# 计量经济学基础

## 第二章　简单线性回归模型

# 主要内容

- 一、模型的建立及其假定条件
- 二、简单线性回归模型的参数估计
- 三、简单线性回归模型的统计检验
- 四、回归预测

# 第一节 模型的建立及其假定条件

1、变量间的关系及回归分析的基本概念

2、简单线性回归模型

3、随机误差项的假定条件

# 1、变量间的关系及回归分析的基本概念

## ■1、变量间的关系

确定性关系或函数关系：如圆面积$S = \pi r^2$

特点：给定自变量的一个值，因变量有惟一的值与其对应。

统计依赖关系或相关关系：

如收入与消费支出之间的关系：$C_i = \alpha + \beta Y_i + u_i$

特点：给定自变量的一个值，因变量可能有多个的值与其对应。

# 变量间的关系

1、如果变量之间存在确定性的函数关系，应使用函数分析工具进行分析，绝对不能使用计量分析工具。

2、统计依赖关系（相关关系）的分析工具主要包括相关分析与回归分析

(1)相关分析用于揭示随机变量之间线性相关关系的方向和密切程度，但不能揭示变量间的数量依存关系。相关分析不要求变量必须具有因果联系。

(2)回归分析用于分析变量间的数量依存关系，同时也能揭示变量间相关关系的方向，但不能揭示相关关系的密切程度。回归分析要求变量间必须具有因果联系。即变量间存在统计依赖关系且具有因果联系时才能使用回归分析工具。

(3)因果关系是计量分析的核心，但计量分析本身并不具备确定因果关系的功能，因果关系只能由（经济）理论确定。

# 回归分析的基本概念

- “回归(regression )”一词最早由英国生理学家高尔顿（Galton,1886）提出，用以指儿女的身高有回复到同龄人口总体平均身高的趋势
- 回归分析(regression analysis)是研究一个变量关于另一个（些）变量的具体依赖关系的计算方法和理论。其目的在于通过后者的已知或设定值，去估计和（或）预测前者的（总体）均值（Gujarati，1995）
- 被解释变量（Explained Variable）或因/应变量、左端变量（Dependent Variable）
- 解释变量（Explanatory Variable）或自变量、右端变量（Independent Variable）

# 回归分析的基本概念

- 回归分析构成计量经济学的方法论基础。其主要内容包括:
  - 模型建立
  - 参数估计
  - 模型检验：计量经济检验、统计推断检验、经济意义检验、预测检验
  - 模型应用：结构分析、政策评价、经济预测、检验和发展经济理论

# 2、简单线性回归模型

假定两变量间的统计依赖关系如下：

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

上式表示变量$Y_i$和$X_i$之间的真实数量关系。其中：

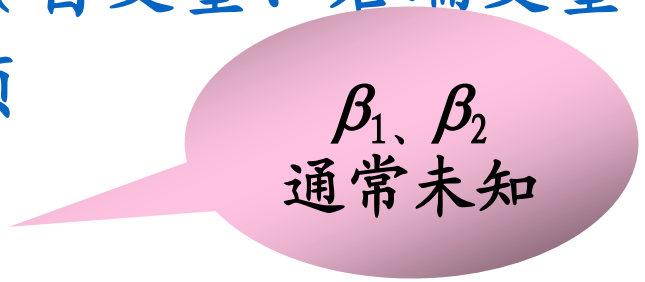
$Y$称被解释变量（因/应变量、左端变量）

$X$称解释变量（自变量、右端变量）

$u$称随机误差项

$\beta_1$称常数项

$\beta_2$称回归系数




称模型 $Y_i = \beta_1 + \beta_2 X_i + u_i$ 为简单线性回归模型

# （一）线性的含义

## ➢变量线性

因变量的条件均值是自变量的线性函数。在这种情况下，解释变量的单位变动所引起的因变量的变化率为一常数，即斜率为常数。

## ➢参数线性

因变量的条件均值是参数（自变量前的回归系数）的线性函数，而变量之间并不一定是线性的。也就是说这个参数是一次方的。

# “线性”的判断

$Y_i = \beta_1 + \beta_2 X_i + u_i$　　变量、参数均为“线性”

$Y_i = \beta_1 + \beta_2 X_i^2 + u_i$　　参数“线性”，变量”非线性”

$Y_i = \beta_1 + \sqrt{\beta_2} X_i + u_i$　　变量“线性”，参数”非线性”

计量经济学中：

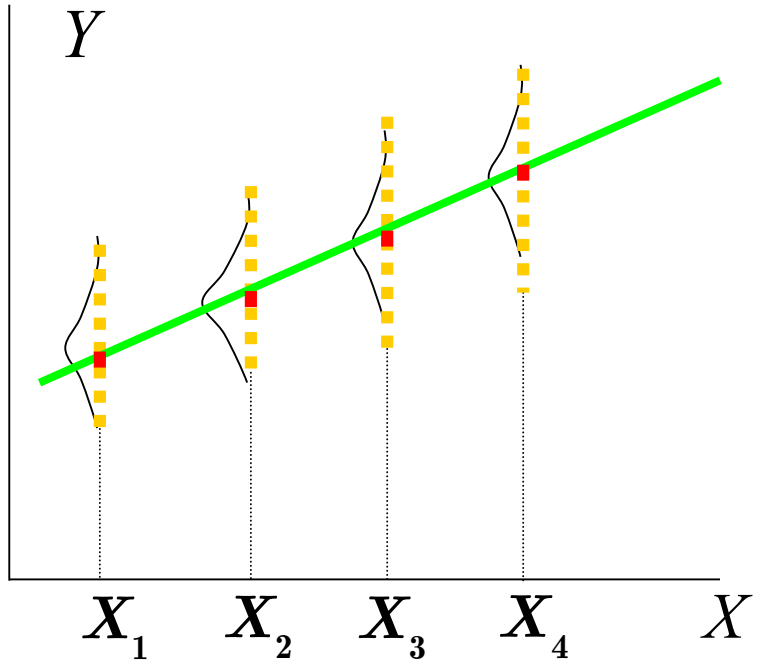
线性回归模型主要指就参数而言是“线性”，因为只要对参数而言是线性的，都可以用类似的方法估计其参数。

# （二）几个概念：

## 1、条件分布和条件期望（均值）

$Y$的条件分布：

当解释变量$X$取某个固定值$X_i$时（条件），$Y$的不同取值所形成的分布称之为$Y$的条件分布，记为$F\left(Y \mid X = X_i\right)$或$f\left(Y \mid X = X_i\right)$。有时也进一步简记为$F\left(Y \mid X_i\right)$或$f\left(Y \mid X_i\right)$




$Y$的条件均值：

对于$X$的某个取值$X_i$，根据其对应的条件分布$f\left(Y \mid X_i\right)$，可确定其条件期望或（条件均值）$E\left(Y \mid X = x_i\right) = \int_\Omega y f\left(Y \mid X_i\right) dy$，通常简记为$E\left(Y \mid X_i\right)$

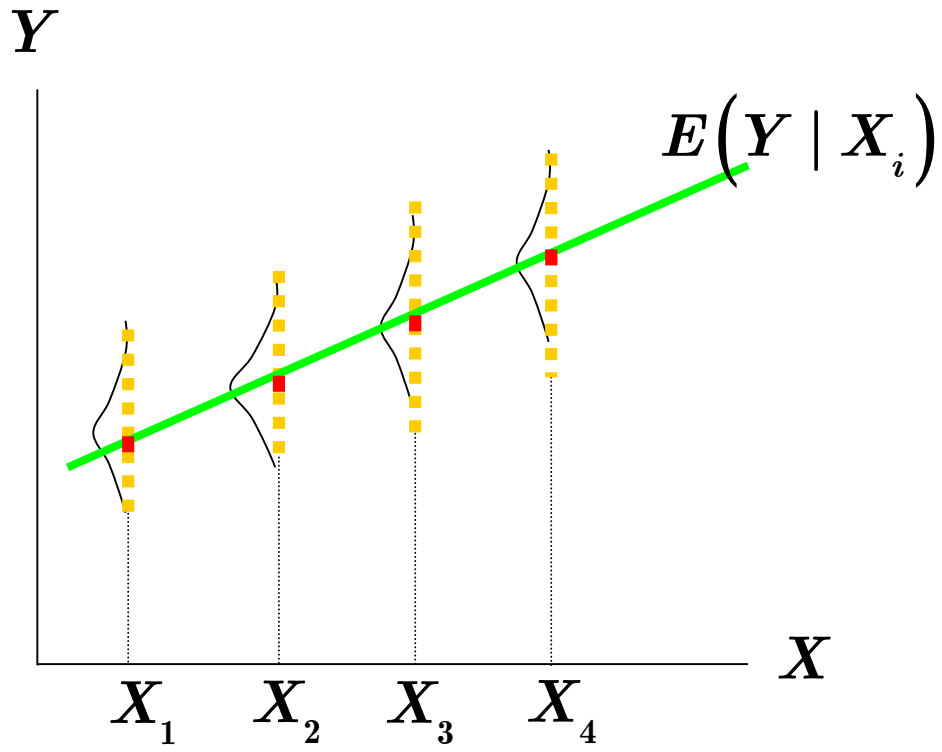
**表 2.1.1　某社区家庭每月收入与消费支出统计表**

| | 每月家庭可支配收入X（元） | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 800 | 1100 | 1400 | 1700 | 2000 | 2300 | 2600 | 2900 | 3200 | 3500 |
| 每月家庭消费支出Y（元） | 561 | 638 | 869 | 1023 | 1254 | 1408 | 1650 | 1969 | 2090 | 2299 |
| | 594 | 748 | 913 | 1100 | 1309 | 1452 | 1738 | 1991 | 2134 | 2321 |
| | 627 | 814 | 924 | 1144 | 1364 | 1551 | 1749 | 2046 | 2178 | 2530 |
| | 638 | 847 | 979 | 1155 | 1397 | 1595 | 1804 | 2068 | 2266 | 2629 |
| | | 935 | 1012 | 1210 | 1408 | 1650 | 1848 | 2101 | 2354 | 2860 |
| | | 968 | 1045 | 1243 | 1474 | 1672 | 1881 | 2189 | 2486 | 2871 |
| | | | 1078 | 1254 | 1496 | 1683 | 1925 | 2233 | 2552 | |
| | | | 1122 | 1298 | 1496 | 1716 | 1969 | 2244 | 2585 | |
| | | | 1155 | 1331 | 1562 | 1749 | 2013 | 2299 | 2640 | |
| | | | 1188 | 1364 | 1573 | 1771 | 2035 | 2310 | | |
| | | | 1210 | 1408 | 1606 | 1804 | 2101 | | | |
| | | | | 1430 | 1650 | 1870 | 2112 | | | |
| | | | | 1485 | 1716 | 1947 | 2200 | | | |
| | | | | | | 2002 | | | | |
| 共计 | 2420 | 4950 | 11495 | 16445 | 19305 | 23870 | 25025 | 21450 | 21285 | 15510 |
| $E(Y_i\|X_i)$ | 605 | 825 | 1045 | 1265 | 1485 | 1705 | 1925 | 2145 | 2365 | 2585 |

# （三）总体回归线与总体回归函数

## 总体回归线

对解释变量$X$的每个取值$X_i\left(i=1,2,\cdots,N\right)$，都有被解释变量$Y$的条件均值$E\left(Y\mid X_i\right)$与之对应，由$Y$的这些条件均值所形成的轨线（直线或曲线）称之为总体回归线




## 总体回归函数

总体回归线本质上反映了$Y$的条件均值$E\left(Y\mid X_i\right)$与$X_i$之间的数量对应关系：给定$X$的任一取值$X_i$，必有惟一的一个条件均值$E\left(Y\mid X_i\right)$与之对应。用于描述二者之间关系的函数$E\left(Y\mid X_i\right)=f\left(X_i\right)$称为总体回归函数

1、最简单的总体回归函数为$E\left(Y \mid X_i\right)=\beta_1+\beta_2 X_i$

2、如果我们拥有总体信息，那么$X_i$和$E\left(Y \mid X_i\right)$均将是已知的，通过求解一个二元线性方程组即可解出总体参数$\beta_1$ 和$\beta_2$ 的值，无须学习计量经济学。

3、总体参数$\left(\text{或结构参数}\right)\beta_2$ 的经济含义是：自变量$X$ 每增加1单位，因变量$Y$平均将增加$\beta_2$个单位$\left(\text{等同于微观经济学中的}\text{边际量}\right)$。

# 总体回归函数（PRF）

表 2.1.1　某社区家庭每月收入与消费支出统计表

| | 每月家庭可支配收入X（元） | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 800 | 1100 | 1400 | 1700 | 2000 | 2300 | 2600 | 2900 | 3200 | 3500 |
| 每月家庭消费支出Y（元） | 561 | 638 | 869 | 1023 | 1254 | 1408 | 1650 | 1969 | 2090 | 2299 |
| | 594 | 748 | 913 | 1100 | 1309 | 1452 | 1738 | 1991 | 2134 | 2321 |
| | 627 | 814 | 924 | 1144 | 1364 | 1551 | 1749 | 2046 | 2178 | 2530 |
| | 638 | 847 | 979 | 1155 | 1397 | 1595 | 1804 | 2068 | 2266 | 2629 |
| | | 935 | 1012 | 1210 | 1408 | 1650 | 1848 | 2101 | 2354 | 2860 |
| | | 968 | 1045 | 1243 | 1474 | 1672 | 1881 | 2189 | 2486 | 2871 |
| | | | 1078 | 1254 | 1496 | 1683 | 1925 | 2233 | 2552 | |
| | | | 1122 | 1298 | 1496 | 1716 | 1969 | 2244 | 2585 | |
| | | | 1155 | 1331 | 1562 | 1749 | 2013 | 2299 | 2640 | |
| | | | 1188 | 1364 | 1573 | 1771 | 2035 | 2310 | | |
| | | | 1210 | 1408 | 1606 | 1804 | 2101 | | | |
| | | | | 1430 | 1650 | 1870 | 2112 | | | |
| | | | | 1485 | 1716 | 1947 | 2200 | | | |
| | | | | | | 2002 | | | | |
| 共计 | 2420 | 4950 | 11495 | 16445 | 19305 | 23870 | 25025 | 21450 | 21285 | 15510 |
| $E(Y_i\|X_i)$ | 605 | 825 | 1045 | 1265 | 1485 | 1705 | 1925 | 2145 | 2365 | 2585 |

