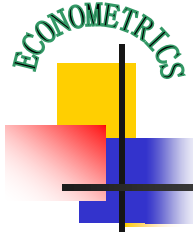


第二章
一元线性回归模型

一元线性回归模型要点

一、单方程一元线性计量经济学模型的特征

二、经典线性模型的基本假设

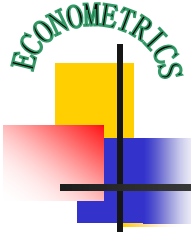
三、最小二乘原理（OLS）与最大似然原理（ML）

四、一元线性模型最小二乘估计量统计性质的证明

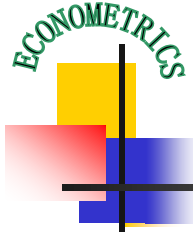
五、一元线性回归模型的统计检验

（拟合优度检验、变量的显著性检验、参数的置信区间）

六、预测问题：总体条件均值与个值预测值的置信区间



一元线性回归模型



§2.1 回归分析概述

一、变量间的关系及回归分析的基本概念

二、总体回归函数（PRF）

三、随机扰动项

四、样本回归函数（SRF）

一、变量间的关系及回归分析的基本概念

1. 变量间的关系

（1）**确定性关系**或**函数关系：**研究的是确定现象非随机变量间的关系。

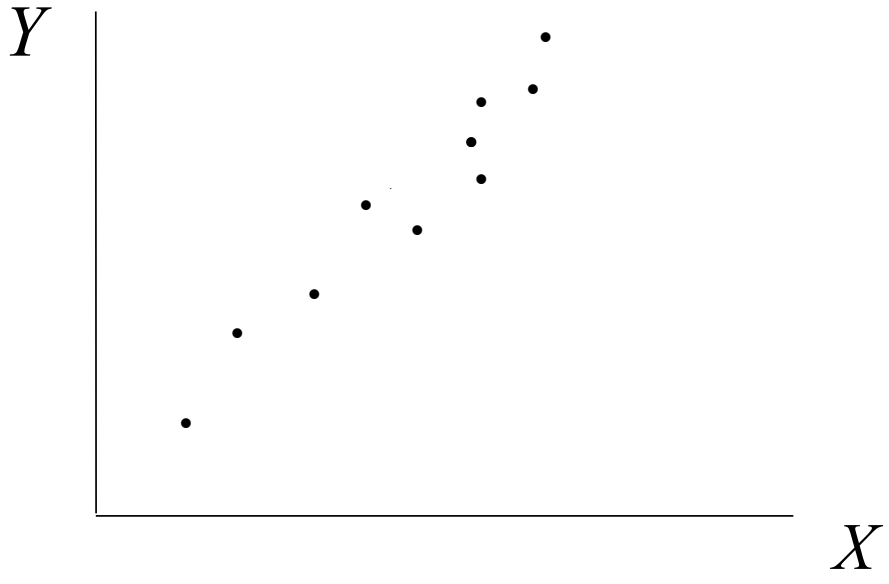
$$\text{圆面积} = f(\pi, \text{半径}) = \pi \cdot \text{半径}^2$$

（2）统计依赖或**相关关系：**研究的是非确定现象随机变量间的关系。

$$\text{农作物产量} = f(\text{气温}, \text{降雨量}, \text{阳光}, \text{施肥量})$$

◆ 相关关系的描述

相关关系最直观的描述方式——坐标图（散布图）



相关程度的度量—相关系数

总体线性相关系数:

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

其中：$\mathrm{Var}(X)$——*X* 的方差；$\mathrm{Var}(Y)$ ——*Y*的方差

$\mathrm{Cov}(X,Y)$ ——*X*和*Y*的协方差

样本线性相关系数:

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

其中：X_i 和 分别是变量 X 和 Y 的样本观测值

$\bar{X}$ 和 分别是变量 X 和 Y 样本值的平均值

ECONOMETRICS

◆相关关系的类型

- 从涉及的变量数量看

  简单相关（两个变量）

  多重相关（复相关）（三个以上变量）

- 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（曲线相关）

- 从变量相关关系变化的方向看

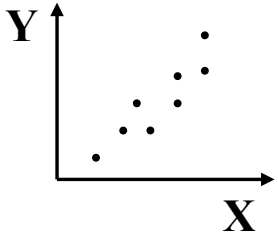
  正相关——变量向同方向变化，即同增同减

  负相关——变量向相反方向变化，即一增一减

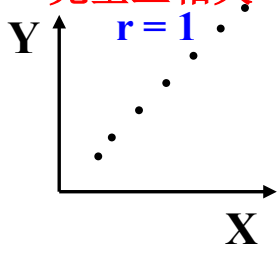
  不相关—— 两个变量的变化相互没有关系。

特别指出：相关关系是变量之间所表现出来的一种纯数学关系，判断变量之间是否具有相关关系的依据只有数据。一般用相关系数来度量变量之间线性相关的程度。

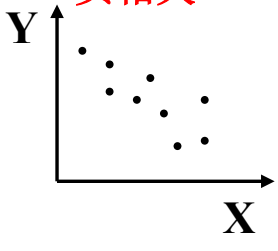
正相关



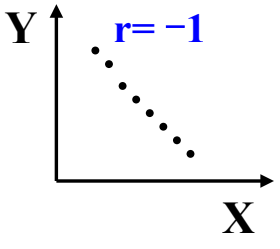
完全正相关



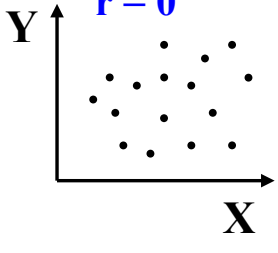
负相关



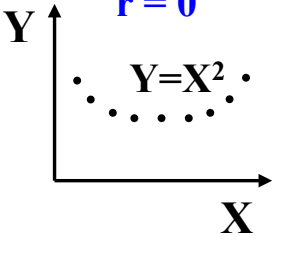
完全负相关



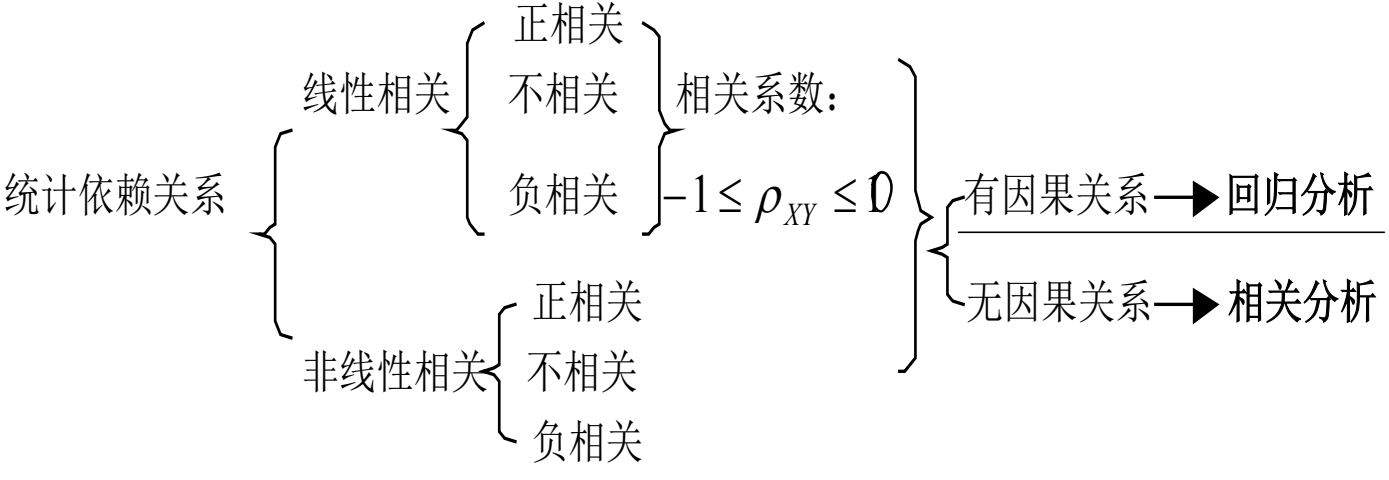
不相关



非线性相关



- 对变量间统计依赖关系的考察主要是通过相关分析(correlation analysis)或回归分析(regression analysis)来完成的



- **注意：**

①不线性相关并不意味着不相关。

②有相关关系并不意味着一定有因果关系。

③回归分析/相关分析研究一个变量对另一个（些）变量的统计依赖关系，但它们并不意味着一定有因果关系。

④相关分析对称地对待任何（两个）变量，两个变量都被看作是随机的。回归分析对变量的处理方法存在不对称性，即区分应变量（被解释变量）和自变量（解释变量）：前者是随机变量，后者不是。

2. 回归分析的基本概念

- **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**。
- **被解释变量**（Explained Variable）或**应变量**（Dependent Variable）。
- **解释变量**（Explanatory Variable）或**自变量**（Independent Variable）。

见P24

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

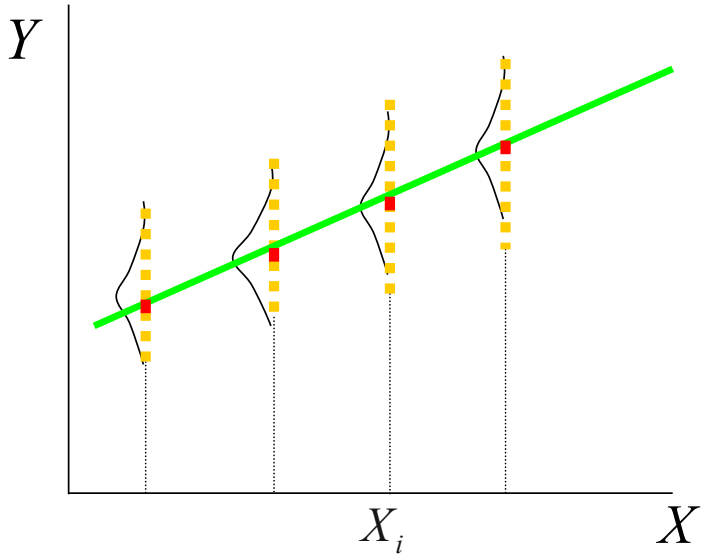
概念：条件分布和条件期望

- Y 的条件分布

  当解释变量X 取某固定值时（条件），Y 的值不确定，Y 的不同取值形成一定的分布，即 Y 的条件分布。

- 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 $\mathrm{E}(Y|X_i)$

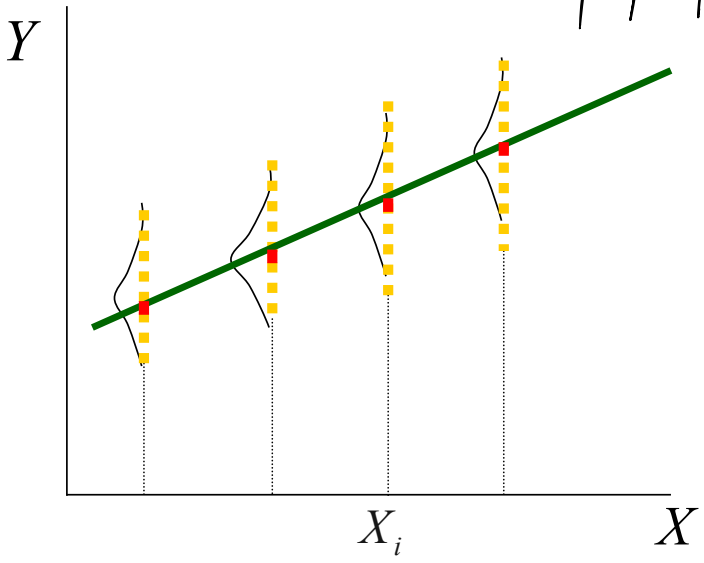


回归线与回归函数

总体回归函数 PRF

- **回归线:**

对于每一个 X 的取值，都有 Y 的条件期望 $\mathrm{E}(Y|X_i)$ 与之对应，代表这些 Y 的条件期望的点的轨迹所形成的直线或曲线，称为回归线。



回归线与回归函数

回归函数：应变量 Y 的条件期望 $\mathrm{E}(Y|X_i)$ 随解释变量 X 的的变化而有规律的变化，如果把 Y 的条件期望 $\mathrm{E}(Y|X_i)$ 表现为 X 的某种函数

$$\mathrm{E}(Y|X_i) = f(X_i)$$

这个函数称为回归函数。

回归函数分为：总体回归函数和样本回归函数

- **回归分析构成计量经济学的方法论基础，其主要内容包括：**

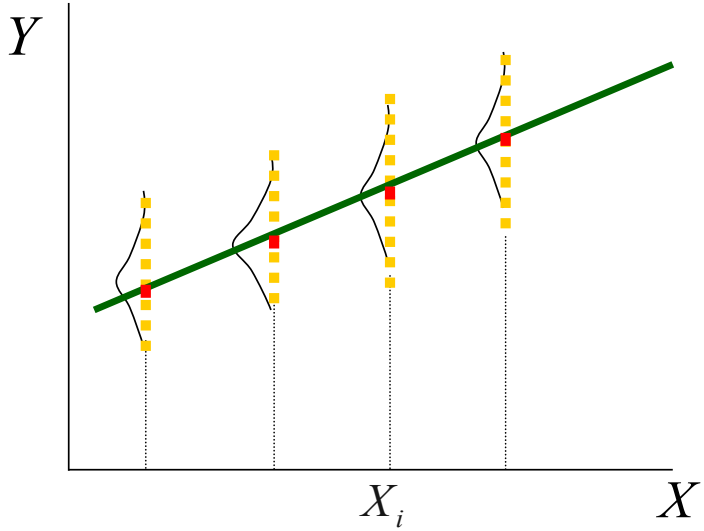
（1）根据样本观察值对经济计量模型参数进行估计，求得回归方程；

（2）对回归方程、参数估计值进行显著性检验；

（3）利用回归方程进行分析、评价及预测。

二、总体回归函数

- **回归分析**关心的是根据解释变量的已知或给定值，考察被解释变量的总体均值，即当解释变量取某个确定值时，与之统计相关的被解释变量所有可能出现的对应值的平均值。



