $X_f = 54000$ 时居民交通通信消费支出的点预测值。

回归预测

区间预测：

平均值区间预测上下限：

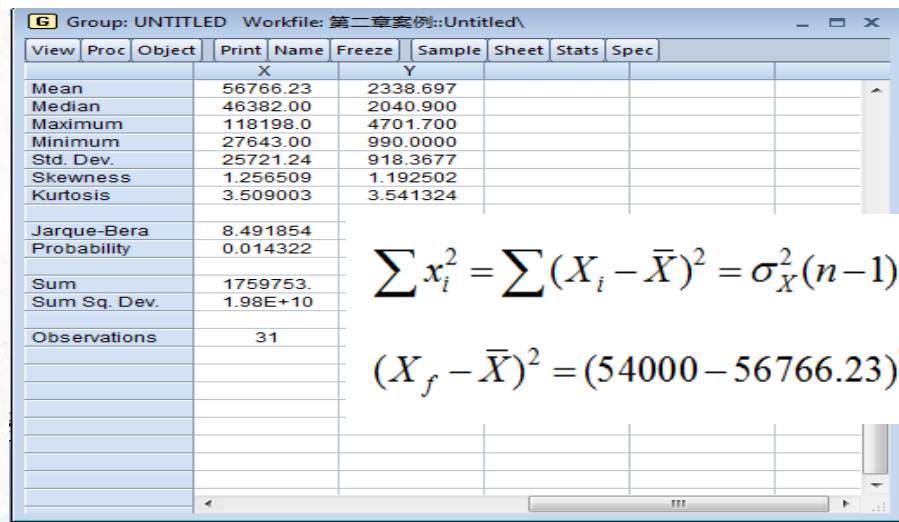
$$Y_F = \hat{Y}_F \mp t_{\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_F - \bar{X})^2}{\sum x_i^2}}$$

个别值区间预测上下限：

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

回归预测

为获得相关数据，在用 EViews 作回归分析中，已经得到 $Y_f = 2250.145$ 、 $t_{0.025}(29) = 2.045$ 、 $\hat{\sigma} = 413.7048$ 、 $n = 31$ 。在 X 和 Y 的数据表中，点击“View”选“Descriptive Stats\Comon Sample”，则得到 X 和 Y 的描述统计结果。



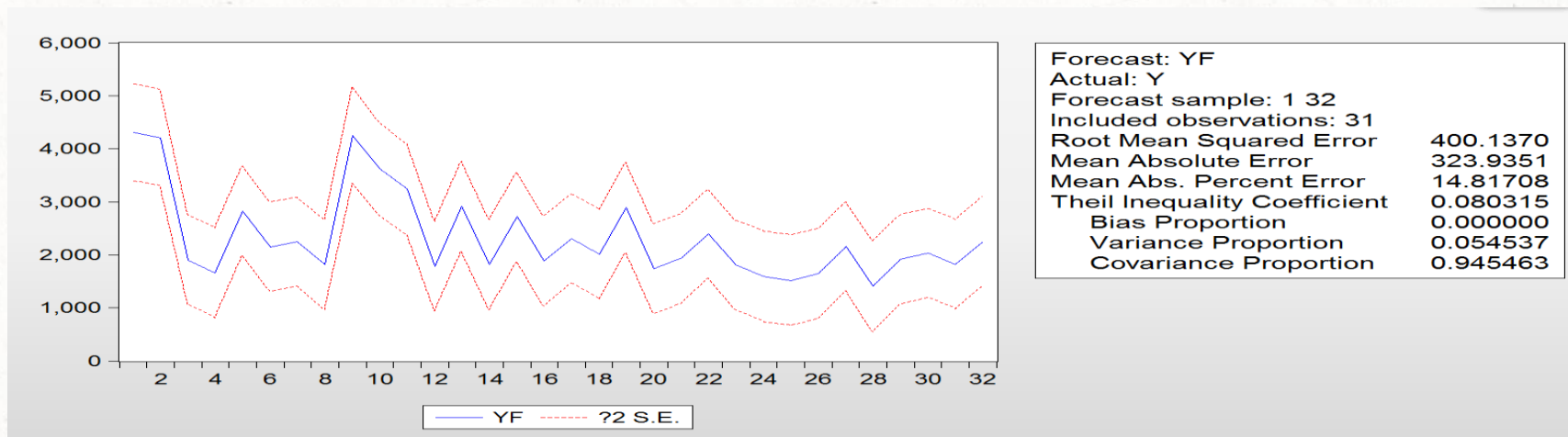
| | X | Y |
|--------------|----------|----------|
| Mean | 56766.23 | 2338.697 |
| Median | 46382.00 | 2040.900 |
| Maximum | 118198.0 | 4701.700 |
| Minimum | 27643.00 | 990.0000 |
| Std. Dev. | 25721.24 | 918.3677 |
| Skewness | 1.256509 | 1.192502 |
| Kurtosis | 3.509003 | 3.541324 |
| Jarque-Bera | 8.491854 | |
| Probability | 0.014322 | |
| Sum | 1759753. | |
| Sum Sq. Dev. | 1.98E+10 | |
| Observations | 31 | |

