第二章 单方程计量经济学模型理论与方法

Theory and Methodology of Single-Equation Econometric Model

第一节 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

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

一. 变量之间的关系

现象(事物)之间的关系 —统计上 数学上→ 变量之间的关系

变量关系
- 函数关系(确定性关系)：Y=f(X)；
- 统计(或相关)关系(不确定关系)：X~Y.

1. 函数关系：一个变量(X)的变化能完全决定另外一个变量的变化，即它们之间有精确的函数关系式 Y=f(X).

例如：

- Taxi收费(Y)和行驶里程(X)的关系：Y=aX+b

- 某种商品的销售额(y)与销售量(x)之间的关系可表示为 $y = p\,x$ (p 为单价)
- 圆的面积（S）与半径之间非关系可表示为 $S = \pi R^2$
- 企业的原材料消耗额(y)与产量(x_1)、单位产量消耗(x_2)、原材料价格(x_3)之间的关系可表示为 $y = x_1\,x_2\,x_3$

二. 统计(相关)关系：*变量之间有密切关系，但密切程度并没有达到由一个完全确定另一个的程度，即不能找到一个精确的函数关系式来描述这种关系，这种关系就称为统计关系(或相关关系).*

比如：

- 商品的消费量(y)与居民收入(x)之间的关系
- 商品销售额(y)与广告费支出(x)之间的关系
- 粮食亩产量(y)与施肥量(x_1)、降雨量(x_2)、温度(x_3)之间的关系
- 收入水平(y)与受教育程度之间的关系(x)
- 父亲身高(y)与子女身高(x)之间的关系

统计关系的研究

相关分析(Correlation analysis)

回归分析(Regression analysis)

▲注意：

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，即区分因变量（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些变量的具体统计依赖关系的计算方法和理论。

其用意：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。

例2.1：一个社区有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。

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

如果该社区全体家庭月收入和消费支出情况未知。现从该社区家庭中抽样10个家庭调查，获得如下数据

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

1）该社区每月**家庭消费支出Y**与每月**可支配收入X**的关系？

2）如果知道了家庭的月收入，预测该社区家庭的平均月消费支出水平？

$$Y = \beta_0 + \beta_1 X + u$$

这里：

前一个变量被称为**被解释变量（Explained Variable）**或**因变量（Dependent Variable）**，后一个（些）变量被称为**解释变量（Explanatory Variable）**或**自变量（Independent Variable）**。

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

在给定解释变量X_1, X_2, …, X_k条件下被解释变量Y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：

$$E(Y \mid X_1, X_2, \cdots, X_k) = f(X_1, X_2, \cdots, X_k)$$

称为**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：

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

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

总体回归函数确定了在给定解释变量X_1，X_2，…，X_k取值的条件下，被解释变量Y的期望值（平均水平）。

但被解释变量Y的观察值Y_i可能与该平均水平有偏差。

记 $$\mu_i = Y_i - E(Y \mid X_{1i}, X_{2i}, \cdots, X_{ki})$$

$$Y_i = E(Y \mid X_{1i}, X_{2i}, \cdots, X_{ki}) + \mu_i \qquad (*)$$

称μ_i为观察值Y_i围绕它的期望值$E(Y \mid X_{1i}, X_{2i}, \ldots, X_{ki})$的**离差**（**deviation**），是一个不可预测的随机变量，又称为**随机干扰项**（**stochastic disturbance**）或**随机误差项**（**stochastic error**）。

例2.1中，个别家庭的消费支出为：

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

即，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。

（*）式称为**总体回归函数（方程，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的影响：

1）在解释变量中被忽略的因素的影响；

2）变量观测值的观测误差的影响；

3）模型关系的设定误差的影响；

4）其它随机因素的影响。

四、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？

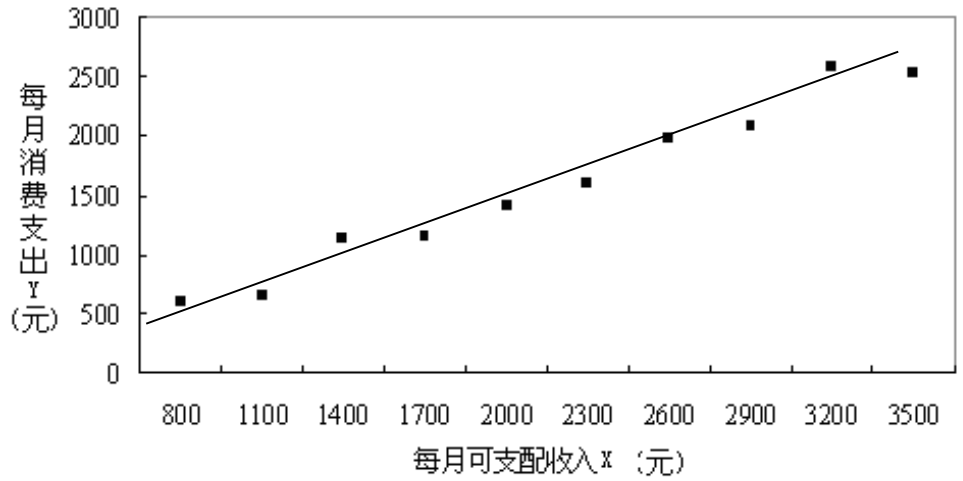
例：一个社区有100户家庭组成，要研究该社区每月家庭消费支出Y与每月家庭可支配收入X的关系。现从该总体中得到如下一个样本，

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

问：能否从该样本估计总体回归函数PRF？

核样本的**散点图**（scatter diagram)：



样本散点图近似于一条函数曲线（直线），画一条函数曲线（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）**。

记样本回归线的函数形式为：

$$\hat{Y}_i = \hat{f}(X_{1i}, X_{2i}, \cdots, X_{ki}) \overset{\text{线性函数}}{=} \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}$$

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

$$\hat{Y}_i = \hat{f}(X_{1i}, X_{2i}, \cdots, X_{ki}) \overset{\text{线性函数}}{=} \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}$$

$$\Downarrow$$

$$E(Y \mid X_{1i}, X_{2i}, \cdots, X_{ki}) = f(X_{1i}, X_{2i}, \cdots, X_{ki})$$

$$\overset{\text{线性函数}}{=} \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$$

即 $\hat{Y}_i$ 为 $E(Y \mid X_{1i}, X_{2i}, \cdots, X_{ki})$的估计

$\hat{\beta}_i$ 为 β_i 的估计

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：

$$Y_i = \hat{Y}_i + \hat{\mu}_i \overset{\Delta}{=} \hat{f}(X_{1i}, X_{2i}, \cdots, X_{ki}) + e_i$$

$$\overset{\text{线性函数}}{=} \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki} + e_i$$

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 $Y_i = \hat{Y}_i + e_i \overset{\text{线性}}{=} \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki} + e_i$

估计 $Y_i = E(Y \mid X_{1i}, X_{2i}, \cdots, X_{ki}) + \mu_i \overset{\text{线性}}{=} \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \mu_i$

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 $\hat{\beta}_i$ 尽可能接近 β_i 。

注意：这里PRF可能永远无法知道。

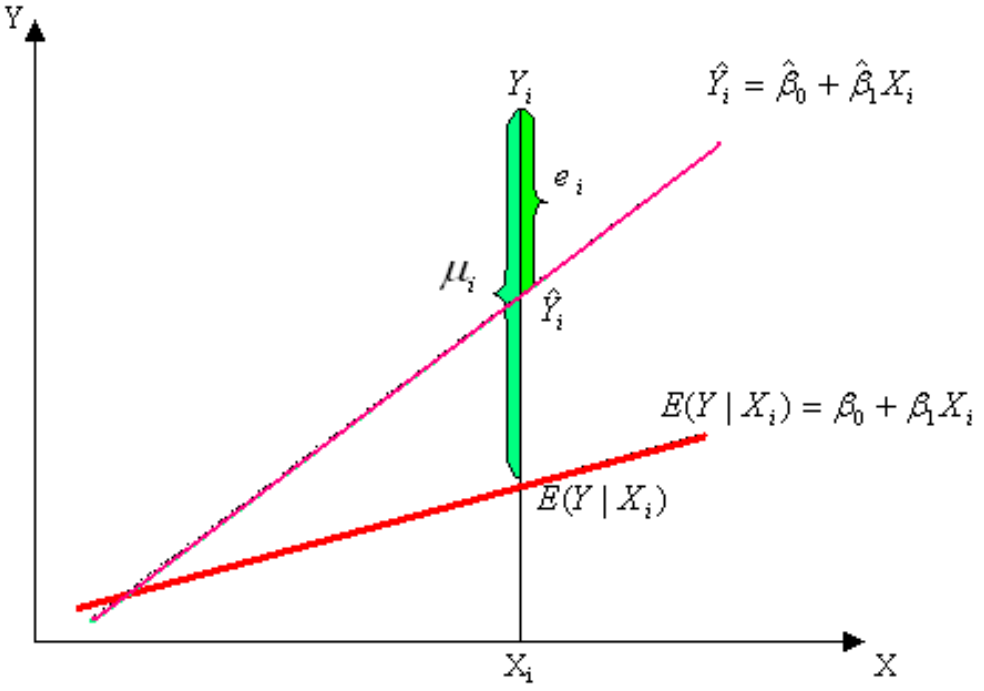


图 2.1.3　总体回归线与样本回归线的基本关系

第二节 一元线性回归模型

- 一元线性回归模型的参数估计
- 一元线性回归模型检验
- 一元线性回归模型预测
- 实例

（一）一元线性回归模型的参数估计

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

$$Y = \beta_0 + \beta_1 X + \mu$$

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

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，其种最广泛使用的是**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。

注：实际这些假设与所采用的估计方法紧密相关。

一、线性回归模型的基本假设

假设1、 随机误差项μ_i不序列相关性：

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

假设2、 随机误差项μ与解释变量X之间不相关：

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

假设3、 μ_i服从零均值同方差的正态分布

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

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）。

另外，在进行模型回归时，还有两个暗含的假设：

假设4：随着样本容量的无限增加，解释变量X的样本方差趋于一有限常数。即

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

假设5：回归模型是正确设定的

假设4旨在排除时间序列数据出现持续上升或下降的变量作为解释变量，因为这类数据不仅使大样本统计推断变得无效，而且往往产生所谓的**伪回归问题**（spurious regression problem）。

假设5也被称为模型没有**设定偏误**（specification error）

二、参数的普通最小二乘估计（OLS）

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。

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：

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

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

方程组（*）称为**正规方程组（normal equations）**。

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

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**）。

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

则有

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

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

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。

三、参数估计的最大似然法(ML)

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

基本原理:

对于**最大似然法**，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。

在满足基本假设条件下，对一元线性回归模型：

$$Y_i = \beta_0 + \beta_1 X_i + \mu_i$$

随机抽取n组样本观测值（X_i, Y_i）（i=1,2,…n）。

那么Y_i服从如下的正态分布：

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

于是，Y的概率函数为

$$P(Y_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2} \qquad (i=1,2,\ldots n)$$

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**似然函数**(likelihood function)为：

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

$$= \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} e^{-\frac{1}{2\sigma^2}\Sigma(Y_i - \beta_0 - \beta_1 X_i)^2}$$

将该似然函数极大化，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：

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

$$= -n\ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\Sigma(Y_i - \beta_0 - \beta_1 X_i)^2$$