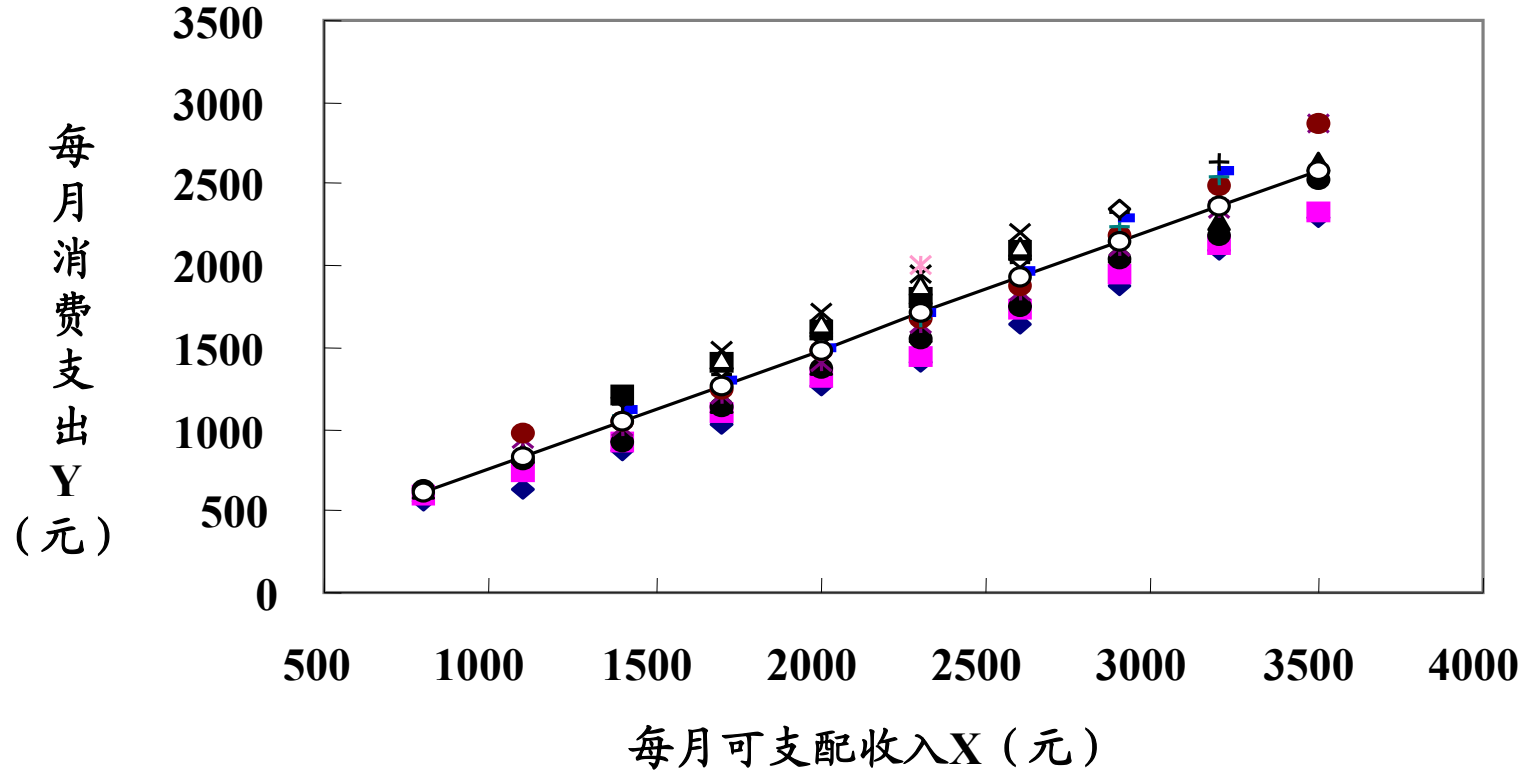
# 分　析:

- （1）由于不确定因素的影响，对同一收入水平X，不同家庭的消费支出不完全相同;

- （2）但由于调查的完备性，给定收入X后，消费支出Y的分布是确定的，即Y的条件条件分布是已知的，如：P（Y=638|X=800=1/4）

因此，给定收入X的任一值$X_i$，可得消费支出Y的条件均值或条件期望：E（Y|X=$X_i$）。在该例中：E（Y | X=800）=605

# 描出散点图:




随着收入的增加，“平均而言”消费也在增加，且Y的条件均值均落在一根正斜率的直线上。这条直线称为总体回归线。

根据总体回归线，可将同一收入水平下居民的平均消费支出看成是其可支配收入的线性函数，即总体回归函数为：

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

任意代入两组$X_i$和$E(Y|X_i)$的值，即可解出$\beta_1$=18.333，$\beta_2$=0.733，即总体回归函数为：

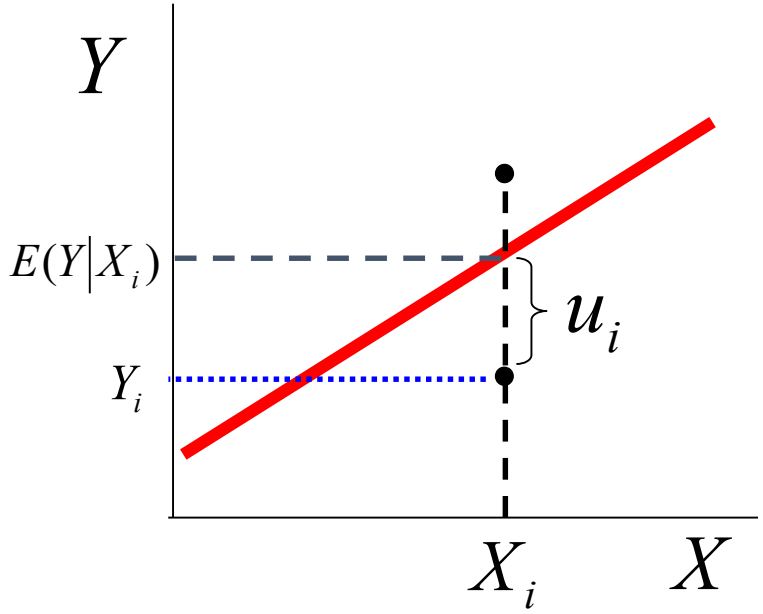
$$E(Y|X_i) = 18.333 + 0.733 X_i$$

其中$\beta_2$=0.733的含义是：居民人均可支配收入每增加1元，消费支出平均将增加0.733元。即边际消费倾向为0.733。

# 随机误差项

总体回归函数说明在给定的收入水平$X_i$下，该社区家庭平均的消费支出水平。

但对某一个别的家庭，其消费支出可能与该平均水平有偏差。




$$\text{记}\quad u_i = Y_i - E(Y \mid X_i)$$

称$u_i$为$Y_i$围绕它的期望值$E(Y|X_i)$的离差，是一个不可观测的随机变量，又称为随机干扰项或随机误差项

由$u_i = Y_i - E\left(Y \mid X_i\right) \Rightarrow Y_i = E\left(Y \mid X_i\right) + u_i \Rightarrow$

$$Y_i = E\left(Y \mid X_i\right) + u_i = \beta_1 + \beta_2 X_i + u_i \qquad (1)$$

式（1）称为总体回归模型，它刻画了每户家庭收入（$X_i$）与消费支出（$Y_i$）之间的真实数量关系

式（1）表明，家庭消费支出受两方面因素的共同影响：

（1）受收入水平$X_i$的影响，这部分影响对所有家庭而言都是确定的。若不同家庭收入相同（都为$X_i$），则其消费支出也相同（都为$E\left(Y \mid X_i\right) = \beta_1 + \beta_2 X_i$）

（2）受其它随机因素$u_i$的影响。由于$u_i$对不同家庭的影响不一样，因此尽管一些家庭收入相同，但其消费支出却不同。

# 为什么回归模型中必须包含随机误差项？

- 随机误差项主要包括下列因素：
  - 在解释变量中被忽略的因素的影响；
    - 影响不显著的因素
    - 未知的影响因素
    - 无法获得数据的因素
  - 变量观测值的观测误差的影响；
  - 模型关系的设定误差的影响；
  - 经济现象的内在随机性。

# 样本回归函数（SRF）

总体的信息往往无法掌握，现实的情况只能是在一次观测中得到总体的一个样本。

问题：能从一次抽样中获得总体的近似的信息吗？如果可以，如何从抽样中获得总体的近似信息？

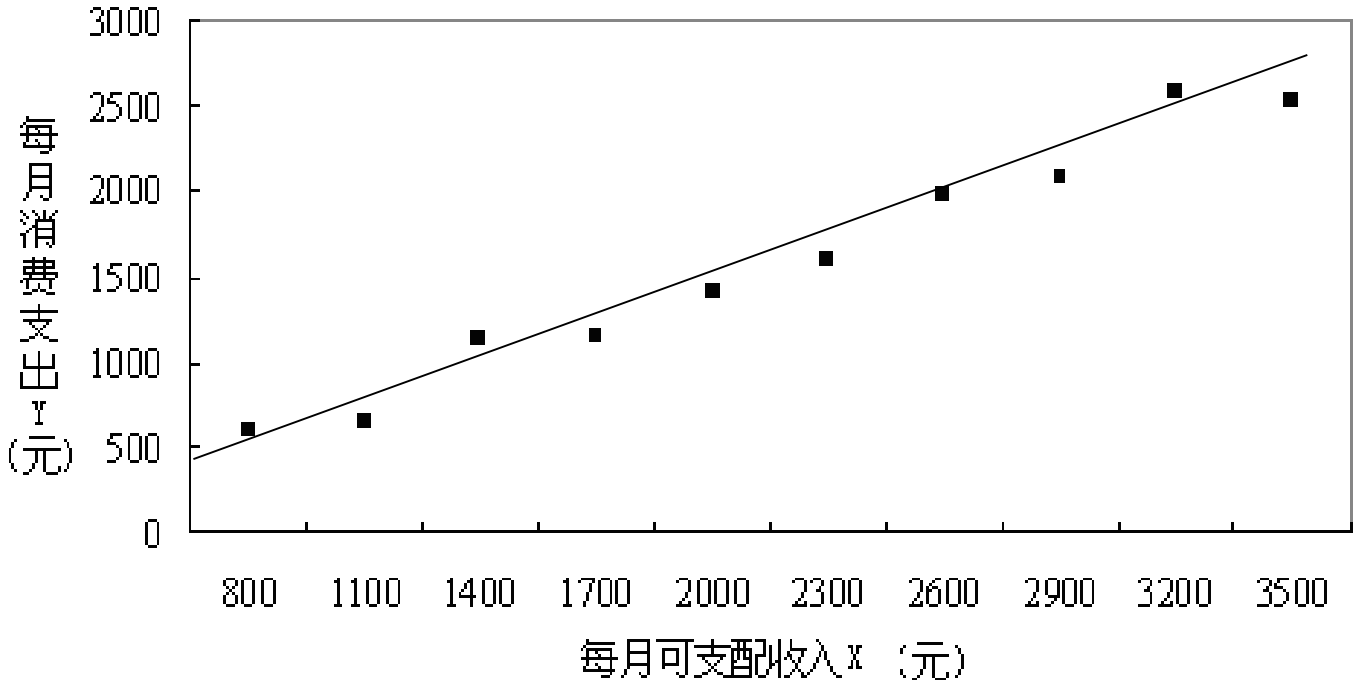
例2.1.2：在例2.1.1的总体中有如下一个样本，问：能否从该样本估计总体回归函数PRF？

