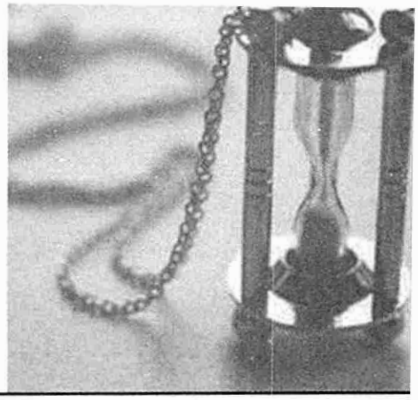

第二章　经典单方程计量经济学模型：一元线性回归模型

单方程计量经济学模型是相对于联立方程计量经济学模型而言的，它以单一经济现象为研究对象，模型中只包括一个方程，是应用最为普遍的计量经济学模型。经典单方程计量经济学模型的理论与方法，不仅是计量经济学内容体系中最重要的组成部分，也是联立方程计量经济学模型理论与方法的基础。本章首先从简单的一元线性回归模型入手，介绍经典单方程计量经济学模型的设定与估计问题，为以后各章的学习打下基础。

§2.1　回归分析概述

一、回归分析基本概念

1. 变量间的相互关系

无论是自然现象之间还是社会经济现象之间，大都存在着不同程度的联系。计量经济学的主要问题之一就是要探寻各种经济变量之间的相互联系程度、联系方式及其运动规律。各种经济变量间的关系可分为两类：一类是确定的函数关系，另一类是不确定的统计相关关系。

确定性现象间的关系常常表现为函数关系。例如，圆面积 S 与圆半径 r 间的关系，只要给定半径值 r，与之对应的圆面积 S 也就随之确定：$S=\pi r^2$。

非确定性现象间的关系常常表现为统计相关关系。例如，农作物产量 Y 与施肥量 X 间的关系，其特点是：农作物产量 Y 随着施肥量 X 的变化呈现某种规律性的变化，在适当的范围内，随着 X 的增加，Y 也增加。但与前述函数关系不同的是，给定施肥量 X，与之对应的农作物产量 Y 并不能确定。主要原因在于，除了施肥量，还有诸如阳光、气温、降雨等其他许多因素都在影响着农作物的产量。这时，我们无法确定农作物产量与施肥量间确定的函数关系，但却能通过统计计量等方法研究它们间的统计相关关系。农作物产量 Y 作为非确定性变量，也被称为随机变量。

当然，变量间的函数关系与相关关系并不是绝对的，在一定条件下两者可相互转化。例如，在对确定性现象的观测中，往往存在测量误差，这时函数关系常会通过相关关系表现出来；反之，如果对非确定性现象的影响因素能够一一辨认出来，并全部纳入到变量间的依存关系式中，则变量间的相关关系就会向函数关系转化。相关分析与回归分析

主要研究非确定性现象间的统计相关关系。

2．相关分析与回归分析

变量间的统计相关关系可以通过相关分析与回归分析来研究。相关分析(correlation analysis)主要研究随机变量间的相关形式及相关程度。

从变量间相关的表现形式看，有线性相关与非线性相关之分，前者往往表现为变量的散点图接近于一条直线。变量间线性相关程度的大小可通过相关系数来测量，两个变量 X 和 Y 的总体相关系数为

$$\rho_{XY}=\frac{\mathrm{Cov}(X,\ Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \tag{2.1.1}$$

其中，Cov(X, Y)是变量 X 和 Y 的协方差，Var(X)和 Var(Y)分别是变量 X 和 Y 的方差。

如果给出 X 与 Y 的一组样本 $(X_i,\ Y_i)$，$i=1,\ 2,\ \cdots,\ n$，则样本相关系数为

$$r_{XY}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\cdot\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{2.1.2}$$

其中，$\bar{X}$ 与 $\bar{Y}$ 分别是变量 X 与 Y 的样本均值。

多个变量间的线性相关程度，可用复相关系数与偏相关系数来度量。

具有相关关系的变量间有时存在着因果关系，这时，我们可以通过回归分析(regression analysis)来研究它们间的具体依存关系。例如，根据经济学理论，消费支出与可支配收入之间不但密切相关，而且有着因果关系，即可支配收入的变化往往是消费支出变化的原因。这时，不仅可以通过相关分析研究两者间的相关程度，而且可以通过回归分析研究两者间的具体依存关系，即考察可支配收入每 1 元的变化所引起的消费支出的平均变化。

回归分析是研究一个变量关于另一个(些)变量的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。前一个变量称为被解释变量(explained variable)或因变量(dependent variable)，后一个变量称为解释变量(explanatory variable)或自变量(independent variable)。

相关分析与回归分析既有联系又有区别。首先，两者都是研究非确定性变量间的统计依赖关系，并能度量线性依赖程度的大小。其次，两者间又有明显的区别。相关分析仅仅是从统计数据上测度变量间的相关程度，而无需考察两者间是否有因果关系，因此，变量的地位在相关分析中是对称的，而且都是随机变量；回归分析则更关注具有统计相关关系的变量间的因果关系分析，变量的地位是不对称的，有解释变量与被解释变量之分，而且解释变量也可以被假设为非随机变量。再次，相关分析只关注变量间的联系程度，不关注具体的依赖关系；而回归分析则更加关注变量间的具体依赖关系，因此可以

进一步通过解释变量的变化来估计或预测被解释变量的变化，达到深入分析变量间依存关系、掌握其运动规律的目的。

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）根据样本观察值对计量经济学模型参数进行估计，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

由于统计相关的随机性，回归分析关心的是根据解释变量的已知值或给定值，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 2.1.1

一个假想的社区是由 99 户家庭组成的总体，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，即根据家庭的每月可支配收入，考察该社区家庭每月消费支出的平均水平。为研究方便，将该 99 户家庭组成的总体按可支配收入水平划分为 10 组，并分别分析每一组的家庭消费支出(表 2.1.1)。

表 2.1.1 某社区家庭每月可支配收入与消费支出统计表 单位：元

每月家庭可支配收入 X	800	1 100	1 400	1 700	2 000	2 300	2 600	2 900	3 200	3 500
每月家庭消费支出 Y	561	638	869	1 023	1 254	1 408	1 650	1 969	2 089	2 299
	594	748	913	1 100	1 309	1 452	1 738	1 991	2 134	2 321
	627	814	924	1 144	1 364	1 551	1 749	2 046	2 178	2 530
	638	847	979	1 155	1 397	1 595	1 804	2 068	2 267	2 629
		935	1 012	1 210	1 408	1 650	1 848	2 101	2 354	2 860
		968	1 045	1 243	1 474	1 672	1 881	2 189	2 486	2 871
			1 078	1 254	1 496	1 683	1 925	2 233	2 552	
			1 122	1 298	1 496	1 716	1 969	2 244	2 585	
			1 155	1 331	1 562	1 749	2 013	2 299	2 640	
			1 188	1 364	1 573	1 771	2 035	2 310		
			1 210	1 408	1 606	1 804	2 101			
				1 430	1 650	1 870	2 112			
				1 485	1 716	1 947	2 200			
						2 002				
共计	2 420	4 950	11 495	16 445	19 305	23 870	25 025	21 450	21 285	15 510