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:

$$\begin{cases} \dfrac{\partial}{\partial \hat{\beta}_0}\Sigma(Y_i - \beta_0 - \beta_1 X_i)^2 = 0 \\ \dfrac{\partial}{\partial \hat{\beta}_1}\Sigma(Y_i - \beta_0 - \beta_1 X_i)^2 = 0 \end{cases}$$

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1：在家庭**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 2.1.3　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

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	−1350	−973	1314090	1822500	947508	640000	352836
2	1100	638	−1050	−929	975870	1102500	863784	1210000	407044
3	1400	1122	−750	−445	334050	562500	198381	1960000	1258884
4	1700	1155	−450	−412	185580	202500	170074	2890000	1334025
5	2000	1408	−150	−159	23910	22500	25408	4000000	1982464
6	2300	1595	150	28	4140	22500	762	5290000	2544025
7	2600	1969	450	402	180720	202500	161283	6760000	3876961
8	2900	2078	750	511	382950	562500	260712	8410000	4318084
9	3200	2585	1050	1018	1068480	1102500	1035510	10240000	6682225
10	3500	2530	1350	963	1299510	1822500	926599	12250000	6400900
求和	21500	15674			5769300	7425000	4590020	53650000	29157448
平均	2150	1567							

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

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

因此，由该样本估计的回归方程为：

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

四、最小二乘估计量的性质

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(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

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

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

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

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。

(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

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

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

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

五、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

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

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

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**。

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。

由于随机项$\boldsymbol{\mu}_i$不可观测，只能从$\boldsymbol{\mu}_i$的估计——残差$\boldsymbol{e}_i$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

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

它是关于σ^2的无偏估计量。

在**最大似然估计法**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：　$S^2_{\hat{\beta}_1} = \hat{\sigma}^2 \Big/ \sum x_i^2$

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

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

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

（二）一元线性回归模型的统计检验

- 通过参数估计得到总体回归函数的估计表达式样本回归函数以后，还必须对样本回归函数能否代表总体回归函数进行统计推断，即进行所谓的统计检验。。
- 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。

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？

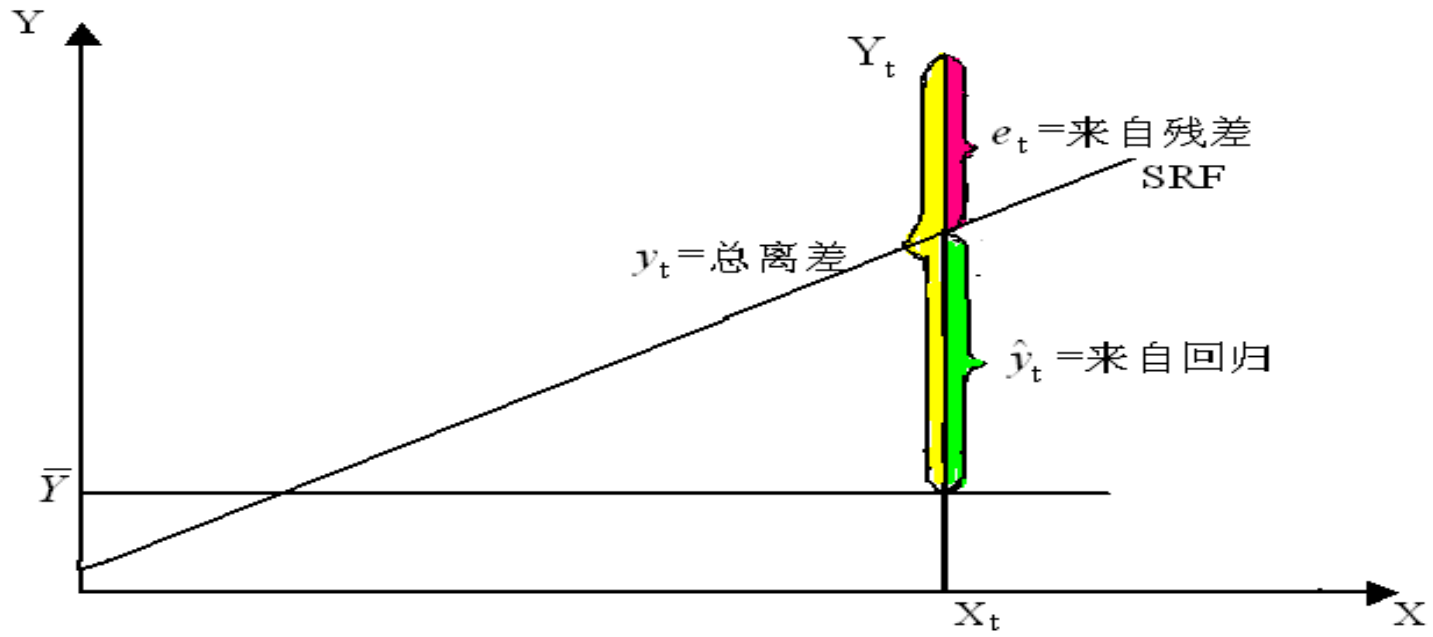
1、总离差平方和的分解

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

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



$\hat{y}_i = (\hat{Y}_i - \bar{Y})$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。

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,可以证明：

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

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

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

Y的观测值围绕其均值的**总离差(total variation)**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。

在实际计算可决系数时，在$\hat{\beta}_1$已经估计出后：

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

在例2.1.1的**收入-消费支出**例中，

$$R^2 = \hat{\beta}_1^2 \frac{\sum x_i^2}{\sum y_i^2} = \frac{(0.777)^2 \times 7425000}{4590020} = 0.9766$$

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。

计量经计学中，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。

1、假设检验

- 所谓**假设检验**，就是事先对总体作出一个假设，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。**

即在一次抽样中，小概率事件（$P<\alpha$） 不可能发生。如果在原假设下发生了小概率事件，则认为原假设是不合理的；反之，则认为原假设是合理的。

• **假设检验是基于样本资料来推断总体特征的，而这种推断是在一定概率置信度下（α）进行的，而非严格的逻辑证明。**

因此，置信度（ α ）大小的不同，有可能做出不同的判断。

• **检验假设H_0（H_1）作出统计推断时，假设检验可能会犯两类错误.**

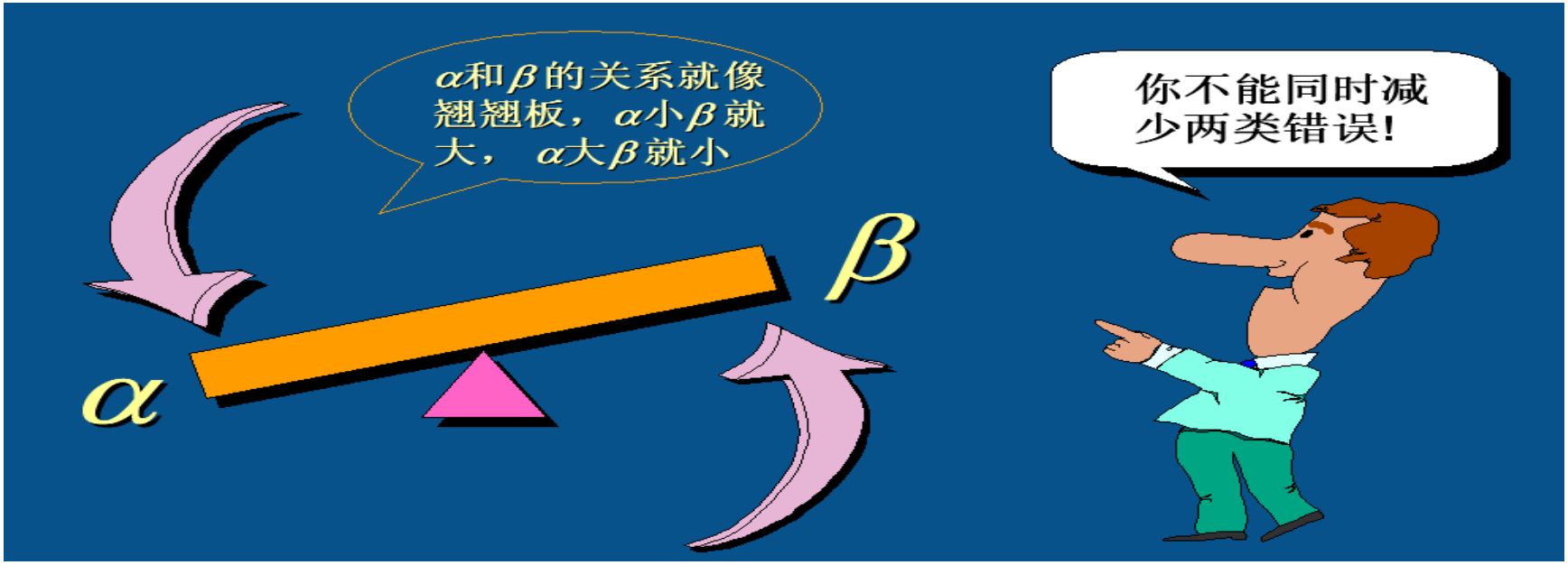
原假设 / 行动决策 / 状态	H_0真实	H_0不真实
接受H_0	正确	取伪错误
拒绝H_0	弃真错误	正确

$$\alpha=P(\text{弃真错误})=P(\text{拒绝}H0 \mid H0)$$
$$\beta=P(\text{取伪错误})=P(\text{接受}H0 \mid H1)$$

- 假设检验原则上规定只控制 α .这种作法可能会夸大H_0的可信程度.

$$P(\text{弃真错误})=P(\text{拒绝}H_0 \mid H_0) = \alpha \leq \alpha$$

α 就是显著性水平（即所允许犯弃真错误的概率），也记其为α。



Example 2. An economist estimates that the average Canadian household saves 15% of its income. In a random sample of 64 households, the average saving rate is found to be 14%, and the standard deviation is 7%. Do we have enough evidence to refute the economist's claim(α=0.05)?

Solution. Hypotheses: $H_0 : \mu = 15; \quad H_1 : \mu \neq 15$

$$\text{Test Statistic: } Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} = \frac{0.14 - 0.15}{0.7/\sqrt{64}} = -1.14$$

Level of significance: $\alpha = 0.05$

Critical z values: $Z_{\alpha/2} = Z_{0.025} = 1.96$

Conclusion: Since the value of $|Z|$ is less than $Z_{0.025} = 1.96$, we cannot reject H_0. We can not refute the economist's claim.