表 2.1.2　家庭消费支出与可支配收入的一个随机样本

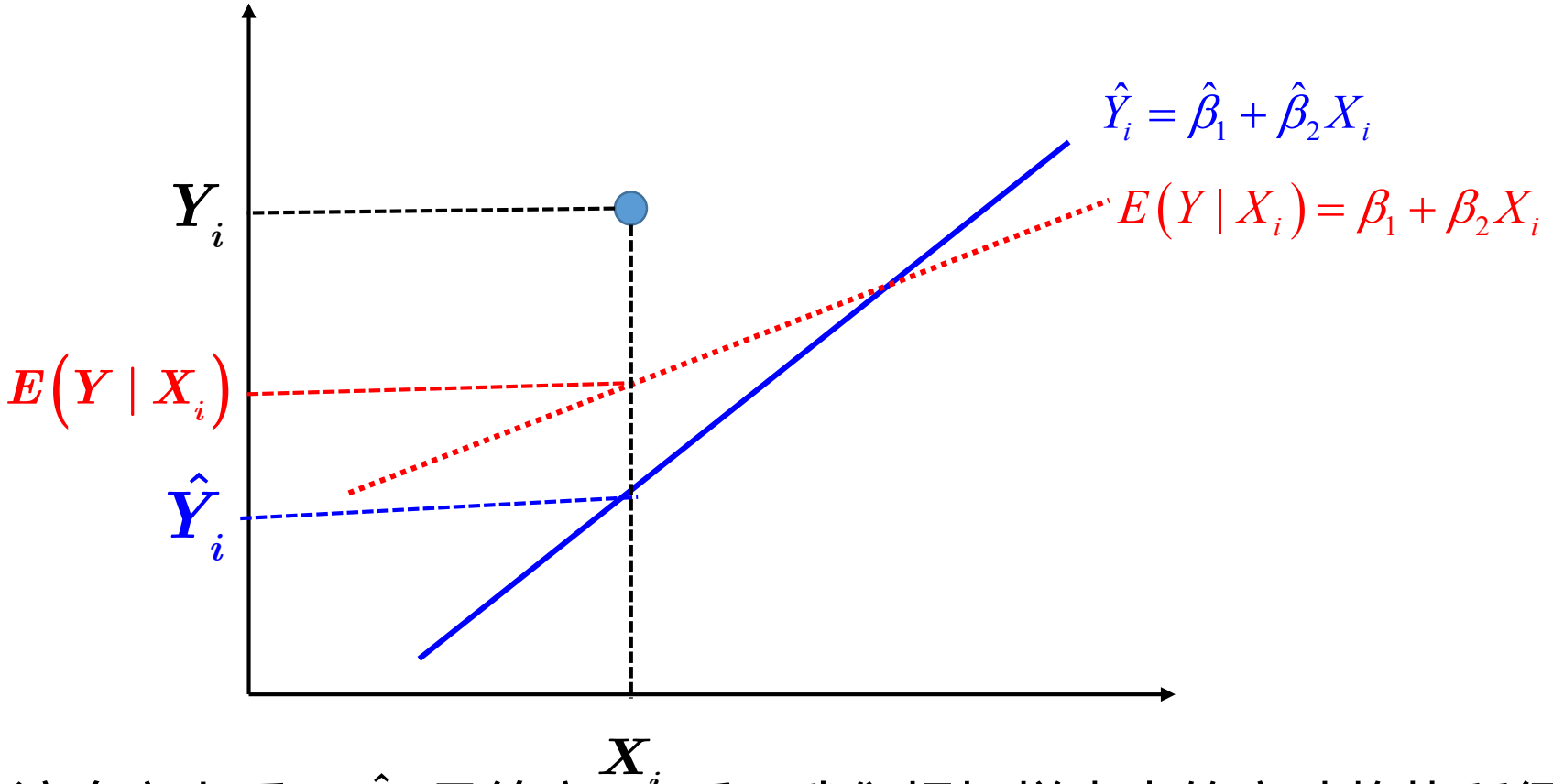
| X | 800 | 1100 | 1400 | 1700 | 2000 | 2300 | 2600 | 2900 | 3200 | 3500 |
|---|---|---|---|---|---|---|---|---|---|---|
| Y | 594 | 638 | 1122 | 1155 | 1408 | 1595 | 1969 | 2078 | 2585 | 2530 |

回答：能

# 描绘出核样本的散点图（scatter diagram):




样本散点图近似于一条直线，画一条直线以尽好地拟合该散点图，由于样本取自总体，可以该线近似地代表总体回归线。该线称为样本回归线（sample regression lines）。




从经济含义上看，$\hat{Y}_i$ 是给定$X_i$ 后，我们根据样本点的变动趋势所得到的因变量的一个估计值，它既不是因变量的真实值 $Y_i$，也不是因变量的条件均值$E(Y \mid X_i)$

$\hat{Y}_i$在真实世界中并不存在。

但仅从数量关系上看，则 $\hat{Y}_i$ 既是 $Y_i$ 的无偏估计量，也是$E(Y \mid X_i)$ 的无偏估计量

若用$\hat{Y}_i$ 来预测$Y_i$ ，称之为个值预测。若用$\hat{Y}_i$ 来预测$E(Y \mid X_i)$ ，则称之为均值预测

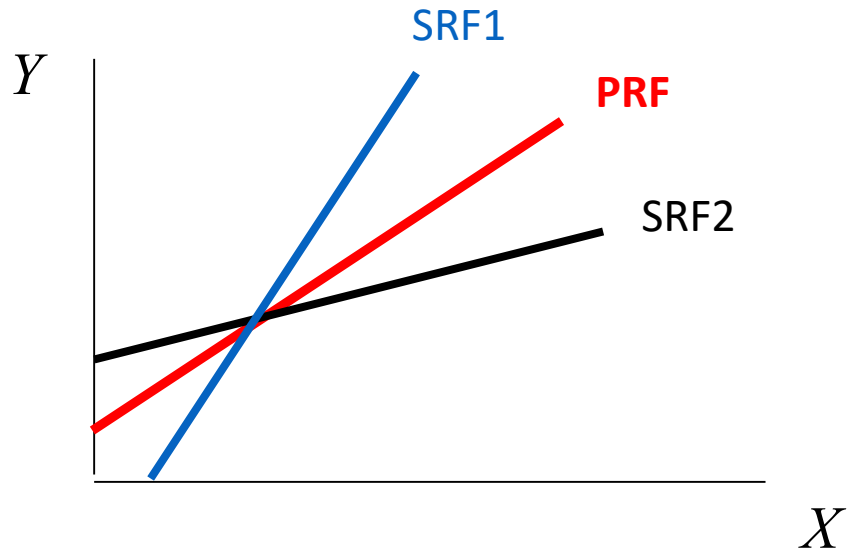
# 样本回归函数SRF

记样本回归线的函数形式为：$\hat{Y}_i = f(X_i) = \hat{\beta}_1 + \hat{\beta}_2 X_i$

则称其样本回归函数（sample regression function，SRF）

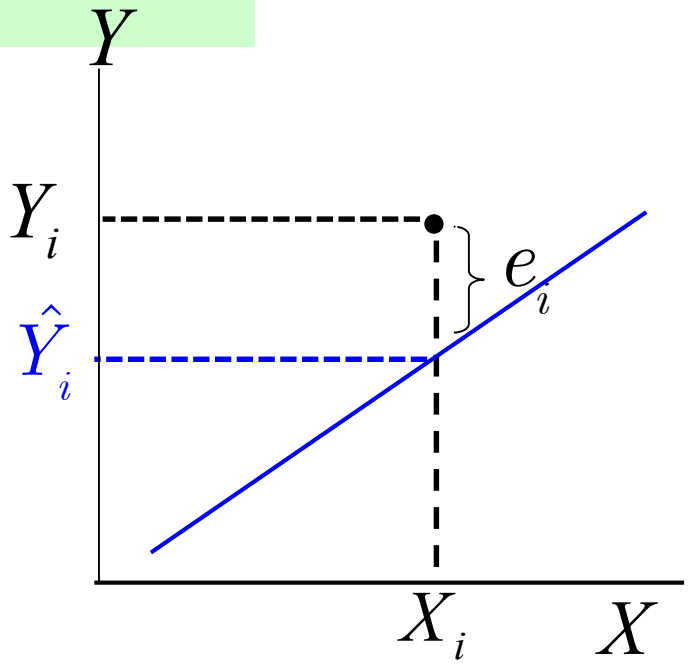
- 每次抽样都能获得一个样本，就可以拟合一条样本回 归线，所以样本回归线随抽样变化而变化，可以有许多条（SRF不唯一）

- 每一条样本回 归线与总体回归线都有偏离，但若对多条样本回归线的位置进行平均，则平均后的样本回归线将与总体回归线重合。



# 样本回归模型：

定义$e_i = Y_i - \hat{Y}_i$，$e_i$称为残差或拟合误差，
它表示影响因变量的其它随机影响因素，
也衡量了样本点对样本回归线的偏离程度。




$$由 e_i = Y_i - \hat{Y}_i \Rightarrow Y_i = \hat{Y}_i + e_i$$

$$\Downarrow$$

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i \qquad (2)$$

式（2）称为样本回归模型。它刻画了样本家庭收入与消费支出的真实数量对应关系。

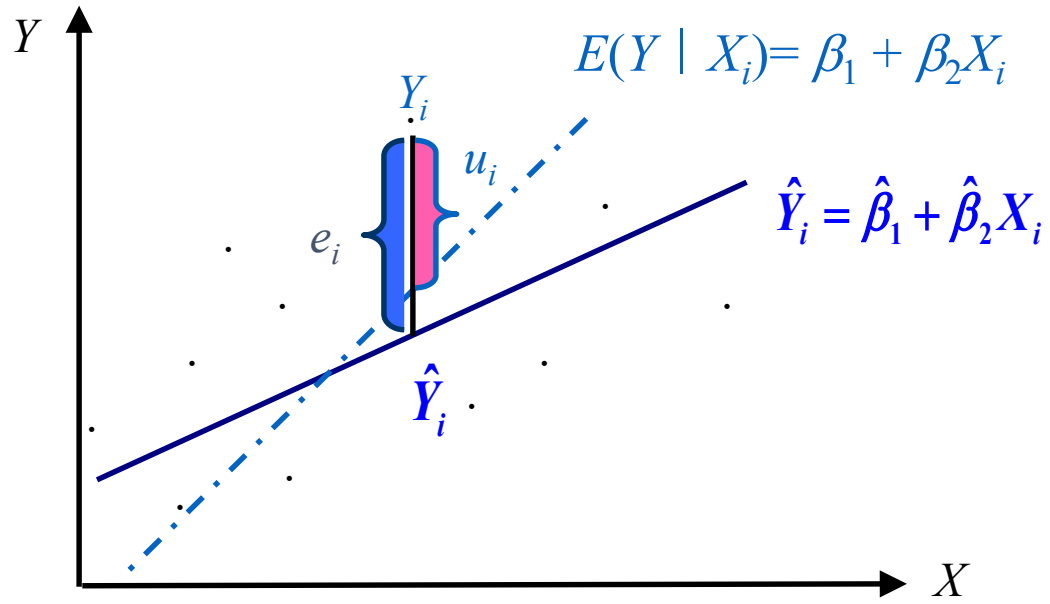
总体回归模型：$Y_i = \beta_1 + \beta_2 X_i + u_i$

样本回归模型：$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$

# 回归分析的主要目的：

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

显然，样本回归线与总体回归线之间总存在差异，回归分析的目的就是设计出一种方法，以使SRF尽可能“接近”PRF；或者说使 $\hat{\beta}_i$ 尽可能“接近” $\beta_i$




总体回归线与样本回归线之间的关系

注意：这里PRF可能永远无法知道。

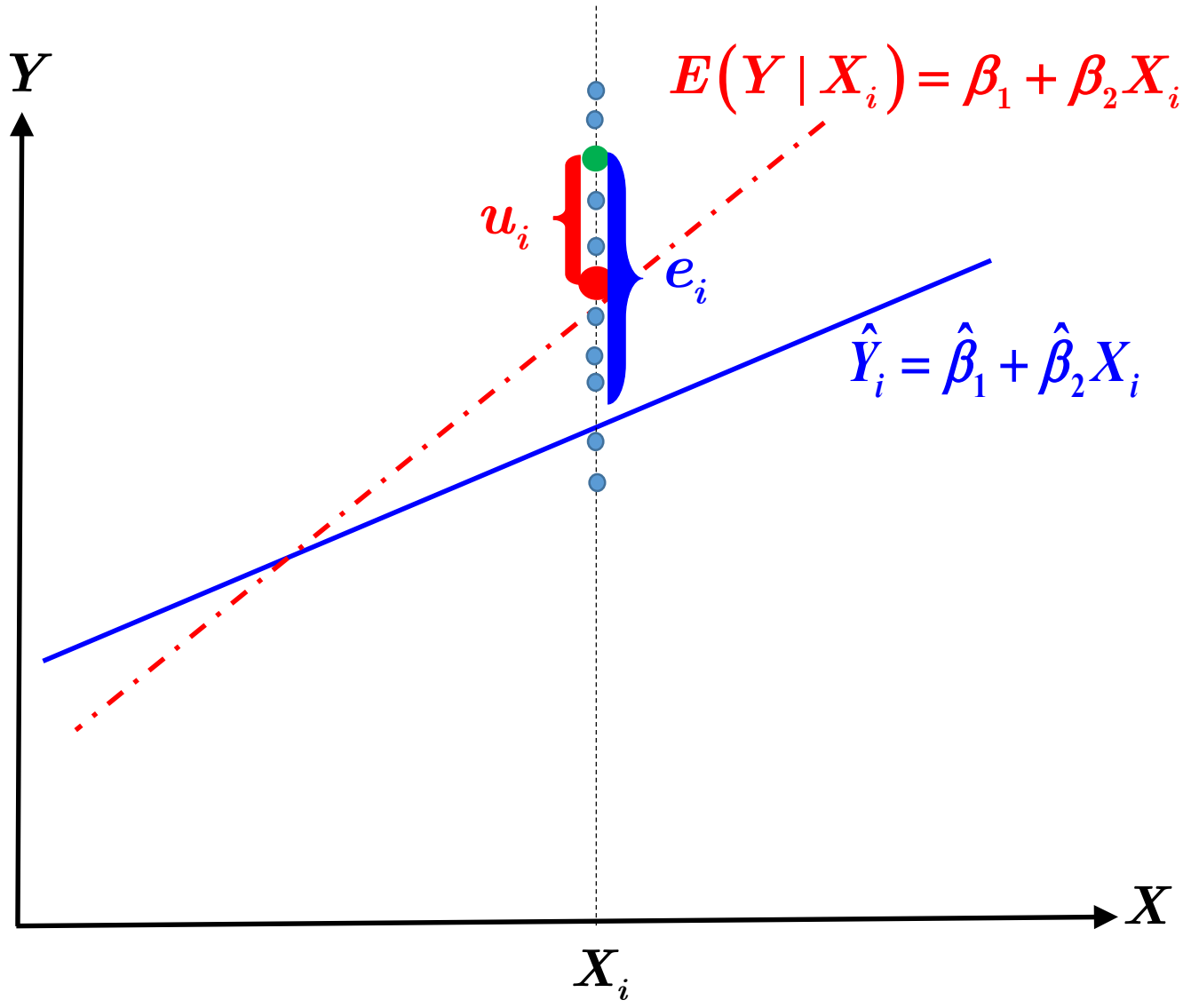
# 残差与随机扰动项的关系：

在计量分析中，由于总体信息未知，因此给定$X_i$后对应的$\{u_i|X_i\}_{i=1}^{10}$未知，我们只能用$e_i$近似代替$\{u_i|X_i\}_{i=1}^{10}$。此时$e_i$甚至不是$\{u_i|X_i\}_{i=1}^{10}$中的某一个$u_i$