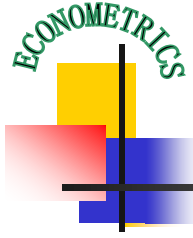
- **例2.1：**一个假想的社区有100户家庭组成，要研究该社区每月**家庭消费支出Y**与每月**家庭可支配收入X**的关系。 即如果知道了家庭的月收入，能否预测该社区家庭的平均月消费支出水平。

  为达到此目的，将该100户家庭划分为组内收入差不多的10组，以分析每一收入组的家庭消费支出。

见P24

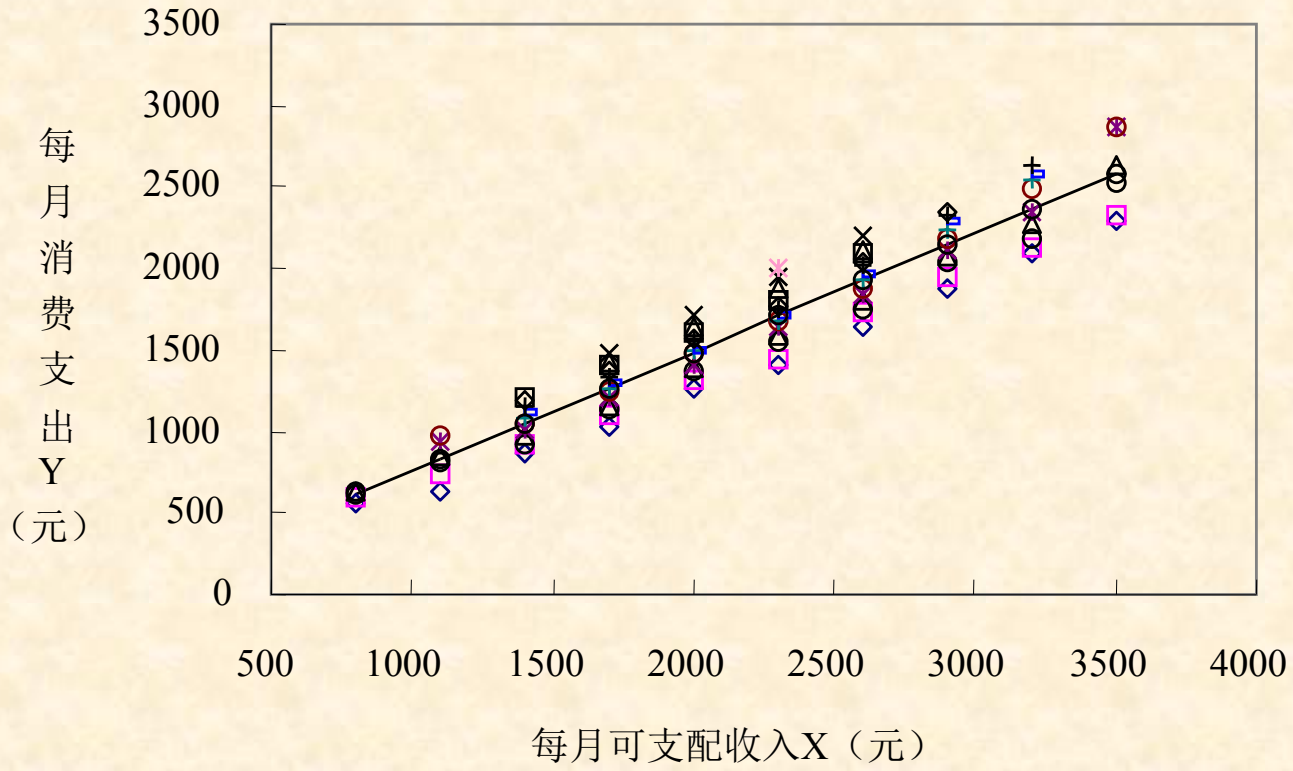
表 2.1.1　某社区家庭每月收入与消费支出统计表

	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



- 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；
- 但由于调查的完备性，给定收入水平X的消费支出Y的分布是确定的，即以X的给定值为条件的Y的**条件分布**（**Conditional distribution**）是已知的，例如：P(Y=561|X=800）=1/4。

- 因此，给定收入X的值Xi，可得消费支出Y的**条件均值**（conditional mean）或**条件期望**（conditional expectation）：E(Y|X=Xi)。
- 该例中：E(Y | X=800)=605
- 描出散点图发现：随着收入的增加，消费“**平均地说**”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**。


3500
3000
2500
2000
1500
1000
500
0
每月消费支出Y（元）
500
1000
1500
2000
2500
3000
3500
4000
每月可支配收入X（元）

- 在给定解释变量X_i条件下被解释变量Y_i的期望轨迹称为**总体回归线**（population regression line），或更一般地称为**总体回归曲线**（population regression curve）。

- 相应的函数：

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

称为（双变量）**总体回归函数**（population regression function, **PRF**）。

- **含义：** 回归函数（PRF）说明被解释变量Y的平均状态（总体条件期望）随解释变量X变化的规律。

- **函数形式：** 可以是线性或非线性的。

- 例2.1中，将居民消费支出看成是其可支配收入的线性函数时：

$$E(Y \mid X_i) = \beta_0 + \beta_1 X_i$$

为一**线性函数**。其中，β_0，β_1是未知参数，称为**回归系数**（regression coefficients）。

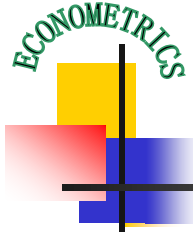
三、随机扰动项

- 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。
- 称为观察值围绕它的期望值的**离差**（**deviation**），是一个不可观测的随机变量，又称为**随机干扰项**（**stochastic disturbance**）或**随机误差项**（**stochastic error**）。

$$\mu_i = Y_i - E(Y \mid X_i)$$

- 例2.1中，给定收入水平X_i，个别家庭的支出可表示为两部分之和：（1）该收入水平下所有家庭的平均消费支出$E(Y|X_i)$，称为**系统性（systematic）**或**确定性（deterministic)部分**；（2）其他**随机或非确定性（nonsystematic)部分**μ_i。

$$Y_i = E(Y \mid X_i) + \mu_i = \beta_0 + \beta_1 X_i + \mu_i$$



- 称为总体回归函数（PRF）的随机设定形式。表明被解释变量除了受解释变量的系统性影响外，还受其他因素的随机性影响。由于方程中引入了随机项，成为计量经济学模型，因此也称为总体回归模型。

- 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；
  - 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；
  - 其他随机因素的影响。
- 产生并设计随机误差项的主要原因：
  - 理论的含糊性；
  - 数据的欠缺；
  - 节省原则。

四、样本回归函数（SRF）

- **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？

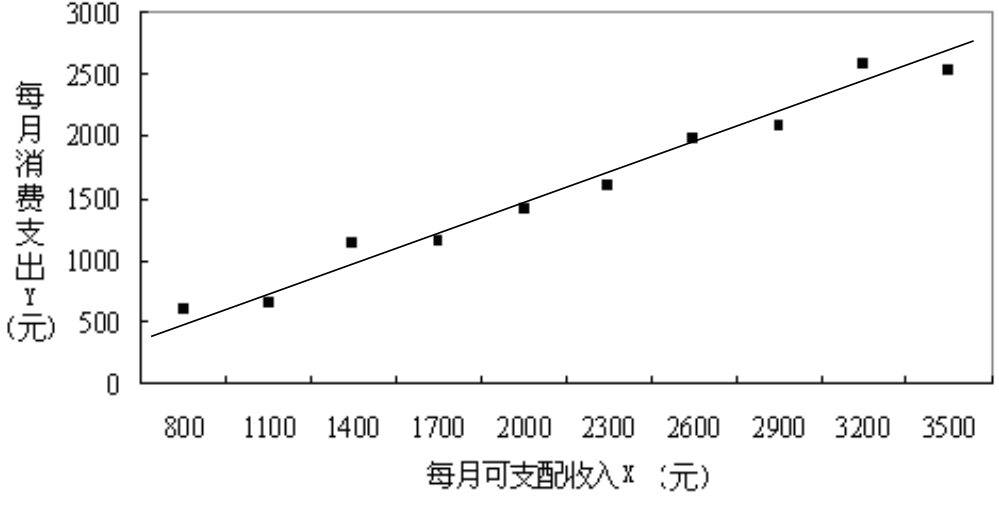
- **例2.2：**在例2.1的总体中有如下一个样本，能否从该样本估计总体回归函数PRF？

表 2.1.3　家庭消费支出与可支配收入的一个随机样本

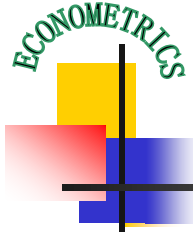
Y	800	1100	1400	1700	2000	2300	2600	2900	3200	3500
X	594	638	1122	1155	1408	1595	1969	2078	2585	2530

回答：能

- 该样本的**散点图**（scatter diagram）：



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



- 记样本回归线的函数形式为：

$$\hat{Y}_i = f(X_i) = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

称为**样本回归函数**（**sample regression function，SRF**）。

注意：这里将样本回归线看成总体回归线的近似替代

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$\Downarrow$$

$$\begin{aligned} Y_i &= E(Y \mid X_i) + \mu_i \\ &= \beta_0 + \beta_1 X_i + \mu_i \end{aligned}$$

则

$\hat{Y}_i$ 为 $E(Y \mid X_i)$ 的估计量；

$\hat{\beta}_i$ 为 β_i 的估计量，$i = (0,1)$

ECONOMETRICS

样本回归函数的随机形式/样本回归模型：

同样地，样本回归函数也有如下的随机形式：

$$Y_i = \hat{Y}_i + \hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$$

式中，e_i 称为**（样本）残差**（或**剩余**）**项**（residual），代表了其他影响 Y_i 的随机因素的集合，可看成是 μ_i 的估计量 $\hat{\mu}_i$。

由于方程中引入了随机项，成为计量经济模型，因此也称为**样本回归模型**（**sample regression model**）。

▼**回归分析的主要目的**：根据样本回归函数SRF，估计总体回归函数PRF。

即，根据 $Y_i = \hat{Y}_i + e_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$ **SRF**

估计 $Y_i = E(Y \mid X_i) + \mu_i = \beta_0 + \beta_1 X_i + \mu_i$ **PRF**

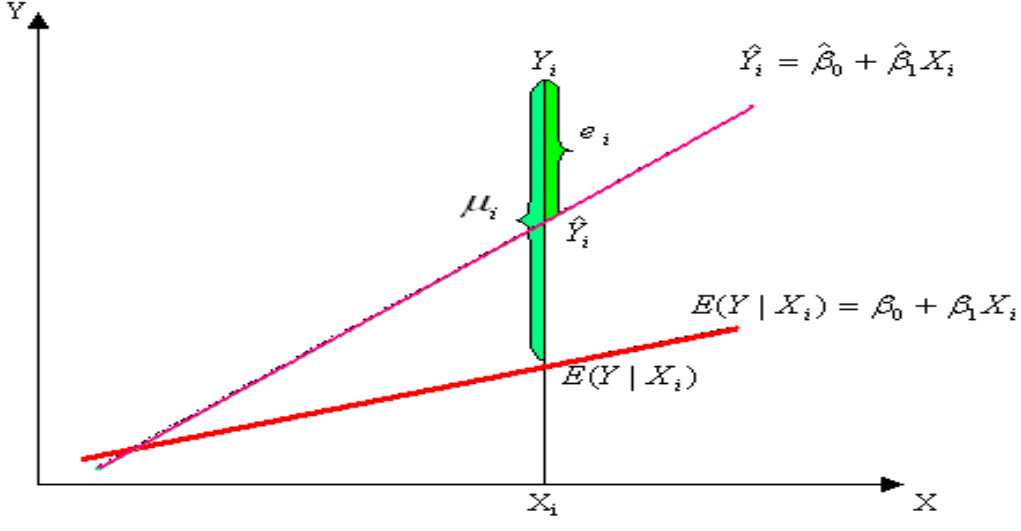


图 2.1.3 总体回归线与样本回归线的基本关系

这就要求：

设计一“方法”构造SRF，以使SRF尽可能“接近”PRF，或者说使$\hat{\beta}_i(i=0,1)$尽可能接近$\beta_i(i=0,1)$。

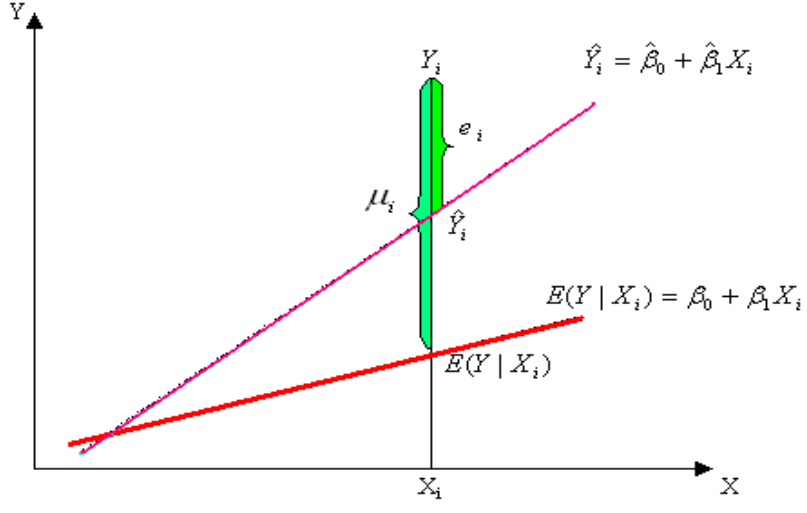


图 2.1.3　总体回归线与样本回归线的基本关系

注意：这里总体回归函数PRF可能永远无法知道。

一元线性回归模型的参数估计

- 一元线性回归模型的基本假设
- 参数的普通最小二乘估计（OLS）
- 参数估计的最大或然法(ML)
- 最小二乘估计量的性质
- 参数估计量的概率分布及随机干扰项方差的估计