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：　$\beta_1=0$，　　H_1：$\beta_1\neq0$

（2）以原假设H_0构造t统计量，并由样本计算其值

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

（3）给定显著性水平α，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　$|t|\leq t_{\alpha/2}(n-2)$，则拒绝$H_1$，接受$H_0$；

对于一元线性回归方程中的β_0，可构造如下t统计量进行显著性检验：

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

在上述收入-消费支出例中，首先计算σ_2的估计值

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

$$\hat{\sigma}^2=\frac{\sum e_i^2}{n-2}=\frac{\sum y_i^2-\hat{\beta}_1^2\sum x_i^2}{n-2}=\frac{4590020-0.777^2\times 7425000}{10-2}=13402$$

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

$$S_{\hat{\beta}_1}=\sqrt{\hat{\sigma}^2/\sum x_i^2}=\sqrt{13402/7425000}=\sqrt{0.0018}=0.0425$$

$$S_{\hat{\beta}_0}=\sqrt{\hat{\sigma}^2\sum X_i^2/n\sum x_i^2}=\sqrt{13402\times 53650000/10\times 7425000}=98.41$$

t统计量的计算结果分别为：

$$t_1 = \hat{\beta}_1 \Big/ S_{\hat{\beta}_1} = 0.777/0.0425 = 18.29$$

$$t_0 = \hat{\beta}_0 \Big/ S_{\hat{\beta}_0} = -103.17/98.41 = -1.048$$

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；**

$|t_2|<2.306$, 表明在95%的置信度下，无法拒绝截距项为零的假设。

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：

$$t = \frac{\hat{\beta}_i - \beta_i}{s_{\hat{\beta}_i}} \sim t(n-2)$$

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：

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

即

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

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

于是得到:(1−α)的置信度下，β_i的置信区间是

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

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 $S_{\hat{\beta}_1} = 0.042$ $S_{\hat{\beta}_0} = 98.41$

于是，β_1、β_0的置信区间分别为:

(0.6345, 0.9195)

(−433.32, 226.98)

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

（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。

(三). 一元线性回归分析的应用：预测问题

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。

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

对总体回归函数$E(Y|X=X_0)=\beta_0+\beta_1 X$，$X=X_0$时

$$E(Y|X=X_0)=\beta_0+\beta_1 X_0$$

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

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

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

可见，$\hat{Y}_0$是条件均值$E(Y|X=X_0)$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 $$Var(\hat{Y}_0)=\frac{\sigma^2\sum X_i^2}{n\sum x_i^2}-\frac{2X_0\bar{X}\sigma^2}{\sum x_i^2}+\frac{X_0^2\sigma^2}{\sum x_i^2}$$

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

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

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

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

$$t=\frac{\hat{Y}_0-(\beta_0+\beta_1X_0)}{S_{\hat{Y}_0}}\sim t(n-2)$$

其中 $$S_{\hat{Y}_0}=\sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum x_i^2}\right)}$$

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

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

从而在1-α的置信度下，**Y_0的置信区间**为

$$\hat{Y}_0-t_{\frac{\alpha}{2}}\times S_{\hat{Y}_0-Y_0}<Y_0<\hat{Y}_0+t_{\frac{\alpha}{2}}\times S_{\hat{Y}_0-Y_0}$$

在上述**收入-消费支出**例中，得到的样本回归函数为

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

则在 X_0=1000处， $\hat{Y}_0 = -103.172+0.777\times 1000=673.84$

而 $$Var(\hat{Y}_0) = 13402\left[\frac{1}{10} + \frac{(1000-2150)^2}{7425000}\right] = 3727.29$$

$$S(\hat{Y}_0) = 61.05$$

因此，**总体均值**E(Y|X=1000)的95%的置信区间为：

673.84-2.306×61.05< E(Y|X=1000) <673.84+2.306×61.05

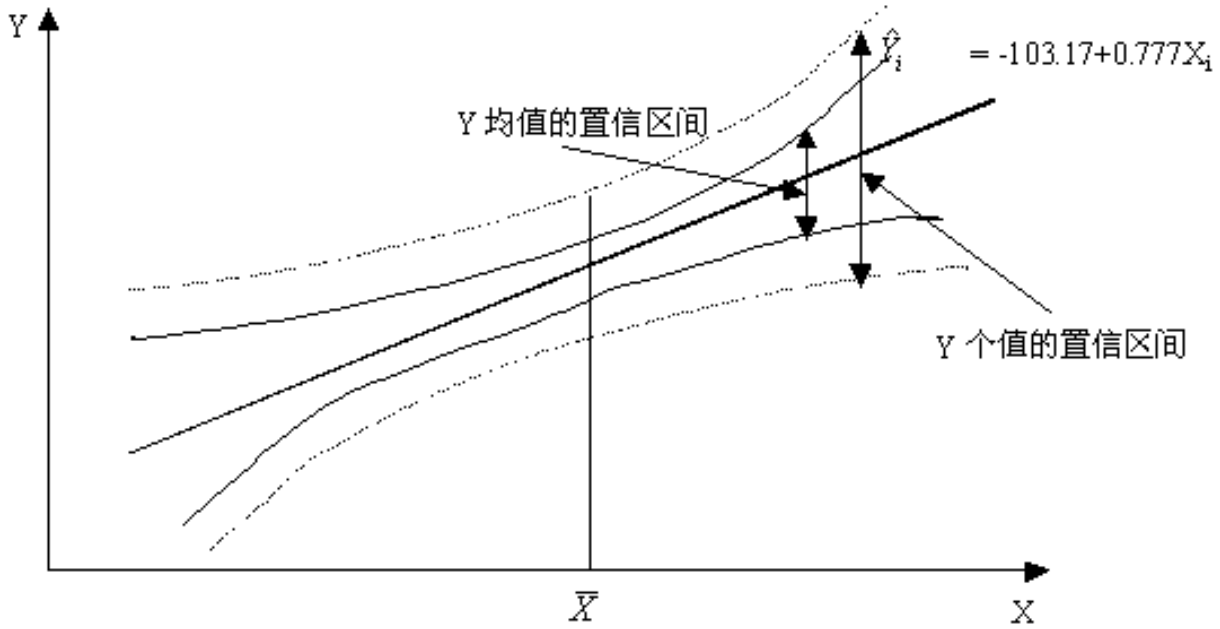
或 （533.05, 814.62）

同样地，对于Y在X=1000的**个体值**，其95%的置信区间为：

$$673.84 - 2.306\times61.05<Y_{x=1000}<673.84 + 2.306\times61.05$$

或 $(372.03, 975.65)$

- 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。

(四) 实例：时间序列问题

一、中国居民人均消费模型

例2 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）（90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%。

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：

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

=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：

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

=1758.7±86.57

或　（1672.1, 1845.3）

二、时间序列问题

上述实例表明，时间序列完全可以进行类似于截面数据的回归分析。

然而，在时间序列回归分析中，有两个需注意的问题：

第一，关于抽样分布的理解问题。

能把表2.5.1中的数据理解为是从某个总体中抽出的一个样本吗？

第二，关于“伪回归问题”（spurious regression problem）。

在现实经济问题中，对时间序列数据作回归，即使两个变量间没有任何的实际联系，也往往会得到较高的可决系数，尤其对于**具有相同变化趋势（同时上升或下降）的变量**，更是如此。

这种现象被称为“**伪回归**”或“**虚假回归**”。

第三节 经典单方程计量经济学模型：多元线性回归

- 多元线性回归模型
- 多元线性回归模型的参数估计
- 多元线性回归模型的统计检验
- 多元线性回归模型的预测
- 回归模型的其他形式
- 回归模型的参数约束

（一） 多元线性回归模型

一、多元线性回归模型

多元线性回归模型:表现在线性回归模型中的解释变量有多个。

一般表现形式:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \mu_i \qquad i=1,2\ldots,n$$

其中:k为解释变量的数目，β_j称为**回归参数**（regression coefficient）。

习惯上：把**常数项**看成为一**虚变量**的系数，该虚变量的样本观测值始终取1。这样：

模型中解释变量的数目为（k+1）

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \mu_i$$

也被称为总体回归函数的随机表达形式。它 的非随机表达式为:

$$E(Y_i \mid X_{1i}, X_{2i}, \cdots X_{ki}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

方程表示：各变量X值固定时Y的平均响应。

β_j也被称为偏回归系数，表示在其他解释变量保持不变的情况下，X_j每变化1个单位时，Y的均值E(Y)的变化;

或者说β_j给出了X_j的单位变化对Y均值的“直接”或“净”（不含其他变量）影响。

总体回归模型n个随机方程的矩阵表达式为

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mu}$$

其中

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix}_{n\times(k+1)}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1)\times 1} \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}_{n\times 1}$$

样本回归函数：用来估计总体回归函数

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_{ki} X_{ki}$$

其随机表示式：

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_{ki} X_{ki} + e_i$$

e_i称为残差或剩余项(residuals)，可看成是总体回归函数中随机扰动项μ_i的近似替代。

样本回归函数的矩阵表达：

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \qquad \text{或} \qquad \mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}$$

其中：

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} \qquad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

二、多元线性回归模型的基本假定

假设1，解释变量是非随机的或固定的，且各X之间互不相关（无多重共线性）。

假设2，随机误差项具有零均值、同方差及不序列相关性

$$E(\mu_i) = 0$$

$$Var(\mu_i) = E(\mu_i^2) = \sigma^2 \qquad i \neq j \quad i, j = 1,2,\cdots,n$$

$$Cov(\mu_i, \mu_j) = E(\mu_i \mu_j) = 0$$

假设3，解释变量与随机项不相关

$$Cov(X_{ji}, \mu_i) = 0 \qquad j = 1,2\cdots,k$$

假设4，随机项满足正态分布

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

上述假设的矩阵符号表示 式：

假设1，$n\times(k+1)$矩阵$\mathbf{X}$是非随机的，且$\mathbf{X}$的秩$\rho=k+1$，即$\mathbf{X}$满秩。

假设2，

$$E(\boldsymbol{\mu})=E\begin{pmatrix}\mu_1\\ \vdots\\ \mu_n\end{pmatrix}=\begin{pmatrix}E(\mu_1)\\ \vdots\\ E(\mu_n)\end{pmatrix}=0$$

$$E(\boldsymbol{\mu}\boldsymbol{\mu}')=E\left(\begin{pmatrix}\mu_1\\ \vdots\\ \mu_n\end{pmatrix}\begin{pmatrix}\mu_1 & \cdots & \mu_n\end{pmatrix}\right)=E\begin{pmatrix}\mu_1^2 & \cdots & \mu_1\mu_n\\ \vdots & \ddots & \vdots\\ \mu_n\mu_1 & \cdots & \mu_n^2\end{pmatrix}$$

$$=\begin{pmatrix}\mathrm{var}(\mu_1) & \cdots & \mathrm{cov}(\mu_1,\mu_n)\\ \vdots & \ddots & \vdots\\ \mathrm{cov}(\mu_n,\mu_1) & \cdots & \mathrm{var}(\mu_n)\end{pmatrix}=\begin{pmatrix}\sigma^2 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sigma^2\end{pmatrix}=\sigma^2\mathbf{I}$$

假设4，向量μ有一多维正态分布，即

$$\mu \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

同一元回归一样，多元回归还具有如下两个重要假设：

假设5，样本容量趋于无穷时，各解释变量的方差趋于有界常数，即$n \to \infty$时，

$$\frac{1}{n}\sum x_{ji}^2 = \frac{1}{n}\sum (X_{ji} - \overline{X}_j)^2 \to Q_j \qquad \text{或} \qquad \frac{1}{n}\mathbf{x}'\mathbf{x} \to \mathbf{Q}$$

其中：**Q**为一非奇异固定矩阵，矩阵**x**是由各解释变量的离差为元素组成的$n \times k$阶矩阵

$$\mathbf{x} = \begin{pmatrix} x_{11} & \cdots & x_{k1} \\ \vdots & \cdots & \vdots \\ x_{1n} & \cdots & x_{kn} \end{pmatrix}$$