$$\sum x_i^2 = \sum (X_i - \bar{X})^2 = \sigma_X^2(n-1) = 25721.24^2 \times (31-1) = 19847465614.128$$

$$(X_f - \bar{X})^2 = (54000 - 56766.23)^2 = 7652028.4129$$

回归预测

在“Equation”框中，点击“Forecast”可得预测值及标准误差的图形：



拓展1: 线性回归模型的若干延伸

一、过原点的回归

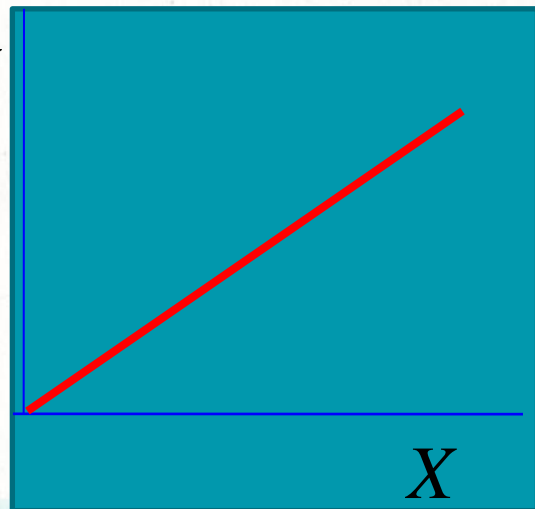
有时根据理论判断模型可能没有截距项(常数项), 例如:

- 弗瑞德曼永久收入假说: 永久消费正比于永久收入。

- 成本分析理论: 生产的可变成本正比于产出。

- 货币主义理论某些假说: 价格变化率(通货膨胀率)
正比于货币供给变化率。

这时总体回归函数可设定为: $Y_i = \beta_2 X_i + u_i$



过原点的回归的OLS估计量

没有截距项的过原点回归模型为: $Y_i = \beta_2 X_i + u_i$

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

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

(注意:正规方程只有一个方程)

令其为零得

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

对比有截距时:

可以证明

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

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-1}$$

$$\begin{aligned} \hat{\beta}_2 &= \frac{\sum x_i y_i}{\sum x_i^2} \\ \text{Var}(\hat{\beta}_2) &= \frac{\sigma^2}{\sum x_i^2} \\ \hat{\sigma}^2 &= \frac{\sum e_i^2}{n-2} \end{aligned}$$

注意:过原点回归的特点

在运用过原点回归模型时应注意以下特点:

1) 在有截距的模型中, 根据最小二乘原理的正规方程有:

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

但在截距项不存在时, 因为正规方程中只有一个方程,

而没有 $\frac{\partial(\sum e_i^2)}{\partial \hat{\beta}_1} = -2 \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$ 这样的关系,

则有可能 $\sum e_i \neq 0$ 从而 $\bar{e}_i \neq 0$

2) 回归线不通过样本均值

过原点回归模型

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

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

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

因为 $\bar{e}_i \neq 0$ 有

$$\bar{Y} \neq \hat{\beta}_2 \bar{X}$$

3) 估计值 \hat{Y}_i 的均值不等于实际观测值 Y_i 的均值

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

这说明过原点回归最小二乘法的数学性质不一定成立

4) 有时零均值假定 $E(u_i|X_i)=0$ 不一定满足

例如对于 $Y_i = \beta_1 + \beta_2 X_{2i} + u_i$

如果:

$$E(u_i|X_i) = \mu \neq 0$$

假如已知 μ , 对于有截距的模型, 此时模型可变换为

$$Y_i = (\beta_1 + \mu) + \beta_2 X_{2i} + (u_i - \mu)$$

令 $\varepsilon_i = u_i - \mu$ 则 $E(\varepsilon_i|X_i) = E[(u_i - \mu)|X_i] = 0$

可见, 有截距的模型可使得随机扰动具有零均值。而不含截距的模型

$Y_i = \beta_2 X_i + u_i$ 变换后成为有截距模型, 若坚持用无截距模型, 则随机项零均值不一定能保证。

5) 有截距模型中 $\sum e_i^2 = \sum y_i^2 - \hat{\beta}_2^2 \sum x_i^2 \leq \sum y_i^2$, R^2 总为正