ECONOMETRICS

说 明：

- 单方程计量经济学模型分为两大类：线性模型和非线性模型
- 线性模型中，变量之间的关系呈线性关系
- 非线性模型中，变量之间的关系呈非线性关系
- 一元线性回归模型：只有一个解释变量

$$Y_i = \beta_0 + \beta_1 X_i + \mu_i \qquad i=1,2,\ldots,n$$

Y为被解释变量，X为解释变量，β_0与β_1为**待估参数**，μ为**随机干扰项**

- **回归分析的主要目的**是要通过样本回归函数（模型）SRF尽可能准确地估计总体回归函数（模型）PRF。
- **估计方法**有多种，其中最广泛使用的是**普通最小二乘法**（ordinary least squares, OLS）。
- 为保证参数估计量具有良好的性质，通常对模型提出若干基本假设。
- 实际这些假设与所采用的估计方法紧密相关。

§2.2一元线性回归模型的基本假设

假设1.回归模型是正确设定的

假设2.解释变量X在所抽取的样本中具有变异性，而且随着样本容量的无限增加，解释变量X的样本方差趋于一个非零的有限常数，即

$$\sum (X_i - \overline{X})^2 / n \to Q, \qquad n \to \infty$$

假设3. 给定解释变量X的任何值，随机误差项μ_i的均值为零，即

$$E(\mu_i|X) = 0 \qquad i=1,2,\ldots,n$$

于是：随机误差项μ_i的无条件期望 $E(\mu_i) = 0$

同时，随机误差项μ_i与解释变量X之间不相关：

$$Cov(X_i, \mu_i) = 0 \qquad i = 1,2, \ldots, n$$

此时，也称X是同期外生的，或X与μ同期不相关。

假设4.随机误差项μ具有给定任何条件下同方差和不序列相关性：

$$Var(\mu_i|X) = \sigma_\mu^2 \qquad i = 1,2, \ldots, n$$

$$Cov(\mu_i, \mu_j|X) = 0 \qquad i \neq j \quad i,j = 1,2, \ldots, n$$

于是，随机误差项μ无条件同方差和不序列相关性也满足，即

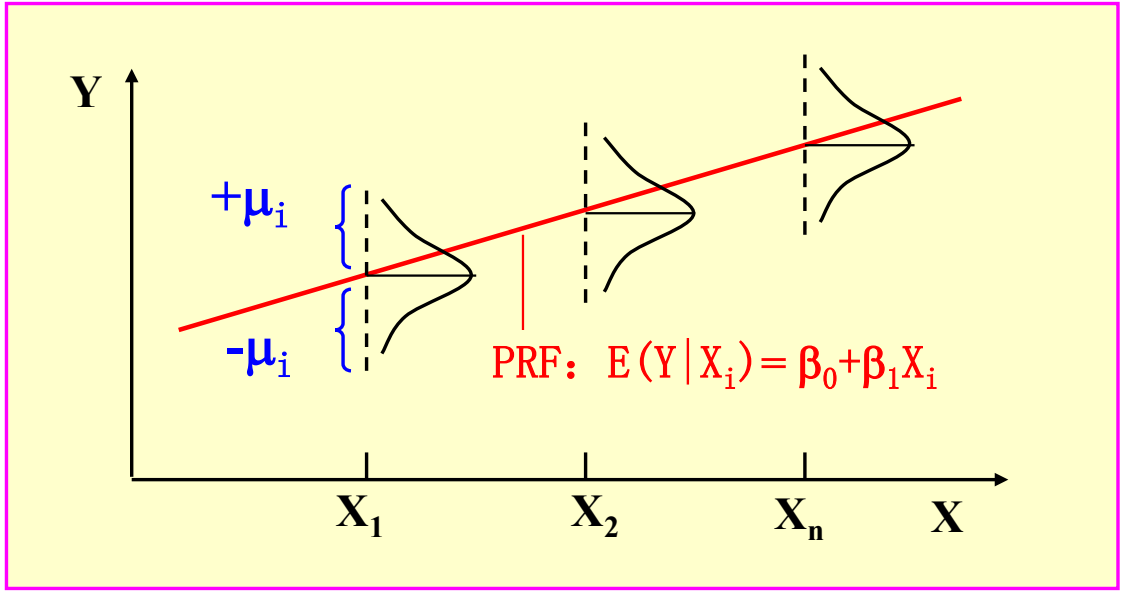
$$Var(\mu_i)=\sigma_\mu^2 \quad , \quad Cov(\mu_i, \mu_j)=0$$

假设5. 随机干扰项μ服从零均值、同方差的正态分布

$$\mu_i|X \sim N(0, \sigma_\mu^2) \qquad i=1,2,\ldots,n$$

以上假设也称为线性回归模型的经典假设或高斯（Gauss）假设，满足该假设的线性回归模型，也称为**经典线性回归模型**（Classical Linear Regression Model, CLRM）。

由假设3，4. 随机误差项μ具有零均值、同方差和不序列相关性：$E(\mu_i)=0$, $Var(\mu_i)=\sigma_\mu^2$,$Cov(\mu_i, \mu_j)=0$　　$i\neq j$



Y_i与μ_i同分布：

模型的一般形式：$Y_i=\beta_0+\beta_1X_i+\mu_i \quad i=1,2,\cdots,n$

$E(\mu_i)=0 \longrightarrow E(Y_i)=\beta_0+\beta_1X_i$

$Var(\mu_i)=\sigma^2 \longrightarrow Var(Y_i)=\sigma^2$

$Cov(\mu_i,\ \mu_j)=0 \longrightarrow Cov(Y_i,\ Y_j)=0$

$\mu_i\sim N(0,\ \sigma^2) \longrightarrow Y_i\sim N(\beta_0+\beta_1X_i,\ \sigma^2)$

事实上：$Var(\mu_i)=E\left[\mu_i-E(\mu_i)\right]^2=E(\mu_i^2)=\sigma^2$

$$Var(Y_i)=E\left[Y_i-E(Y_i)\right]^2=E\left[\beta_0+\beta_1X_i+\mu_i-(\beta_0+\beta_1X_i)\right]^2$$

$$=E(\mu_i^2)=\sigma^2$$

$$Cov(Y_i,Y_j)=E[Y_i-E(Y_i)][Y_j-E(Y_j)]=E(\mu_i\mu_j)=E(\mu_i)E(\mu_j)=0$$

§2.3一元线性回归模型的参数估计

给定一组样本观测值（X_i, Y_i）（i=1,2,...n）要求样本回归函数尽可能好地拟合这组值.

普通最小二乘法（Ordinary least squares, OLS）给出的判断标准是：二者之差的平方和

$$Q=\sum_{1}^{n}(Y_i-\hat{Y}_i)^2=\sum_{1}^{n}(Y_i-(\hat{\beta}_0+\hat{\beta}_1X_i))^2$$ 最小。

即在给定样本观测值之下，选择出$\hat{\beta}_0$、$\hat{\beta}_1$ 能使 Y_i 与$\hat{Y}_i$之差的平方和最小。

$$Q=\sum_{1}^{n}(Y_i-\hat{Y}_i)^2=\sum_{1}^{n}(Y_i-(\hat{\beta}_0+\hat{\beta}_1X_i))^2$$

根据微分运算，可推得用于估计$\hat{\beta}_0$、$\hat{\beta}_1$的下列方程组：

$$\begin{cases}\sum(\hat{\beta}_0+\hat{\beta}_1X_i-Y_i)=0\\ \sum(\hat{\beta}_0+\hat{\beta}_1X_i-Y_i)X_i=0\end{cases}\qquad(*)$$

$$\sum e_i=0$$

$$\sum X_ie_i=0$$

或

$$\begin{cases}\Sigma Y_i=n\hat{\beta}_0+\hat{\beta}_1\Sigma X_i\\ \Sigma Y_iX_i=\hat{\beta}_0\Sigma X_i+\hat{\beta}_1\Sigma X_i^2\end{cases}$$

解得：

$$\begin{cases}\hat{\beta}_0=\dfrac{\Sigma X_i^2\Sigma Y_i-\Sigma X_i\Sigma Y_iX_i}{n\Sigma X_i^2-(\Sigma X_i)^2}\\ \hat{\beta}_1=\dfrac{n\Sigma Y_iX_i-\Sigma Y_i\Sigma X_i}{n\Sigma X_i^2-(\Sigma X_i)^2}\end{cases}$$

方程组（*）称为**正规方程组**（**normal equations**）。

$$\Sigma Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \Sigma X_i$$

$$\begin{cases} \hat{\beta}_1 = \dfrac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} \\ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \end{cases}$$

记

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

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

上述参数估计量可以写成：

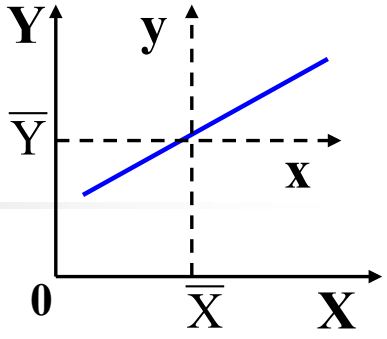
$$\begin{cases} \hat{\beta}_1 = \dfrac{\Sigma x_i y_i}{\Sigma x_i^2} \\ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \end{cases}$$

协方差

方差

称为OLS估计量的**离差形式**（**deviation form**）。

由于参数的估计结果是通过最小二乘法得到的，故称为**普通最小二乘估计量**（**ordinary least squares estimators**）**（点估计）**。



顺便指出，记 $\hat{y}_i = \hat{Y}_i - \overline{Y}$

则有 $\hat{y}_i = (\hat{\beta}_0 + \hat{\beta}_1 X_i) - (\hat{\beta}_0 + \hat{\beta}_1 \overline{X} + \overline{e})$

$$= \hat{\beta}_1 (X_i - \overline{X}) - \frac{1}{n}\sum e_i$$

可得 $\hat{y}_i = \hat{\beta}_1 x_i$ （**）

（**）式也称为**样本回归函数**的**离差形式**。

注意：

在计量经济学中，往往以**小写字母表示对均值的离差**。

参数估计的最大或然法(ML)

最大或然法(Maximum Likelihood,简称ML)，也称**最大似然法**，是不同于最小二乘法的另一种参数估计方法，是从最大或然原理出发发展起来的其他估计方法的基础。

基本原理：

对于**最大或然法**，当从模型总体随机抽取n组样本观测值后，最合理的参数估计量应该使得从模型中抽取该n组样本观测值的概率最大。

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

在满足基本假设条件下，对一元线性回归模型：

$$Y_i=\beta_0+\beta_1X_i+\mu_i$$

随机抽取n组样本观测值（X_i, Y_i）（i=1,2,…n）。

假如模型的参数估计量已经求得，为 $\hat{\beta}_0$ 、$\hat{\beta}_1$

那么Y_i服从如下的正态分布：

$$Y_i\sim N(\hat{\beta}_0+\hat{\beta}_1X_i,\sigma^2)$$

于是，Y的概率函数为

$$P(Y_i)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)^2}\qquad (i=1,2,\ldots n)$$

因为Y_i是相互独立的，所以的所有样本观测值的联合概率，也即**或然函数**(likelihood function)为：

$$L(\hat{\beta}_0,\hat{\beta}_1,\sigma^2)=P(Y_1,Y_2,\cdots,Y_n)$$

$$=\frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}e^{-\frac{1}{2\sigma^2}\Sigma(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)^2}$$

将该或然函数极大化，即可求得到模型参数的极大或然估计量。

由于或然函数的极大化与或然函数的对数的极大化是等价的，所以，取对数或然函数如下：

$$L^* = \ln(L)$$

$$= -n\ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

对L^*求极大值，等价于对$\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$求极小值:

$$\begin{cases} \dfrac{\partial}{\partial \hat{\beta}_0}\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = 0 \\ \dfrac{\partial}{\partial \hat{\beta}_1}\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = 0 \end{cases}$$

解得模型的参数估计量为：

$$\begin{cases} \hat{\beta}_0 = \dfrac{\Sigma X_i^2 \Sigma Y_i - \Sigma X_i \Sigma Y_i X_i}{n\Sigma X_i^2 - (\Sigma X_i)^2} \\ \hat{\beta}_1 = \dfrac{n\Sigma Y_i X_i - \Sigma Y_i \Sigma X_i}{n\Sigma X_i^2 - (\Sigma X_i)^2} \end{cases}$$

可见，在满足一系列基本假设的情况下，模型结构参数的**最大或然估计量**与**普通最小二乘估计量**是相同的。

例2.2.1：在上述家庭**可支配收入-消费支出**例中，对于所抽出的一组样本数，参数估计的计算可通过下面的表2.2.1进行。

表 2.2.1　　参数估计的计算表

	X_i	Y_i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2	X_i^2	Y_i^2
1	800	594	-1			1822500	947508	640000	352836
2	1100	638	-1			1102500	863784	1210000	407044
3	1400	1122	-			562500	198381	1960000	1258884
4	1700	1155	-			202500	170074	2890000	1334025
5	2000	1408	-			22500	25408	4000000	1982464
6	2300	1595	150	28	4140	22500	762	5290000	2544025
7	2600	1969	450	402	1807				376961
8	2900	2078	750	511	3829				318084
9	3200	2585	1050	1018	10684				382225
10	3500	2530	1350	963	12995				400900
求和	21500	15674			5769300	7425000	4590020	53650000	29157448
平均	2150	1567							