假设6，回归模型的设定是正确的。

（二） 多元线性回归模型的估计

估计目标：结构参数$\hat{\beta}_j$及随机误差项的方差$\hat{\sigma}^2$

估计方法：OLS、ML或者MM

一、普通最小二乘估计

*二、最大或然估计

*三、矩估计

四、参数估计量的性质

五、样本容量问题

六、估计实例

一、普通最小二乘估计

对于随机抽取的n组观测值 $(Y_i, X_{ji}), i = 1,2,\cdots,n,\ j = 0,1,2,\cdots k$

如果**样本函数**的参数估计值已经得到，则有：

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_{ki} X_{Ki} \qquad \text{i=1,2…n}$$

根据**最小二乘原理**，参数估计值应该是下列方程组的解

$$\begin{cases} \dfrac{\partial}{\partial \hat{\beta}_0} Q = 0 \\ \dfrac{\partial}{\partial \hat{\beta}_1} Q = 0 \\ \dfrac{\partial}{\partial \hat{\beta}_2} Q = 0 \\ \vdots \\ \dfrac{\partial}{\partial \hat{\beta}_k} Q = 0 \end{cases}$$

其中 $Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$

$$= \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}))^2$$

于是得到关于待估参数估计值的**正规方程组**：

$$\begin{cases} \Sigma(\hat{\beta}_0+\hat{\beta}_1X_{1i}+\hat{\beta}_2X_{2i}+\cdots+\hat{\beta}_kX_{ki})=\Sigma Y_i \\ \Sigma(\hat{\beta}_0+\hat{\beta}_1X_{1i}+\hat{\beta}_2X_{2i}+\cdots+\hat{\beta}_kX_{ki})X_{1i}=\Sigma Y_iX_{1i} \\ \Sigma(\hat{\beta}_0+\hat{\beta}_1X_{1i}+\hat{\beta}_{2i}X_{2i}+\cdots+\hat{\beta}_kX_{ki})X_{2i}=\Sigma Y_iX_{2i} \\ \vdots \\ \Sigma(\hat{\beta}_0+\hat{\beta}_1X_{1i}+\hat{\beta}_2X_{2i}+\cdots+\hat{\beta}_kX_{ki})X_{ki}=\Sigma Y_iX_{ki} \end{cases}$$

解该（k+1）个方程组成的线性代数方程组，即可得到(k+1)个待估参数的估计值 $\hat{\beta}_j, j=0,1,2,\cdots,k$ 。

正规方程组的矩阵形式

$$\begin{pmatrix} n & \sum X_{1i} & \cdots & \sum X_{ki} \\ \sum X_{1i} & \sum X_{1i}^2 & \cdots & \sum X_{1i}X_{ki} \\ \cdots & \cdots & \cdots & \cdots \\ \sum X_{ki} & \sum X_{ki}X_{1i} & \cdots & \sum X_{ki}^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \cdots \\ \hat{\beta}_k \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{12} & \cdots & X_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ X_{k1} & X_{k2} & \cdots & X_{kn} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{pmatrix}$$

即

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$$

由于**X'X**满秩，故有

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

将上述过程用矩阵表示如下：

寻找一组参数估计值 $\hat{\boldsymbol{\beta}}$，使得残差平方和

$$Q=\sum_{i=1}^{n}e_i^2=\mathbf{e}'\mathbf{e}=(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})$$

最小。

即求解方程组：$\dfrac{\partial}{\partial\hat{\boldsymbol{\beta}}}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})=0$

$$\frac{\partial}{\partial\hat{\boldsymbol{\beta}}}(\mathbf{Y}'\mathbf{Y}-\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}-\mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}}+\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}})=0$$

$$\frac{\partial}{\partial\hat{\boldsymbol{\beta}}}(\mathbf{Y}'\mathbf{Y}-2\mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}}+\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}})=0$$

$$-\mathbf{X}'\mathbf{Y}+\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}=0$$

得到：$\mathbf{X}'\mathbf{Y}=\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$

于是：$\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

例3：在例2的**家庭收入-消费支出**例中，

$$(\mathbf{X}'\mathbf{X})=\begin{pmatrix}1 & 1 & \cdots & 1\\ X_1 & X_2 & \cdots & X_n\end{pmatrix}\begin{pmatrix}1 & X_1\\ 1 & X_2\\ \cdots & \cdots\\ 1 & X_n\end{pmatrix}=\begin{pmatrix}n & \sum X_i\\ \sum X_i & \sum X_i^2\end{pmatrix}=\begin{pmatrix}10 & 21500\\ 21500 & 53650000\end{pmatrix}$$

$$\mathbf{X}'\mathbf{Y}=\begin{pmatrix}1 & 1 & \cdots & 1\\ X_1 & X_2 & \cdots & X_n\end{pmatrix}\begin{pmatrix}Y_1\\ Y_2\\ \cdots\\ Y_n\end{pmatrix}=\begin{pmatrix}\sum Y_i\\ \sum X_iY_i\end{pmatrix}=\begin{pmatrix}15674\\ 39468400\end{pmatrix}$$

可求得

$$(\mathbf{X}'\mathbf{X})^{-1}=\begin{pmatrix}0.7226 & -0.0003\\ -0.0003 & 1.35E-07\end{pmatrix}$$

于是

$$\hat{\boldsymbol{\beta}}=\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix}=\begin{pmatrix}0.7226 & -0.0003\\ -0.0003 & 1.35E-07\end{pmatrix}\begin{pmatrix}15674\\ 39648400\end{pmatrix}=\begin{pmatrix}-103.172\\ 0.7770\end{pmatrix}$$

正规方程组 的另一种写法

对于正规方程组

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

将 $\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}$ 代入得

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{e} = \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

于是

$$\mathbf{X}'\mathbf{e} = \mathbf{0} \qquad (*)$$

或

$$\begin{cases} \sum e_i = 0 \\ \sum_i X_{ji} e_i = 0 \end{cases} \qquad (**)$$

(*)或（**）是多元线性回归模型正规方程组的另一种写法

样本回归函数的离差形式

$$y_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki} + e_i \qquad \text{i=1,2…n}$$

其矩阵形式为

$$\mathbf{y} = \mathbf{x}\hat{\boldsymbol{\beta}} + \mathbf{e}$$

其中：

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{x} = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{k1} \\ x_{12} & x_{22} & \cdots & x_{k2} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix} \quad \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$$

在离差形式下，参数的最小二乘估计结果为

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}_1 - \cdots - \hat{\beta}_k \overline{X}_k$$

随机误差项μ的方差σ的无偏估计

可以证明，随机误差项μ的方差的无偏估计量为

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k-1} = \frac{\mathbf{e}'\mathbf{e}}{n-k-1}$$

*二、最大或然估计

对于多元线性回归模型

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \mu_i$$

易知 $$Y_i \sim N(\mathbf{X_i\beta}, \sigma^2)$$

Y的随机抽取的n组样本观测值的联合概率

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

$$= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} e^{-\frac{1}{2\sigma^2}\Sigma(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}))^2}$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} e^{-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}$$

即为变量Y的似然函数

对数或然函数为

$$L^* = Ln(L)$$

$$= -nLn(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})\,(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

对对数或然函数求极大值，也就是对

$$(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

求极小值。

因此，参数的最大或然估计为

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

结果与参数的普通最小二乘估计相同

*三、矩估计（Moment Method, MM）

OLS估计是通过得到一个关于参数估计值的正规方程组

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$$

并对它进行求解而完成的。

该正规方程组 可以从另外一种思路来导：

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mu}$$

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{X}'\boldsymbol{\mu}$$

$$\mathbf{X}'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{X}'\boldsymbol{\mu}$$

求期望：

$$E(\mathbf{X}'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$$

$$E(\mathbf{X}'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})=\mathbf{0}$$

称为原总体回归方程的一组矩条件，表明了原总体回归方程所具有的内在特征。

如果随机抽出原总体的一个样本，估计出的样本回归方程

$$\hat{\mathbf{Y}}=\mathbf{X}\hat{\boldsymbol{\beta}}$$

能够近似代表总体回归方程的话，则应成立：

$$\frac{1}{n}\mathbf{X}'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})=0$$

由此得到正规方程组

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}=\mathbf{X}'\mathbf{Y}$$

解此正规方程组即得参数的MM估计量。

易知MM估计量与OLS、ML估计量等价。

矩方法是工具变量方法(Instrumental Variables,IV)和广义矩估计方法(Generalized Moment Method, GMM)的基础

- 在矩方法中关键是利用了

$$E(X'\mu)=0$$

- 如果某个解释变量与随机项相关，只要能找到1个工具变量，仍然可以构成一组矩条件。这就是IV。

- 如果存在＞k+1个变量与随机项不相关，可以构成一组包含＞k+1方程的矩条件。这就是GMM。

四、参数估计量的性质

在满足基本假设的情况下，其结构参数β的普通最小二乘估计、**最大或然估计**及**矩估计**仍具有：

线性性、无偏性、有效性。

同时，随着样本容量增加，参数估计量具有：

渐近无偏性、渐近有效性、一致性。

1、线性性

$$\hat{\beta} = (X'X)^{-1}X'Y = CY$$

其中, $C=(X'X)^{-1}X'$ 为一仅与固定的X有关的行向量

2、无偏性

$$
\begin{aligned}
E(\hat{\boldsymbol{\beta}}) &= E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) \\
&= E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\mu})) \\
&= \boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}E(\mathbf{X}'\boldsymbol{\mu}) \\
&= \boldsymbol{\beta}
\end{aligned}
$$

这里利用了假设：$E(X'\mu)=0$

3、有效性（最小方差性）

参数估计量 $\hat{\beta}$ 的方差-协方差矩阵

$$
\begin{aligned}
Cov(\hat{\beta}) &= E(\hat{\beta}-E(\hat{\beta}))(\hat{\beta}-E(\hat{\beta}))' \\
&= E(\hat{\beta}-\beta)(\hat{\beta}-\beta)'
\end{aligned}
$$

$$= E((XX')^{-1}X'\,\mu\,\mu'X(X'X)^{-1})$$
$$= (X'X)^{-1}X'E(\mu\,\mu')X(X'X)^{-1}$$
$$= E(\mu\,\mu')(X'X)^{-1}$$
$$= \sigma^2\mathbf{I}(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1}$$

其中利用了

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$
$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mu})$$
$$= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\mu}$$

和 $E(\boldsymbol{\mu}\,\boldsymbol{\mu}') = \sigma^2\mathbf{I}$

根据高斯—马尔可夫定理，$Cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$ 在所有无偏估计量的方差中是最小的。

五、样本容量问题

1. 最小样本容量

所谓“**最小样本容量**”，即从最小二乘原理和最大或然原理出发，欲得到参数估计量，不管其质量如何，所要求的样本容量的下限。

样本最小容量必须不少于模型中解释变量的数目（包括常数项），即

$$n \geq k+1$$

因为，无多重共线性要求：秩$(\mathbf{X})=k+1$

2、满足基本要求的样本容量

从统计检验的角度：

n>30 时，Z检验才能应用；

n-k≥8时，t分布较为稳定

一般经验认为：

当n≥30或者至少n≥3(k+1)时，才能说满足模型估计的基本要求。

模型的良好性质只有在大样本下才能得到理论上的证明

六、多元线性回归模型的参数估计实例

例3 在例2中，已建立了**中国居民人均消费**一元线性模型。这里我们再考虑建立多元线性模型。

解释变量：人均GDP：GDPP

前期消费：CONSP(-1)

