2-2

总体回归函数与样本回归函数

主讲人：范国斌

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。

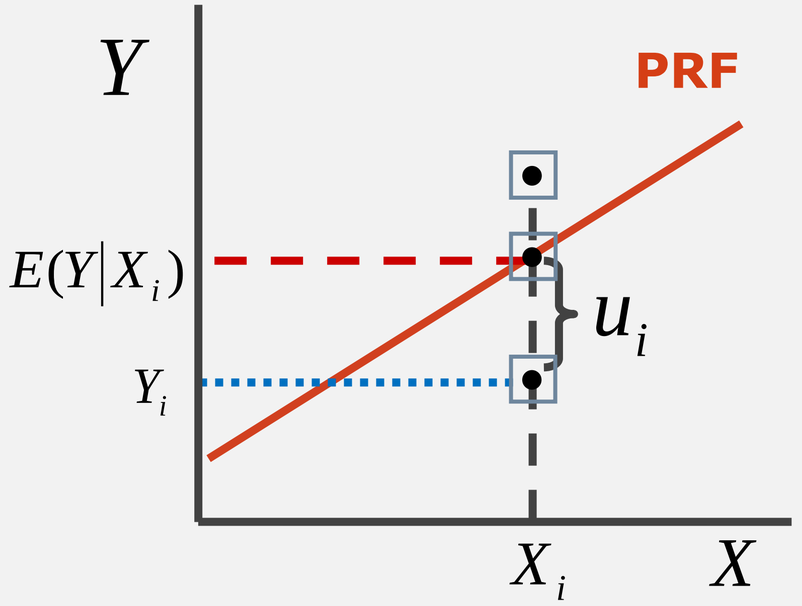
计量经济学的根本目的是要探寻变量间数量关系的规律,也就是要去寻求总体回归函数。

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

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

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”

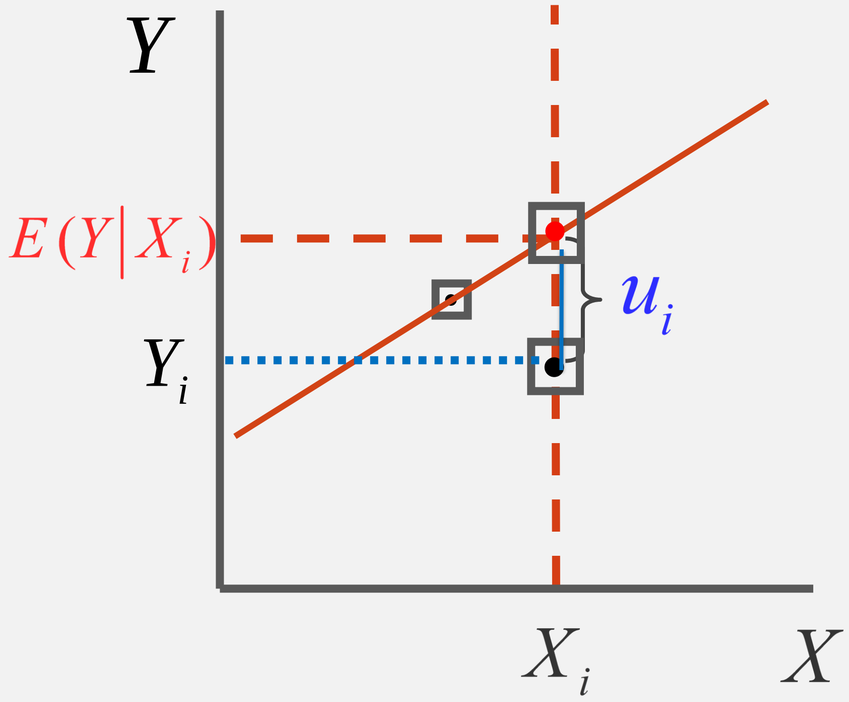
$E(Y_i|X_i)=\beta_1+\beta_2X_i^2$ 对参数“线性”，对变量“非线性”

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。

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)