由于不确定因素的影响，对同一可支配收入水平 X，不同家庭的消费支出不完全相

同，但由于调查的完备性，给定可支配收入水平 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，如 $P(Y=561|X=800)=1/4$。因此，给定收入 X 的值，可得消费支出 Y 的**条件均值**(conditional mean)或**条件期望**(conditional expectation)，如 $E(Y|X=800)=605$。表 2.1.2 给出了 10 组可支配收入水平下相应家庭消费支出的条件概率，以及各可支配收入水平组家庭消费支出的条件均值。

表 2.1.2 各可支配收入水平组相应家庭消费支出的条件概率与条件均值 单位：元

收入水平	800	1 100	1 400	1 700	2 000	2 300	2 600	2 900	3 200	3 500
条件概率	1/4	1/6	1/11	1/13	1/13	1/14	1/13	1/10	1/9	1/6
条件均值	605	825	1 045	1 265	1 485	1 705	1 925	2 145	2 365	2 585

以表 2.1.1 中的数据绘出可支配收入 X 与家庭消费支出 Y 的散点图(图 2.1.1)。从该散点图可以看出，虽然不同的家庭消费支出存在差异，但平均来说，随着可支配收入的增加，家庭消费支出也在增加。进一步，这个例子中 Y 的条件均值恰好落在一根正斜率的直线上，这条直线称为**总体回归线**。

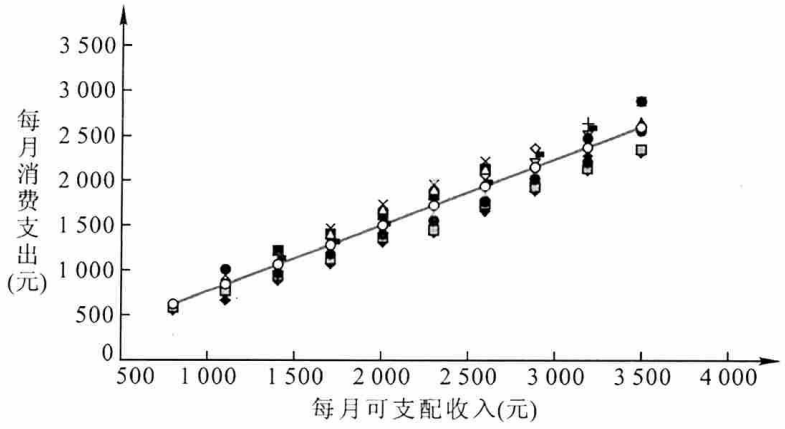

图 2.1.1 不同可支配收入水平组家庭消费支出的条件分布图

在给定解释变量 X 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

$$E(Y|X)=f(X) \tag{2.1.3}$$

称为(双变量)**总体回归函数**(Population Regression Function, PRF)。

总体回归函数表明被解释变量 Y 的平均状态(总体条件期望)随解释变量 X 变化的规律。至于具体的函数形式，是由所考察总体固有的特征来决定的。由于实践中总体往往无法全部考察到，因此总体回归函数形式的选择就是一个经验方面的问题，这时经济学等相关学科的理论就显得很重要。例如，生产函数常以 Cobb-Douglas 幂函数的形式出现，U 形边际成本函数以二次多项式的形式出现，等等。将居民消费支出看成是其可支配收

入的线性函数时，(2.1.3)式可进一步写成

$$E(Y \mid X) = \beta_0 + \beta_1 X \tag{2.1.4}$$

其中，β_0，β_1 是未知参数，称为**回归系数**(regression coefficients)。(2.1.4)式也称为**线性总体回归函数**。线性函数形式最为简单，其中参数的估计与检验也相对容易，而且多数非线性函数可转换为线性形式，因此，为了研究的方便，计量经济学中总体回归函数常设定成线性形式。

需注意的是，经典计量经济方法中所涉及的线性函数指回归系数是线性的，即回归系数只以它的一次方出现，对解释变量则可以不是线性的。

三、随机干扰项

在上述家庭可支配收入-消费支出的例子中，总体回归函数描述了所考察总体的家庭消费支出平均说来随可支配收入变化的规律，但对某一个个别家庭，其消费支出 Y 不一定恰好就是给定可支配收入水平 X 下的消费的平均值 $E(Y \mid X)$。图 2.1.1 显示，个别家庭消费支出 Y 聚集在给定可支配收入水平 X 下所有家庭平均消费支出 $E(Y \mid X)$ 的周围。

对每个个别家庭，记

$$\mu = Y - E(Y \mid X) \tag{2.1.5}$$

称 μ 为观察值 Y 围绕它的期望值 $E(Y \mid X)$ 的**离差**(deviation)，它是一个不可观测的随机变量，称为**随机误差项**(stochastic error)，通常又不加区别地称为**随机干扰项**（stochastic disturbance）。

由(2.1.5)式，个别家庭的消费支出为

$$Y = E(Y \mid X) + \mu \tag{2.1.6}$$

或者在线性假设下

$$Y = \beta_0 + \beta_1 X + \mu \tag{2.1.7}$$

即给定可支配收入水平 X，个别家庭的消费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 $E(Y \mid X)$，称为**系统性**(systematic)**部分**或**确定性**(deterministic)**部分**；(2)其他**随机部分**或**非系统性**(nonsystematic)**部分** μ。

(2.1.6)式或(2.1.7)式称为**总体回归函数的随机设定形式**，它表明被解释变量 Y 除了受解释变量 X 的系统性影响外，还受其他未包括在模型中的诸多因素的随机性影响，μ 即为这些影响因素的综合代表。由于方程中引入了随机干扰项，成为计量经济学模型，因此也称为**总体回归模型**(population regression model)。

在总体回归函数中引入随机干扰项，主要有以下六方面的原因。

(1) 代表未知的影响因素。由于对所考察总体认识上的非完备性，许多未知的影响因素还无法引入模型，因此，只能用随机干扰项代表这些未知的影响因素。

(2) 代表残缺数据。即使所有的影响变量都能被包括在模型中，也会有某些变量的

数据无法取得。例如，经济理论指出，居民消费支出除受可支配收入的影响外，还受财富拥有量的影响，但后者在实践中往往是无法收集到的。这时，模型中不得不省略这一变量，而将其归入随机干扰项。

(3) 代表众多细小影响因素。有一些影响因素已经被认识，而且其数据也可以收集到，但它们对被解释变量的影响却是细小的。考虑到模型的简洁性，以及取得诸多变量数据可能带来的较大成本，建模时往往省掉这些细小变量，而将它们的影响综合到随机干扰项中。

(4) 代表数据观测误差。由于某些主客观的原因，在取得观测数据时，往往存在测量误差，这些观测误差也被归入随机干扰项。

(5) 代表模型设定误差。由于经济现象的复杂性，模型的真实函数形式往往是未知的，因此，实际设定的模型可能与真实的模型有偏差。随机干扰项包含了这种模型设定误差。

(6) 变量的内在随机性。即使模型没有设定误差，也不存在数据观测误差，由于某些变量所固有的内在随机性，也会对被解释变量产生随机性影响。这种影响只能被归入随机干扰项中。