进一步，序列$\{u_i|X_i\}_{i=1}^{10}$的方差为

$$Var(u_i|X_i)=\frac{\sum_{i=1}^{10}u_i^2}{10}\approx e_i^2$$

因此以后我们将用$e_i^2$作为$\{u_i|X_i\}$方差的近似估计。



# 总结：

总体回归方程：$E(Y \mid X_i) = \beta_1 + \beta_2 X_i$

样本回归方程：$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

总体回归模型：$Y_i = \beta_1 + \beta_2 X_i + u_i$

样本回归模型：$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$

# 第二节：简单线性回归模型的参数估计

本节基本内容:

- 简单线性回归模型的基本假定
- 普通最小二乘法
- OLS回归线的性质

# 一、简单线性回归模型的基本假定

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有一定的适用条件——假定条件

## 1. 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，参数的估计量也是一个随机变量，只有对随机扰动的分布作出假定，才能确定所估计参数的分布性质，也才可能进行假设检验和区间估计。

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

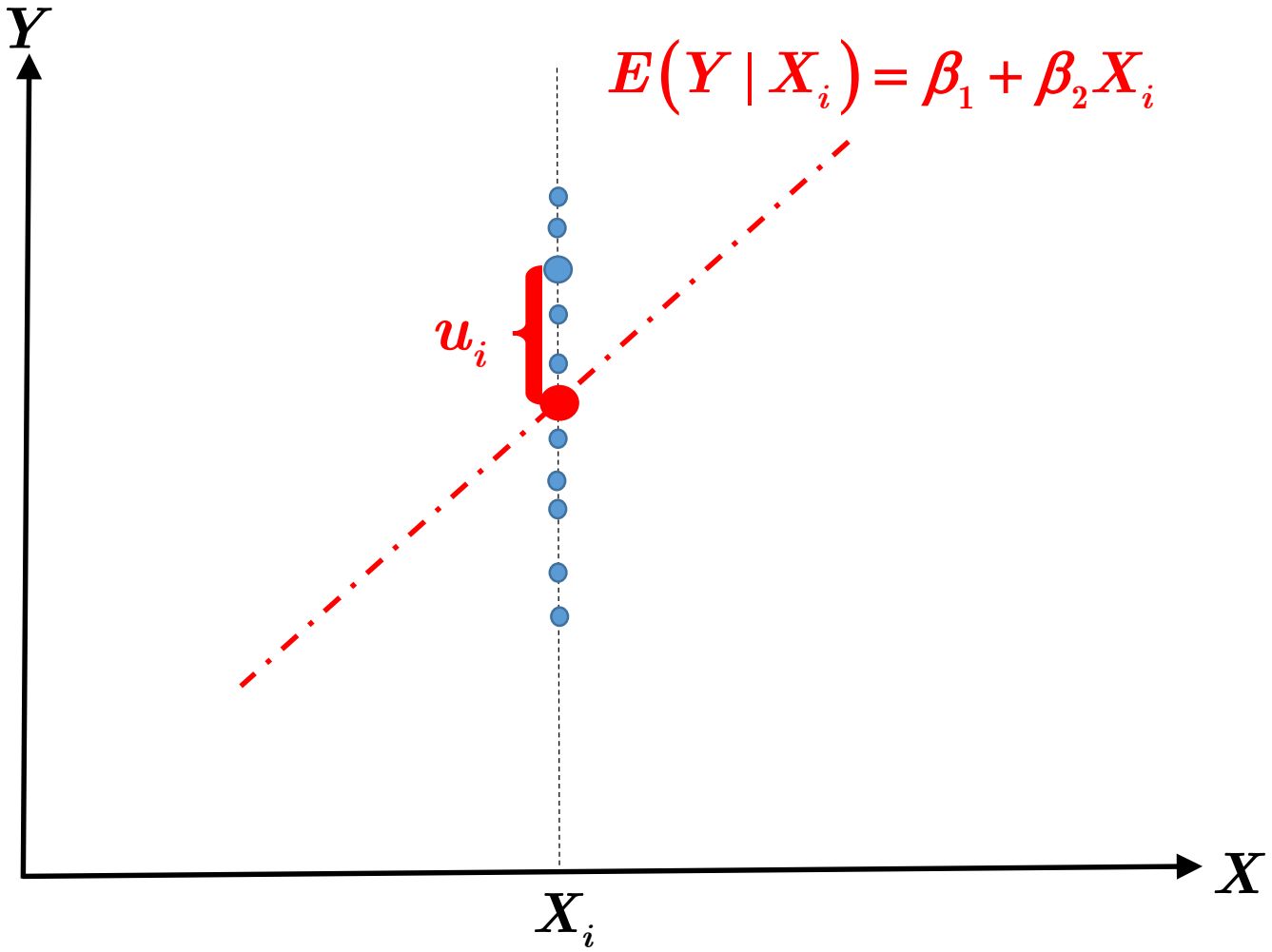
例如对于 $Y_i = \beta_1 + \beta_2 X_i + u_i$

- 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是非随机的，即在重复抽样中取固定值。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

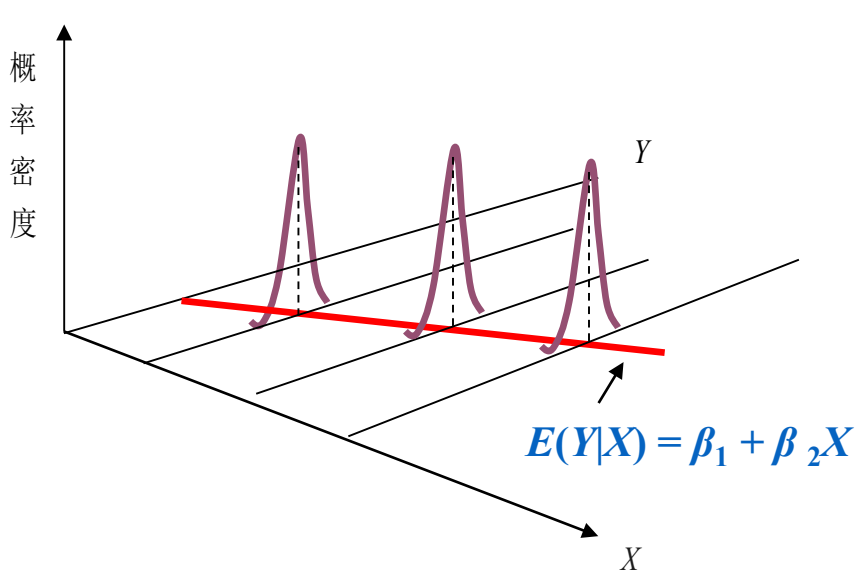
**假定1：零均值假定:** 在给定$X_i$的条件下，$u_i$的条件期望为零，即$E(u_i|X_i)=0$。



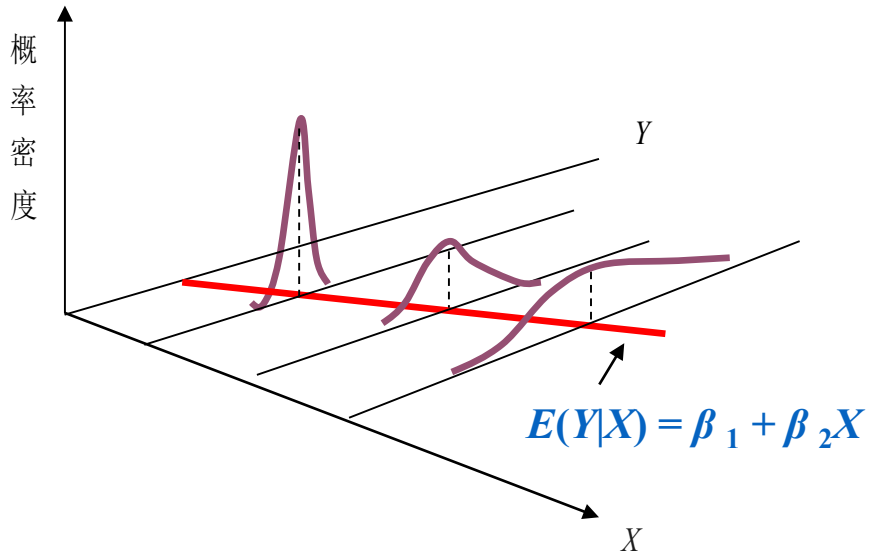
# 2、对随机扰动项u的假定

**假定2：同方差假定:**在给定$X_i$的条件下，$u_i$的条件方差为某个常数，即

$$Var\left(u_i|X_i\right)=E\left(u_i^2\right)-\left[E\left(u_i|X_i\right)\right]^2=E\left(u_i^2\right)=\frac{\sum_{i=1}^{s}u_i^2}{s}=\sigma^2$$




同方差



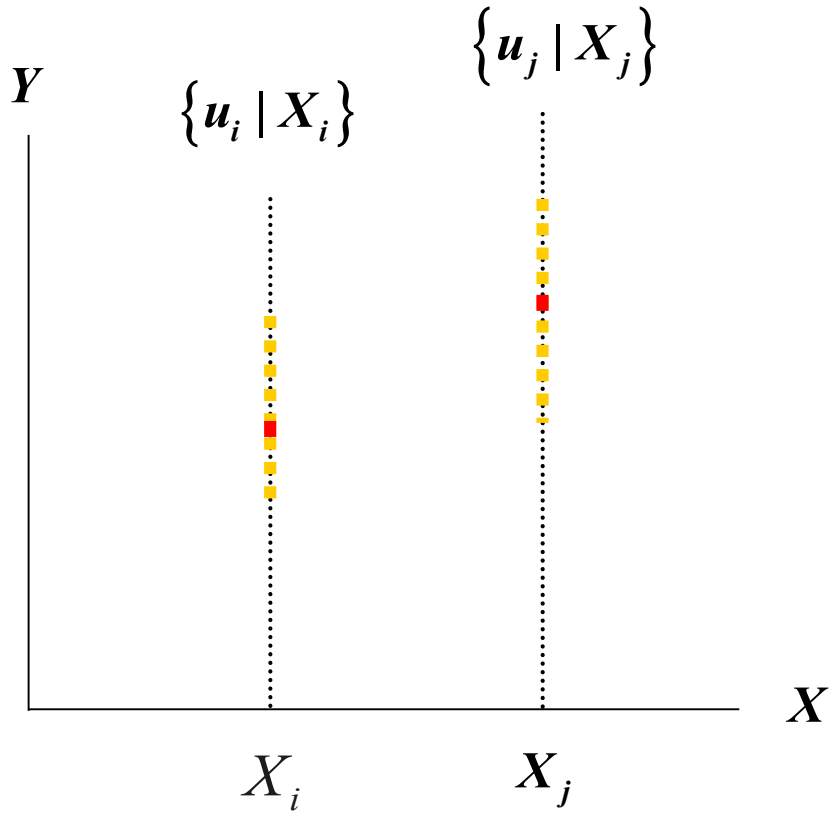

异方差

**假定3：无自相关假定：** 随机扰动项$u_i$的逐次值互不相关，即：

$$Cov\left(u_i,u_j\middle|X_i,X_j\right)=0\left(i\neq j\right)$$

**注：** 当随机扰动项满足同方差和无自相关假定时，其方差-协方差矩阵为：

$$Var\left(U\mid X\right)=E\left(UU'\mid X\right)=\begin{array}{c}\\ u_1\\ u_2\\ \vdots\\ u_n\end{array}\begin{array}{c}\begin{array}{cccc}u_1 & u_2 & \cdots & u_n\end{array}\\ \begin{pmatrix}\sigma^2 & 0 & \cdots & 0\\ 0 & \sigma^2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \sigma^2\end{pmatrix}\end{array}=\sigma^2 I$$



假定4：解释变量$X$是非随机的，或者虽然$X$是随机的但与扰动项$u$不相关，即：$Cov(u_i, X_i) = E(u_i X_i) - E(u_i)E(X_i) = 0$

假定5：正态性假定：随机扰动项服从均值为零，方差为$\sigma^2$的正态分布

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

（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）

# 被解释变量Y的分布性质

由于 $Y_i=\beta_1+\beta_2X_i+u_i$ 且$X$为确定性变量，因此$u_i$ 的分布性质也就决定了 $Y_i$ 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：