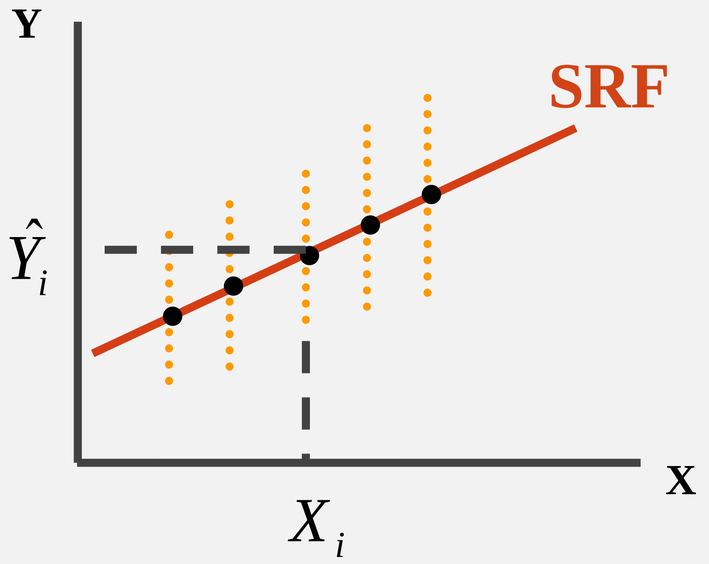
样本回归函数（SRF）

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）。



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：

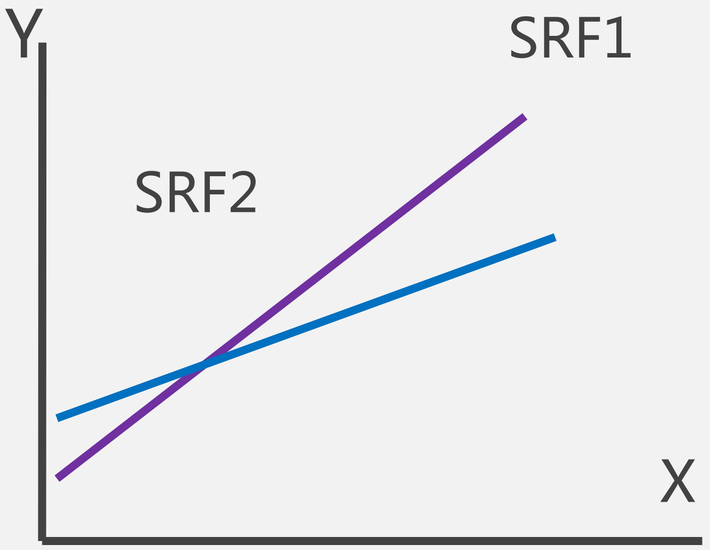
则 $e_i = Y_i - \hat{Y}_i$ 或 $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$

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。

对样本回归的理解

对比：

总体回归函数	样本回归函数
$E(Y_i \mid X_i) = \beta_1 + \beta_2 X_i$	$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$
$Y_i = \beta_1 + \beta_2 X_i + u_i$	$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$

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 $E(Y_i \mid X_i)$ 的估计
- e_i 在概念上类似总体回归函数中的 u_i，可视为对 u_i 的估计。

多元线性回归模型

1、多元线性回归模型的意义

一般形式：对于有K-1个解释变量的线性回归模型

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i$$
$$(i = 1, 2, \cdots n)$$

注意：模型中的 β_j（j=2,3,---k）是**偏回归系数**

样本容量为n

偏回归系数：

控制其它解释量不变的条件下，第 j 个解释变量的单位变动对被解释变量平均值的影响，即对Y平均值的“直接”或“净”影响。

多元总体回归函数

条件期望表现形式：

将Y的总体条件期望表示为多个解释变量的函数，如:

$$E(Y_i|X_{2i},X_{3i},\cdots X_{ki})=\beta_1+\beta_2X_{2i}+\beta_3X_{3i}+\cdots+\beta_kX_{ki}$$