总之，随机干扰项具有非常丰富的内容，在计量经济学模型的建立中起着重要的作用。如果进一步分析，可以发现，当随机干扰项仅包含上述(3)和(6)时，称之为“原生”的随机干扰，是模型所固有的；当随机干扰项包含上述(1)、(2)、(4)、(5)时，称之为“衍生”的随机误差，是在模型设定过程中产生的，是可以避免的。尽管本书对此不加区别，但认识这一点是重要的。

四、样本回归函数

尽管总体回归函数揭示了所考察总体被解释变量与解释变量间的平均变化规律，但总体的信息往往无法全部获得，因此，总体回归函数实际上是未知的。现实的情况往往是，通过抽样得到总体的样本，再通过样本的信息来估计总体回归函数。

仍以例 2.1.1 中社区家庭可支配收入与消费支出的关系为例，假设从该总体中按每组可支配收入水平各取一个家庭进行观测，得到表 2.1.3 所示的一个样本。问题归结为：能否从该样本中预测整个总体对应于选定 X 的平均每月消费支出，即能否从该样本估计总体回归函数？

表 2.1.3 家庭消费支出与可支配收入的一个随机样本 单位：元

X	800	1 100	1 400	1 700	2 000	2 300	2 600	2 900	3 200	3 500
Y	638	935	1 155	1 254	1 408	1 650	1 925	2 068	2 267	2 530

该样本的散点图如图 2.1.2 所示，可以看出，该样本散点图近似于一条直线。画一条直线尽可能地拟合该散点图。由于样本取自总体，可用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 line)，其函数形式记为

$$\hat{Y} = f(X) = \hat{\beta}_0 + \hat{\beta}_1 X \tag{2.1.8}$$

称之为**样本回归函数**(Sample Regression Function，SRF)。

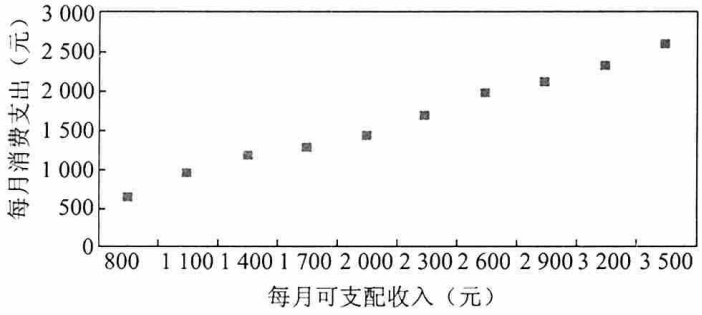

图 2.1.2　家庭可支配收入与消费支出的样本散点图

将(2.1.8)式看成(2.1.7)式的近似替代，则 $\hat{Y}$ 就为 $\mathrm{E}(Y \mid X)$ 的估计量，$\hat{\beta}_0$ 为 β_0 的估计量，$\hat{\beta}_1$ 为 β_1 的估计量。

同样地，样本回归函数也有如下的随机形式：

$$Y = \hat{Y} + \hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 X + e \tag{2.1.9}$$

其中，e 称为(**样本**)**残差**(或**剩余**)**项**(residual)，代表了其他影响 Y 的随机因素的集合，可看成是 μ 的估计量 $\hat{\mu}$。由于方程中引入了随机项，成为计量经济学模型，因此也称之为**样本回归模型**(sample regression model)。

回归分析的主要目的，就是根据样本回归函数，估计总体回归函数。也就是根据

$$Y = \hat{Y} + e = \hat{\beta}_0 + \hat{\beta}_1 X + e$$

估计

$$Y = \mathrm{E}(Y \mid X) + \mu = \beta_0 + \beta_1 X + \mu$$

即设计一种“方法”构造 SRF，以使 SRF 尽可能“接近”PRF，或者说使 $\hat{\beta}_j (j = 0,1)$ 尽可能接近 $\beta_j (j = 0,1)$。图 2.1.3 绘出了总体回归线与样本回归线的基本关系。

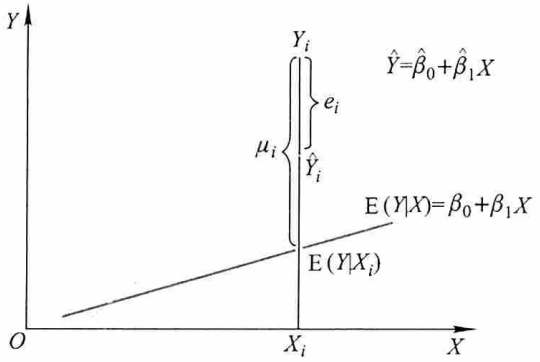

图 2.1.3　总体回归线与样本回归线的基本关系

§2.2 一元线性回归模型的基本假设

单方程计量经济学模型分为线性模型和非线性模型两大类。在线性模型中，变量之间的关系呈线性关系；在非线性模型中，变量之间的关系呈非线性关系。线性回归模型是线性模型的一种，它的数学基础是回归分析，即用回归分析方法建立的线性模型，用以揭示经济现象中的因果关系。

一元线性回归模型是最简单的计量经济学模型，在模型中只有一个解释变量，其一般形式是

$$Y = \beta_0 + \beta_1 X + \mu \tag{2.2.1}$$

其中，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**。在有 n 个样本观测点 $\{(X_i, Y_i): i = 1,2,\cdots,n\}$ 的情况下，(2.2.1)式也可写成如下形式：

$$Y_i = \beta_0 + \beta_1 X_i + \mu_i, \quad i = 1,2,\cdots,n \tag{2.2.2}$$

回归分析的主要目的是要通过样本回归函数(模型)尽可能准确地估计总体回归函数(模型)。**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。**

对模型(2.2.1)或(2.2.2)，基本假设包括对模型设定的假设、对解释变量 X 的假设以及对随机干扰项 μ 的假设。

一、对模型设定的假设

假设 1：回归模型是正确设定的。

计量经济模型是对所关注经济现象或经济理论进行经验研究的基本工具，因此刻画经济现象或描述经济理论的计量模型的正确设定最为重要。模型的正确设定主要包括两方面的内容：（1）模型选择了正确的变量；（2）模型选择了正确的函数形式。

模型选择了正确的变量指在设定总体回归函数时，既没有遗漏重要的相关变量，也没有多选无关的变量。模型选择了正确的函数形式是指当被解释量与解释变量间呈现某种函数形式时，我们所设定的总体回归方程恰为该函数形式。例如在生产函数的设定中，如果产出量与资本投入及劳动投入的关系呈现幂函数的形式，我们在总体回归模型的设定中就设定了该幂函数的形式。