1、零均值假设：$E(u_i\mid X_i)=0\Rightarrow E(Y_i\mid X_i)=\beta_1+\beta_2X_i$

2、同方差假设：$Var(u_i\mid X_i)=\sigma^2\Rightarrow Var(Y_i\mid X_i)=\sigma^2$

3、无自相关假设：$Cov(u_i,u_j\mid X_i,X_j)=0\Rightarrow Cov(Y_i,Y_j\mid X_i,X_j)=0$

证明：$Cov(Y_i,Y_j\mid X_i,X_j)=Cov(\beta_1+\beta_2X_i+u_i,\beta_1+\beta_2X_j+u_j\mid X_i,X_j)$

$=Cov(\beta_1+\beta_2X_i,\beta_1+\beta_2X_j\mid X_i,X_j)+Cov(\beta_1+\beta_2X_i,u_j\mid X_i,X_j)$

$+Cov(u_i,\beta_1+\beta_2X_j\mid X_i,X_j)+Cov(u_i,u_j\mid X_i,X_j)=Cov(u_i,u_j\mid X_i,X_j)$

4、正态性假设：$u_i|X_i\sim N(0,\sigma^2)\Rightarrow Y_i\mid X_i\sim N(\beta_1+\beta_2X_i,\sigma^2)$

正态性意味着：样本$Y_1,Y_2,\cdots,Y_n$为相互独立的正态随机变量，故其任意线性组合$\sum\lambda_iY_i$ 必然服从正态分布。

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

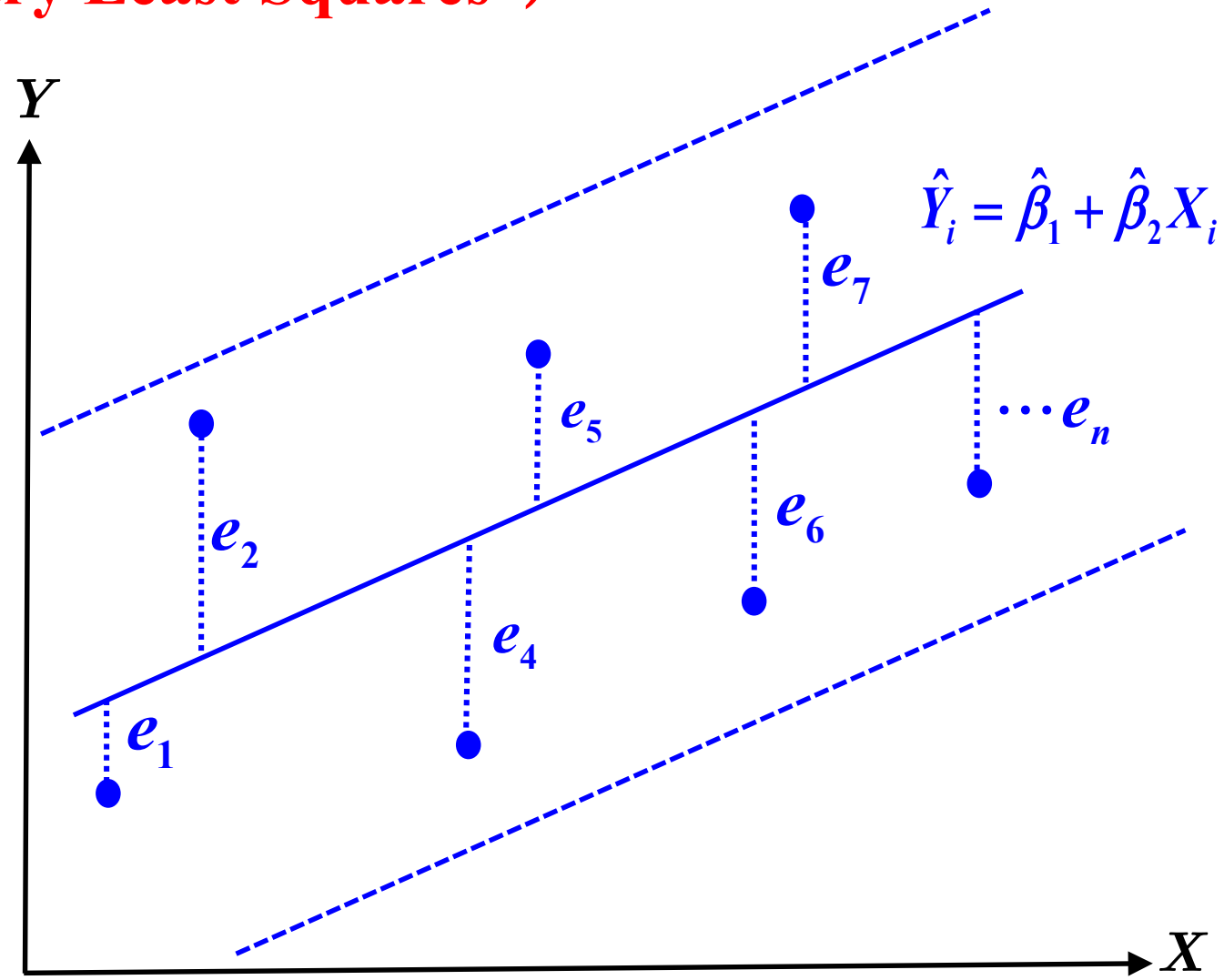
### ◆OLS的基本思想

如何刻画样本点偏离样本回归线的平均偏离程度?

$$\frac{\sum_{i=1}^{n}|e_i|}{n} \text{或} \frac{\sum_{i=1}^{n}e_i^2}{n}$$

因此，最小二乘法的原理是:
使样本点偏离样本回归线的平均偏离程度最小:$\min \frac{\sum_{i=1}^{n}e_i^2}{n} \Leftrightarrow \min \sum_{i=1}^{n}e_i^2$

即：使残差平方和最小化



故参数估计问题就转化为求解如下无约束极值问题：

$$\min_{\hat{\beta}_1,\ \hat{\beta}_2} \sum_{i=1}^{n} e_i^2$$

$$\Downarrow$$

$$\min_{\hat{\beta}_1,\ \hat{\beta}_2} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \Leftrightarrow \min_{\hat{\beta}_1,\ \hat{\beta}_2} Q = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)^2$$

由无约束极值问题的一阶条件知：

$$\begin{cases} \dfrac{\partial Q}{\partial \hat{\beta}_1} = 0 \\ \dfrac{\partial Q}{\partial \hat{\beta}_2} = 0 \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right) = 0 \Rightarrow \sum_{i=1}^{n} e_i = 0 \\ \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right) X_i = 0 \Rightarrow \sum_{i=1}^{n} e_i X_i = 0 \end{cases}$$

$$\text{注：} \sum_{i=1}^{n} e_i X_i = 0 \Leftrightarrow \left(e_1\ \ e_2 \cdots e_n\right) \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = 0 \Leftrightarrow e \cdot X = 0 \Leftrightarrow e\text{与}X\text{正交} \Leftrightarrow Cov\left(e, X\right) = 0$$

# OLS估计原理

即
$$\begin{cases} n\hat{\beta}_1 + (\sum_{i=1}^{n} X_i)\hat{\beta}_2 = \sum_{i=1}^{n} Y_i \\ (\sum_{i=1}^{n} X_i)\hat{\beta}_1 + (\sum_{i=1}^{n} X_i^2)\hat{\beta}_2 = \sum_{i=1}^{n} X_i Y_i \end{cases} \tag{1}$$

式(1)称为正规方程组（或简称为正规方程）

使用消元法可求得总体回归系数的估计值为：

$$\begin{cases} \hat{\beta}_2 = \dfrac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} \\ \hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} \end{cases} \tag{2}$$

令$\bar{X}=\frac{1}{n}\sum X_i$，$\bar{Y}=\frac{1}{n}\sum Y_i$，$x_i=X_i-\bar{X}$，$y_i=Y_i-\bar{Y}$

(其中$x_i$，$y_i$分别称为对应的样本值与其均值的离差)

由于：$\sum x_i^2=\sum(X_i-\bar{X})^2=\sum X_i^2-\frac{1}{n}\left(\sum X_i\right)^2$

$$\sum x_i y_i=\sum(X_i-\bar{X})(Y_i-\bar{Y})=\sum X_iY_i-\frac{1}{n}\sum X_i\sum Y_i$$

将其代入与方程组（2），则得到OLS估计的离差形式为：

$$\begin{cases}\hat{\beta}_2=\dfrac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2}=\dfrac{\sum x_i y_i}{\sum x_i^2}\\ \hat{\beta}_1=\bar{Y}-\hat{\beta}_2\bar{X}\end{cases}$$

牢记！

☞ 注意：在计量经济学中，往往以小写字母表示变量观测值对其均值的离差！

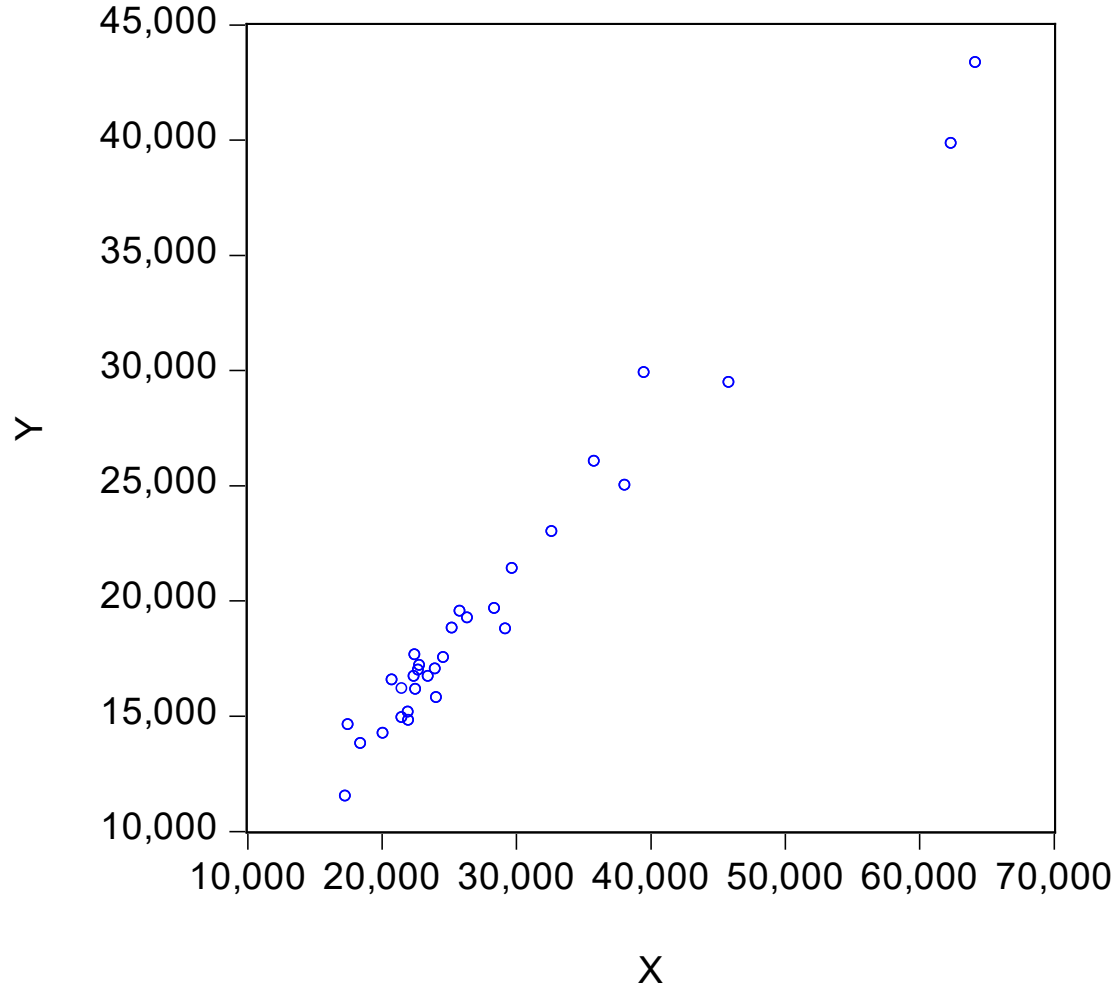
# 一元线性回归模型举例

案例2.1　右表给出了中国31个省市自治区2018年底居民人均可支配收入（元）和人均消费支出（元）的相关统计数据。

试估计Y与X之间的关系

| 地区 | 人均可支配收入X | 人均消费支出Y | 地区 | 人均可支配收入X | 人均消费支出Y |
|---|---|---|---|---|---|
| 北京市 | 62361.22 | 39842.69 | 湖北省 | 25814.54 | 19537.79 |
| 天津市 | 39506.15 | 29902.91 | 湖南省 | 25240.75 | 18807.94 |
| 河北省 | 23445.65 | 16722 | 广东省 | 35809.9 | 26053.98 |
| 山西省 | 21990.14 | 14810.12 | 广西 | 21485.03 | 14934.75 |
| 内蒙古 | 28375.65 | 19665.22 | 海南省 | 24579.04 | 17528.44 |
| 辽宁省 | 29701.45 | 21398.31 | 重庆市 | 26385.84 | 19248.47 |
| 吉林省 | 22798.37 | 17200.41 | 四川省 | 22460.55 | 17663.55 |
| 黑龙江 | 22725.85 | 16993.96 | 贵州省 | 18430.18 | 13798.06 |
| 上海市 | 64182.65 | 43351.3 | 云南省 | 20084.19 | 14249.93 |
| 江苏省 | 38095.79 | 25007.44 | 西藏 | 17286.06 | 11520.23 |
| 浙江省 | 45839.84 | 29470.68 | 陕西省 | 22528.26 | 16159.69 |
| 安徽省 | 23983.58 | 17044.64 | 甘肃省 | 17488.39 | 14623.95 |
| 福建省 | 32643.93 | 22996.04 | 青海省 | 20757.26 | 16557.19 |
| 江西省 | 24079.68 | 15792.02 | 宁夏 | 22400.42 | 16715.09 |
| 山东省 | 29204.61 | 18779.77 | 新疆 | 21500.24 | 16189.09 |
| 河南省 | 21963.54 | 15168.5 | | | |