估计区间：1979~2000年

Eviews软件估计结果

LS // Dependent Variable is CONS

Sample(adjusted): 1979 2000

Included observations: 22 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	120.7000	36.51036	3.305912	0.0037
GDPP	0.221327	0.060969	3.630145	0.0018
CONSP(-1)	0.451507	0.170308	2.651125	0.0158

R-squared	0.995403	Mean dependent var	928.4946
Adjusted R-squared	0.994920	S.D. dependent var	372.6424
S.E. of regression	26.56078	Akaike info criterion	6.684995
Sum squared resid	13404.02	Schwarz criterion	6.833774
Log likelihood	-101.7516	F-statistic	2057.271
Durbin-Watson stat	1.278500	Prob(F-statistic)	0.000000

随机误差项的方差的估计值为$\hat{\sigma}^2$=13404.02/(22-3)=705.47

（三） 多元线性回归模型的统计检验

一、拟合优度检验

二、方程的显著性检验(F检验)

三、变量的显著性检验（t检验）

四、参数的置信区间

一、拟合优度检验

1、可决系数与调整的可决系数

总离差平方和的分解

记 $TSS=\sum(Y_i-\bar{Y})^2$ 总离差平方和

$ESS=\sum(\hat{Y}_i-\bar{Y})^2$ 回归平方和

$RSS=\sum(Y_i-\hat{Y}_i)^2$ 剩余平方和

则
$$\begin{aligned} TSS &= \Sigma(Y_i-\bar{Y})^2 \\ &= \Sigma((Y_i-\hat{Y}_i)+(\hat{Y}_i-\bar{Y}))^2 \\ &= \Sigma(Y_i-\hat{Y}_i)^2+2\Sigma(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y})+\Sigma(\hat{Y}_i-\bar{Y})^2 \end{aligned}$$

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

$$=\hat{\beta}_0\sum e_i+\hat{\beta}_1\sum e_iX_{1i}+\cdots+\hat{\beta}_k\sum e_iX_{ki}+\bar{Y}\sum e_i$$

$$=0$$

所以有：

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

注意：一个有趣的现象

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

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

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

可决系数

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

该统计量越接近于1，模型的拟合优度越高。

问题：

在应用过程中发现，如果在模型中增加一个解释变量，R^2往往增大（Why?）

这就给人一个错觉：**要使得模型拟合得好，只要增加解释变量即可。**

但是，现实情况往往是，由增加解释变量个数引起的R^2的增大与拟合好坏无关，**R^2需调整。**

调整的可决系数（adjusted coefficient of determination）

在样本容量一定的情况下，增加解释变量必定使得自由度减少，所以调整的思路是:**将残差平方和与总离差平方和分别除以各自的自由度，以剔除变量个数对拟合优度的影响:**

$$\overline{R}^2 = 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)}$$

其中：n-k-1为残差平方和的自由度，n-1为总体平方和的自由度。

$\overline{R}^2$与R^2之间存在如下关系：

$$\overline{R}^2 = 1-(1-\overline{R}^2)\frac{n-1}{n-k-1}$$

在中国居民消费支出的二元模型例中，$\overline{R}^2$=0.9954

在中国居民消费支出的一元模型例中，$\overline{R}^2$=0.9927

问题：$\overline{R}^2$多大才算通过拟合优度检验？

*2、赤池信息准则和施瓦茨准则

为了比较所含解释变量个数不同的多元回归模型的拟合优度，常用的标准还有：

赤池信息准则（Akaike information criterion, AIC）

$$AIC = \ln \frac{\mathbf{e}'\mathbf{e}}{n} + \frac{2(k+1)}{n}$$

施瓦茨准则（Schwarz criterion，SC）

$$AC = \ln \frac{\mathbf{e}'\mathbf{e}}{n} + \frac{k}{n} \ln n$$

这两准则均要求仅当所增加的解释变量能够减少AIC值或AC值时才在原模型中增加该解释变量。

Eviews的估计结果显示：

中国居民消费二元例中：

AIC=6.68　　AC=6.83

中国居民消费一元例中：

AIC=7.09　　AC=7.19

从这点看，可以说前期人均居民消费CONSP(-1)应包括在模型中。

二、方程的显著性检验(F检验)

方程的显著性检验，旨在对模型中被解释变量与解释变量之间的线性关系在总体上是否显著成立作出推断。

1、方程显著性的F检验

即检验模型

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\ \dots\ +\beta_kX_{ki}+\mu_i \qquad i=1,2,\ \dots,n$$

是否显著。

可提出如下原假设与备择假设：

H_0： $\beta_0=\beta_1=\beta_2=\ \dots\ =\beta_k=0$

H_1： β_j不全为0

F检验的思想来自于总离差平方和的分解式：

TSS=ESS+RSS

由于回归平方和 $ESS=\sum \hat{y}_i^2$ 是解释变量 X 的联合体对被解释变量 Y 的线性作用的结果，考虑比值

$$ESS / RSS=\sum \hat{y}_i^2 \Big/ \sum e_i^2$$

如果这个比值较大，则X的联合体对Y的解释程度高，可认为总体存在线性关系，反之总体上可能不存在线性关系。

因此, 可通过该比值的大小对总体线性关系进行推断。

根据数理统计学中的知识，在原假设H_0成立的条件下，统计量

$$F = \frac{ESS/k}{RSS/(n-k-1)}$$

服从自由度为$(k, n\text{-}k\text{-}1)$的F分布

给定显著性水平α，可得到临界值$F_\alpha(k,n\text{-}k\text{-}1)$，由样本求出统计量**F**的数值，通过

$F > F_\alpha(k,n\text{-}k\text{-}1)$　　或　　$F \le F_\alpha(k,n\text{-}k\text{-}1)$

来拒绝或接受原假设H_0，以判定原方程总体上的线性关系是否显著成立。

对于中国居民人均消费支出的例子：

一元模型：F=285.92

二元模型：F=2057.3

给定显著性水平α =0.05，查分布表，得到临界值：

一元例：$F_\alpha(1,21)=4.32$

二元例：$F_\alpha(2,19)=3.52$

显然有 $F> F_\alpha(k,n-k-1)$

即二个模型的线性关系在显著性水平 α =0.05下显著成立。

2、关于拟合优度检验与方程显著性检验关系的讨论

由 $\bar{R}^2 = 1 - \dfrac{RSS/(n-k-1)}{TSS/(n-1)}$ 与 $F = \dfrac{ESS/k}{RSS/(n-k-1)}$

可推出： $\bar{R}^2 = 1 - \dfrac{n-1}{n-k-1+kF}$

或 $F = \dfrac{\bar{R}^2/k}{(1-\bar{R}^2)/(n-k-1)}$

F 与 $\bar{R}^2$ 同向变化：当 $\bar{R}^2 = 0$ 时，F = 0；

$\bar{R}^2$ 越大，F 值也越大；

当 $\bar{R}^2 = 1$ 时，F 为无穷大。

因此，F检验是所估计回归的总显著性的一个度量，也是$\bar{R}^2$的一个显著性检验。亦即

检验 $H_0: \beta_1 = 0, \beta_2 = 0, \cdots, \beta_k = 0$ 等价于检验 $\bar{R}^2 = 0$

回答前面的问题：$\bar{R}^2$多大才算通过拟合优度检验

- 在中国居民人均收入-消费一元模型中，

F>4.32 → $\bar{R}^2$>0.131 → 模型在95%的水平下显著成立

- 在中国居民人均收入-消费二元模型中，

F>3.52 → $\bar{R}^2$>0.194 → 模型在95%的水平下显著成立

三、变量的显著性检验（t检验）

方程的**总体线性**关系显著≠**每个解释变量**对被解释变量的影响都是显著的

因此，必须对每个解释变量进行显著性检验，以决定是否作为解释变量被保留在模型中。

这一检验是由对变量的 t 检验完成的。

1、t统计量

由于 $$Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

以c_{ii}表示矩阵$(X'X)^{-1}$主对角线上的第i个元素，于是参数估计量的方差为：

$$Var(\hat{\beta}_i) = \sigma^2 c_{ii}$$

其中σ^2为随机误差项的方差，在实际计算时，用它的估计量代替：

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k-1} = \frac{\mathbf{e}'\mathbf{e}}{n-k-1}$$

易知 $\hat{\beta}$ 服从如下正态分布

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 c_{ii})$$

因此，可构造如下t统计量

$$t = \frac{\hat{\beta}_i - \beta_i}{S_{\hat{\beta}_i}} \frac{\hat{\beta}_i - \beta_i}{\sqrt{c_{ii} \dfrac{\mathbf{e}'\mathbf{e}}{n-k-1}}} \sim t(n-k-1)$$

2、t检验

设计原假设与备择假设：

H_0：$\beta_i=0$　　（i=1,2…k）

H_1：$\beta_i\neq0$

给定显著性水平α，可得到临界值$t_{\alpha/2}(n-k-1)$，由样本求出统计量**t**的数值，通过

$|t|>t_{\alpha/2}(n-k-1)$　　或　　$|t|\leq t_{\alpha/2}(n-k-1)$

来拒绝或接受原假设H_0，从而**判定对应的解释变量是否应包括在模型中。**

注意：一元线性回归中，t检验与F检验一致

一方面，t检验与F检验都是对相同的原假设 H_0：$\beta_1=0$ 进行检验；

另一方面，两个统计量之间有如下关系：

$$F=\frac{\sum \hat{y}_i^2}{\sum e_i^2/(n-2)}=\frac{\hat{\beta}_1^2\sum x_i^2}{\sum e_i^2/(n-2)}=\frac{\hat{\beta}_1^2}{\sum e_i^2/(n-2)\sum x_i^2}$$

$$=\left(\frac{\hat{\beta}_1}{\sqrt{\sum e_i^2/(n-2)\sum x_i^2}}\right)^2=\left(\hat{\beta}_1\Big/\sqrt{\frac{\sum e_i^2}{n-2}\cdot\frac{1}{\sum x_i^2}}\right)^2=t^2$$

在中国居民人均收入-消费支出二元模型例中，由应用软件计算出参数的t值：

$$|t_0| = 3.306 \qquad |t_1| = 3.630 \qquad |t_2| = 2.651$$

给定显著性水平α=0.05，查得相应临界值：$t_{0.025}(19)$ =2.093。

可见，计算的所有t值都大于该临界值，所以拒绝原假设。即:

包括常数项在内的3个解释变量都在95%的水平下显著，都通过了变量显著性检验。

四、参数的置信区间

参数的置信区间用来考察：在一次抽样中所估计的参数值离参数的真实值有多“近”。

在变量的显著性检验中已经知道：

$$t=\frac{\hat{\beta}_i-\beta_i}{S_{\hat{\beta}_i}}\frac{\hat{\beta}_i-\beta_i}{\sqrt{c_{ii}\dfrac{\mathbf{e}'\mathbf{e}}{n-k-1}}}\sim t(n-k-1)$$

容易推出：在$(1-\alpha)$的置信水平下β_i的置信区间是

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

其中，$t_{\alpha/2}$为显著性水平为α 、自由度为n-k-1的临界值。

在中国居民人均收入-消费支出二元模型例中，

给定α=0.05，查表得临界值：$t_{0.025}(19)$=2.093

从回归计算中已得到：

$$\hat{\beta}_0 = 120.70 \qquad s_{\hat{\beta}_0} = 36.51$$

$$\hat{\beta}_1 = 0.2213 \qquad s_{\hat{\beta}_1} = 0.061$$

$$\hat{\beta}_2 = 0.4515 \qquad s_{\hat{\beta}_2} = 0.170$$

计算得参数的置信区间：

β_0： (44.284, 197.116)