值, 即 模型可决系数总是非负的。但对无截距的模型

$Y_i = \hat{\beta}_2 X_i + e_i$, 可决系数可能出现负值, 因此计算可

决系数的公式不一定适合于过原点的回归模型。

一般规则: 除非有充分的理由特别说明, 否则模型还是应当包含常数项为好。

截距项的讨论

在某些特定的例子中，截距项是有具体含义的。例如在分析销售收入 Y 与广告投入 X （单位：百万）间关系时，可建立如下回归模型：

$$Y = \beta_1 + \beta_2 X + u$$

截距项 $\beta_1 = E(Y|X=0)$ ，代表着：不做广告投入的情况下的基本销售收入。

而斜率 β_2 代表着：每增加一个单位（100万）的 X 的投入，在销售收入 Y 上能有多少增加？

截距项的讨论

- ▶ 在很多实际分析中， X 不可能为0（例如人的身高）。因此，截距项也就失去了被解读的意义，它的存在主要是理论意义和计算意义。因此在预测的时候，我们需要计算它，但大多数实证分析中不需要对它做具体解读。
- ▶ 一般并不关注截距项的经济含义，但无论截距项是否显著都应保留。
- ▶ 斜率系数才是整个回归的核心，关心三个问题：
 - 是否为0？
 - 如果不为0，是正是负？
 - 如果为正，大小如何？

二、变量度量单位对回归的影响

变量的度量单位对估计的参数数值会有什么影响？

例如美国1988年-1997年国内总投资(Y)与GDP的回归

(数据略): $Y_i = \beta_1 + \beta_2 X_i + u_i$

A.当总投资(Y)与GDP都以10亿美元为度量单位时，估计

结果为: $\hat{Y}_t = -1026.498 + 0.3016GDP$
 $SE = (257.5874) \quad (0.0399) \quad R^2 = 0.8772$

B.当总投资(Y)仍以10亿美元计，而GDP以百万美元计时

估计结果为: $\hat{Y}_t = -1026.498 + 0.0003016GDP$
 $SE = (257.5874) \quad (0.0000399) \quad R^2 = 0.8772$

C.当总投资(Y)与GDP都以百万美元（缩小1000倍)计时

估计结果为: $\hat{Y}_t = -1026498 + 0.3016GDP$

$$SE = (257587.4) \quad (0.0399) \quad R^2 = 0.8772$$

D.当总投资(Y)以百万美元计，而GDP以10亿美元计时，

估计结果为: $\hat{Y}_t = -1026498 + 301.58266GDP$

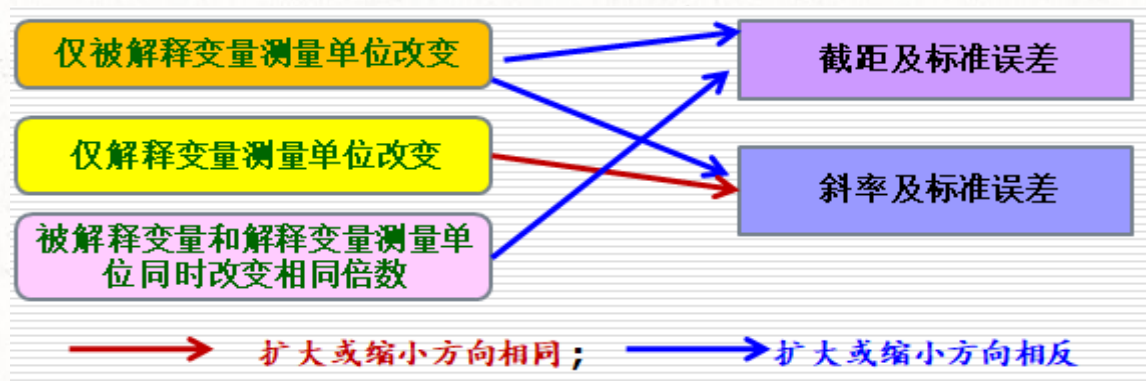
$$SE = (257587.4) \quad (39.899899) \quad R^2 = 0.8772$$