$$(i=1,2,\cdots n)$$

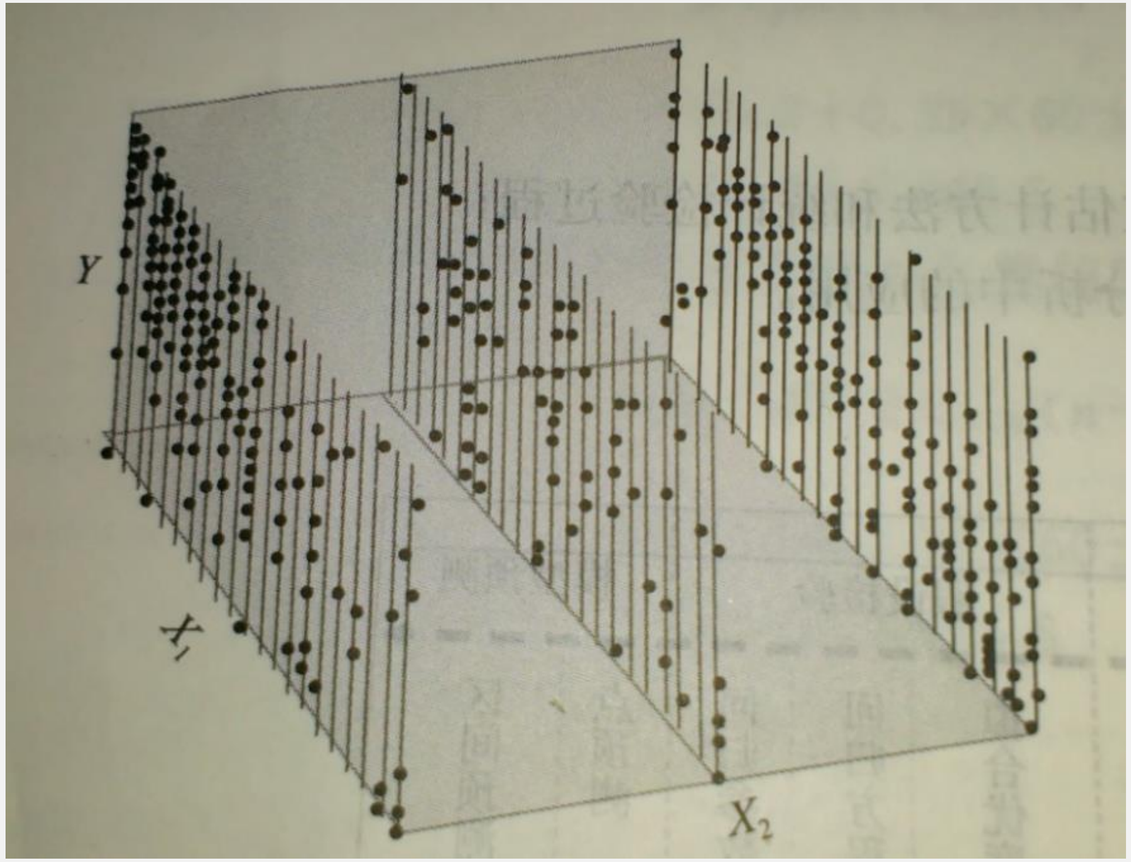
注意：这时Y总体条件期望的轨迹是K维空间的一条线

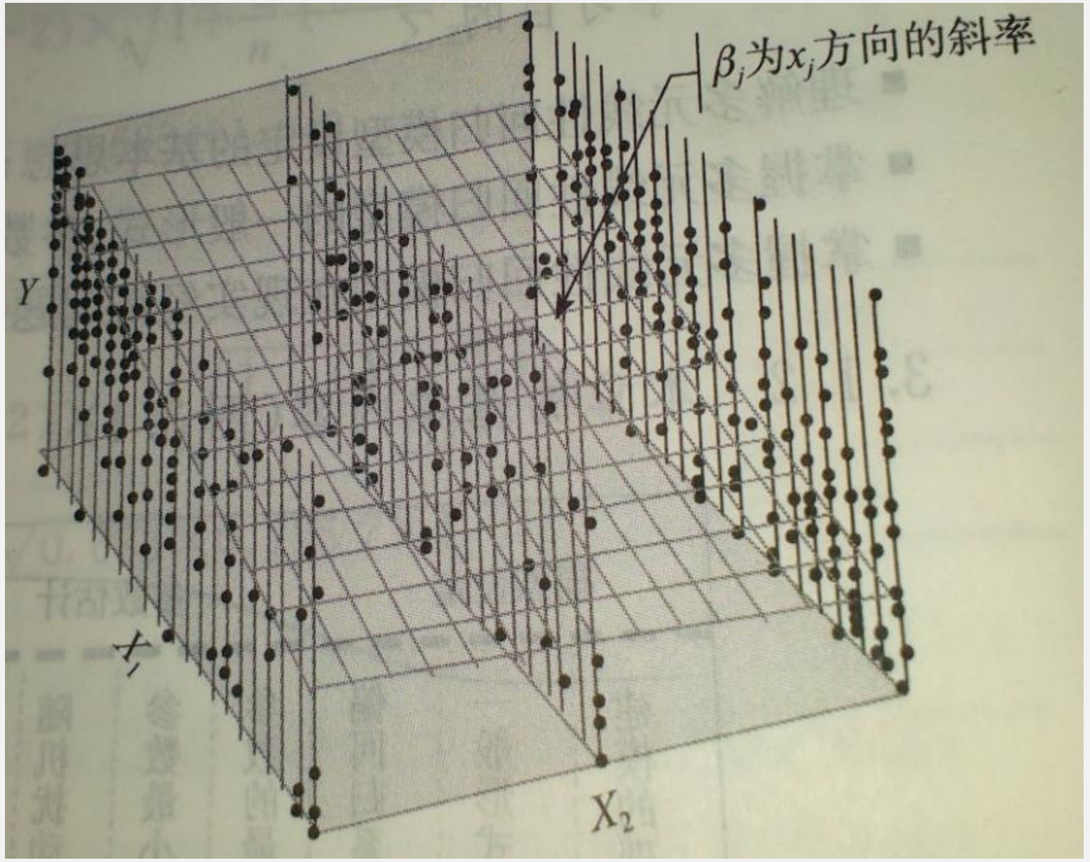
个别值表现形式：

引入随机扰动项 $u_i=Y_i-E(Y_i|X_{2i},X_{3i}\cdots X_{ki})$

或表示为 $Y_i=\beta_1+\beta_2X_{2i}+\beta_3X_{3i}+\cdots+\beta_kX_{ki}+u_i \quad (i=1,2,\cdots n)$

直观理解（供参考）





多元样本回归函数

Y 的样本条件均值可表示为多个解释变量的函数

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki}$$

或回归剩余（残差）： $e_i = Y_i - \hat{Y}_i$

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki} + e_i$$

其中 $i = 1, 2, \cdots n$

多元线性回归模型的矩阵表示

多个解释变量的多元线性回归模型的n组样本观测值，可表示为

$$Y_1 = \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \cdots + \beta_k X_{k1} + u_1$$
$$Y_2 = \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \cdots + \beta_k X_{k2} + u_2$$

$$Y_n = \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + \cdots + \beta_k X_{kn} + u_n$$

用矩阵表示

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & \cdots & X_{k1} \\ 1 & X_{22} & \cdots & X_{k2} \\ \vdots & \cdots & \cdots & \vdots \\ 1 & X_{2n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$$\underset{n\times 1}{\boldsymbol{Y}} \qquad \underset{n\times k}{\boldsymbol{X}} \qquad \underset{k\times 1}{\boldsymbol{\beta}} \qquad \underset{n\times 1}{\boldsymbol{u}}$$

矩阵表示方式

总体回归函数 $E(Y)=X\beta$ 或 $Y=X\beta+u$

样本回归函数 $\hat{Y}=X\hat{\beta}$ 或 $Y=X\hat{\beta}+e$

其中：$Y,\hat{Y},u,e$ 都是有 n 个元素的列向量

$\beta,\hat{\beta}$ 是有 k 个元素的列向量

（ k = 解释变量个数 + 1 ）

X是第一列为1的 n×k 阶解释变量数据矩阵 。

(截距项可视为解释变量总是取值为1)

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 的参数估计尽可能“接近”总体回归函数中的参数的真实值。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。