β_1： (0.0937, 0.3489)

β_2： (0.0951, 0.8080)

如何才能缩小置信区间？

- 增大样本容量n，因为在同样的样本容量下，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；

- 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，模型优度越高，残差平方和应越小。

- 提高样本观测值的分散度，一般情况下，样本观测值越分散，$(X'X)^{-1}$的分母的$|X'X|$的值越大，致使区间缩小。

（四）多元线性回归模型的预测

一、$E(Y_0)$的置信区间

二、Y_0的置信区间

对于模型

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

给定样本以外的解释变量的观测值$\mathbf{X}_0=(1,X_{10},X_{20},\ldots,X_{k0})$，可以得到被解释变量的预测值：

$$\hat{Y}_0 = \mathbf{X}_0\hat{\boldsymbol{\beta}}$$

它可以是总体均值$E(Y_0)$或个值Y_0的预测。

但严格地说，这只是被解释变量的预测值的估计值，而不是预测值。

为了进行科学预测，还需求出预测值的置信区间，包括$E(Y_0)$和Y_0的置信区间。

一、$E(Y_0)$的置信区间

易知

$$E(\hat{Y}_0) = E(\mathbf{X}_0\hat{\boldsymbol{\beta}}) = \mathbf{X}_0 E(\hat{\boldsymbol{\beta}}) = \mathbf{X}_0\boldsymbol{\beta} = E(Y_0)$$

$$Var(\hat{Y}_0) = E(\mathbf{X}_0\hat{\boldsymbol{\beta}} - \mathbf{X}_0\boldsymbol{\beta})^2 = E(\mathbf{X}_0(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\mathbf{X}_0(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))$$

由于$\mathbf{X}_0(\hat{\beta}-\beta)$为标量，因此

$$\begin{aligned} Var(\hat{Y}_0) &= E(\mathbf{X}_0(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}_0') \\ &= \mathbf{X}_0 E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}_0' \\ &= \sigma^2 \mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0' \end{aligned}$$

容易证明

$$\hat{Y}_0 \sim N(\mathbf{X_0}\boldsymbol{\beta}, \sigma^2 \mathbf{X_0}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0')$$

取随机扰动项的样本估计量$\hat{\sigma}^2$，构造如下t统计量

$$\frac{\hat{Y}_0 - E(Y_0)}{\hat{\sigma}\sqrt{\mathbf{X_0}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'}} \sim t(n-k-1)$$

于是，得到$(1-\alpha)$的置信水平下$E(Y_0)$的**置信区间**：

$$\hat{Y}_0 - t_{\frac{\alpha}{2}} \times \hat{\sigma}\sqrt{\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'} < E(Y_0) < \hat{Y}_0 + t_{\frac{\alpha}{2}} \times \hat{\sigma}\sqrt{\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'}$$

其中，$t_{\alpha/2}$为$(1-\alpha)$的置信水平下的**临界值**。

二、Y_0的置信区间

如果已经知道实际的预测值Y_0，那么预测误差为：

$$e_0 = Y_0 - \hat{Y}_0$$

容易证明

$$\begin{aligned} E(e_0) &= E(\mathbf{X}_0\boldsymbol{\beta} + \mu_0 - \mathbf{X}_0\hat{\boldsymbol{\beta}}) \\ &= E(\mu_0 - \mathbf{X}_0(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})) \\ &= E(\mu_0 - \mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\mu}) \\ &= 0 \end{aligned}$$

$$\begin{aligned} Var(e_0) &= E(e_0^2) \\ &= E(\mu_0 - \mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\mu})^2 \\ &= \sigma^2(1 + \mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0') \end{aligned}$$

e_0服从正态分布，即

$$e_0 \sim N(0, \sigma^2(1+\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'))$$

取随机扰动项的样本估计量$\hat{\sigma}^2$，可得e_0的方差的估计量

$$\hat{\sigma}_{e_0}^2 = \hat{\sigma}^2(1+\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0')$$

构造t统计量

$$t = \frac{\hat{Y}_0 - Y_0}{\hat{\sigma}_{e_0}} \sim t(n-k-1)$$

可得给定$(1-\alpha)$的置信水平下Y_0的**置信区间**：

$$\hat{Y}_0 - t_{\frac{\alpha}{2}} \times \hat{\sigma}\sqrt{1+\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'} < Y_0 < \hat{Y}_0 + t_{\frac{\alpha}{2}} \times \hat{\sigma}\sqrt{1+\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'}$$

中国居民人均收入-消费支出二元模型例中：

2001年人均GDP：4033.1元，

于是人均居民消费的预测值为

$\hat{Y}_{2001}=120.7+0.2213\times 4033.1+0.4515\times 1690.8=1776.8$（元）

实测值（90年价）=1782.2元，**相对误差**：-0.31%

预测的置信区间：

在95%的置信度下，临界值$t_{\alpha/2}(19)=2.093$，$\hat{\sigma}^2=705.5$，

$$(\mathbf{X}'\mathbf{X})^{-1}=\begin{pmatrix} 1.88952 & 0.00285 & -0.00828 \\ 0.00285 & 0.00001 & -0.00001 \\ -0.00828 & -0.00001 & 0.00004 \end{pmatrix}$$

$$\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0 = 0.3938$$

于是$E(\hat{Y}_{2001})$的95%的置信区间为:

$$1776.8 \pm 2.093 \times \sqrt{705.5} \times \sqrt{0.3938}$$

或　　（1741.8，1811.7）

同样，易得$\hat{Y}_{2001}$的95%的置信区间为

$$1776.8 \pm 2.093 \times \sqrt{705.5} \times \sqrt{1.3938}$$

或　　（1711.1, 1842.4）

第四节 回归模型的其他函数形式

一、模型的类型与变换

二、非线性回归实例

在实际经济活动中，经济变量的关系是复杂的，直接表现为线性关系的情况并不多见。

如著名的**恩格尔曲线**(Engle curves)表现为**幂函数曲线**形式、宏观经济学中的**菲利普斯曲线**（Pillips cuves）表现为**双曲线**形式等。

但是，大部分非线性关系又可以通过一些简单的数学处理，使之化为数学上的线性关系，从而可以运用线性回归的方法进行计量经济学方面的处理。

一、模型的类型与变换

1、倒数模型、多项式模型与变量的直接置换法

例如，描述税收与税率关系的拉弗曲线：抛物线

$$s = a + b\,r + c\,r^2 \qquad c<0$$

s：税收；　r：税率

设$X_1 = r$，$X_2 = r^2$，　则原方程变换为

$$s = a + b\,X_1 + c\,X_2 \qquad c<0$$

倒数模型：

$$\hat{Y} = \frac{1}{\beta_0 + \beta_1 X_1 + \beta_2 X_2}$$

双曲函数模型：

$$\hat{Y} = \frac{X}{\beta_0 + \beta_1 X}$$

2、幂函数模型、指数函数模型与对数变换法

例如，Cobb-Dauglas生产函数：幂函数

$$Q = AK^{\alpha}L^{\beta}$$

Q：产出量，K：投入的资本；L：投入的劳动

方程两边取对数：

$$\ln Q = \ln A + \alpha \ln K + \beta \ln L$$

3、复杂函数模型与级数展开法

例如，常替代弹性CES生产函数

$$Q = A(\delta_1 K^{-\rho} + \delta_2 L^{-\rho})^{-\frac{1}{\rho}} e^{\mu} \qquad (\delta_1+\delta_2=1)$$

Q:产出量，K：资本投入，L：劳动投入

ρ：替代参数，δ_1、δ_2：分配参数

方程两边取对数后，得到：

$$LnQ = LnA - \frac{1}{\rho} Ln(\delta_1 K^{-\rho} + \delta_2 L^{-\rho}) + \mu$$

将式中$\ln(\delta_1 K^{-\rho} + \delta_2 L^{-\rho})$在$\rho=0$处展开台劳级数,取关于ρ的线性项，即得到一个线性近似式。

如取0阶、1阶、2阶项，可得

$$\ln Y = \ln A + \delta_1 m \ln K + \delta_2 m \ln L - \frac{1}{2}\rho\, m\delta_1\delta_2 \left(\ln\left(\frac{K}{L}\right)\right)^2$$

并非所有的函数形式都可以线性化

无法线性化模型的一般形式为:

$$Y = f(X_1, X_2, \cdots, X_k) + \mu$$

其中，$f(x_1,x_2,\ldots,X_k)$为非线性函数。如：

$$Q = AK^{\alpha}L^{\beta} + \mu$$

二、非线性回归实例

例5. 建立中国城镇居民食品消费需求函数模型。

根据需求理论，居民对食品的消费需求函数大致为

$$Q = f(X, P_1, P_0) \qquad (*)$$

Q:居民对食品的需求量，X：消费者的消费支出总额

P_1：食品价格指数，P_0：居民消费价格总指数。

零阶齐次性，当所有商品和消费者货币支出总额按同一比例变动时，需求量保持不变

$$Q = f(X / P_0, P_1 / P_0) \qquad (**)$$

为了进行比较，将同时估计（*）式与（**）式。

首先, 确定具体的函数形式

根据**恩格尔定律**，居民对食品的消费支出与居民的总支出间呈**幂函数**的变化关系：

$$Q = AX^{\beta_1} P_1^{\beta_2} P_0^{\beta_3}$$

对数变换：

$$\ln(Q) = \beta_0 + \beta_1 \ln X + \beta_2 \ln P_1 + \beta_3 \ln P_0 + \mu \qquad (***)$$

考虑到**零阶齐次性**时

$$\ln(Q) = \beta_0 + \beta_1 \ln(X / P_0) + \beta_2 \ln(P_1 / P_0) + \mu \qquad (****)$$

(****)式也可看成是对（***）式施加如下约束而得

$$\beta_1 + \beta_2 + \beta_3 = 0$$

因此，**对（****）式进行回归，就意味着原需求函数满足零阶齐次性条件。**

表 3.5.1　中国城镇居民消费支出（元）及价格指数

	X (当年价)	X1 (当年价)	GP (上年=100)	FP (上年=100)	XC (1990年价)	Q (1990年价)	P0 (1990=100)	P1 (1990=100)
1981	456.8	420.4	102.5	102.7	646.1	318.3	70.7	132.1
1982	471.0	432.1	102.0	102.1	659.1	325.0	71.5	132.9
1983	505.9	464.0	102.0	103.7	672.2	337.0	75.3	137.7
1984	559.4	514.3	102.7	104.0	690.4	350.5	81.0	146.7
1985	673.2	351.4	111.9	116.5	772.6	408.4	87.1	86.1
1986	799.0	418.9	107.0	107.2	826.6	437.8	96.7	95.7
1987	884.4	472.9	108.8	112.0	899.4	490.3	98.3	96.5
1988	1104.0	567.0	120.7	125.2	1085.5	613.8	101.7	92.4
1989	1211.0	660.0	116.3	114.4	1262.5	702.2	95.9	94.0
1990	1278.9	693.8	101.3	98.8	1278.9	693.8	100.0	100.0
1991	1453.8	782.5	105.1	105.4	1344.1	731.3	108.2	107.0
1992	1671.7	884.8	108.6	110.7	1459.7	809.5	114.5	109.3
1993	2110.8	1058.2	116.1	116.5	1694.7	943.1	124.6	112.2
1994	2851.3	1422.5	125.0	134.2	2118.4	1265.6	134.6	112.4
1995	3537.6	1766.0	116.8	123.6	2474.3	1564.3	143.0	112.9
1996	3919.5	1904.7	108.8	107.9	2692.0	1687.9	145.6	112.8
1997	4185.6	1942.6	103.1	100.1	2775.5	1689.6	150.8	115.0
1998	4331.6	1926.9	99.4	96.9	2758.9	1637.2	157.0	117.7
1999	4615.9	1932.1	98.7	95.7	2723.0	1566.8	169.5	123.3
2000	4998.0	1958.3	100.8	97.6	2744.8	1529.2	182.1	128.1
2001	5309.0	2014.0	100.7	100.7	2764.0	1539.9	192.1	130.8