当假设 1 满足时，称为模型没有**设定偏误**(specification error)，否则就会出现模型的设定偏误。第四章将详细讨论模型的设定偏误问题。

二、对解释变量的假设

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

$$\sum_{i=1}^{n}(X_i-\bar{X})^2/n\to Q,\ n\to\infty \tag{2.2.3}$$

在以因果关系为基础的回归分析中，往往就是通过解释变量 X 的变化来解释被解释变量 Y 的变化的，因此，解释变量 X 要有足够的变异性。而对其样本方差的极限为非零的有限常数的假设，则旨在排除数据取值出现无界的变量作为解释变量，因为这类数据将使大样本统计推断变得无效。

需要说明的是，大多数初、中级教材还假设了 X 是固定的非随机变量，在实验或可控条件下，X 的非随机性往往能得到满足，但对社会调查数据则基本不具有这种特点，尤其通过随机抽样调查获得的数据，被解释变量与解释变量更具有随机特征。因此，本书认为解释变量 X 是随机变量，不再假设它是固定的非随机变量。

三、对随机干扰项的假设

假设 3：给定解释变量 X 的任何值，随机干扰项 μ_i 的均值为零，即

$$\mathrm{E}(\mu_i\mid X)=0 \tag{2.2.4}$$

随机干扰项 μ 的条件零均值假设意味着 μ 的期望不依赖于 X 的任何观测点取值的变化而变化，且总为常数零。该假设表明 μ 与 X 不存在任何形式的相关性，因此该假设成立时也往往称 X 为**外生解释变量**(exogenous explanatory variable)，或称 X 是**严格外生**的(strictly exogenous)，否则称 X 为**内生解释变量**(endogenous explanatory variable)。该假设最为重要，只有该假设成立时，总体回归函数的随机形式(2.1.7)式才能等价于非随机形式(2.1.4)式。

需要注意的是，当随机干扰项 μ 的条件零均值假设成立时，根据期望迭代法则（law of iterated expectations）一定有如下非条件零均值性质：

$$\mathrm{E}(\mu_i)=\mathrm{E}(\mathrm{E}(\mu_i\mid X))=\mathrm{E}(0)=0 \tag{2.2.5}$$

同时，当随机干扰项 μ 的条件零均值假设成立时，一定可得到随机干扰项与解释变量之间的不相关性，即

$$\mathrm{Cov}(X,\mu_i)=\mathrm{E}(X\mu_i)-\mathrm{E}(X)\mathrm{E}(\mu_i)=\mathrm{E}(X\mu_i)=0$$

其中最后一个等式仍可通过期望迭代法则推出。这一性质意味着任何观测点处的 X 都与 μ_i 不相关，当然也包括第 i 个观测点处的 X_i 与 μ_i 的不相关性，即有

$$\mathrm{Cov}(X_i,\mu_i)=\mathrm{E}(X_i\mu_i)=0 \tag{2.2.6}$$

这时，也称 X 是**同期外生**的（contemporaneously exogenous）或称 X 与 μ **同期不相关**（contemporaneously uncorrelated）。这一特征在回归分析中十分重要，尤其是在模型参数的估计中扮演着重要的角色。

假设 4：随机干扰项 μ 具有给定 X 任何值条件下的同方差性及不序列相关性，即

$$\mathrm{Var}(\mu_i\mid X)=\sigma^2 \qquad i=1,2,\cdots,n \tag{2.2.7}$$

$$\mathrm{Cov}(\mu_i, \mu_j \mid X) = 0 \qquad i \neq j \tag{2.2.8}$$

随机干扰项 μ 的条件同方差假设意味着 μ 的方差不依赖于 X 的变化而变化，且总为常数 σ^2。在 μ 的条件零均值与条件同方差假设下，总体回归函数可显示为图 2.2.1。

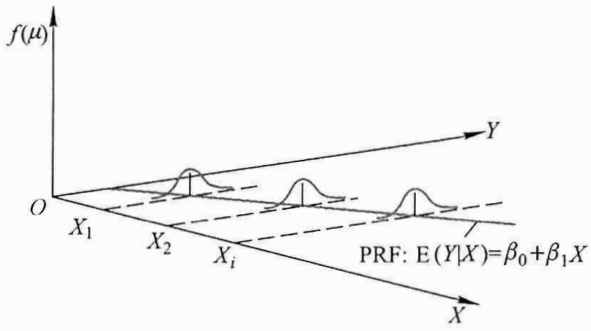

图 2.2.1 μ 的条件零均值与条件同方差假设下的总体回归函数

同样地，随机干扰项 μ 的条件同方差假设成立时，根据期望迭代法则一定有如下非条件同方差性质：

$$\mathrm{Var}(\mu_i) = \sigma^2 \tag{2.2.9}$$

另外，在随机干扰项零均值的假设下，同方差还可写成如下的表达式：

$$\begin{aligned} \mathrm{Var}(\mu_i \mid X) &= \mathrm{E}(\mu_i^2 \mid X) - [\mathrm{E}(\mu_i \mid X)]^2 \\ &= \mathrm{E}(\mu_i^2 \mid X) = \sigma^2 \end{aligned} \tag{2.2.10}$$

或

$$\mathrm{Var}(\mu_i) = \mathrm{E}(\mu_i^2) - [\mathrm{E}(\mu_i)]^2 = \mathrm{E}(\mu_i^2) = \sigma^2 \tag{2.2.11}$$

随机干扰项 μ 的条件不序列相关性表明在给定解释变量的任何值时，任意两个不同观测点的随机干扰项不相关。同样地，(2.2.8)式可等价地表示为

$$\mathrm{Cov}(\mu_i, \mu_j \mid X) = \mathrm{E}[(\mu_i \mid X)(\mu_j \mid X)] = 0 \tag{2.2.12}$$

假设 5：随机干扰项服从零均值、同方差的正态分布，即

$$\mu_i \mid X \sim N(0, \sigma^2) \tag{2.2.13}$$

假设 5 是为通过样本回归函数推断总体回归函数的需要而提出的，尤其是在小样本下，该假设显得十分重要。在大样本的情况下，正态性假设可以放松，因为根据中心极限定理，当样本容量趋于无穷大时，在大多数情况下，随机干扰项的分布会越来越接近正态分布。

以上假设也称为线性回归模型的**经典假设**(classical assumption)，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)。而前 4 个假设也专门称为**高斯-马尔可夫假设**(Gauss-Markov assumption)，这些假设能够保证下节介绍的估计方法具有良好的效果。