- 作出$X$、$Y$之间的散点图：




- 建立回归模型

由散点图可知，$Y$随$X$的增加而线性增加，因此建立线性回归模型：

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

## ➢估计回归系数

| 地区 | 人均可支配收入X | 人均消费支出Y | xi | yi | xiyi | xi^2 |
|---|---|---|---|---|---|---|
| 海南省 | 24579.04 | 17528.44 | -3587.04871 | -2398.47 | 8603422.933 | 12866918.4 |
| 重庆市 | 26385.84 | 19248.47 | -1780.24871 | -678.438 | 1207789.063 | 3169285.47 |
| 四川省 | 22460.55 | 17663.55 | -5705.53871 | -2263.36 | 12913678.89 | 32553172 |
| 贵州省 | 18430.18 | 13798.06 | -9735.90871 | -6128.85 | 59669908.39 | 94787918.4 |
| 云南省 | 20084.19 | 14249.93 | -8081.89871 | -5676.98 | 45880764.3 | 65317086.8 |
| 西藏 | 17286.06 | 11520.23 | -10880.02871 | -8406.68 | 91464902.2 | 118375025 |
| 陕西省 | 22528.26 | 16159.69 | -5637.82871 | -3767.22 | 21238931.98 | 31785112.6 |
| 甘肃省 | 17488.39 | 14623.95 | -10677.69871 | -5302.96 | 56623391.93 | 114013250 |
| 青海省 | 20757.26 | 16557.19 | -7408.82871 | -3369.72 | 24965666.33 | 54890742.8 |
| 宁夏 | 22400.42 | 16715.09 | -5765.66871 | -3211.82 | 18518280.78 | 33242935.7 |
| 新疆 | 21500.24 | 16189.09 | -6665.84871 | -3737.82 | 24915731.87 | 44433539 |
| 均值 | **28166.0887** | **19926.9084** | | 求和 | **2457860502** | **3943671954** |
| | | | | **Beta1^=** | **2372.629599** | |
| | | | | **Beta2^=** | **0.623241621** | |

➢得出样本回归方程 $\hat{Y}_i = 2372.630 + 0.623X_i$

## ➢附： EViews回归结果

Equation: UNTITLED Workfile: UNTITLED::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y
Method: Least Squares
Date: 04/02/20 Time: 15:46
Sample: 1 31
Included observations: 31

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 2372.630 | 546.0272 | 4.345259 | 0.0002 |
| X | 0.623242 | 0.017997 | 34.63095 | 0.0000 |

| | | | |
|---|---|---|---|
| R-squared | 0.976390 | Mean dependent var | 19926.91 |
| Adjusted R-squared | 0.975576 | S.D. dependent var | 7231.602 |
| S.E. of regression | 1130.167 | Akaike info criterion | 16.96046 |
| Sum squared resid | 37041016 | Schwarz criterion | 17.05297 |
| Log likelihood | -260.8871 | Hannan-Quinn criter. | 16.99062 |
| F-statistic | 1199.303 | Durbin-Watson stat | 1.652205 |
| Prob(F-statistic) | 0.000000 | | |

# 几个常用的结果

1、残差之和等于零：$\sum_{i=1}^{n} e_i = 0$

2、残差与解释变量不相关：$\sum_{i=1}^{n} e_i X_i = 0$

3、$(\bar{X}, \bar{Y})$必然位于样本回归线 $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ 上

证明：令 $X_i = \bar{X}$，则 $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 \bar{X} = \bar{Y} - \hat{\beta}_2 \bar{X} + \hat{\beta}_2 \bar{X} = \bar{Y} \Rightarrow$ 结论

4、 $\bar{Y} = \bar{\hat{Y}}$

证明：由 $e_i = Y_i - \hat{Y}_i \Rightarrow \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{Y}_i \Rightarrow \frac{\sum_{i=1}^{n} e_i}{n} = \frac{\sum_{i=1}^{n} Y_i}{n} - \frac{\sum_{i=1}^{n} \hat{Y}_i}{n}$

$\Rightarrow 0 = \bar{Y} - \bar{\hat{Y}} \Rightarrow$ 结论

## 5、样本回归方程 $\hat{Y}_i=\hat{\beta}_1+\hat{\beta}_2 X_i$ 可由变量的离差形式重新表述为 $\hat{y}_i=\hat{\beta}_2 x_i$

证明：因样本回归方程为 $\hat{Y}_i=\hat{\beta}_1+\hat{\beta}_2 X_i$ $(1)$

由3知 $\bar{Y}=\hat{\beta}_1+\hat{\beta}_2\bar{X}$，再由4可得 $\bar{\hat{Y}}=\hat{\beta}_1+\hat{\beta}_2\bar{X}$ $(2)$

$(1)-(2)$ 可得 $\hat{Y}_i-\bar{\hat{Y}}=\hat{\beta}_2\left(X_i-\bar{X}\right)\Rightarrow \hat{y}_i=\hat{\beta}_2 x_i \Rightarrow$ 结论

## 6、$\sum_{i=1}^{n} e_i\hat{y}_i=0$

## 7、$\sum_{i=1}^{n} e_i\hat{Y}_i=0$

证明：$\sum_{i=1}^{n} e_i\hat{y}_i=\sum_{i=1}^{n} e_i\left(\hat{Y}_i-\bar{\hat{Y}}\right)=\sum_{i=1}^{n} e_i\hat{Y}_i-\sum_{i=1}^{n} e_i\bar{\hat{Y}}=\sum_{i=1}^{n} e_i\hat{Y}_i-\bar{\hat{Y}}\sum_{i=1}^{n} e_i=\sum_{i=1}^{n} e_i\hat{Y}_i$

而 $\sum_{i=1}^{n} e_i\hat{Y}_i=\sum_{i=1}^{n} e_i\left(\hat{\beta}_1+\hat{\beta}_2 X_i\right)=\hat{\beta}_1\sum_{i=1}^{n} e_i+\hat{\beta}_2\sum_{i=1}^{n} e_i X_i=0+0=0\Rightarrow$ 结论

$$8、\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2$$

证明：由于 $Y_i - \bar{Y} = Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y} = \left(Y_i - \hat{Y}_i\right) + \left(\hat{Y}_i - \bar{\hat{Y}}\right)$，即 $y_i = e_i + \hat{y}_i \Rightarrow y_i^2 = \left(e_i + \hat{y}_i\right)^2$

$$\Rightarrow \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \left(e_i + \hat{y}_i\right)^2 = \sum_{i=1}^{n} \left(e_i^2 + \hat{y}_i^2 + 2e_i\hat{y}_i\right) = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2 + 2\sum_{i=1}^{n} e_i\hat{y}_i = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2$$

$$9、\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} \hat{Y}_i^2 + \sum_{i=1}^{n} e_i^2$$

10、对简单线性回归模型 $Y_i = \beta_1 + \beta_2 X_i + u_i$ ，若 $X$ 和 $Y$ 的值同时扩大 $k$ 倍，则 $\hat{\beta}_2$ 的值不变而 $\hat{\beta}_1$ 的值将扩大 $k$ 倍;若 $X$ 的值不变而将 $Y$ 的值扩大 $k$ 倍，则 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的值均将扩大 $k$ 倍；若 $Y$ 的值不变而将 $X$ 的值扩大 $k$ 倍，则 $\hat{\beta}_1$ 的值不变，而 $\hat{\beta}_2$ 将变为原来的 $1/k$

证明：由于$\begin{cases} X_i' = kX_i \\ Y_i' = kY_i \end{cases} \Rightarrow \begin{cases} \bar{X}' = k\bar{X} \\ \bar{Y}' = k\bar{Y} \end{cases} \Rightarrow \begin{cases} x_i' = X_i' - \bar{X}' = kX_i - k\bar{X} = k\left(X_i - \bar{X}\right) = kx_i \\ y_i' = Y_i' - \bar{Y}' = kY_i - k\bar{Y} = k\left(Y_i - \bar{Y}\right) = ky_i \end{cases}$

$$\Rightarrow \hat{\beta}_2' = \frac{\sum_{i=1}^{n} x_i' y_i'}{\sum_{i=1}^{n} x_i'^2} = \frac{\sum_{i=1}^{n} kx_i ky_i}{\sum_{i=1}^{n} \left(kx_i\right)^2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \hat{\beta}_2$$

$$\Rightarrow \hat{\beta}_1' = \bar{Y}' - \hat{\beta}_2' \bar{X}' = k\bar{Y} - \hat{\beta}_2 k\bar{X} = k\left(\bar{Y} - \hat{\beta}_2 \bar{X}\right) = k\hat{\beta}_1$$

# 第三节　最小二乘估计量的统计性质

- 线性性
- 无偏性
- 有效性（最小方差性）

# 线性性

➢所谓线性性是指参数估计量可以表示为被解释变量观测值的线性组合；即 $\hat{\beta}_1, \hat{\beta}_2$ 均为 $Y_i$（$i=1,2,\cdots,n$）的线性函数

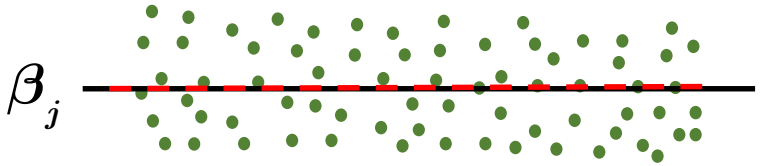
$$\begin{cases} \hat{\beta}_2 = \dfrac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} = \dfrac{\sum x_i y_i}{\sum x_i^2} = \sum \lambda_i Y_i \\ \hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = \sum \tau_i Y_i \end{cases}$$

线性性表明：结构参数$\beta_i$ 的最小二乘估计量$\hat{\beta}_i$ 是$n$个独立正态随机变量的某一线性组合，因此$\hat{\beta}_i$ 必然服从正态分布。

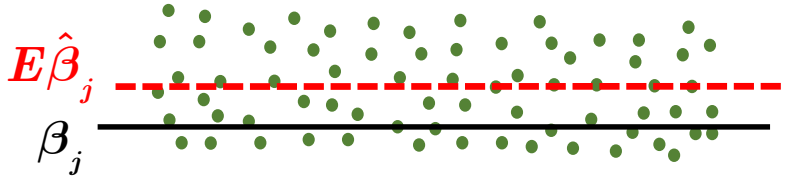
# 无偏性

➢无偏性，是指即OLS估计量 $\hat{\beta}_1$，$\hat{\beta}_2$ 的数学期望（均值）等于总体回归参数 $\beta_1$，$\beta_2$ 真实值。即： $E(\hat{\beta}_1)=\beta_1$，$E(\hat{\beta}_2)=\beta_2$

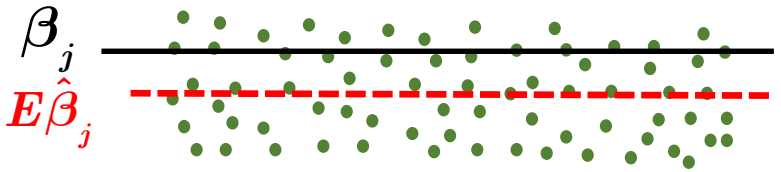
➢无偏性表明：样本回归模型的参数估计量是以总体参数的真实值为分布中心的随机变量，反复进行抽样估计即可得总体参数的真实值。




$E\hat{\beta}_j = \beta_j$（无偏估计）




$E\hat{\beta}_j > \beta_j$（系统性偏高）




$E\hat{\beta}_j < \beta_j$（系统性偏低）

# 有效性

- 所谓有效性，是指在$\beta_1$，$\beta_2$的所有线性无偏估计量中，最小二乘估计量$\hat{\beta}_1$，$\hat{\beta}_2$具有最小的方差

- 有效性表明：在一元线性回归模型的各种线性无偏估计量中，OLS估计量的效果最好。

在有效性的证明过程中，可以求解出OLS估计量的方差为：

$$Var(\hat{\beta}_1)=\frac{\sigma^2\sum X_i^2}{n\sum x_i^2}\xrightarrow{\text{简记为}}\sigma^2_{\hat{\beta}_1} \qquad Var(\hat{\beta}_2)=\frac{\sigma^2}{\sum x_i^2}\xrightarrow{\text{简记为}}\sigma^2_{\hat{\beta}_2}$$