注意:与A相比较，截距、斜率系数、标准误差、可决系数的变化。

变量度量单位对回归影响的一般规律

1. 当被解释变量测量单位改变（扩大或缩小常数 c 倍），而解释变量测量单位不变时：OLS截距和斜率的估计值及标准误差都缩小或扩大为原来的 c 倍。（如D的情况）
2. 当解释变量测量单位改变（扩大或缩小常数 c 倍），而被解释变量测量单位不变时：OLS斜率的估计值及标准误差扩大或缩小为原来的 c 倍，但不影响截距的估计。（如B的情况）
3. 当被解释变量和解释变量测量单位同时改变相同倍数时，OLS的截距估计值及标准误差扩大为原来的 c 倍，但不影响斜率的估计。（如C的情况）

4. 当被解释变量和解释变量测量单位改变时，不会影响拟合优度.可决系数是纯数没有维度，所以不随计量单位而变化。（如B、C、D的情况）



当被解释变量和解释变量测量单位同时改变相同倍数时，对斜率系数的影响相互抵消了。

量纲变化的后果

- ▶ 改变解释变量与被解释变量的测量单位会导致回归系数大小的变化，但不会影响参数**OLS**估计量的显著性。
- ▶ 切勿在一个回归中轻易比较不同系数的大小，而应比较不同系数的显著性。
- ▶ 若要比​​较不同系数的大小，应首先将不同的解释变量统一到相同的量纲水平（如标准化变换），然后进行正式的统计检验。

拓展2: 简单线性回归模型的极大似然估计

一、极大似然估计的思想:

举例: 对一种药物, 药剂师认为有效率为70%。生产该药物的公司声称: 有效率为90%, 谁的说法更可信呢?

统计学家抽取10个病人, 发现有8人被治愈

●若真实概率为 $P=0.7$ 时: 产生“10个病人有8个治愈”

结果的概率为:(实验结果只有“治愈”和“未治愈”是二项分布)

$$\frac{10!}{8! \times 2!} \times 0.7^8 \times 0.3^2 = 0.2335$$

●若真实概率为 $P=0.9$ 时, 产生“10个病人有8个治愈”

结果的概率为:

$$\frac{10!}{8! \times 2!} \times 0.9^8 \times 0.1^2 = 0.1937$$

统计学家判断: 有效率为0.7作为真实有效率的估计值比0.9更为可信。
(为什么?)

极大似然原理：“一个事件由于与实际最近似而发生”

原理：一个事件之所以会发生，是因为存在着产生这一事件概率最大的客观现实（总体）。

总体的分布规律是由其**分布性质**和**参数**决定的。

样本观测值是从总体中抽取而得到的，从总体中随机抽取容量为 n 的样本观测值时，这 n 组样本观测值会以一定的概率出现。

当从总体中随机抽取 n 组样本观测值后，要寻求最可能产生该 n 组样本的那个总体的参数。

最合理的参数估计量应该是能够**使得从总体中抽取出该 n 组样本观测值的概率最大**。

二、简单线性回归模型的极大似然估计

在满足基本假设的条件下，对简单线性回归模型

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

若随机抽取n组样本观测值 (X_i, Y_i) ($i=1,2,\dots,n$)

Y_i 为随机变量，其分布特征与参数 β_1 和 β_2 及 σ^2 有关，

$$\text{已知 } E(Y_i | X_i) = \beta_1 + \beta_2 X_i \quad \text{Var}(Y_i | X_i) = \sigma^2$$

假定 Y_i 服从正态分布且是独立分布的，则：

$$Y_i \sim N(\beta_1 + \beta_2 X_i, \sigma^2)$$

于是，每个 Y_i 的概率密度函数为 $P(Y_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i - \beta_1 - \beta_2 X_i)^2}$ ($i=1,2,\dots,n$)

1. 似然函数 (likelihood function)

因为各个 Y_i 相互独立，因此获得所有 n 组样本观测值

的联合概率(即似然函数)为:

$$L(\beta_1, \beta_2, \sigma^2) = P(Y_1, Y_2, \dots, Y_n) = \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} e^{-\frac{1}{2\sigma^2} \sum (Y_i - \beta_1 - \beta_2 X_i)^2}$$