最后需要指出，在上述经典假设下，线性回归模型(2.2.1)中被解释变量 Y 具有如下条件分布特征：

$$Y \mid X \sim N(\beta_0 + \beta_1 X, \sigma^2) \tag{2.2.14}$$

图 2.2.2 描绘出了总体回归线与 Y 的条件分布状况。

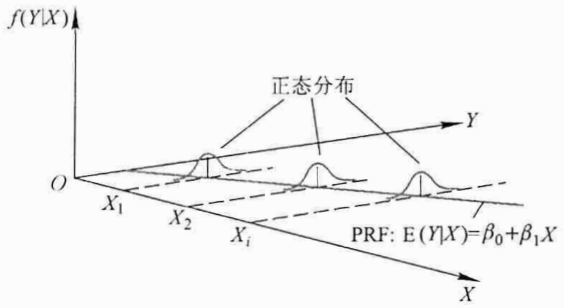

图 2.2.2　总体回归函数与 Y 的条件分布图

在实际建立模型的过程中，除了随机干扰项的正态性假设外，对模型是否满足其他假设都要进行检验。这就是“建立计量经济学模型步骤”中“计量经济学检验”的任务。对于随机干扰项的正态性假设，根据中心极限定理，如果仅包括源生性的随机干扰，当样本容量趋于无穷大时，都是满足的。如果包括衍生的随机误差，即使样本容量趋于无穷大，正态性假设也经常是不满足的。但是在初、中级教材中，一般将它忽略。

§2.3　一元线性回归模型的参数估计

一元线性回归模型的参数估计，是在一组样本观测值 $\{(X_i, Y_i): i = 1, 2, \cdots, n\}$ 下，通过一定的参数估计方法，估计出样本回归线。常见的估计方法有三种：普通最小二乘法(OLS)、最大似然法(ML)与矩估计法(MM)。

一、参数估计的普通最小二乘法(OLS)

已知一组样本观测值 $\{(X_i, Y_i): i = 1, 2, \cdots, n\}$，普通最小二乘法(Ordinary Least Squares, OLS)要求样本回归函数尽可能好地拟合这组值，即样本回归线上的点 $\hat{Y}_i$ 与真实观测点 Y_i 的“总体误差”尽可能地小。普通最小二乘法给出的判断标准是：被解释变量的估计值与实际观测值之差的平方和

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2 \tag{2.3.1}$$

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

为什么用平方和？因为样本回归线上的点 $\hat{Y}_i$ 与真实观测点 Y_i 之差可正可负，简单求和可能将很大的误差抵消掉，只有平方和才能反映二者在总体上的接近程度，这就是**最**

小二乘原理。

根据微积分学的运算，当Q对$\hat{\beta}_0$，$\hat{\beta}_1$的一阶偏导数为0时，Q达到最小，即

$$\begin{cases}\dfrac{\partial Q}{\partial \hat{\beta}_0}=0\\[2ex]\dfrac{\partial Q}{\partial \hat{\beta}_1}=0\end{cases}$$

可推得用于估计$\hat{\beta}_0,\hat{\beta}_1$的下列方程组：

$$\begin{cases}\sum(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)=0\\\sum(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)X_i=0\end{cases}\tag{2.3.2}$$

或

$$\begin{cases}\sum Y_i=n\hat{\beta}_0+\hat{\beta}_1\sum X_i\\\sum Y_iX_i=\hat{\beta}_0\sum X_i+\hat{\beta}_1\sum X_i^2\end{cases}\tag{2.3.3}$$

解得

$$\begin{cases}\hat{\beta}_0=\dfrac{\sum X_i^2\sum Y_i-\sum X_i\sum Y_iX_i}{n\sum X_i^2-\left(\sum X_i\right)^2}\\[2ex]\hat{\beta}_1=\dfrac{n\sum Y_iX_i-\sum Y_i\sum X_i}{n\sum X_i^2-\left(\sum X_i\right)^2}\end{cases}\tag{2.3.4}$$

方程组(2.3.2)或(2.3.3)称为**正规方程组**(normal equations)。记

$$\begin{aligned}\sum x_i^2&=\sum(X_i-\bar{X})^2\\&=\sum X_i^2-\frac{1}{n}\left(\sum X_i\right)^2\\\sum x_iy_i&=\sum(X_i-\bar{X})(Y_i-\bar{Y})\\&=\sum X_iY_i-\frac{1}{n}\sum X_i\sum Y_i\end{aligned}$$

方程组(2.3.4)的参数估计量可以写成

$$\begin{cases}\hat{\beta}_1=\dfrac{\sum x_iy_i}{\sum x_i^2}\\[2ex]\hat{\beta}_0=\bar{Y}-\hat{\beta}_1\bar{X}\end{cases}\tag{2.3.5}$$

称为普通最小二乘法估计量的**离差形式**(deviation form)。在本书中，往往以小写字母表示对均值的离差。由于$\hat{\beta}_0$，$\hat{\beta}_1$的估计结果是从最小二乘原理得到的，故称为**普通最小二乘估计量**(ordinary least squares estimator)。

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

$$\begin{aligned}\hat{y}_i&=(\hat{\beta}_0+\hat{\beta}_1X_i)-(\hat{\beta}_0+\hat{\beta}_1\bar{X}+\bar{e})\\&=\hat{\beta}_1(X_i-\bar{X})-\frac{1}{n}\sum e_i\end{aligned}$$

可得

$$\hat{y}_i = \hat{\beta}_1 x_i \tag{2.3.6}$$

其中，用到了正规方程组的第一个方程

$$\sum e_i = \sum [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)] = 0$$

(2.3.6)式也称为样本回归函数的**离差形式**。

最后，需要交代一个重要的概念，即“估计量”(estimator)和“估计值”(estimate)的区别。由(2.3.4)式或(2.3.5)式给出的参数估计结果是由一个具体样本资料计算出来的，它是一个“估计值”，或者“点估计”，是参数估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的一个具体数值；但从另一个角度，仅仅把(2.3.4)式或(2.3.5)式看成 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的一个表达式，那么，它就成为 Y_i 的函数，而 Y_i 是随机变量，所以 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 也是随机变量，因而从这个角度考虑，就称之为“估计量”。在本章后续内容中，有时把 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 作为随机变量，有时又把 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 作为确定的数值，道理就在于此。

二、参数估计的最大似然法(ML)

最大似然法(Maximum Likelihood, ML)，也称**最大或然法**，是不同于普通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他估计方法的基础。虽然其应用没有普通最小二乘法普遍，但在计量经济学理论上占据很重要的地位，因为最大似然原理比最小二乘原理更本质地揭示了通过样本估计总体参数的内在机理。计量经济学理论的发展，更多的是以最大似然原理为基础的，对于一些特殊的计量经济学模型，只有最大似然方法才是很成功的估计方法。

对于普通最小二乘法，当从模型总体随机抽取容量为 n 的样本观测值后，最合理的参数估计量应该使得模型能最好地拟合样本数据；而对于最大似然法，当从模型总体随机抽取容量为 n 的样本观测值后，最合理的参数估计量应该使得从模型中抽取该样本观测值的概率最大。显然，这是从不同原理出发的两种参数估计方法。

从总体中经过 n 次随机抽取得到样本容量为 n 的样本观测值，在任一次随机抽取中，样本观测值都以一定的概率出现。如果已经知道总体的参数，当然由变量的频率函数可以计算其概率。如果只知道总体服从某种分布，但不知道其分布参数，通过随机样本可以求出总体的参数估计量。以正态分布的总体为例，每个总体都有自己的分布参数的期望和方差，如果已经得到容量为 n 的样本观测值，在这些可供选择的总体中，哪个总体最可能产生已经得到的样本观测值呢？显然，要对每个可能的正态总体估计取得容量为 n 的样本观测值的联合概率，然后选择其参数能使观测值的联合概率为最大的那个总体。将样本观测值联合概率函数称为变量的**似然函数**。在已经取得样本观测值的情况下，使似然函数取极大值的总体分布参数所代表的总体具有最大的概率取得这些样本观测值，