# 总结：高斯－马尔科斯定理

在满足经典假设的条件下，线性回归模型的最小二乘估计量是具有最小方差的线性无偏估计量

称线性回归模型的OLS估计量为最优线性无偏估计量(The Best Linear Unbiased Estimator)，简称BLUE性质

# 参数估计量的概率分布及随机误差项方差的估计

## 1、参数估计量的概率分布

在解释变量非随机及随机扰动项服从正态分布的假定下，被解释变量$Y_i$服从正态分布，且彼此相互独立；

OLS估计量的线性性意味着 $\hat{\beta}_1$ 、$\hat{\beta}_2$分别是$Y_i$的线性组合，故必然服从正态分布。

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2 \sum X_i^2}{n\sum x_i^2}\right) \xrightarrow{\text{简记为}} N\left(\beta_1, \sigma_{\hat{\beta}_1}^2\right) \quad \text{若}\sigma^2\text{已知} \Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim N(0,1)$$

$$\hat{\beta}_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum x_i^2}\right) \xrightarrow{\text{简记为}} N\left(\beta_2, \sigma_{\hat{\beta}_2}^2\right) \quad \text{若}\sigma^2\text{已知} \Rightarrow \frac{\hat{\beta}_2 - \beta_2}{\sigma_{\hat{\beta}_2}} \sim N(0,1)$$

# 2、随机误差项$u$的方差$\sigma^2$的估计

上式中的$\sigma^2$称为总体方差，在没有总体信息时，随机误差项$u_i$ 不可观测，因此$\sigma^2$的值未知（厌恶参数），只能利用$u_i$ 的样本估计——即残差$e_i$来估计$\sigma^2$

可以证明，$\sigma^2$ 的无偏估计量为：$\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} e_i^2}{n-k}$

式中$k$为模型中待估参数个数，简单线性回归模型中的$k=2$;

$n-k$为残差平方和的自由度，在简单线性回归模型中为$n-2$ 。因为在计算$\sum_{i=1}^{n} e_i^2$ 时要受到$\sum_{i=1}^{n} e_i = 0$ 和$\sum_{i=1}^{n} e_i X_i = 0$共2个约束，故其自由度相应减少2

$\hat{\sigma} = \sqrt{\dfrac{\sum_{i=1}^{n} e_i^2}{n-k}}$称为回归标准误（$S.E.\ of\ Regression$）,它度量样本点偏离样本回归线的平均偏离程度（拟合误差）。

# 3、参数估计量的方差和标准差

由于$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2 \sum X_i^2}{n\sum x_i^2}\right)$，$\hat{\beta}_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum x_i^2}\right)$，且$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k}$，故有：

$$S_{\hat{\beta}_1}^2 = \frac{\hat{\sigma}^2 \sum X_i^2}{n\sum x_i^2} \Rightarrow S_{\hat{\beta}_1} = \hat{\sigma}\sqrt{\frac{\sum X_i^2}{n\sum x_i^2}}$$

$$S_{\hat{\beta}_2}^2 = \frac{\hat{\sigma}^2}{\sum x_i^2} \Rightarrow S_{\hat{\beta}_2} = \frac{\hat{\sigma}}{\sqrt{\sum x_i^2}}$$

故若总体方差已知:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2 \sum X_i^2}{n\sum x_i^2}\right) \xrightarrow{\text{简记为}} N\left(\beta_1, \sigma_{\hat{\beta}_1}^2\right) \Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim N(0,1)$$

$$\hat{\beta}_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum x_i^2}\right) \xrightarrow{\text{简记为}} N\left(\beta_2, \sigma_{\hat{\beta}_2}^2\right) \Rightarrow \frac{\hat{\beta}_2 - \beta_2}{\sigma_{\hat{\beta}_2}} \sim N(0,1)$$

若总体方差未知: $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k}, \quad E\left(\hat{\sigma}^2\right) = \sigma^2$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\hat{\sigma}^2 \sum X_i^2}{n\sum x_i^2}\right) \xrightarrow{\text{简记为}} N\left(\beta_1, S_{\hat{\beta}_1}^2\right) \Rightarrow \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \sim t(n-k)$$

$$\hat{\beta}_2 \sim N\left(\beta_2, \frac{\hat{\sigma}^2}{\sum x_i^2}\right) \xrightarrow{\text{简记为}} N\left(\beta_2, S_{\hat{\beta}_2}^2\right) \Rightarrow \frac{\hat{\beta}_2 - \beta_2}{S_{\hat{\beta}_2}} \sim t(n-k)$$

# 第五节、参数的显著性检验与区间估计

•回归分析的目的之一是要判断解释变量$X$是否是被解释变量$Y$的一个显著性的影响因素。

•在一元线性回归模型中，就是要判断自变量$X$是否对因变量$Y$具有显著的线性影响。这就需要进行变量的显著性检验。

■变量的显著性检验所应用的方法是数理统计学中的假设检验。

■在计量经计学中，主要是针对参数真值是否为零来进行显著性检验的。

# 1）参数的显著性检验

当我们只有样本信息时，前述分析表明$\hat{\beta}_j \sim N\left(\beta_j, S_{\hat{\beta}_j}^2\right) \Rightarrow t=\dfrac{\hat{\beta}_j-\beta_j}{S_{\hat{\beta}_j}} \sim t(n-k)$

$(j=1,2)$，由此即可进行参数的显著性检验：

1: 提出假设

$H_0$: $\beta_j=0(j=1,2)$

$H_1$: $\beta_j \neq 0$

2: 在$H_0$成立的前提下构造检验统计量：

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

3: 给定显著性水平$\alpha$，查表获得临界值$t_{\alpha/2}(n-k)$，确定拒绝域 $|t|>t_{\alpha/2}(n-k)$

4: 代入样本信息，计算出检验统计量值$t=\hat{\beta}_j / S_{\hat{\beta}_j}$，并与临界值比较：

(1)若$|t|>t_{\alpha/2}(n-k)$，落入拒绝域，拒绝$H_0$，即$\beta_j$ 显著异于零

(2)若$|t|<t_{\alpha/2}(n-k)$，未落入拒绝域，不能拒绝$H_0$，即$\beta_j$ 不显著

# 2）参数的显著性检验

在案例2.1中：

$\hat{\beta}_1$=2372.630，$S_{\hat{\beta}_1} = 546.0272 \Rightarrow t = 2372.630 / 546.0272 = 4.3453$

$\hat{\beta}_2$=0.6232，$S_{\hat{\beta}_2} = 0.0180 \Rightarrow t = 0.6232 / 0.0180 = 34.6310$

而$t_{0.025}(29)$=2.045，故$|t_1| > t_{0.025}(29)$，$|t_2| > t_{0.025}(29)$，即回归系数$\beta_1$、$\beta_2$均显著异于零

# 参数显著性检验的$p$值检验准则（双侧检验）

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$t$的伴随概率（也称为边际显著性水平），它是拒绝原假设的最低显著性水平。统计分析软件中通常都给出了相关检验的 $p$ 值

方法: 将给定的显著性水平 $\alpha$ 与$p$值进行比较:

准则: $p > \alpha$, 接受原假设　　$p < \alpha$, 拒绝原假设

含义: $p$表示拒绝原假设实际犯错误的概率

$\alpha$表示拒绝原假设容许犯错误的概率

若$p > \alpha$, 则拒绝原假设犯错误的概率超过了容许限度，故不能拒绝$H_0$

若$p < \alpha$, 则拒绝原假设犯错误的概率在容许限度内，可放心拒绝$H_0$

# 3）参数的置信区间估计

显著性检验可以通过一次抽样的结果检验总体参数可能的假设值（如是否为零），但它并没有指出在一次抽样中样本参数值到底离总体参数的真值有多“近”。

要判断样本参数的估计值在多大程度上可以“近似”地替代总体参数的真值，往往需要通过构造一个以样本参数的估计值为中心的“区间”，来考察它以多大的可能性（概率）包含着真实的参数值。这种方法就是参数检验的置信区间估计。

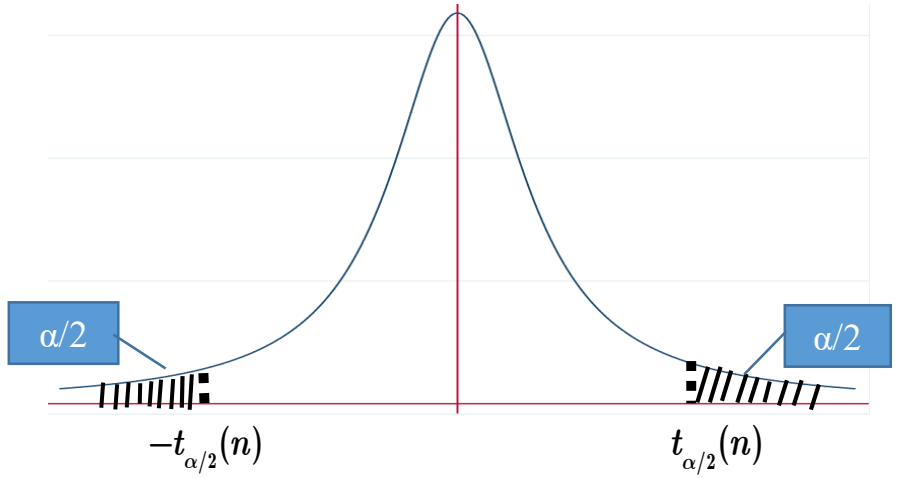
# 3）回归系数的置信区间

在最小二乘估计量统计性质的讨论中，已知$\hat{\beta}_j \sim N\left(\beta_j, S^2_{\hat{\beta}_j}\right) \Rightarrow t = \dfrac{\hat{\beta}_j - \beta_j}{S_{\hat{\beta}_j}} \sim t(n-k)$

根据t分布的对称性，易知：$P\left\{-t_{\alpha/2} < t < t_{\alpha/2}\right\} = 1-\alpha$

$$\Leftrightarrow P\left\{-t_{\alpha/2} < \frac{\hat{\beta}_j - \beta_j}{S_{\hat{\beta}_j}} < t_{\alpha/2}\right\} = 1-\alpha$$

$$\Leftrightarrow P\left\{\hat{\beta}_j - t_{\alpha/2}S_{\hat{\beta}_j} < \beta_j < \hat{\beta}_j + t_{\alpha/2}S_{\hat{\beta}_j}\right\} = 1-\alpha$$




故$\left(\hat{\beta}_j - t_{\alpha/2}S_{\hat{\beta}_j},\ \hat{\beta}_j + t_{\alpha/2}S_{\hat{\beta}_j}\right)$即为$\beta_j$的置信度为$1-\alpha$的置信区间，可简记为：

$$\hat{\beta}_j \mp t_{\alpha/2}(n-k)S_{\hat{\beta}_j}$$

在案例2.1中：

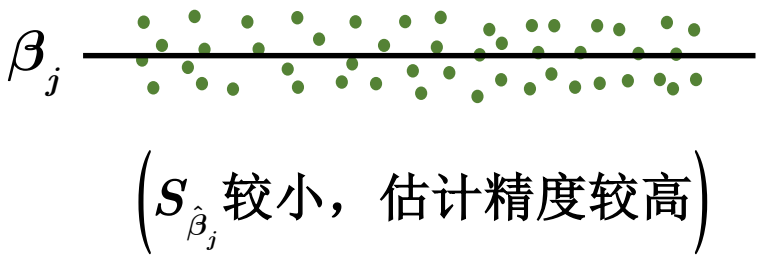
$\hat{\beta}_1$=2372.630，$S_{\hat{\beta}_1}=546.0272$，$t_{0.025}(29)=2.045$，故$\beta_1$的置信度为95%的置信区间为：

$$(2372.630-2.045\times546.0272,2372.630+2.045\times546.0272)=(1256.004,3489.256)$$

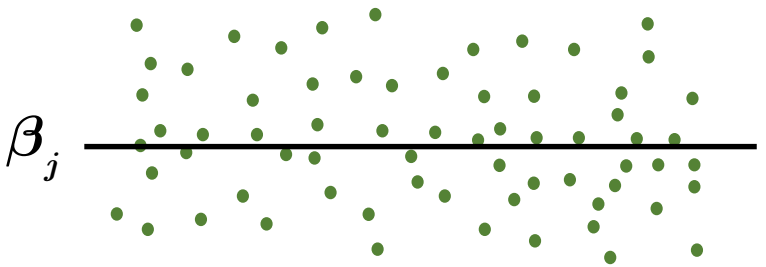
$\hat{\beta}_2$=0.6232，$S_{\hat{\beta}_2}=0.0180$，$t_{0.025}(29)=2.045$，故$\beta_2$的置信度为95%的置信区间为：

$$(0.6232-2.045\times0.0180,0.6232+2.045\times0.0180)=(0.5864,0.6600)$$

进一步，因$\beta_j$的置信度为$1-\alpha$的置信区间为$\left(\hat{\beta}_j-t_{\alpha/2}S_{\hat{\beta}_j},\hat{\beta}_j+t_{\alpha/2}S_{\hat{\beta}_j}\right)$，该区间的长度(估计精度)为$2t_{\alpha/2}(n-k)S_{\hat{\beta}_j}$。显然在给定$n$和$k$后，$S_{\hat{\beta}_j}$就刻画了参数$\beta_j$的估计精度，$S_{\hat{\beta}_j}$越小，估计精度越高。




($S_{\hat{\beta}_j}$较小，估计精度较高)