X：人均消费

X1：人均食品消费

GP：居民消费价格指数

FP：居民食品消费价格指数

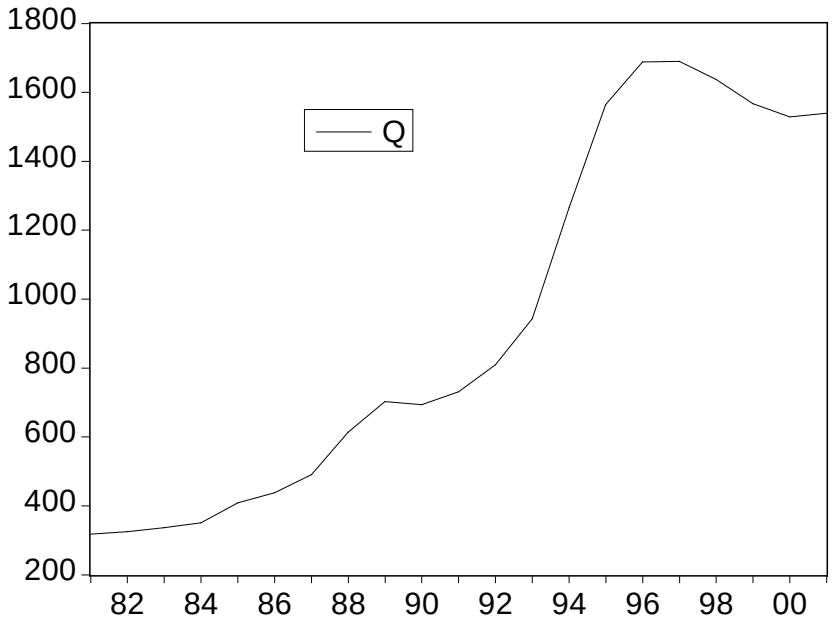
XC：人均消费（90年价）

Q：人均食品消费（90年价）

P0：居民消费价格缩减指数（1990=100）

P：居民食品消费价格缩减指数（1990=100

中国城镇居民人均食品消费



特征：

消费行为在1981~1995年间表现出较强的一致性

1995年之后呈现出另外一种变动特征。

建立1981~1994年中国城镇居民对食品的消费需求模型：

$$\ln(\hat{Q}) = 3.63 + 1.05\ln(X) - 0.08\ln(P_1) - 0.92\ln(P_0)$$

(9.03) (25.35) (-2.28) (-7.34)

R^2=0.9987 $\overline{R}^2$=0.9983 DW=1.50 F=2583.28

各变量的弹性和 $\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$ =0.05，比较接近于零，但不为零。

按零阶齐次性表达式回归：

$$\ln(\hat{Q}) = 3.83 + 1.07\ln(X/P_0) - 0.09\ln(P_1/P_0)$$

（75.86）(52.66)　　(-3.62)

$R^2=0.9986$　　$\bar{R}^2=0.9984$　　DW=1.51　　F=4166.3

为了比较，改写该式为：

$$\ln\hat{Q} = 3.83 + 1.07(\ln X - \ln P_0) - 0.09(\ln P_1 - \ln P_0)$$
$$= 3.83 + 1.07\ln X - 0.09\ln P_1 - 0.98\ln P_0$$

发现与

$$\ln(\hat{Q}) = 3.63 + 1.05\ln(X) - 0.08\ln(P_1) - 0.92\ln(P_0)$$

接近。

意味着：**所建立的食品需求函数满足零阶齐次性特征**

第五节 受约束回归

在建立回归模型时，有时根据经济理论需对模型中变量的参数施加一定的约束条件。

如：　**0阶齐次性** 条件的消费需求函数
　　1阶齐次性 条件的C-D生产函数

模型施加约束条件后进行回归，称为受约束回归（restricted regression）；

不加任何约束的回归称为无约束回归（unrestricted regression）。

受约束回归

一、模型参数的线性约束

二、对回归模型增加或减少解释变量

三、参数的稳定性

*四、非线性约束

一、模型参数的线性约束

对模型

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k+\mu \qquad (*)$$

施加约束

$$\beta_1+\beta_2=1 \qquad \beta_{k-1}=\beta_k$$

得

$$Y=\beta_0+\beta_1X_1+(1-\beta_1)X_2+\cdots+\beta_{k-1}X_{k-1}+\beta_{k-1}X_k+\mu^*$$

或

$$Y^*=\beta_0+\beta_1X_1^*+\beta_3X_3+\cdots+\beta_{k-1}X_{k-1}^*+\mu^* \qquad (**)$$

如果对（**）式回归得出 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_3,\cdots,\hat{\beta}_{k-1}$

则由约束条件可得： $\hat{\beta}_2=1-\hat{\beta}_1$ $\hat{\beta}_k=\hat{\beta}_{k-1}$

然而，对所考查的具体问题**能否施加约束**？需进一步进行相应的检验。**常用的检验有**：

F检验、x^2检验与t检验，

主要介绍F检验

在同一样本下，记**无约束**样本回归模型为

$$\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}$$

受约束样本回归模型为

$$\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}}_* + \mathbf{e}_*$$

于是

$$\mathbf{e}_* = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e} - \mathbf{X}\hat{\boldsymbol{\beta}}_* = \mathbf{e} - \mathbf{X}(\hat{\boldsymbol{\beta}}_* - \hat{\boldsymbol{\beta}})$$

受约束样本回归模型的残差平方和RSS_R

$$e_*'e_* = e'e + (\hat{\beta}_* - \hat{\beta})'X'X(\hat{\beta}_* - \hat{\beta})$$

于是

$$e_*'e_* \geq e'e \qquad (*)$$

e'e为无约束样本回归模型的残差平方和RSS_U

受约束与无约束模型都有相同的TSS

由（*）式 $RSS_R \geq RSS_U$

从而 $ESS_R \leq ESS_U$

这意味着，通常情况下，对模型施加约束条件会降低模型的解释能力。

但是，如果约束条件为真，则受约束回归模型与无约束回归模型具有相同的解释能力，RSS_R 与 RSS_U的差异变小。

可用RSS_R - RSS_U的大小来检验约束的真实性

根据数理统计学的知识：

$$RSS_U / \sigma^2 \sim \chi^2(n - k_U - 1)$$

$$RSS_R / \sigma^2 \sim \chi^2(n - k_R - 1)$$

$$(RSS_R - RSS_U) / \sigma^2 \sim \chi^2(k_U - k_R)$$

于是：

$$F = \frac{(RSS_R - RSS_U)/(k_U - k_R)}{RSS_U/(n - k_U - 1)} \sim F(k_U - k_R, n - k_U - 1)$$

讨论：

如果约束条件无效，RSS_R 与 RSS_U的差异较大，计算的F值也较大。

于是，可用计算的F统计量的值与所给定的显著性水平下的临界值作比较，对约束条件的真实性进行检验。

注意，$k_U - k_R$恰为约束条件的个数。

例6 中国城镇居民对食品的人均消费需求实例中，对零阶齐次性检验：

无约束回归：$RSS_U=0.00324$，　　$k_U=3$

受约束回归：$RSS_R=0.00332$，　　$K_R=2$

样本容量$n=14$，　约束条件个数$k_U-k_R=3-2=1$

$$F=\frac{(0.003315-0.003240)/1}{0.003240/10}=0.231$$

取$\alpha=5\%$，查得临界值$F_{0.05}(1,10)=4.96$

判断：不能拒绝中国城镇居民对食品的人均消费需求函数具有零阶齐次特性这一假设。

这里的F检验适合所有关于参数线性约束的检验

如：多元回归中对方程总体线性性的F检验：

H_0： $\beta_j=0$ $\quad\quad$ j=1,2,…,k

这里：受约束回归模型为

$$Y=\beta_0+\mu_*$$

$$F=\frac{(RSS_R-RSS_U)/(k_U-k_R)}{RSS_U/(n-k_U-1)}=\frac{(TSS-ESS_R-RSS_U)/k}{RSS_U/(n-k-1)}$$

$$=\frac{(TSS-RSS_U)/k}{RSS_U/(n-k-1)}=\frac{ESS_U/k}{RSS_U/(n-k-1)}$$

这里，运用了$ESS_R=0$。

二、对回归模型增加或减少解释变量

考虑如下两个回归模型

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \mu \qquad (*)$$

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \beta_{k+1} X_{k+1} + \cdots \beta_{k+q} X_{k+q} + \mu \qquad (**)$$

(*)式可看成是（**）式的**受约束回归**：

H0：$\beta_{k+1} = \beta_{k+2} = \cdots = \beta_{k+q} = 0$

相应的 F 统计量为：

$$F = \frac{(RSS_R - RSS_U)/q}{RSS_U/(n-(k+q+1))}$$

$$= \frac{(ESS_U - ESS_R)/q}{RSS_U/(n-(k+q+1))} \sim F(q, n-(k+q+1))$$

讨论：

如果约束条件为真，即额外的变量X_{k+1}，…，X_{k+q}对Y没有解释能力，则F统计量较小；

否则，约束条件为假，意味着额外的变量对Y有较强的解释能力，则F统计量较大。

因此，可通过F的**计算值**与**临界值**的比较，来判断额外变量是否应包括在模型中。

F统计量的另一个等价式

$$F = \frac{(R_U^2 - R_R^2)/q}{(1-R_U^2)/(n-(k+q+1))}$$

三、参数的稳定性

建立模型时往往希望模型的参数是稳定的，即所谓的**结构不变**，这将提高模型的预测与分析功能。**如何检验？**

1、邹氏参数稳定性检验

假设**需要建立的模型**为

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \mu$$

在两个连续的时间序列（1,2,…，n_1）与（n_1+1,…，n_1+n_2）中，相应的模型分别为：

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \mu_1$$

$$Y = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_k X_k + \mu_2$$

合并两个时间序列为(1,2,…， n_1 ， n_1+1,…， n_1+n_2)，则可写出如下**无约束回**归模型

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \beta \\ \alpha \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad (*)$$

如果α=β，表示没有发生结构变化，因此可针对如下假设进行检验：

H_0: α=β

(*)式施加上述约束后变换为**受约束**回归模型

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \beta + \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad (**)$$