(n个密度函数的乘积)

其中未知参数为 $\beta_1, \beta_2, \sigma^2$ ，为使产生 n 个样本观测值的联合概率最大，可寻求能使该似然函数极大化的参数值，即可求得模型参数的极大似然估计量。为便于取最大化，**取对数**似然函数，因为似然函数的极大化与似然函数的对数的极大化是等价的，所以，

$$L^* = \ln(L) = -n \ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum (Y_i - \beta_1 - \beta_2 X_i)^2$$

产生n组样本观测值的联合概率的对数（对数似然函数）为：

$$L^* = \ln(L) = -n \ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$
$$= -\frac{n}{2} \ln \sigma^2 - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

$$\frac{\partial L^*}{\partial \beta_1} = -\frac{1}{\sigma^2} \sum (Y_i - \beta_1 - \beta_2 X_i)(-1)$$

$$\frac{\partial L^*}{\partial \beta_2} = -\frac{1}{\sigma^2} \sum (Y_i - \beta_1 - \beta_2 X_i)(-X_i)$$

$$\frac{\partial L^*}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (Y_i - \beta_1 - \beta_2 X_i)^2$$

令各方程为0，记参数估计量为 $\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\sigma}^2$ 可得：

注意到：

使 L^* 最大化
等价于使

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

最小化

$$\frac{1}{\sigma^2} \sum (Y_i - \beta_1 - \beta_2 X_i) = 0 \quad (\text{A})$$

$$\frac{1}{\sigma^2} \sum (Y_i - \beta_1 - \beta_2 X_i) X_i = 0 \quad (\text{B})$$

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (Y_i - \beta_1 - \beta_2 X_i)^2 = 0 \quad (\text{C})$$

经简化，由 (A) (B) 式有：

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

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

这与OLS正规方程相同

2. 简单线性回归模型的极大似然估计量

对 L^* 求极大值，等价于对 $\sum (Y_i - \beta_1 - \beta_2 X_i)^2$ 求极小值：

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

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

解方程得参数估计量：

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

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

可见，在满足基本假设的情况下，模型参数的最大似然估计量与普通最小二乘估计量是相同的。

3. σ^2 的极大似然估计(ML)

把参数估计量 $\tilde{\beta}_1, \tilde{\beta}_2$ 代入(C)式并简化, 得 σ^2 的极大似然计:

因为

$$\frac{1}{2\sigma^4} \sum (Y_i - \beta_1 - \beta_2 X_i)^2 = \frac{n}{2\sigma^2}$$
$$\tilde{\sigma}^2 = \frac{1}{n} \sum (Y_i - \tilde{\beta}_1 - \tilde{\beta}_2 X_i)^2 = \frac{1}{n} \sum e_i^2$$

所以 σ^2 的ML估计为:

对比 σ^2 的OLS估计:

$$\hat{\sigma}^2 = \sum e_i^2 / (n-2)$$

在OLS无偏性证明中有

$$E(\sum e_i^2) = (n-2)\sigma^2$$

所以

$$E(\tilde{\sigma}^2) = \frac{1}{n} E(\sum e_i^2) = \left(\frac{n-2}{n}\right)\sigma^2 = \sigma^2 - \frac{2}{n}\sigma^2$$

结论: σ^2 的极大似然估计 (ML) 是有偏的。其偏误因子 $\frac{2}{n}\sigma^2$

是随 $n \rightarrow \infty$ 而趋于0的, 因此 σ^2 的ML估计只是一致估计量。

4. 极大似然估计与最小二乘估计的比较

1. 在满足基本假设的情况下，模型参数的极大似然估计量与普通最小二乘估计量是相同的。

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

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

2. σ^2 的普通最小二乘估计是无偏估计。

σ^2 的极大似然估计 (ML) 是有偏的。

但 $E(\tilde{\sigma}^2) = \sigma^2 - \frac{2}{n}\sigma^2$ 随 $n \rightarrow \infty$, $\tilde{\sigma}^2$ 是渐近无偏的, 即

$$\lim_{n \rightarrow \infty} E(\tilde{\sigma}^2) = \sigma^2$$

本章作业

本科教材练习题2.2（P56-57，电子版，要求附Eviews输出结果）