该总体参数即是所要求的参数。通过似然函数最大化以求得总体参数估计量的方法称为最大似然法。

在满足基本假设条件下，对一元线性回归模型

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

随机抽取容量为 n 的样本观测值 $\{(X_i, Y_i): i = 1, 2, \cdots, n\}$，由于 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_i 的概率函数为

$$P(Y_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2}, \qquad i = 1, 2, \cdots, n$$

因为 Y_i 是相互独立的，所以 Y 的所有样本观测值的联合概率，也即**似然函数**为

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2) &= P(Y_1, Y_2, \cdots, Y_n) \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} e^{-\frac{1}{2\sigma^2}\sum(Y_i - \beta_0 - \beta_1 X_i)^2} \end{aligned} \tag{2.3.7}$$

将该似然函数最大化，即可求得模型参数的**最大似然估计量**。

由于似然函数的最大化与似然函数对数的最大化是等价的，所以取对数似然函数如下：

$$L^* = \ln L = -n\ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum(Y_i - \beta_0 - \beta_1 X_i)^2 \tag{2.3.8}$$

对 L^* 求最大值，等价于对 $\sum(Y_i - \beta_0 - \beta_1 X_i)^2$ 求最小值。设 $\hat{\beta}_0, \hat{\beta}_1$ 满足该最值条件，即

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

解得模型的参数估计量为

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

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似然估计量与普通最小二乘估计量是相同的。**

三、参数估计的矩估计法(MM)

普通最小二乘法是通过得到一个关于参数估计值的**正规方程组**并对它进行求解而完成的。正规方程组(2.3.2)或(2.3.3)可以通过矩估计（Method of Moment, MM）的思想来导

出。矩估计的基本原理是用相应的样本矩来估计总体矩。

在本章 § 2.2 对一元回归模型的假设中，通过随机干扰项的条件零均值假设可得到它的非条件零均值性以及它与解释变量的同期不相关性，意味着存在如下两个总体矩条件：

$$\mathrm{E}(\mu_i)=0$$

$$\mathrm{Cov}(X_i,\mu_i)=\mathrm{E}(X_i\mu_i)=0$$

于是，相应的样本矩条件可写成

$$\frac{1}{n}\sum\left(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i\right)=0 \tag{2.3.9}$$

$$\frac{1}{n}\sum\left(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i\right)X_i=0 \tag{2.3.10}$$

以上述方程组成的方程组，各自去掉 $\frac{1}{n}$ 后不改变该方程组的解，而去掉 $\frac{1}{n}$ 后该方程组恰为普通最小二乘法中的正规方程组(2.3.2)式。因此，解与普通最小二乘法以及最大似然法的结果相同。这种估计样本回归函数的方法称为**矩估计法**。

例 2.3.1

在上述家庭可支配收入-消费支出例子中，对于所抽出的一组样本数，参数估计的计算可通过表 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_iy_i	x_i^2	y_i^2	X_i^2	Y_i^2
1	800	638	–1 350	–945	1 275 750	1 822 500	893 025	640 000	407 044
2	1 100	935	–1 050	–648	680 400	1 102 500	419 904	1 210 000	874 225
3	1 400	1 155	–750	–428	321 000	562 500	183 184	1 960 000	1 334 025
4	1 700	1 254	–450	–329	148 050	202 500	108 241	2 890 000	1 572 516
5	2 000	1 408	–150	–175	26 250	22 500	30 625	4 000 000	1 982 464
6	2 300	1 650	150	67	10 050	22 500	4 489	5 290 000	2 722 500
7	2 600	1 925	450	342	153 900	202 500	116 964	6 760 000	3 705 625
8	2 900	2 068	750	485	363 750	562 500	235 225	8 410 000	4 276 624
9	3 200	2 267	1 050	684	718 200	1 102 500	467 856	10 240 000	5 139 289
10	3 500	2 530	1 350	947	1 278 450	1 822 500	896 809	12 250 000	6 400 900
求和	21 500	15 830			4 975 800	7 425 000	3 356 322	53 650 000	28 415 212
平均	2 150	1 583							

由(2.3.5)式计算得

$$\hat{\beta}_1=\frac{\sum x_iy_i}{\sum x_i^2}=\frac{4\ 975\ 800}{7\ 425\ 000}=0.670\,1$$

$$\hat{\beta}_0=\overline{Y}-\hat{\beta}_1\overline{X}=1\ 583-0.670\,1\times 2\ 150=142.28$$

因此，由该样本估计的回归方程为

$$\hat{Y}_i=142.28+0.670\,1X_i$$

四、最小二乘估计量的统计性质

当估计出模型参数后，需考虑参数估计值的精度，即是否能代表总体参数的真值。一般地，由于抽样波动的存在，以及所选估计方法的不同，都会使估计的参数与总体参数的真值有差距，因此考察参数估计量的统计性质就成了衡量该估计量“好坏”的主要准则。

一个用于考察总体的估计量，可从如下六个方面考察其优劣性：（1）线性性，即它是否是另一个随机变量的线性函数；（2）无偏性，即它的均值或期望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；（4）渐近无偏性，即样本容量趋于无穷大时，它的均值序列是否趋于总体真值；（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，它在所有的一致估计量中是否具有最小的渐近方差。

这里，前三个准则也称作估计量的**有限样本性质或小样本性质**(small-sample properties)，因为一旦某估计量具有该类性质，它是不以样本的大小而改变的。拥有这类性质的估计量称为**最佳线性无偏估计量**(Best Linear Unbiased Estimator，BLUE)。当然，在有限样本情形下，有时很难找到最佳线性无偏估计量，这时就需要考察样本容量无限增大时估计量的渐近性质。后三个准则称为估计量的**无限样本性质或大样本渐近性质**(large-sample asymptotic properties)。如果有限样本情况下不能满足估计的准则，则应扩大样本容量，考察参数估计量的大样本性质。

需要说明的是，从估计量统计性质的角度看，无偏性与有效性是小样本性质中最为重要的两个性质，线性性并不是必须的；而在大样本性质中，由于问题较为复杂，人们更多地关注一致性。

可以证明，在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

线性性，即估计量 $\hat{\beta}_0$，$\hat{\beta}_1$ 是 Y_i 的线性组合。由(2.3.5)式知

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

$$=\frac{\sum x_i Y_i}{\sum x_i^2}-\frac{\bar{Y}\sum x_i}{\sum x_i^2}=\sum k_i Y_i$$

其中，$k_i=\dfrac{x_i}{\sum x_i^2}$。同样可得

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

$$=\sum(\frac{1}{n}-\bar{X}k_i)Y_i=\sum w_i Y_i$$

其中，$w_i=\dfrac{1}{n}-\bar{X}k_i$。

2．无偏性

无偏性，即以 X 的所有样本值为条件，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。由线性性得

$$\begin{aligned}\hat{\beta}_1&=\sum k_iY_i=\sum k_i(\beta_0+\beta_1X_i+\mu_i)\\&=\beta_0\sum k_i+\beta_1\sum k_iX_i+\sum k_i\mu_i\end{aligned}$$