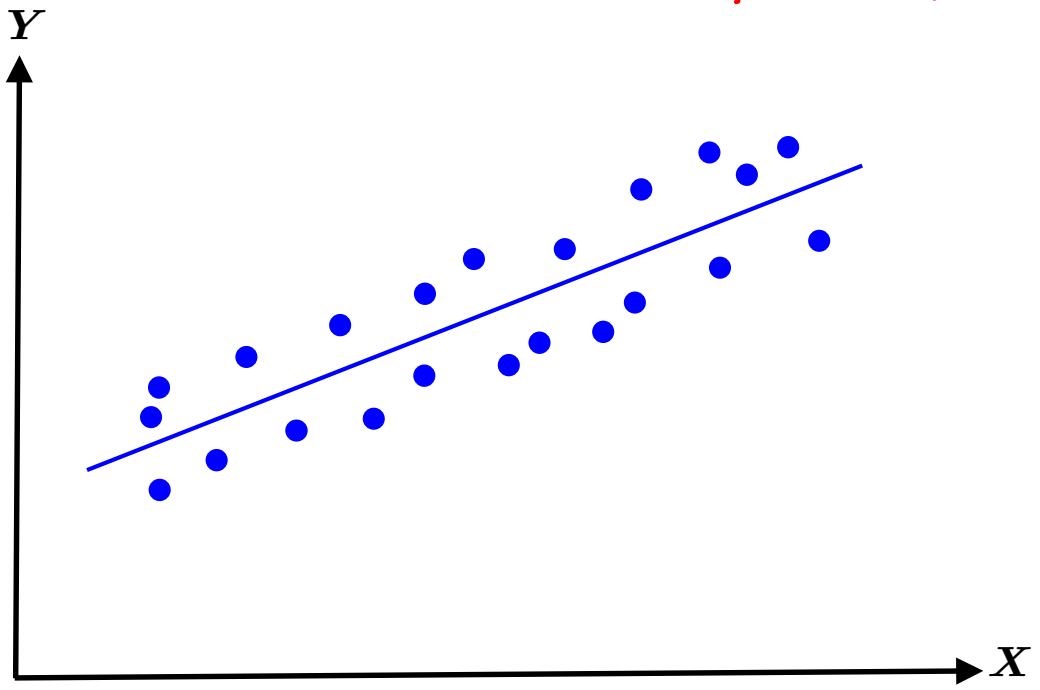

($S_{\hat{\beta}_j}$较大，估计精度较低)

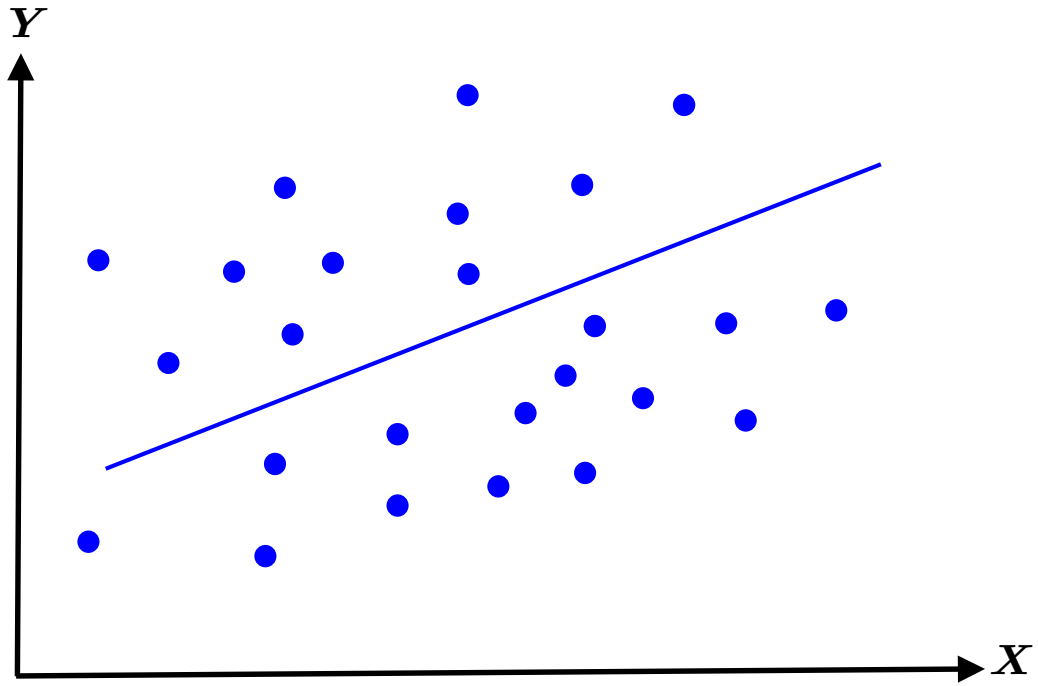
# 第四节、回归方程的拟合优度检验

本节基本内容:

- 什么是拟合优度
- 总变差的分解
- 可决系数

# 一、为什么要进行拟合优度检验?

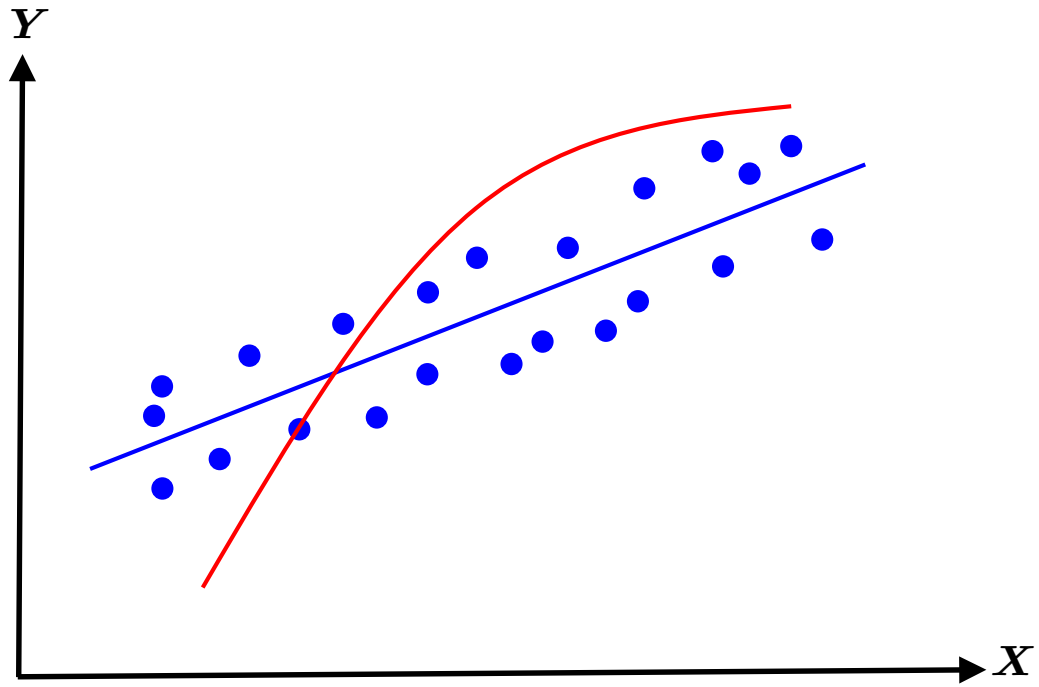



(1) 样本回归线对样本点的拟合程度究竟有多高？即衡量样本回归线对样本观测数据拟合的优劣程度。

样本回归线对样本点的拟合程度等价于回归模型中自变量对因变量的解释能力。

# 一、为什么要进行拟合优度检验?



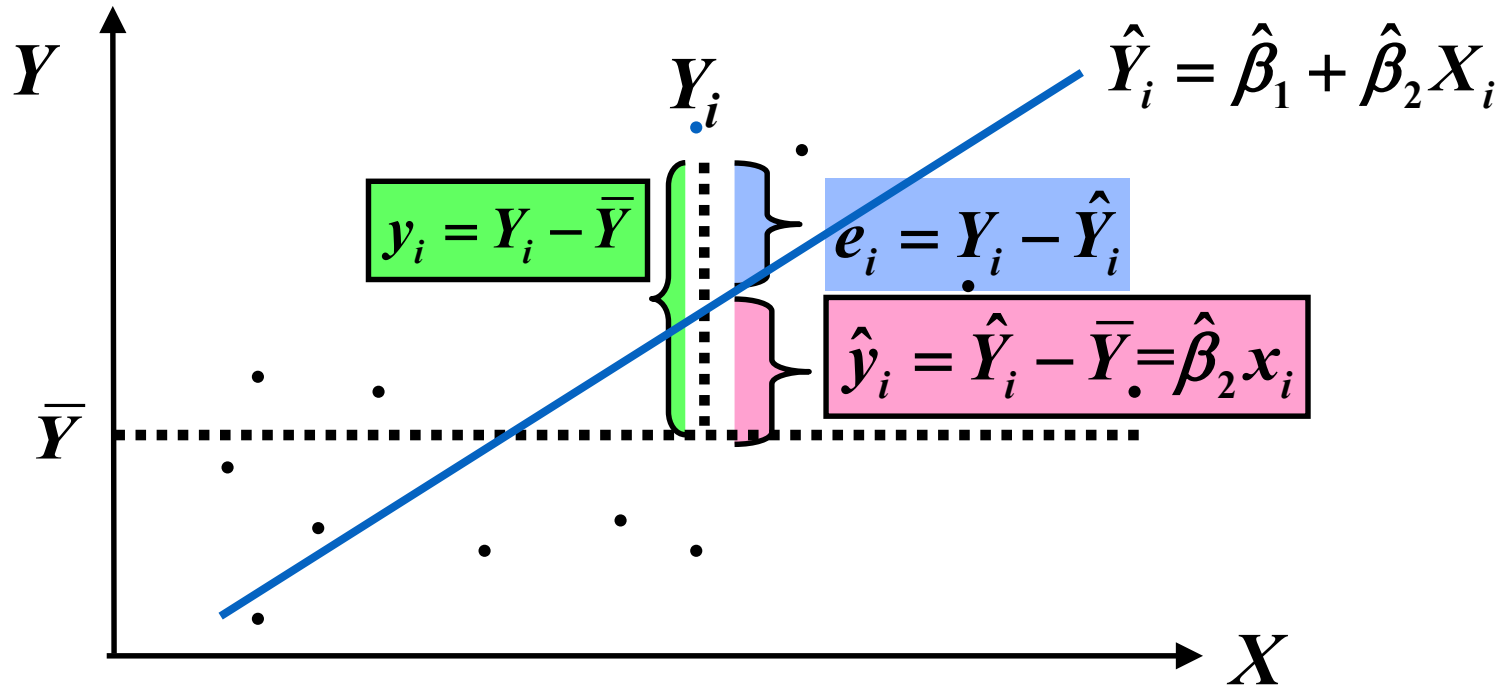

(2) 同一样本数据集，不同回归模型的拟合优度不一样，从而可以通过拟合优度的比较进行模型选择。（待比较的模型必须具有相同的研究对象，即相同的因变量）

## 拟合优度的度量建立在总离差平方和分解的基础上

# 1）总离差平方和的分解

已知由一组样本观测值（$X_i$，$Y_i$），$i=1,2,\ldots,n$ 得到如下样本回归线：

其关系如图：$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$



# 1）总离差平方和的分解

样本第$i$个观测值$Y_i$与样本均值$\bar{Y}$的离差$y_i$以分解为 $\hat{y}_i$和$e_i$ 两部分:

$$y_i = \hat{y}_i + e_i$$

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

为综合考虑样本回归线对所有样本点的拟合程度，进行如下运算:

$$y_i = \hat{y}_i + e_i \xrightarrow{\text{消除负号}} y_i^2 = (\hat{y}_i + e_i)^2 \xrightarrow{\text{所有样本点加总}} \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} (\hat{y}_i + e_i)^2$$

$$\xrightarrow{\text{展开}} = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2 + 2\sum_{i=1}^{n} e_i \hat{y}_i \xrightarrow{\text{利用性质6}} = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2$$

# 1）总离差平方和的分解

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 \overset{\hat{y}=\hat{\beta}_2 x_i}{=} \sum_{i=1}^{n} \hat{\beta}_2^2 x_i^2 + \sum_{i=1}^{n} e_i^2$$

$$\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

$$TSS = ESS + RSS$$

总离差平方和 = 回归平方和 + 残差平方和

衡量了因变量（围绕均值取值）的波动性

衡量了在因变量的波动性中，能够由自变量的波动（$x_i^2$）进行解释的部分

衡量了在因变量的波动性中，不能由自变量的波动，只能由其它随机影响因素的波动进行解释的部分

# 2）样本可决系数

- 在给定样本中，*TSS*不变。如果*ESS*在*TSS*中占的比重越大，这说明样本回归线对样本值的拟合优度越好，自变量对因变量的解释能力越强。令:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

即，$$R^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = 1 - \frac{\sum e_i^2}{\sum y_i^2}$$

或，$$R^2 = \hat{\beta}_1^2 \frac{\sum x_i^2}{\sum y_i^2} = \frac{\left(\sum x_i y_i\right)^2}{\sum x_i^2 \sum y_i^2}$$

## 2）样本可决系数

则称$R^2$为样本可决系数，或决定系数、判定系数

由$R^2$的定义知：其取值范围为[0，1]。$R^2$越接近1，说明被解释变量的实际观测值离样本回归线越近，即样本回归线对样本数据的拟合程度越高，也即解释变量$X$对被解释变量$Y$的解释力越强。

$R^2$的经济含义是：在被解释变量$Y$的波动中，有$R^2$（%）的部分可由自变量$X$的波动得到解释。

在简单线性回归模型中，$R^2$ 在数值上等于自变量$X$ 与因变量 $Y$ 之间简单线性相关系数$\rho_{XY}$的平方，同时也等于因变量 $Y$ 与其拟合值 $\hat{Y}$ 之间简单线性相关系数$\rho_{Y\hat{Y}}$的平方，即$R^2=\rho_{XY}^2=\rho_{Y\hat{Y}}^2$