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum x_i y_i}{\sum x_i^2} \\ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \end{cases}$$

$$\begin{cases} \hat{\beta}_1 = \dfrac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} \\ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \end{cases}$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{5769300}{7425000} = 0.777$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_0 \overline{X} = 1567 - 0.777 \times 2150 = -103.172$$

因此，由该样本估计的回归方程为：

$$\hat{Y}_i = -103.172 + 0.777 X_i$$

Dependent Variable: Y
Method: Least Squares
Date: 02/02/09 Time: 17:04
Sample: 1 10
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X	0.777010	0.042485	18.28900	0.0000
C	-103.1717	98.40598	-1.048429	0.3251

R-squared	0.976641	Mean dependent var	1567.400
Adjusted R-squared	0.973722	S.D. dependent var	714.1444
S.E. of regression	115.7670	Akaike info criterion	12.51789
Sum squared resid	107216.0	Schwarz criterion	12.57841
Log likelihood	-60.58946	Hannan-Quinn criter.	12.45151
F-statistic	334.4876	Durbin-Watson stat	3.120320
Prob(F-statistic)	0.000000		

最小二乘估计量的性质(高斯—马尔可夫定理)

当模型参数估计出后，需考虑参数估计值的精度，即是否能代表总体参数的真值，或者说需考察参数估计量的统计性质。

一个用于考察总体的估计量，可从如下几个方面考察其优劣性：

（1）线性性，即它是否是另一随机变量的线性函数；

（2）无偏性，即它的均值或期望值是否等于总体的真实值；

（3）有效性，即它是否在所有线性无偏估计量中具有最小方差。

这三个准则也称作估计量的小样本性质。

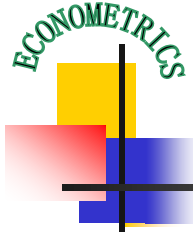
拥有这类性质的估计量称为最佳线性无偏估计量（**best liner unbiased estimator, BLUE**）。

当不满足小样本性质时，需进一步考察估计量的**大样本或渐近性质**：

（4）渐近无偏性，即样本容量趋于无穷大时，是否它的均值序列趋于总体真值；

（5）一致性，即样本容量趋于无穷大时，它是否依概率收敛于总体的真值；

（6）渐近有效性，即样本容量趋于无穷大时，是否它在所有的一致估计量中具有最小的渐近方差。



高斯—马尔可夫定理

高斯—马尔可夫定理(Gauss-Markov theorem)

在给定经典线性回归的假定下，最小二乘估计量是具有最小方差的线性无偏估计量。

@

1、线性性，即估计量$\hat{\beta}_0$ 、$\hat{\beta}_1$ 是Y_i 的线性组合。

证：$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i (Y_i - \bar{Y})}{\sum x_i^2} = \frac{\sum x_i Y_i}{\sum x_i^2} - \frac{\bar{Y}\sum x_i}{\sum x_i^2}$$

令 $k_i = \frac{x_i}{\sum x_i^2}$，因 $\sum x_i = \sum (X_i - \bar{X}) = 0$，故有

$$\hat{\beta}_1 = \sum \frac{x_i}{\sum x_i^2} Y_i = \sum k_i Y_i$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{1}{n}\sum Y_i - \sum k_i Y_i \bar{X} = \sum (\frac{1}{n} - \bar{X} k_i) Y_i = \sum w_i Y_i$$

$$\sum k_i X_i = \sum \frac{x_i}{\sum x_i^2} X_i = \sum \frac{x_i(X_i - \bar{X} + \bar{X})}{\sum x_i^2} = \frac{\sum x_i^2 + \bar{X}\sum x_i}{\sum x_i^2} = 1$$

2、无偏性，即估计量 $\hat{\beta}_0$、$\hat{\beta}_1$ 的均值（期望）等于总体回归参数真值 β_0 与 β_1

证：

$$\hat{\beta}_1 = \sum k_i Y_i = \sum k_i(\beta_0 + \beta_1 X_i + \mu_i) = \beta_0 \sum k_i + \beta_1 \sum k_i X_i + \sum k_i \mu_i$$

易知 $\sum k_i = \frac{\sum x_i}{\sum x_i^2} = 0$ $\sum k_i X_i = 1$

故 $\hat{\beta}_1 = \beta_1 + \sum k_i \mu_i$

$$E(\hat{\beta}_1) = E(\beta_1 + \sum k_i \mu_i) = \beta_1 + \sum k_i E(\mu_i) = \beta_1$$

同样地，容易得出

$$E(\hat{\beta}_0) = E(\beta_0 + \sum w_i \mu_i) = E(\beta_0) + \sum w_i E(\mu_i) = \beta_0$$

3、有效性（最小方差性），即在所有线性无偏估计量中，最小二乘估计量 $\hat{\beta}_0$、$\hat{\beta}_1$ 具有最小方差。

证：(1)先求 $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 的方差

$$\mathrm{var}(\hat{\beta}_1)=\mathrm{var}(\sum k_iY_i)=\sum k_i^2\ \mathrm{var}(\beta_0+\beta_1X_i+\mu_i)=\sum k_i^2\ \mathrm{var}(\mu_i)$$

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

$$\mathrm{var}(\hat{\beta}_0)=\mathrm{var}(\sum w_iY_i)=\sum w_i^2\ \mathrm{var}(\beta_0+\beta_1X_i+\mu_i)=\sum(1/n-\bar{X}k_i)^2\sigma^2$$

$$=\sum\left[\left(\frac{1}{n}\right)^2-2\frac{1}{n}\bar{X}k_i+\bar{X}^2k_i^2\right]\sigma^2=\left(\frac{1}{n}-\frac{2}{n}\bar{X}\sum k_i+\bar{X}^2\sum\left(\frac{x_i}{\sum x_i^2}\right)^2\right)\sigma^2$$

$$=\left(\frac{1}{n}+\frac{\bar{X}^2}{\sum x_i^2}\right)\sigma^2=\frac{\sum x_i^2+n\bar{X}^2}{n\sum x_i^2}\sigma^2=\frac{\sum X_i^2}{n\sum x_i^2}\sigma^2$$

（2）证明最小方差性

假设 $\hat{\beta}_1^*$ 是其他估计方法得到的关于 β_1 的线性无偏估计量：

$$\hat{\beta}_1^* = \sum c_i Y_i$$

其中，$c_i=k_i+d_i$，d_i为不全为零的常数

则容易证明　$\text{var}(\hat{\beta}_1^*) \geq \text{var}(\hat{\beta}_1)$

同理，可证明β_0的最小二乘估计量 $\hat{\beta}_0$具有最的小方差

普通最小二乘估计量（ordinary least Squares Estimators）称为**最佳线性无偏估计量（best linear unbiased estimator, BLUE）**

由于最小二乘估计量拥有一个“好”的估计量所应具备的小样本特性，它自然也拥有大样本特性。

如考察$\hat{\beta}_1$的一致性

$$P\lim(\hat{\beta}_1) = P\lim(\beta_1 + \sum k_i \mu_i) = P\lim(\beta_1) + P\lim(\frac{\sum x_i \mu_i}{\sum x_i^2})$$

$$= \beta_1 + \frac{P\lim(\sum x_i \mu_i / n)}{P\lim(\sum x_i^2 / n)}$$

$$= \beta_1 + \frac{Cov(X, \mu)}{Q} = \beta_1 + \frac{0}{Q} = \beta_1$$

高斯—马尔可夫定理

OLS参数估计量的分布

在基本假定条件下，假定μ_i服从正态分布，因此Y_i也服从正态分布，而$\hat{\beta}_0$和$\hat{\beta}_1$均是Y_i的线性函数，所以，$\hat{\beta}_0$和$\hat{\beta}_1$也服从正态分布。可表示为：

$$\hat{\beta}_0 \sim N(\beta_0, \frac{\sigma^2 \sum X_i^2}{n\sum(X_i - \overline{X})^2}) \qquad \hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum(X_i - \overline{X})^2})$$

参数估计量$\hat{\beta}_0$和$\hat{\beta}_1$的标准差(standard error)为：

$$se(\hat{\beta}_0) = \sigma\sqrt{\frac{\sum X_i^2}{n\sum(X_i - \overline{X})^2}} \qquad se(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum(X_i - \overline{X})^2}}$$

参数估计量的概率分布及随机干扰项方差的估计

1、参数估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的概率分布

普通最小二乘估计量 $\hat{\beta}_0$、$\hat{\beta}_1$ 分别是 Y_i 的线性组合，因此，$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的概率分布取决于 Y 的分布特征

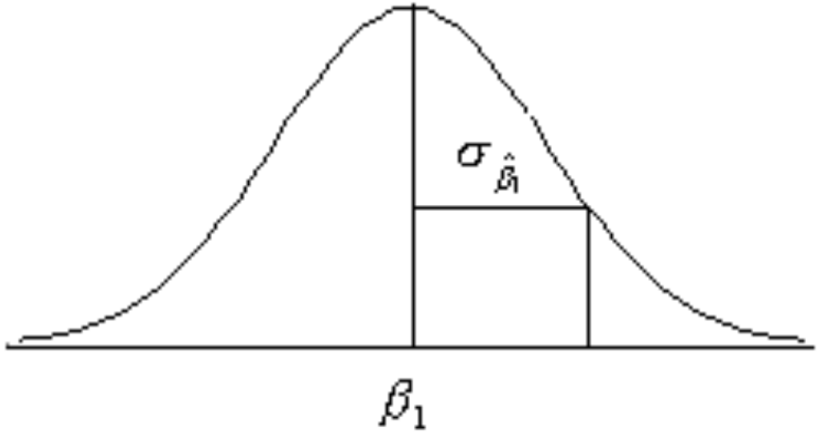
在 μ 是正态分布的假设下，Y 是正态分布，则 $\hat{\beta}_0$、$\hat{\beta}_1$ 也服从正态分布，因此

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum x_i^2}) \qquad \hat{\beta}_0 \sim N(\beta_0, \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2)$$

$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的标准差

$$\sigma_{\hat{\beta}_1} = \sqrt{\sigma^2 / \sum x_i^2}$$

$$\sigma_{\hat{\beta}_0} = \sqrt{\frac{\sigma^2 \sum X_i^2}{n \sum x_i^2}}$$



$$\sigma_{\hat{\beta}_1} = \sqrt{\sigma^2 / \sum x_i^2} \qquad \sigma_{\hat{\beta}_0} = \sqrt{\frac{\sigma^2 \sum X_i^2}{n \sum x_i^2}}$$

2. 随机误差项μ的方差σ^2的估计

在估计的参数 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的方差表达式中，都含有随机扰动项 μ 的方差 σ^2。 σ^2又称为**总体方差**。

$$Y_i = \hat{Y}_i + e_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i \qquad \text{SRF}$$

$$Y_i = E(Y \mid X_i) + \mu_i = \beta_0 + \beta_1 X_i + \mu_i \qquad \text{PRF}$$

由于σ^2实际上是未知的，因此 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的方差实际上无法计算，这就需要对其进行估计。

$$\sigma_{\hat{\beta}_1} = \sqrt{\sigma^2 / \sum x_i^2} \qquad \sigma_{\hat{\beta}_0} = \sqrt{\frac{\sigma^2 \sum X_i^2}{n \sum x_i^2}}$$

由于随机项μ_i不可观测，只能从μ_i的估计——残差e_i出发，对总体方差进行估计。

可以证明， σ^2的最小二乘估计量为 $\hat{\sigma}^2 = \dfrac{\sum e_i^2}{n-2}$

$E(\hat{\sigma}^2) = E(\dfrac{\sum e_i^2}{n-2}) = \sigma^2$ 它是关于σ^2的无偏估计量。

替代后参数估计的标准差为：

$$se(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{\sum X_i^2}{n\sum (X_i - \overline{X})^2}} \qquad se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum (X_i - \overline{X})^2}}$$

最大或然估计法σ^2的估计量

在最大或然估计法中，

解或然方程

$$\frac{\partial}{\partial \sigma^2} L^* = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \Sigma (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = 0$$

即可得到σ^2的最大或然估计量为：

$$\hat{\sigma}^2 = \frac{1}{n} \Sigma (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \frac{\sum e_i^2}{n}$$

因此，σ^2的最大或然估计量不具无偏性，但却具有一致性。

在随机误差项μ的方差σ^2估计出后，参数$\hat{\beta}_0$和$\hat{\beta}_1$的**方差**和**标准差**的估计量分别是：

$\hat{\beta}_1$的样本方差：　$S^2_{\hat{\beta}_1} = \hat{\sigma}^2 / \sum x_i^2$

$\hat{\beta}_1$的样本标准差：　$S_{\hat{\beta}_1} = \hat{\sigma} / \sqrt{\sum x_i^2}$

$\hat{\beta}_0$的样本方差：　$S^2_{\hat{\beta}_0} = \hat{\sigma}^2 \sum X_i^2 / n \sum x_i^2$

$\hat{\beta}_0$的样本标准差：　$S_{\hat{\beta}_0} = \hat{\sigma} \sqrt{\sum X_i^2 / n \sum x_i^2}$

★请大家注意以上关于方差的几种表示方法

$$se(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{\sum X_i^2}{n\sum(X_i - \overline{X})^2}} \qquad se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum(X_i - \overline{X})^2}}$$

Equation: EQ01 Workfile: 孔伟杰计量经济学课堂教学练习(...