易知

$$\sum k_i=\frac{\sum x_i}{\sum x_i^2}=0,\qquad \sum k_iX_i=1$$

故

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

$$\mathrm{E}(\hat{\beta}_1\mid X)=\mathrm{E}[(\beta_1+\sum k_i\mu_i)\mid X]=\beta_1+\sum k_i\mathrm{E}(\mu_i\mid X)=\beta_1$$

同样地，容易得出

$$\mathrm{E}(\hat{\beta}_0\mid X)=\mathrm{E}[(\beta_0+\sum w_i\mu_i)\mid X]=\beta_0+\sum w_i\mathrm{E}(\mu_i\mid X)=\beta_0$$

3．有效性(最小方差性)

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

首先，由 $\hat{\beta}_1,\hat{\beta}_0$ 是关于 Y_i 的线性函数，可求得它们的条件方差为

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_1\mid X)&=\mathrm{Var}(\sum k_iY_i\mid X)=\sum k_i^2\mathrm{Var}[(\beta_0+\beta_1X_i+\mu_i)\mid X]\\&=\sum k_i^2\mathrm{Var}(\mu_i\mid X)=\sum\left(\frac{x_i}{\sum x_i^2}\right)^2\sigma^2=\frac{\sigma^2}{\sum x_i^2}\end{aligned}\tag{2.3.11}$$

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_0\mid X)&=\mathrm{Var}(\sum w_iY_i\mid X)=\sum w_i^2\mathrm{Var}[(\beta_0+\beta_1X_i+\mu_i)]\\&=\sum\left(\frac{1}{n}-\bar{X}k_i\right)^2\sigma^2=\sum\left[\left(\frac{1}{n}\right)^2-2\frac{1}{n}\bar{X}k_i+\bar{X}^2k_i^2\right]\sigma^2\\&=\left[\frac{1}{n}-\frac{2}{n}\bar{X}\sum k_i+\bar{X}^2\sum\left(\frac{x_i}{\sum x_i^2}\right)^2\right]\sigma^2\\&=\left(\frac{1}{n}+\frac{\bar{X}^2}{\sum x_i^2}\right)\sigma^2=\frac{\sum x_i^2+n\bar{X}^2}{n\sum x_i^2}\sigma^2=\frac{\sum X_i^2}{n\sum x_i^2}\sigma^2\end{aligned}\tag{2.3.12}$$

其次，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：

$$\hat{\beta}_1^*=\sum c_iY_i$$

其中，$c_i=k_i+d_i$，d_i 为不全为零的常数，则容易证明（参见《计量经济学学习指南与练习（第二版）》，潘文卿，李子奈编著，高等教育出版社，2015）

$$\mathrm{Var}(\hat{\beta}_1^*) \geqslant \mathrm{Var}(\hat{\beta}_1)$$

同理，设 $\hat{\beta}_0^*$ 是其他估计方法得到的关于 β_0 的线性无偏估计量，则有

$$\mathrm{Var}(\hat{\beta}_0^*) \geqslant \mathrm{Var}(\hat{\beta}_0)$$

由以上分析可以看出，普通最小二乘估计量具有线性性、无偏性和有效性等优良性质，是**最佳线性无偏估计量**，这就是著名的**高斯-马尔可夫定理**(Gauss- Markov theorem)。显然这些优良的性质依赖于对模型的基本假设。

对于线性回归模型的普通最小二乘估计量，除了拥有一个“好”的估计量所应具备的小样本性质外，它也拥有“好”的大样本性质。例如，对 $\hat{\beta}_1$ 的一致性来说，易知

$$\begin{aligned}\mathrm{P\lim}(\hat{\beta}_1) &= \mathrm{P\lim}(\beta_1 + \sum k_i \mu_i) \\ &= \mathrm{P\lim}(\beta_1) + \mathrm{P\lim}\left(\frac{\sum x_i \mu_i}{\sum x_i^2}\right) \\ &= \beta_1 + \frac{\mathrm{P\lim}\left(\dfrac{\sum x_i \mu_i}{n}\right)}{\mathrm{P\lim}\left(\dfrac{\sum x_i^2}{n}\right)}\end{aligned}$$

等式右边第二项分子是 X 与 μ 的样本协方差的概率极限，它等于总体协方差 $\mathrm{Cov}(X,\mu)$，根据基本假设，其值为 0；而分母是 X 的样本方差的概率极限，它等于 X 的总体方差，由基本假设它为一有限常数 Q，因此有

$$\mathrm{P\lim}(\hat{\beta}_1) = \beta_1 + \frac{0}{Q} = \beta_1$$

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

为了达到对所估计参数精度测定的目的，还需进一步确定参数估计量的概率分布。由于普通最小二乘估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 分别是 Y_i 的线性组合，因此 $\hat{\beta}_0$，$\hat{\beta}_1$ 的概率分布取决于 Y_i。在 μ_i 是正态分布的假设下，Y_i 是正态分布，则 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 也服从正态分布，其分布特征由其均值和方差唯一决定。由此，以 X 的样本值为条件，有

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

于是，$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的标准差分别为

$$\sigma_{\hat{\beta}_0} = \sqrt{\frac{\sigma^2 \sum X_i^2}{n\sum x_i^2}} \tag{2.3.13}$$

$$\sigma_{\hat{\beta}_1} = \sqrt{\frac{\sigma^2}{\sum x_i^2}} \tag{2.3.14}$$

标准差可用来衡量估计量接近其真实值的程度，进而判断估计量的可靠性(图 2.3.1)。

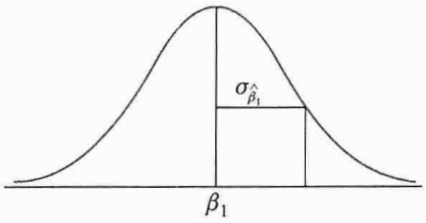

图 2.3.1 判断估计量的可靠性

2．随机干扰项 μ_i 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，都含有随机干扰项的方差 σ^2。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。由于随机干扰项 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 σ^2 进行估计。可以证明 σ^2 的最小二乘估计量为(参见《计量经济学学习指南与练习（第二版）》，潘文卿，李子奈编著，高等教育出版社，2015)

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} \tag{2.3.15}$$

它是关于 σ^2 的无偏估计量。在最大似然估计法中，可通过对对数似然函数

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

关于 σ^2 求偏导，求得 σ^2 的如下最大似然估计量：

$$\hat{\sigma}^2 = \frac{1}{n}\sum(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \frac{\sum e_i^2}{n} \tag{2.3.16}$$

在矩估计法中，由于有总体矩条件

$$\mathrm{Var}(\mu_i) = \mathrm{E}(\mu_i^2) = \sigma^2$$

其对应的样本矩条件即为

$$\hat{\sigma}^2 = \frac{1}{n}\sum(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \frac{\sum e_i^2}{n} \tag{2.3.17}$$