因此，检验的F统计量为：

$$F=\frac{(RSS_R-RSS_U)/k}{RSS_U/[n_1+n_2-2(k+1)]}\sim F[k,n_1+n_2-2(k+1)]$$

记RSS_1与RSS_2为在两时间段上分别回归后所得的残差平方和，容易验证，

$$RSS_U=RSS_1+RSS_2$$

于是

$$F=\frac{[RSS_R-(RSS_1+RSS_2)]/k}{(RSS_1+RSS_2)/[n_1+n_2-2(k+1)]}\sim F[k,n_1+n_2-2(k+1)]$$

参数稳定性的检验步骤：

（1）分别以两连续时间序列作为两个样本进行回归，得到相应的残差平方：RSS_1与RSS_2

（2）将两序列并为一个大样本后进行回归，得到大样本下的残差平方和RSS_R

（3）计算F统计量的值，与临界值比较：

若F值大于临界值，则拒绝原假设，认为发生了结构变化，参数是非稳定的。

该检验也被称为**邹氏参数稳定性检验**（**Chow test for parameter stability**）。

2、邹氏预测检验

上述参数稳定性检验要求$n_2>k$。

如果出现$n_2<k$，则往往进行如下的**邹氏预测检验**（**Chow test for predictive failure**）。

邹氏预测检验的基本思想:

先用前一时间段n_1个样本估计原模型，再用估计出的参数进行后一时间段n_2个样本的预测。

如果预测误差较大，则说明参数发生了变化，否则说明参数是稳定的。

分别以β、α 表示第一与第二时间段的参数，则

$$\begin{cases} \mathbf{Y}_1 = \mathbf{X}_1\beta + \mu_1 \\ \mathbf{Y}_2 = \mathbf{X}_2\alpha + \mu_2 = \mathbf{X}_2\beta + \mathbf{X}_2(\alpha - \beta) + \mu_2 = \mathbf{X}_2\beta + \gamma + \mu_2 \end{cases} \quad (*)$$

其中， $\gamma = \mathbf{X}_2(\alpha - \beta)$

如果γ =0，则 α = β，表明参数在估计期与预测期相同

(*)的矩阵式：

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{X}_2 & \mathbf{I}_{n_2} \end{pmatrix} \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad (**)$$

可见，用前n_1个样本估计可得前k个参数β的估计，而γ不外是用后n_2个样本测算的预测误差$\mathbf{X}_2(\alpha - \beta)$

如果参数没有发生变化，则$\gamma=0$，矩阵式简化为

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \beta + \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad (***)$$

（***）式与（**）式 $\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{X}_2 & \mathbf{I}_{n_2} \end{pmatrix} \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$

分别可看成受约束与无约束回归模型，于是有如下F检验：

$$F = \frac{(RSS_R - RSS_U)/(k_U - k_R)}{RSS_U/(n - k_U - 1)} = \frac{(RSS_R - RSS_1)/n_2}{RSS_1/(n_1 - k - 1)}$$

这里：$K_U - K_R = n_2$

$RSS_U = RSS_1$

邹氏预测检验步骤：

第一步，在两时间段的合成大样本下做OLS回归，得受约束模型的残差平方和RSS_R ；

第二步，对前一时间段的n_1个子样做OLS回归，得残差平方和RSS_1 ；

第三步，计算检验的F统计量，做出判断：

给定显著性水平α，查F分布表，得临界值$F_\alpha(n_2, n_1-k-1)$

如果 $F>F(n_2, n_1-k-1)$ ，则拒绝原假设，认为预测期发生了结构变化。

例3.6.2 中国城镇居民食品人均消费需求的邹氏检验。

1、参数稳定性检验

1981~1994：

$$\ln(\hat{Q}) = 3.63 + 1.05\ln(X) - 0.08\ln(P_1) - 0.92\ln(P_0)$$ RSS_1=0.003240

1995~2001：

$$\ln Q = 13.78 + 0.55\ln X - 3.06\ln P_1 + 0.71\ln P_0$$

(9.96) (7.14) (-5.13) (1.81)

R^2=0.9946 $\overline{R}^2$=0.9893 DW=2.80 F=185.37 RSS_2 =0.000058

1981~2001:

$$\ln Q = 5.00 + 1.21\ln X - 0.14\ln P_1 - 1.39\ln P_0$$

(14.83) (27.26) (-3.24) (-11.17)

R^2=0.9982 $\overline{R}^2$ =0.9979 DW=0.93 F=3228.0 RSS_R=0.013789

$$F = \frac{[0.013789 - (0.003240 + 0.0000580)]/4}{(0.003240 + 0.000058)/(21-8)} = 10.34$$

给定α=5%，查表得临界值$F_{0.05}(4, 13)=3.18$

判断：F值>临界值，拒绝参数稳定的原假设，表明中国城镇居民食品人均消费需求在1994年前后发生了显著变化。

2、邹氏预测检验

$$F = \frac{(0.013789 - 0.003240)/7}{0.003240/(14-3-1)} = 4.65$$

给定α=5%，查表得临界值$F_{0.05}(7, 10)=3.18$

判断： F值>临界值，拒绝参数稳定的原假设

*四、非线性约束

也可对模型参数施加**非线性约束**, 如对模型

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \mu$$

施加非线性约束$\beta_1\beta_2=1$, 得到**受约束回归模型**:

$$Y = \beta_0 + \beta_1 X_1 + \frac{1}{\beta_1} X_2 + \cdots + \beta_k X_k + \mu^*$$

该模型必需采用**非线性最小二乘法**（**nonlinear least squares**）进行估计。

非线性约束检验是建立在**最大似然原理**基础上的, 有**最大似然比检验**、**沃尔德检验**与**拉格朗日乘数检验.**

1、最大似然比检验 (likelihood ratio test, LR)

估计：无约束回归模型与受约束回归模型，

方法：最大似然法，

检验：两个似然函数的值的差异是否“足够”大。

记$L(\beta, \sigma^2)$为一似然函数：

无约束回归： Max: $L(\hat{\beta}, \hat{\sigma}^2)$

受约束回归： Max: $L(\tilde{\beta}, \tilde{\sigma}^2)$ 约束：$g(\beta)=0$

或求极值： $\Phi = L(\beta, \sigma^2) - \lambda' g(\beta)$

$g(\beta)$:以各约束条件为元素的列向量，

λ'：以相应拉格朗日乘数为元素的行向量

受约束的函数值不会超过无约束的函数值，但如果约束条件为真，则两个函数值就非常“接近”由此，定义似然比（likelihood ratio）：

$$L(\tilde{\beta},\ \tilde{\sigma}^2)/L(\hat{\beta},\ \hat{\sigma}^2)$$

如果比值很小，说明两似然函数值差距较大，则应拒绝约束条件为真的假设；

如果比值接近于 1，说明两似然函数值很接近，应接受约束条件为真的假设。

具体检验时，由于大样本下：

$$LR = -2[\ln L(\tilde{\beta},\tilde{\sigma}^2) - \ln L(\hat{\beta},\hat{\sigma}^2)] \sim \chi^2(h)$$

h是约束条件的个数。因此：

通过LR统计量的χ^2分布特性来进行判断。

在中国城镇居民人均食品消费需求例中，对零阶齐次性的检验：

受约束回归模型：$\ln L(\tilde{\beta}, \tilde{\sigma}^2) = 38.57$

无约束回归模型：$\ln L(\hat{\beta}, \hat{\sigma}^2) = 38.73$

$$LR = -2(38.57-38.73)=0.32$$

给出$\alpha=5\%$、查得临界值$\chi^2_{0.05}(1) = 3.84$，

判断：$LR < \chi^2_{0.05}(1)$，不拒绝原约束的假设，

表明：中国城镇居民对食品的人均消费需求函数满足零阶齐次性条件。

2、沃尔德检验（Wald test, W）

沃尔德检验中，只须估计无约束模型。如对

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \mu$$

要检验约束$\beta_1+\beta_2=1$，只须对该模型进行回归，并判断$\hat{\beta}_1+\hat{\beta}_2$与1的差距是否足够大。

在所有古典假设都成立的条件下，容易证明

$$\hat{\beta}_1 + \hat{\beta}_2 \sim N(\beta_1 + \beta_2, \sigma^2_{\hat{\beta}_1+\hat{\beta}_2})$$

因此，在$\beta_1+\beta_2=1$的约束条件下

$$z = \frac{\hat{\beta}_1 + \hat{\beta}_2 - 1}{\sigma_{\hat{\beta}_1+\hat{\beta}_2}} \sim N(0,1)$$

$\sigma^2_{\hat{\beta}_1+\hat{\beta}_2}$是$\hat{\beta}_1+\hat{\beta}_2$的方差，可记为$\sigma^2_{\hat{\beta}_1+\hat{\beta}_2} = \sigma^2 f(X)$

以σ^2的极大似然估计量 $\tilde{\sigma}^2 = \mathbf{e}'\mathbf{e}/n$ 代入 $\sigma^2_{\hat{\beta}_1+\hat{\beta}_2}$

记 $\tilde{\sigma}^2_{\hat{\beta}_1+\hat{\beta}_2} = \tilde{\sigma}^2 f(\mathbf{X})$　　可建立**沃尔德统计量:**

$$W = \frac{(\hat{\beta}_1 + \hat{\beta}_2 - 1)^2}{\tilde{\sigma}^2_{\hat{\beta}_1+\hat{\beta}_2}} \sim \chi^2(1)$$

如果有h个约束条件，可得到h个统计量$z_1, z_2, \ldots, z_h$

约束条件为真时，可建立**大样本**下的服从自由度为h的渐近χ^2 分布统计量

$$W = \mathbf{Z}'\mathbf{C}^{-1}\mathbf{Z} \sim \chi^2(h)$$

其中，**Z**为以z_i为元素的列向量，**C**是**Z**的方差-协方差矩阵。

因此，**W从总体上测量了无约束回归不满足约束条件的程度。**

对非线性约束，沃尔德统计量W的算法描述要复杂得多。

3、拉格朗日乘数检验

拉格朗日乘数检验则只需估计受约束模型.

受约束回归是求最大似然法的极值问题:

$$\Phi = L(\boldsymbol{\beta}, \sigma^2) - \boldsymbol{\lambda}' g(\boldsymbol{\beta})$$

λ'是拉格朗日乘数行向量，衡量各约束条件对最大似然函数值的影响程度。

如果某一约束为真，则该约束条件对最大似然函数值的影响很小，于是，相应的拉格朗日乘数的值应接近于零。

因此，拉格朗日乘数检验就是检验某些拉格朗日乘数的值是否“足够大”，如果“足够大”，则拒绝约束条件为真的假设。

拉格朗日统计量LM本身是一个关于拉格朗日乘数的复杂的函数，在各约束条件为真的情况下，服从一自由度恰为约束条件个数的渐近χ^2分布。

同样地，如果为线性约束，LM服从一精确的χ^2分布：

$$LM = nR^2 \qquad (*)$$

n为样本容量，R^2为如下被称为**辅助回归**（auxiliary regression）的可决系数：

$$\hat{e}_R = \hat{\delta}_0 + \hat{\delta}_1 X_1 + \hat{\delta}_2 X_2 + \cdots + \hat{\delta}_k X_k$$

这里，$\hat{e}_R$为受约束回归模型的残差序列

如果约束是非线性的，辅助回归方程的估计比较复杂，但仍可按（*）式计算LM统计量的值。

最后，一般地有：LM≤LR≤W