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

Dependent Variable: Y
Method: Least Squares
Date: 10/18/16 Time: 14:53
Sample: 1 10
Included observations: 10

$\hat{\beta}_j$ $se(\hat{\beta}_j)$

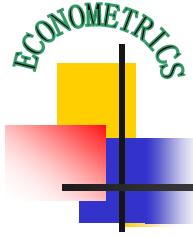
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	142.4000	44.44673	3.203835	0.0125
X	0.670000	0.019189	34.91562	0.0000

R-squared	0.993481	M	
Adjusted R-squared	0.992666	S.D. dependent var	610.5512
S.E. of regression	52.28814	A	10.92827
Sum squared resid	21672.40	S	10.98879
Log likelihood	-52.64136	Hannan-Quinn criter.	10.86189
F-statistic	1219.101	Durbin-Watson stat	1.677411
Prob(F-statistic)	0.000000		

回归标准差$\hat{\sigma} = \sqrt{\sum e_i^2/(n-2)}$

残差平方和$\sum e_i^2$

数据来自课本P41例题



§2.4 一元线性回归模型的统计检验

一、拟合优度检验

二、变量的显著性检验

三、参数的置信区间

说 明

- 回归分析是要通过样本所估计的参数来代替总体的真实参数，或者说是用样本回归线代替总体回归线。
- 尽管从统计性质上已知，如果有足够多的重复 抽样，参数的估计值的期望（均值）就等于其总体的参数真值，但在一次抽样中，估计值不一定就等于该真值。
- 那么，在一次抽样中，参数的估计值与真值的差异有多大，是否显著，这就需要进一步进行**统计检验**。
- 主要包括**拟合优度检验**、变量的**显著性检验**及参数的**区间估计**。

ECONOMETRICS

一、拟合优度检验

拟合优度检验：对样本回归直线与样本观测值之间拟合程度的检验。

度量拟合优度的指标：**判定系数**（**可决系数**）R^2

问题：采用普通最小二乘估计方法，已经保证了模型最好地拟合了样本观测值，为什么还要检验拟合程度？

EViews

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

```
series y1
equation eq_xy1.ls y1 c x
```

Workfile: 孔伟杰计量经济学课堂教学练习(第四版) - (d:\skydrive...

View Proc Object Print Save Details+/- Show Fetch Store Delete Genr Sample

Range: 1 10 -- 10 obs Filter: *
Sample: 1 10 -- 10 obs

- c
- eq01
- eq_xy1
- equation
- graph01
- group
- group01
- resid
- x
- y
- y1

Group: GROUP01 Workfile: 孔伟杰计量经济学课堂教学练习(...

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

obs	X	Y	Y1
1	800.0000	638.0000	627.0000
2	1100.000	935.0000	847.0000
3	1400.000	1155.000	1045.000
4	1700.000	1254.000	1144.000
5	2000.000	1408.000	1474.000
6	2300.000	1650.000	1452.000
7	2600.000	1925.000	1969.000
8	2900.000	2068.000	1969.000
9	3200.000	2266.000	2354.000
10	3500.000	2530.000	2629.000

Equation: EQ01 Workfile: 孔伟杰计量经济学课堂教学练习(第...

View Proc Object Print Name Freeze Estimate Forecast Stats Resids

Dependent Variable: Y
Method: Least Squares
Date: 03/22/11 Time: 13:39
Sample: 1 10
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	142.4000	44.44673	3.203835	0.0125
X	0.670000	0.019189	34.91562	0.0000

R-squared	0.993481	Mean dependent var	1582.900
Adjusted R-squared	0.992666	S.D. dependent var	610.5512
S.E. of regression	52.28814	Akaike info criterion	10.92827
Sum squared resid	21872.40	Schwarz criterion	10.98879
Log likelihood	-52.64136	Hannan-Quinn criter.	10.86189
F-statistic	1219.101	Durbin-Watson stat	1.677411
Prob(F-statistic)	0.000000		

Equation: EQ_XY1 Workfile: 孔伟杰计量经济学课堂教学练习(...

View Proc Object Print Name Freeze Estimate Forecast Stats Resids

Dependent Variable: Y1
Method: Least Squares
Date: 10/18/16 Time: 14:40
Sample: 1 10
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.000000	92.15528	0.032554	0.9748
X	0.720000	0.039786	18.09661	0.0000

R-squared	0.976154	Mean dependent var	1551.000
Adjusted R-squared	0.973173	S.D. dependent var	661.9120
S.E. of regression	108.4136	Akaike info criterion	12.38664
Sum squared resid	94028.00	Schwarz criterion	12.44716
Log likelihood	-59.93320	Hannan-Quinn criter.	12.32025
F-statistic	327.4871	Durbin-Watson stat	2.690348
Prob(F-statistic)	0.000000		

Path = c:\users\kwj\documents DB = none WF = 孔伟杰计量经济学课堂教学练习(第四版)

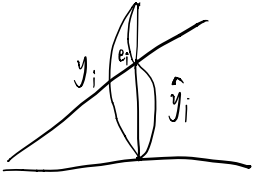
1、总离差平方和的分解

已知由一组样本观测值（X_i, Y_i），i=1,2…,n
得到如下样本回归直线

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

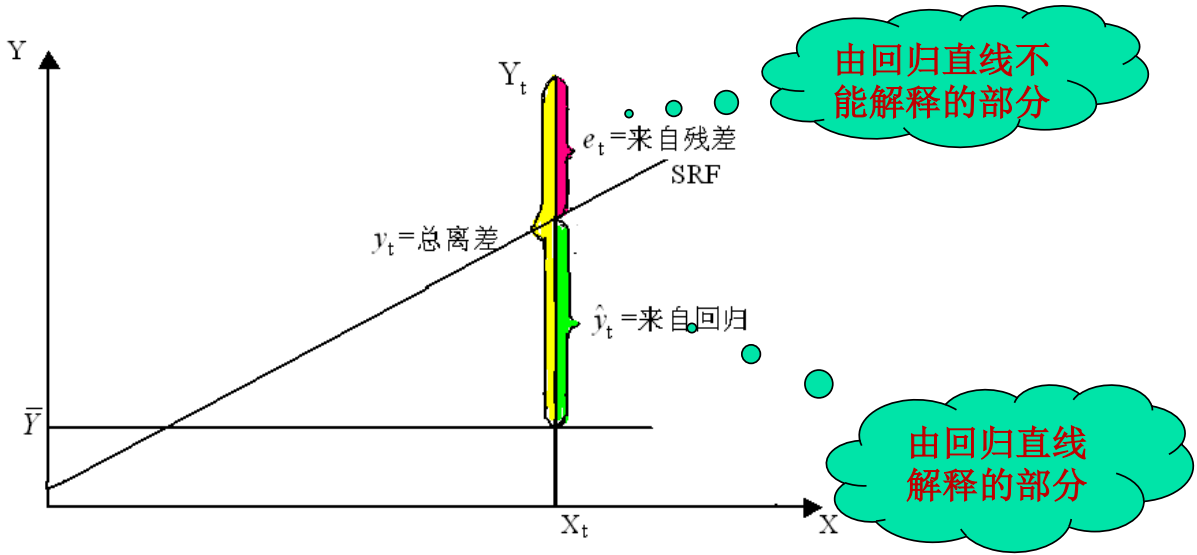
而Y的第*i*个观测值与样本均值的离差 $y_i = (Y_i - \bar{Y})$
可分解为两部分之和

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



总离差：

$$y_i = Y_i - \overline{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y}) = e_i + \hat{y}_i$$



$\hat{y}_i = \hat{Y}_i - \overline{Y}$ 是样本回归拟合值与观测值的平均值之差，可认为是由**回归直线解释的部分**；

$e_i = Y_i - \hat{Y}_i$ 是实际观测值与回归拟合值之差，是**回归直线不能解释的部分**。

如果$Y_i=\hat{Y}_i$ 即实际观测值落在样本回归“线”上，则**拟合最好**。可认为，“离差”全部来自回归线，而与“残差”无关。

事实上：

$$\sum \hat{y}_i e_i = \sum (\hat{Y}_i - \bar{Y}) e_i = \sum (\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y}) e_i$$
$$= \hat{\beta}_0 \sum e_i + \hat{\beta}_1 \sum X_i e_i - \bar{Y} \sum e_i = 0$$

对于所有样本点，则需考虑这些点与样本均值离差的平方和：

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

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

记：

$TSS = \sum y_i^2 = \sum (Y_i - \bar{Y})^2$ 总体平方和（Total Sum of Squares）

$ESS = \sum \hat{y}_i^2 = \sum (\hat{Y}_i - \bar{Y})^2$ 回归平方和（Explained Sum of Squares）

$RSS = \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$ 残差平方和（Residual Sum of Squares ）

TSS=ESS+RSS

TSS=ESS+RSS

Y的观测值围绕其均值的总离差(TSS)可分解为两部分：一部分来自回归线(ESS)，另一部分则来自随机势力(RSS)。

- 在给定样本中，TSS不变，
- 如果实际观测点离样本回归线越近，则**ESS**在**TSS**中占的比重越大，因此
- **拟合优度=回归平方和ESS/Y的总离差TSS**

2、可决系数R^2统计量

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

称 R^2 为（样本）可决系数/判定系数（coefficient of determination)。

可决系数的取值范围：[0，1]

R^2越接近1，说明实际观测点离样本线越近，拟合优度越高。

很显然，可决系数是一个非负的统计量，它也是随着抽样的不同而不同。

在实际计算可决系数时，在 $\hat{\beta}_1$ 已经估计出后：

$$\hat{y}_i = \hat{\beta}_1 x_i$$

$$R^2 = \frac{ESS}{TSS} = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{\sum (\hat{\beta}_1 x_i)^2}{\sum y_i^2} = \hat{\beta}_1^2 \left(\frac{\sum x_i^2}{\sum y_i^2} \right) = \hat{\beta}_1^2 \left(\frac{\sum x_i^2}{\sum y_i^2} \right)$$

例：在上述家庭**可支配收入-消费支出**例中，对于所抽出的一组样本数，参数估计的计算可通过下面的表2.3.1进行。

表 2.3.1　参数估计的计算表

	X_i	Y_i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2	X_i^2	Y_i^2
1	800	638	-1350	-945	1275615	1822500	892836	640000	407044
2	1100	935	-1050	-648	680295	1102500	419774	1210000	874225
3	1400	1155	-750	-428	320925	562500	183098	1960000	1334025
4	1700	1254	-450	-329	148005	202500	108175	2890000	1572516
5	2000	1408	-150	-175	26235	22500	30590	4000000	1982464
6	2300	1650	150	67	10065	22500	4502	5290000	2722500
7	2600	1925	450	342	153945	202500	117032	6760000	3705625
8	2900	2068	750	485	363825	562500	235322	8410000	4276624
9	3200	2266	1050	683	717255	1102500	466626	10240000	5134756
10	3500	2530	1350	947	1278585	1822500	896998	12250000	6400900
求和	21500	15829			4974750	7425000	3354955	53650000	28410679
平均	2150	1583							

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{4975800}{7425000} = 0.67$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 1583 - 0.67 \times 2150 = 142.40$$

因此，由该样本估计的回归方程为：

$$\hat{Y}_i = 142.40 + 0.67 X_i$$

$$R^2 = \hat{\beta}_1^2 \frac{\sum x_i^2}{\sum y_i^2} = \frac{(0.67)^2 \times 7425000}{3356322} = 0.9934$$

ECONOMETRICS

$$se(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{\sum X_i^2}{n\sum(X_i - \overline{X})^2}} \qquad se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum(X_i - \overline{X})^2}}$$

Equation: EQ01 Workfile: 孔伟杰计量经济学课堂教学练习(...

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

Dependent Variable: Y
Method: Least Squares
Date: 10/18/16 Time: 14:53
Sample: 1 10
Included observations: 10

$\hat{\beta}_j$ $se(\hat{\beta}_j)$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	142.4000	44.44673	3.203835	0.0125
X	0.679000	0.019189		

R-squared	0.993481	Mean dependent var	1500.000
Adjusted R-squared	0.992666	S.D. dependent var	289.8851
S.E. of regression	52.28814	Akaike info criterion	10.92827
Sum squared resid	21872.40	Schwarz criterion	10.98879
Log likelihood	-52.64136	Hannan-Quinn criter.	10.86189
F-statistic	1219.101	Durbin-Watson stat	1.677411
Prob(F-statistic)	0.000000		

判定系数R^2

调整的判定系数$\overline{R}^2$

回归标准差$\hat{\sigma} = \sqrt{\sum e_i^2/(n-2)}$

残差平方和$\sum e_i^2$

ECONOMETRICS

二、变量的显著性检验

回归分析是要判断**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**一元线性模型**中，就是要判断***X***是否对***Y***具有显著的线性性影响。这就需要进行**变量的显著性检验。**

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。

计量经济学中，主要是针对变量的参数真值是否为零来进行显著性检验的。

ECONOMETRICS

1、假设检验

所谓**假设检验**，就是事先对**总体参数或总体分布**形式作出一个**假设**，然后**利用样本信息来判断原假设是否合理**，即判断样本信息与原假设是否有显著差异，从而决定是否接受或否定原假设。

- **假设检验采用的逻辑推理方法是反证法**

  先假定原假设正确，然后根据样本信息，观察由此假设而导致的结果是否合理，从而判断是否接受原假设。

- 判断结果合理与否，是基于“小概率事件不易发生”这一原理的，**即构造一个小概率事件，可以认为小概率事件在一次观察中基本不会发生。**

参数的显著性检验

参数的显著性检验的基本思想是：在某种原假设成立的条件下，利用适当的统计量和给定的显著性水平α，构造一个小概率事件，可以认为小概率事件在一次观察中基本不会发生。如果该事件竟然发生了，就认为原假设不真，从而拒绝原假设，接受备择假设。

- 以β_1为例，对总体参数β1提出假设，H_0：$\beta_1=\beta_1^*$（β1*为一具体数值），检验参数估计量 与这个假设值β1*之间的差异是否显著。若差异显著，就不能接受这个假设；若差异不显著，就不能拒绝这个假设。
- 计量经济学中最常用的假设是H_0：$\beta_1=0$，对 进行显著性检验，就是对β_1是否异于零进行检验。

统计量：$t=\dfrac{\hat{\beta}_1-\beta_1}{se(\hat{\beta}_1)}=\dfrac{\hat{\beta}_1}{se(\hat{\beta}_1)}\sim t(n-2)$　　$se(\hat{\beta}_1)$：表示$S_{\hat{\beta}_1}$

ECONOMETRICS

2、变量的显著性检验

对于一元线性回归方程中的 $\hat{\beta}_1$，已经知道它服从正态分布

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum x_i^2})$$

由于真实的σ^2未知，在用它的无偏估计量 $\hat{\sigma}^2 = \sum e_i^2 / (n-2)$ 替代时，可构造如下统计量

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 / \sum x_i^2}} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \sim t(n-2)$$

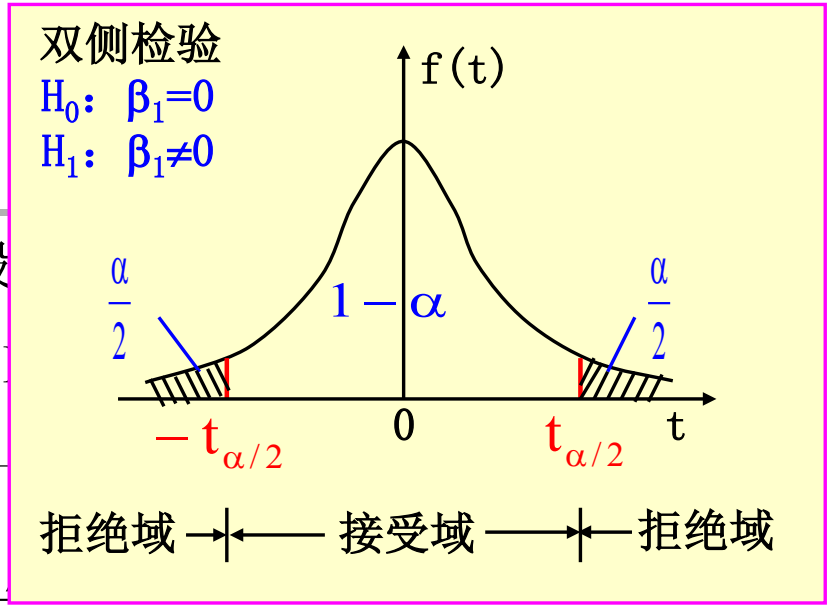
检验步骤：

（1）对总体参数提出假设

H_0： $\beta_1=0$，

（2）以原假设H_0构造t统计

$$t=\frac{}{S_{\hat{\beta}_1}}$$



（3）给定显著性水平α (一般取0.01，0.05，0.1)，查t分布表得临界值$t_{\alpha/2}(n-2)$

(4) 比较，判断

若 $|t|> t_{\alpha/2}(n-2)$，则拒绝H_0，接受H_1；

若 $|t|\le t_{\alpha/2}(n-2)$，则拒绝H_1，接受H_0；

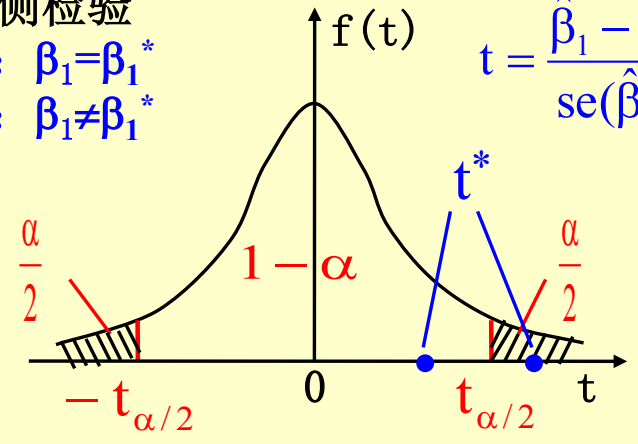
变量的显著性检验

双侧检验

H_0：$\beta_1=\beta_1^*$

H_1：$\beta_1\neq\beta_1^*$

$$t=\frac{\hat{\beta}_1-\beta_1^*}{se(\hat{\beta}_1)}$$



p值—与查表找临界值的一个

p值(p value)即概率值，“t统计值的p值”称为t统计值的显著性概率，也有称零系数概率。

t*的外侧概率的表达式为 1-P(t≤t*)

定义双侧检验中t统计值的p值为:t统计值外侧概率的两倍，即：t*的p值=2×(1-P(t≤t*))

用t统计值的p值与显著性水平

一般规律：p值越小，越能拒绝原假设

（1）若p<α，表明t*落在由α所决定的分界点外侧，应拒绝H_0；

（2）若p>α，表明t*落在由α所决定的分界点内侧，应接受H_0。

比较t*和临界值$t_{\alpha/2}$与比较t*的外侧概率和α/2是等价的。

在上述收入—消费支出例中

家庭可支配收入-消费支出一组样本数据计算表

	X_i	Y_i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2	X_i^2	Y_i^2
1	800	638	-1350	-945	1275615	1822500	892836	640000	407044
2	1100	935	-1050	-648	680295	1102500	419774	1210000	874225
3	1400	1155	-750	-428	320925	562500	183098	1960000	1334025
4	1700	1254	-450	-329	148005	202500	108175	2890000	1572516
5	2000	1408	-150	-175	26235	22500	30590	4000000	1982464
6	2300	1650	150	67	10065	22500	4502	5290000	2722500
7	2600	1925	450	342	153945	202500	117032	6760000	3705625
8	2900	2068	750	485	363825	562500	235322	8410000	4276624
9	3200	2266	1050	683	717255	1102500	466626	10240000	5134756
10	3500	2530	1350	947	1278585	1822500	896998	12250000	6400900
求和	21500	15829			4974750	7425000	3354955	53650000	28410679
平均	2150	1583							

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{4974750}{7425000} = 0.670$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 1583 - 0.670 \times 2150 = 142.4$$

样本回归函数：$\hat{Y}_i = 142.4 + 0.670 X_i$

$$t=\frac{\hat{\beta}_0-\beta_0}{\sqrt{\hat{\sigma}^2\sum X_i^2/n\sum x_i^2}}=\frac{\hat{\beta}_0}{S_{\hat{\beta}_0}}\sim t(n-2)$$

首先计算σ^2的估计值

$$\hat{\sigma}^2=\frac{\sum e_i^2}{n-2}=\frac{\sum y_i^2-\hat{\beta}_1^2\sum x_i^2}{n-2}=\frac{3356322-0.670^2\times 7425000}{10-2}=2780.56$$

于是$\hat{\beta}_1$和$\hat{\beta}_0$的标准差的估计值分别是：

$$S_{\hat{\beta}_1}=\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2}}=\sqrt{\frac{2780.56}{7425000}}=0.019$$

$$S_{\hat{\beta}_0}=\sqrt{\frac{\hat{\sigma}^2\sum X_i^2}{n\sum x_i^2}}=\sqrt{\frac{2780.56\times 53650000}{10\times 7425000}}=44.45$$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	142.4000	44.44673	3.203835	0.0125
Y	0.670000	0.019189	34.91562	0.0000

R-squared	0.993481	Mean dependent var	1582.900
Adjusted R-squared	0.992666	S.D. dependent var	610.5512
S.E. of regression	52.28814	Akaike info criterion	10.92827
Sum squared resid	21872.40	Schwarz criterion	10.98879
Log likelihood	-52.64136	Hannan-Quinn criter.	10.86189
F-statistic	1219.101	Durbin-Watson stat	1.677411
Prob(F-statistic)	0.000000		

$$t_1=\frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}=\frac{0.670}{0.019}=34.92,\quad t_0=\frac{\hat{\beta}_0}{S_{\hat{\beta}_0}}=\frac{142.40}{44.45}=3.20$$

β_0 β_1 非零

变量显著性检验的思想

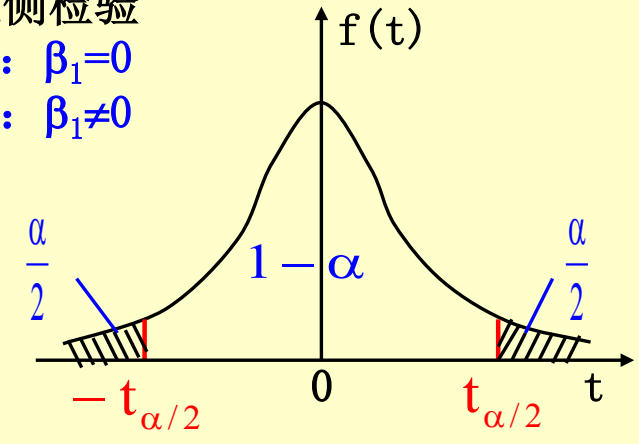
t统计量的计算结果分别为

$$t_1 = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{0.670}{0.019} = 34.92,$$

双侧检验

H_0：$\beta_1=0$

H_1：$\beta_1\neq0$



给定显著性水平α=0.05，查t分布表得临界值

$t_{0.05/2}(8)=2.306$

$|t_1|>2.306$，说明家庭可支配收入在95%的置信度下显著，即是消费支出的主要解释变量;

$|t_0|>2.306$, 表明在95%的置信度下，拒绝截距项为零的假设。

$$se(\hat{\beta}_0)=\hat{\sigma}\sqrt{\frac{\sum X_i^2}{n\sum(X_i-\overline{X})^2}} \qquad se(\hat{\beta}_1)=\frac{\hat{\sigma}}{\sqrt{\sum(X_i-\overline{X})^2}}$$

Equation: EQ01 Workfile: 孔伟杰计量经济学课堂教学练习(...

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

Dependent Variable: Y
Method: Least Squares
Date: 10/18/16 Time: 14:53
Sample: 1 10
Included observations: 10

$\hat{\beta}_j$ $se(\hat{\beta}_j)$ $t(\hat{\beta}_j)$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	142.4000	44.44673	3.203835	0.0125
X	0.679000	0.019189		

R-squared	0.993481	Mean dependent var	1500.000
Adjusted R-squared	0.992666	S.D. dependent var	348.8553
S.E. of regression	52.28814	Akaike info criterion	10.92827
Sum squared resid	21872.40	Schwarz criterion	10.98879
Log likelihood	-52.64136	Hannan-Quinn criter.	10.86189
F-statistic	1219.101	Durbin-Watson stat	1.677411
Prob(F-statistic)	0.000000		

判定系数R^2

调整的判定系数$\overline{R}^2$

回归标准差$\hat{\sigma}=\sqrt{\sum e_i^2/(n-2)}$

残差平方和$\sum e_i^2$

各项统计结果解释如下：

常数和解释变量	参数估计值	参数标准差	t统计量	双侧概率
C	142.4000	44.44673	3.203835	0.0125
X	0.670000	0.019189	34.91562	0.0000

判定系数	0.993481	被解释变量均值	1582.900
调整的判定系数	0.992666	被解释变量标准差	610.5512
回归方程标准差	52.28814	赤池信息准则	10.92827
残差平方和	21872.40	施瓦兹信息准则	10.98879
似然函数的对数	1219.101	汉南-昆信息准则	10.86189
F统计量	334.4876	D-W统计量	1.677411
F统计量的概率	0.000000		

ECONOMETRICS

三、参数的置信区间

回归分析希望通过样本所估计出的参数$\hat{\beta}_1$来代替总体的参数β_1。

假设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的范围（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**。

要判断估计的参数值$\hat{\beta}$离真实的参数值β有多“近”，可预先选择一个概率$\alpha(0<\alpha<1)$，并求一个正数δ，使得随机区间$(\hat{\beta}-\delta,\hat{\beta}+\delta)$包含参数的真值的概率为1-$\alpha$。即：

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

如果存在这样一个区间，称之为**置信区间**（confidence interval）；**1-α称为置信系数**（**置信度**）（confidence coefficient），**α称为显著性水平**（level of significance）；置信区间的端点称为**置信限**（confidence limit）或**临界值**（critical values）。

一元线性模型中，β_i (i=1，2）的置信区间:

在变量的显著性检验中已经知道：$t=\dfrac{\hat{\beta}_i-\beta_i}{s_{\hat{\beta}_i}}\sim t(n-2)$

意味着，如果给定置信度（1-α），从分布表中查得自由度为(n-2)的临界值，那么t值处在($-t_{\alpha/2}$，$t_{\alpha/2}$)的概率是(1-α)。表示为：

即
$$P(-t_{\frac{\alpha}{2}}<t<t_{\frac{\alpha}{2}})=1-\alpha$$

$$P(-t_{\frac{\alpha}{2}}<\frac{\hat{\beta}_i-\beta_i}{s_{\hat{\beta}_i}}<t_{\frac{\alpha}{2}})=1-\alpha$$

$$P(\hat{\beta}_i-t_{\frac{\alpha}{2}}\times s_{\hat{\beta}_i}<\beta_i<\hat{\beta}_i+t_{\frac{\alpha}{2}}\times s_{\hat{\beta}_i})=1-\alpha$$

于是得到:(1-α)的置信度下，β_i的置信区间是

$$(\hat{\beta}_i-t_{\frac{\alpha}{2}}\times s_{\hat{\beta}_i},\hat{\beta}_i+t_{\frac{\alpha}{2}}\times s_{\hat{\beta}_i})$$

$$(\hat{\beta}_i - t_{\frac{\alpha}{2}} \times s_{\hat{\beta}_i}, \hat{\beta}_i + t_{\frac{\alpha}{2}} \times s_{\hat{\beta}_i})$$

在上述收入-消费支出例中，如果给定α =0.01，查表得：

$$t_{\frac{\alpha}{2}}(n-2) = t_{0.005}(8) = 3.355$$

由于 $S_{\hat{\beta}_1} = 0.019$ $S_{\hat{\beta}_0} = 44.82$

于是，β_1、β_0的置信区间分别为：

（0.606, 0.7344）

（-8.09, 292.65）

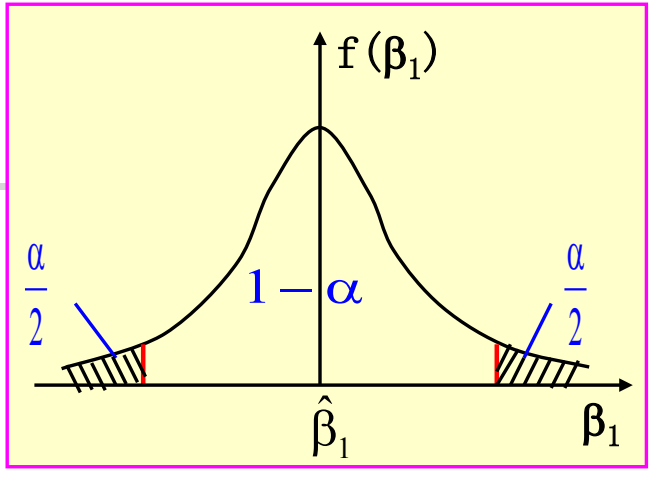
ECONOMETRICS

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum(X_i - \overline{X})^2})$$

$$t = \frac{\hat{\beta}_i - \beta_i}{s_{\hat{\beta}_i}} \sim t(n-2) \qquad \hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$$

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum(X_i - \overline{X})^2}} \qquad t = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_i}} \sim t(n-2)$$

$$(\hat{\beta}_i - t_{\frac{\alpha}{2}} \times s_{\hat{\beta}_i}, \hat{\beta}_i + t_{\frac{\alpha}{2}} \times s_{\hat{\beta}_i})$$



缩小置信区间

（1）增加样本容量n，n越大，临界值$t_{\alpha/2}$越小。

（2）提高模型的拟合程度，以减小残差平方和$\sum e_i^2$，以及$\hat{\sigma}$。

（3）提高样本观测值的分散程度，则$\sum(X_i - \overline{X})^2$越大，$se(\hat{\beta}_1)$越小。

提高置信水平与缩小置信区间是矛盾的：

置信水平越高，1-α越大，α越小，在其它情况不变时，临界值$t_{\alpha/2}$越大，置信区间越大。

如果置信区间为一精确值，则其置信水平为0；而若置信水平为100%，则置信区间为∞。

- 由于置信区间一定程度地给出了样本参数估计值与总体参数真值的“接近”程度，因此置信区间越小越好。
- 要缩小置信区间，需要

（1）增大样本容量n。

因为在同样的置信水平下，n越大，t分布表中的临界值越小；同时，增大样本容量，还可使样本参数估计量的标准差减小；

（2）提高模型的拟合优度。

因为样本参数估计量的标准差与残差平方和呈正比，模型拟合优度越高，残差平方和应越小。

§2.5一元线性回归分析应用：预测问题

一、$\hat{Y}_0$是条件均值$E(Y|X=X_0)$或个值Y_0的一个无偏估计

二、总体条件均值与个值预测值的置信区间

说 明

对于一元线性回归模型

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

给定样本以外的解释变量的观测值X_0，可以得到被解释变量的预测值$\hat{Y}_0$，可以此作为其**条件均值**$E(Y|X=X_0)$或**个别值**Y_0的一个近似估计。

严格地说，这只是被解释变量的预测值的**估计值**，而不是预测值。原因：

（1）参数估计量不确定；

（2）随机项的影响

一、预测值$\hat{Y}_0$是条件均值$E(Y|X=X_0)$或个值Y_0的一个无偏估计

对总体回归函数$E(Y|X=X_0)=\beta_0+\beta_1X$，$X=X_0$时

$$E(Y|X=X_0)=\beta_0+\beta_1X_0$$

通过样本回归函数$\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X$，求得的拟合值为

$$\hat{Y}_0=\hat{\beta}_0+\hat{\beta}_1X_0$$

于是 $$E(\hat{Y}_0)=E(\hat{\beta}_0+\hat{\beta}_1X_0)=E(\hat{\beta}_0)+X_0E(\hat{\beta}_1)=\beta_0+\beta_1X_0$$

可见，$\hat{Y}_0$是条件均值$E(Y|X=X_0)$的无偏估计。

对总体回归模型$Y=\beta_0+\beta_1X+\mu$，当$X=X_0$时

$$Y_0 = \beta_0 + \beta_1 X_0 + \mu$$

于是

$$E(Y_0) = E(\beta_0 + \beta_1 X_0 + \mu) = \beta_0 + \beta_1 X_0 + E(\mu) = \beta_0 + \beta_1 X_0$$

而 通过样本回归函数 $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$， 求得拟合值

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

的期望为

$$E(\hat{Y}_0) = E(\hat{\beta}_0 + \hat{\beta}_1 X_0) = E(\hat{\beta}_0) + X_0 E(\hat{\beta}_1) = \beta_0 + \beta_1 X_0$$

$\hat{Y}_0$是个值Y_0的无偏估计。

二、总体条件均值与个值预测值的置信区间

1、总体均值预测值的置信区间

由于 $$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum x_i^2}) \qquad \hat{\beta}_0 \sim N(\beta_0, \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2)$$

于是 $$E(\hat{Y}_0) = E(\hat{\beta}_0) + X_0 E(\hat{\beta}_1) = \beta_0 + \beta_1 X_0$$

$$Var(\hat{Y}_0) = Var(\hat{\beta}_0) + 2X_0 Cov(\hat{\beta}_0, \hat{\beta}_1) + X_0^2 Var(\hat{\beta}_1)$$

可以证明

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \overline{X} / \sum x_i^2$$

因此 $$Var(\hat{Y}_0)=\frac{\sigma^2\sum X_i^2}{n\sum x_i^2}-\frac{2X_0\bar{X}\sigma^2}{\sum x_i^2}+\frac{X_0^2\sigma^2}{\sum x_i^2}$$

$$=\frac{\sigma^2}{\sum x_i^2}\left(\frac{\sum X_i^2-n\bar{X}^2}{n}+\bar{X}^2-2X_0\bar{X}+X_0^2\right)$$

$$=\sigma^2\left(\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum x_i^2}\right)$$

$$=\frac{\sigma^2}{\sum x_i^2}\left(\frac{\sum x_i^2}{n}+(X_0-\bar{X})^2\right)$$

故 $$\hat{Y}_0\sim N(\beta_0+\beta_1X_0,\sigma^2\left(\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum x_i^2}\right))$$

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

将未知的σ^2代以它的无偏估计量$\hat{\sigma}^2$，可构造t统计量

$$t = \frac{\hat{Y}_0 - (\beta_0 + \beta_1 X_0)}{S_{\hat{Y}_0}} \sim t(n-2) \quad \text{其中} \quad S_{\hat{Y}_0} = \sqrt{\hat{\sigma}^2 (\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum x_i^2})}$$

于是，在1-α的置信度下，**总体均值**$E(Y|X_0)$**的置信区间为**

$$\hat{Y}_0 - t_{\frac{\alpha}{2}} \times S_{\hat{Y}_0} < E(Y \mid X_0) < \hat{Y}_0 + t_{\frac{\alpha}{2}} \times S_{\hat{Y}_0}$$

2、总体个值预测值的预测区间

由 $Y_0=\beta_0+\beta_1X_0+\mu$ 知：$Y_0 \sim N(\beta_0+\beta_1X_0,\sigma^2)$

于是 $\hat{Y}_0-Y_0 \sim N(0,\sigma^2(1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum x_i^2}))$

将未知的σ^2代以它的无偏估计量$\hat{\sigma}^2$，可构造t统计量

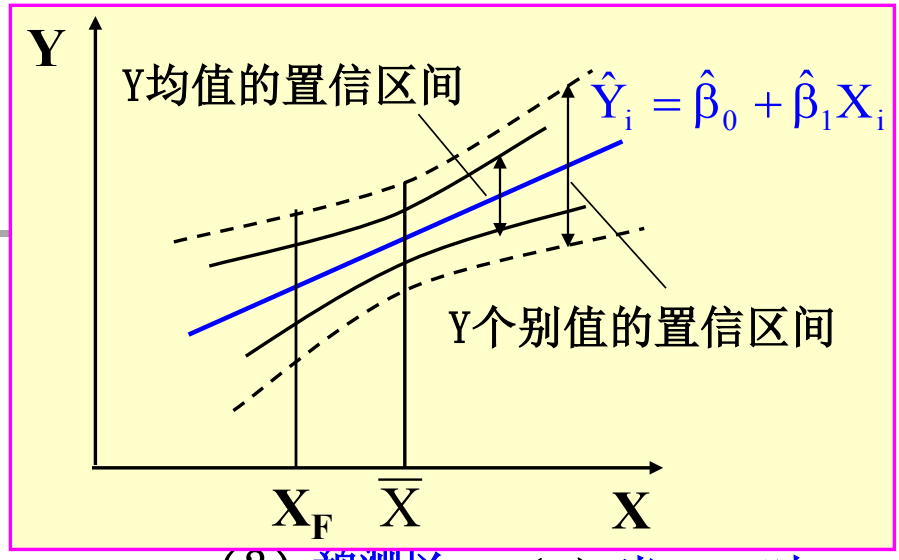
$$t=\frac{\hat{Y}_0-Y_0}{S_{\hat{Y}_0-Y_0}} \sim t(n-2)$$ 式中：$$S_{\hat{Y}_0-Y_0}=\sqrt{\hat{\sigma}^2(1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum x_i^2})}$$

从而给定显著性水平α，查表得临界值$t_{\alpha/2}(n-2)$在1-α的置信度下，

Y_0的置信区间为：$$\hat{Y}_0-t_{\frac{\alpha}{2}}\times S_{\hat{Y}_0-Y_0}<Y_0<\hat{Y}_0+t_{\frac{\alpha}{2}}\times S_{\hat{Y}_0-Y_0}$$

应变量Y个别值的区间预测： $$\hat{Y}_0\pm t_{\frac{\alpha}{2}}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum(X_i-\bar{X})^2}}$$

从应变量Y均值和个别值的置信区间看，有如下特点：



（1）$\hat{Y}_F$作为$E(Y|X_F)$的估计，由于抽样波动而存在抽样误差；而$\hat{Y}_F$作为Y_F的估计，除了抽样误差外，还由于随机误差项μ_i的存在，使得Y_F与$E(Y|X_F)$之间有差异。因此由$\hat{Y}_F$对个别值预测的误差的方差大于对平均值预测的误差的方差，即对个别值预测的置信区间比对平均值预测的置信区间宽。

（2）预测区间不是常数，而是随解释变量预测值X_F而变化。当$X_F = \overline{X}$时，区间最窄。

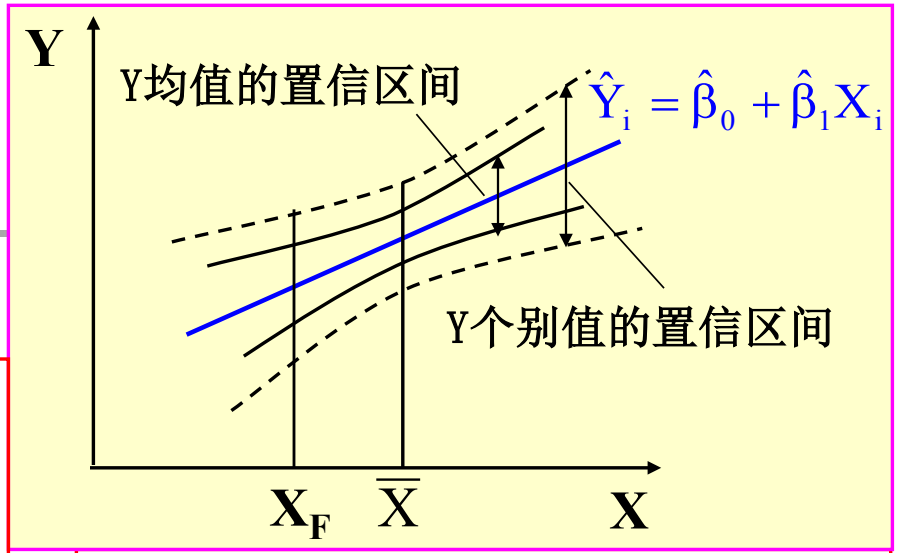
（3）当$n \to \infty$时，不存在抽样误差，对平均值预测的误差趋于0，对个别值预测的误差只决定于随机误差项μ_i的方差。

应变量Y平均值的置信区间：

$$\hat{Y}_F \pm t_{\frac{\alpha}{2}} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_F - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$$

应变量Y个别值的置信区间：

$$\hat{Y}_F \pm t_{\frac{\alpha}{2}} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(X_F - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$$



缩小应变量置信区间：

（1）增加样本容量n，n越大，临界值$t_{\alpha/2}$越小；且$\sum(X_i-\overline{X})^2$越大。

（2）提高模型的拟合程度，以减小残差平方和$\sum e_i^2$，以及$\hat{\sigma}$。

（3）提高样本观测值的分散程度，则$\sum(X_i-\overline{X})^2$越大。

（4）X_F尽可能接近$\overline{X}$。

提高置信水平与缩小置信区间是矛盾的：

置信水平越高，1-α越大，α越小，在其它情况不变时，临界值$t_{\alpha/2}$越大，置信区间越大。

如果置信区间为一精确值，则其置信水平为0；而若置信水平为100%，则置信区间为∞。

应变量Y平均值的置信区间

$$\hat{Y}_F \pm t_{\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(X_F-\overline{X})^2}{\sum(X_i-\overline{X})^2}}$$

应变量Y个别值的置信区间

$$\hat{Y}_F \pm t_{\frac{\alpha}{2}}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(X_F-\overline{X})^2}{\sum(X_i-\overline{X})^2}}$$

在上述**收入—消费支出**例中，得到的样本回归函数为:

$$\hat{Y}_i = 142.4 + 0.670X_i$$

则在 X_0=1000处，

$$\hat{Y}_0 = 142.4 + 0.670 \times 1000 = 812.4$$

而

$$Var(\hat{Y}_0) = 2780 * \left[\frac{1}{10} + \frac{(1000 - 2150)^2}{7425000}\right] = 773.3$$

$$S(\hat{Y}_0) = 27.8$$

$$\hat{Y}_0 - t_{\frac{\alpha}{2}} \times S_{\hat{Y}_0} < E(Y \mid X_0) < \hat{Y}_0 + t_{\frac{\alpha}{2}} \times S_{\hat{Y}_0}$$

因此，**总体均值E(Y|X=1000)的95%的置信区间为：**

$$812.4 - 2.306 \times 27.8 < E(Y \mid X = 1000) < 812.4 + 2.306 \times 27.8$$

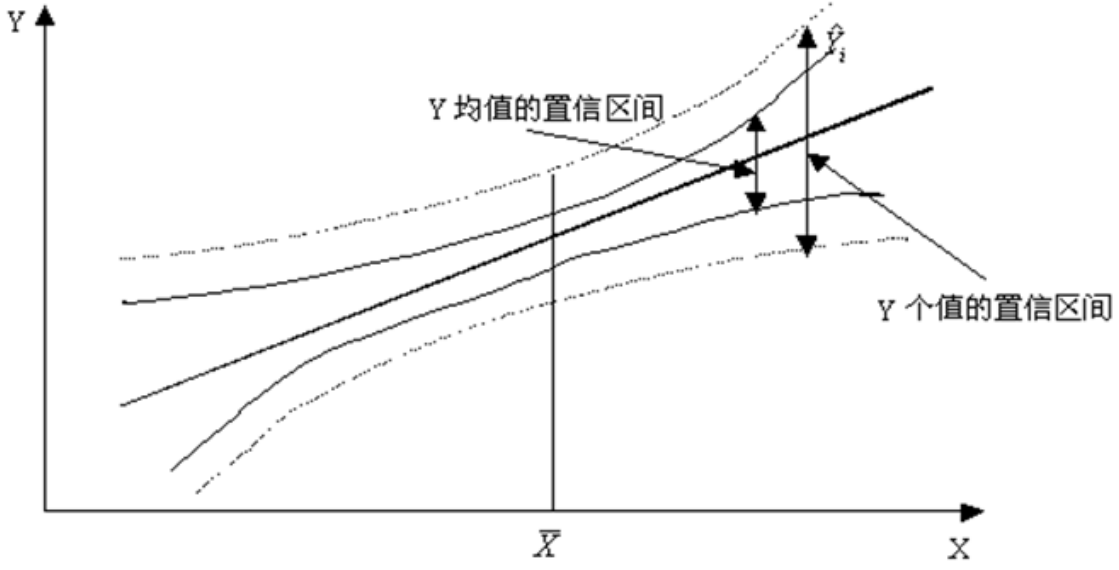
或为：　　（748.8，875.9）

同样地，对于Y在X=1000的个体值，其95%的置信区间为：

$$812.4 - 2.306 \times 59.61 < Y \mid_{X=1000} < 812.4 + 2.306 \times 59.61$$

或为：　　（674.9，949.84）

- 总体回归函数的**置信带（域）**（confidence band）
- 个体的**置信带（域）**



对于Y的总体均值E(Y|X)与个体值的预测区间（置信区间）：

（1）样本容量n越大，预测精度越高，反之预测精度越低；

（2）样本容量一定时，置信带的宽度当在X均值处最小，其附近进行预测（插值预测）精度越大；X越远离其均值，置信带越宽，预测可信度下降。

§2.6 实例

中国居民人均消费模型

表 2.6.1　中国各地区居民家庭人均全年可支配收入与人均全年消费性支出（元）

地区	可支配收入 X	消费支出 Y	地区	可支配收入 X	消费支出 Y
北　京	40830.0	29175.6	湖　北	16472.5	11760.8
天　津	26359.2	20418.7	湖　南	16004.9	11945.9
河　北	15189.6	10872.2	广　东	23420.7	17421.0
山　西	15119.7	10118.3	广　西	14082.3	9596.5
内蒙古	18692.9	14877.7	海　南	15733.3	11192.9
辽　宁	20817.8	14950.2	重　庆	16568.7	12600.2
吉　林	15998.1	12054.3	四　川	14231.0	11054.7
黑龙江	15903.4	12037.2	贵　州	11083.1	8288.0
上　海	42173.6	30399.9	云　南	12577.9	8823.8
江　苏	24775.5	17925.8	西　藏	9746.8	6310.6
浙　江	29775.0	20610.1	陕　西	14371.5	11217.3
安　徽	15154.3	10544.1	甘　肃	10954.4	8943.4
福　建	21217.9	16176.6	青　海	12947.8	11576.5
江　西	15099.7	10052.8	宁　夏	14565.8	11292.0
山　东	19008.3	11896.8	新　疆	13669.6	11391.8
河　南	14203.7	10002.5			

资料来源：《中国统计年鉴》(2014)。

回归结果：

Equation: EQ01_31SAMPLE Workfile: 孔伟杰计量经济学课堂教...

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

Dependent Variable: Y
Method: Least Squares
Date: 10/18/16 Time: 13:45
Sample: 1 31
Included observations: 31

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	477.1229	403.9231	1.181222	0.2471
X	0.707081	0.020419	34.62877	0.0000

R-squared	0.976387	Mean dependent var	13404.14
Adjusted R-squared	0.975573	S.D. dependent var	5495.729
S.E. of regression	858.9343	Akaike info criterion	16.41160
Sum squared resid	21395274	Schwarz criterion	16.50412
Log likelihood	-252.3798	Hannan-Quinn criter.	16.44176
F-statistic	1199.152	Durbin-Watson stat	1.495760
Prob(F-statistic)	0.000000		

一、中国居民人均消费模型

例　考察中国居民收入与消费支出的关系。

GDPP：人均国内生产总值（1990年不变价）

CONSP：人均居民消费（以居民消费价格指数（1990=100）缩减）。

表 2.5.1　中国居民人均消费支出与人均 GDP（元/人）

年份	人均居民消费 CONSP	人均GDP GDPP	年份	人均居民消费 CONSP	人均GDP GDPP
1978	395.8	675.1	1990	797.1	1602.3
1979	437.0	716.9	1991	861.4	1727.2
1980	464.1	763.7	1992	966.6	1949.8
1981	501.9	792.4	1993	1048.6	2187.9
1982	533.5	851.1	1994	1108.7	2436.1
1983	572.8	931.4	1995	1213.1	2663.7
1984	635.6	1059.2	1996	1322.8	2889.1
1985	716.0	1185.2	1997	1380.9	3111.9
1986	746.5	1269.6	1998	1460.6	3323.1
1987	788.3	1393.6	1999	1564.4	3529.3
1988	836.4	1527.0	2000	1690.8	3789.7
1989	779.7	1565.9			

该两组数据是1978—2000年的**时间序列数据**（time series data）；

前述**收入—消费支出例**中的数据是**截面数据**（cross-sectional data）。

1. 建立模型

拟建立如下一元回归模型

$$CONSP = C + \beta\ GDPP + \mu$$

采用**Eviews软件**进行回归分析的结果见下表

表 2.5.2　中国居民人均消费支出对人均 GDP 的回归（1978~2000）

LS // Dependent Variable is CONSP

Sample: 1978 2000

Included observations: 23

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	201.1071	14.88514	13.51060	0.0000
GDPP1	0.386187	0.007222	53.47182	0.0000

R-squared	0.992709	Mean dependent var	905.3331
Adjusted R-squared	0.992362	S.D. dependent var	380.6428
S.E. of regression	33.26711	Akaike info criterion	7.092079
Sum squared resid	23240.71	Schwarz criterion	7.190818
Log likelihood	-112.1945	F-statistic	2859.235
Durbin-Watson stat	0.550288	Prob(F-statistic)	0.000000

一般可写出如下回归分析结果：

$$\widehat{CONSP} = 201.107 + 0.3862GDPP$$

(13.51)　　(53.47)

R^2=0.9927　F=2859.23　DW=0.5503

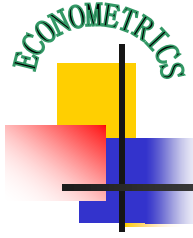
2. 模型检验

R^2=0.9927

T值：C：13.51，　GDPP：53.47

临界值: $t_{0.05/2}(21)=2.08$

斜率项：0<0.3862<1，符合绝对收入假说



3. 预测

2001年：GDPP=4033.1（元）（1990年不变价）

点估计：$CONSP_{2001}$= 201.107 + 0.3862×4033.1

= 1758.7（元）

2001年实测的CONSP（1990年价）:1782.2元，

相对误差: -1.32%。

EViews

File Edit Object View Proc Quick Options Window Help

```
series consp
data consp gdpp
ls consp c gdpp
```

$$E(Y|X_0)=\hat{Y}_0\pm t_{\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum(X_i-\bar{X})^2}}$$

$$Y_0=\hat{Y}_0\pm t_{\frac{\alpha}{2}}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum(X_i-\bar{X})^2}}$$

Workfile: 例221 - (f:\计量经济学\例221.wf1)

View Proc Object Print Save Details+/- Show Fetch Store Delete Genr Sample

Range: 1978 2000 -- 23 obs
Sample: 1978 2000 -- 23 obs
Display Filter: *

c
consp
gdpp
resid

Exp2_2_1 | EXP2_5_1 | New Page

Group: UNTITLED Workfile: 例221:...

View Proc Object Print Name Freeze Default Sort Transpose

obs	CONSP	GDPP
1978	395.8000	675.1000
1979	437.0000	716.9000
1980	464.1000	763.7000
1981	501.9000	792.4000
1982	533.5000	851.1000
1983	572.8000	931.4000
1984	635.6000	1059.200
1985	716.0000	1185.200
1986		

Equation: UNTITLED Workfile: 例221::EXP...

View Proc Object Print Name Freeze Estimate Forecast Stats Resids

Dependent Variable: CONSP
Method: Least Squares
Date: 03/10/09 Time: 13:48
Sample: 1978 2000
Included observations: 23

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	201.1189	14.88402	13.51241	0.0000
GDPP	0.386180	0.007222	53.47471	0.0000

R-squared	0.992710	Mean dependent var	905.3304
Adjusted R-squared	0.992363	S.D. dependent var	380.6334
S.E. of regression	33.26450	Akaike info criterion	9.929800
Sum squared resid	23237.06	Schwarz criterion	10.02854
Log likelihood	-112.1927	Hannan-Quinn criter.	9.954632
F-statistic	2859.544	Durbin-Watson stat	0.550636
Prob(F-statistic)	0.000000		

$\hat{\sigma}=$ 33.26450

Table: UNTITLED Workfile: 例221...

View Proc Object Print Name Edit+/- CellFmt Grid+/- Title Comments+

	A	B	C	D
1	Date: 03/10/09 Time: 13:52			
2	Sample: 1978 2000			
3				
4		CONSP	GDPP	
5				
6	Mean	905.3304	1823.530	
7	Median	797.1000	1565.900	
8	Maximum	1690.800	3789.700	
9	Minimum	395.8000	675.1000	
10	Std. Dev.	380.6334	982.0372	
11	Skewness	0.550085	0.606583	
12	Kurtosis	2.209755	2.093734	
13				
14	Jarque-Bera	1.758410	2.197542	
15	Probability	0.415113	0.333280	
16				
17	Sum	20822.60	41941.20	
18	Sum Sq. Dev.	3187399.	21216735	
19				
20	Observations	23	23	
21				

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

Path = d:\博士学习\孙敬水计量经济学\案例分析 DB = none WF = 例221

2001年人均居民消费的预测区间

人均GDP的样本均值与样本方差：

E(GDPP) =1823.5

Var(GDPP) = 982.04^2=964410.4

在95%的置信度下，$E(CONSP_{2001})$的预测区间为：

$$1758.7 \pm 2.306 \times \sqrt{\frac{23240.71}{23-2} \times \left(\frac{1}{23} + \frac{(4033.1-1823.5)^2}{(23-1) \times 964410.4}\right)}$$

=1758.7±40.13

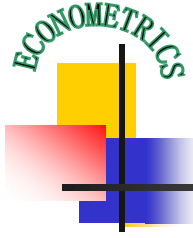
或： （1718.6,1798.8）

同样地，在95%的置信度下，$CONSP_{2001}$的预测区间为：

$$1758.7 \pm 2.306 \times \sqrt{\frac{23240.71}{23-2} \times \left(1+\frac{1}{23}+\frac{(4033.1-1823.5)^2}{(23-1)\times 964410.4}\right)}$$

$$=1758.7 \pm 86.57$$

或　　（1672.1, 1845.3）



谢谢！