对照(2.3.15)式知，σ^2 的最大似然估计量与矩估计量都不具无偏性，但却具有一致性。

在随机干扰项 μ_i 的方差 σ^2 估计出后，参数 $\hat{\beta}_1$ 和 $\hat{\beta}_0$ 的方差和标准差的估计量分别是：

$\hat{\beta}_1$ 的样本方差

$$S_{\hat{\beta}_1}^2 = \frac{\hat{\sigma}^2}{\sum x_i^2} \tag{2.3.18}$$

$\hat{\beta}_1$ 的样本标准差

$$S_{\hat{\beta}_1}=\frac{\hat{\sigma}}{\sqrt{\sum x_i^2}} \tag{2.3.19}$$

$\hat{\beta}_0$ 的样本方差

$$S_{\hat{\beta}_0}^2=\frac{\hat{\sigma}^2\sum X_i^2}{n\sum x_i^2} \tag{2.3.20}$$

$\hat{\beta}_0$ 的样本标准差

$$S_{\hat{\beta}_0}=\hat{\sigma}\sqrt{\frac{\sum X_i^2}{n\sum x_i^2}} \tag{2.3.21}$$

§2.4 一元线性回归模型的统计检验

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检验的置信区间估计。

一、拟合优度检验

拟合优度检验，顾名思义，是检验模型对样本观测值的拟合程度。检验的方法是构造一个可以表征拟合程度的指标，在这里称为统计量，它是样本的函数。从检验对象中计算出该统计量的数值，然后与某一标准进行比较，得出检验结论。有人也许会问，采用普通最小二乘法进行估计，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呢？问题在于，在一个特定的条件下做得最好的并不一定就是高质量的。普通最小二乘法所保证的最好的拟合，是同一个问题内部的比较，拟合优度检验结果所表示的优劣是不同问题之间的比较。例如，图 2.4.1 中的直线方程都是由散点表示的样本观测值用最小二乘法估计的结果，对于每个问题它们都满足残差的平方和最小，但是二者对样本观测值的拟合程度显然是不同的。

1．总离差平方和的分解

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

$$\hat{Y}_i=\hat{\beta}_0+\hat{\beta}_1X_i$$

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

$$y_i=Y_i-\overline{Y}=(Y_i-\hat{Y}_i)+(\hat{Y}_i-\overline{Y})=e_i+\hat{y}_i \tag{2.4.1}$$

图 2.4.2 表示了这种分解，其中，$\hat{y}_i=\hat{Y}_i-\overline{Y}$ 是样本回归线理论值(回归拟合值)与观测值

Y_i 的平均值之差，可认为是由回归线解释的部分；$e_i = Y_i - \hat{Y}_i$ 是实际观测值与回归拟合值之差，是回归线不能解释的部分。显然，如果 Y_i 落在样本回归线上，则 Y 的第 i 个观测值与样本均值的离差，全部来自样本回归拟合值与样本均值的离差，即完全可由样本回归线解释，表明在该点处实现完全拟合。

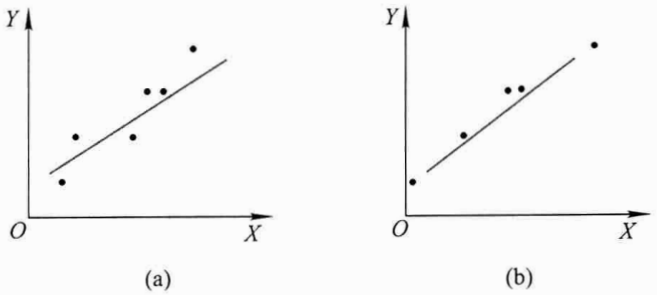

图 2.4.1 OLS 法样本回归直线

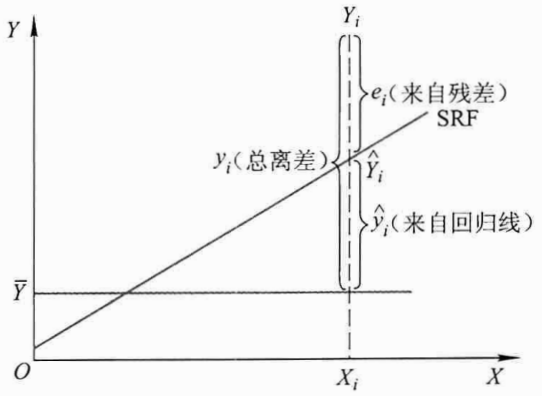

图 2.4.2 离差分解示意图

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。由于

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

可以证明 $\sum \hat{y}_i e_i = 0$，所以有

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2 \tag{2.4.2}$$

记

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

称为**总离差平方和**(total sum of squares)，反映样本观测值总体离差的大小；

$$\sum \hat{y}_i^2 = \sum (\hat{Y}_i - \bar{Y})^2 = \text{ESS}$$

称为**回归平方和**(explained sum of squares)，反映由模型中解释变量所解释的那部分离差的大小；

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

称为**残差平方和**(residual sum of squares)，反映样本观测值与估计值偏离的大小，也是模型中解释变量未解释的那部分离差的大小。

(2.4.2)式表明 Y **的观测值围绕其均值的总离差平方和可分解为两部分：一部分来自回归线，另一部分则来自随机势力。因此，可用来自回归线的回归平方和占 Y 的总离差平方和的比例来判断样本回归线与样本观测值的拟合优度。**

读者也许会问，既然 RSS 反映样本观测值与估计值偏离的大小，可否直接用它作为拟合优度检验的统计量呢？这里提出了一个普遍的问题，即**作为检验统计量的一般应该是相对量，而不能用绝对量**。因为用绝对量作为检验统计量，无法设置标准。在这里，残差平方和与样本容量关系很大，当 n 比较小时，它的值也较小，但不能因此而判断模型的拟合优度就好。

2．可决系数 R^2 统计量

根据上述关系，可以用

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{2.4.3}$$

检验模型的拟合优度，称 R^2 为**可决系数**(coefficient of determination)。显然，在总离差平方和中，回归平方和所占的比重越大，残差平方和所占的比重越小，回归直线与样本点拟合得越好。如果模型与样本观测值完全拟合，则有 $R^2 = 1$。当然，模型与样本观测值完全拟合的情况很少发生，R^2 等于 1 的情况较少。但毫无疑问的是该统计量越接近于 1，模型的拟合优度越高。

实际计算可决系数时，在 $\hat{\beta}_1$ 已经有估计值后，一个较为简单的计算公式为

$$R^2 = \hat{\beta}_1^2 \left(\frac{\sum x_i^2}{\sum y_i^2} \right) \tag{2.4.4}$$

这里用到了样本回归函数的离差形式来计算回归平方和：

$$\text{ESS} = \sum \hat{y}_i^2 = \sum (\hat{\beta}_1 x_i)^2 = \hat{\beta}_1^2 \sum x_i^2$$

在例 2.1.1 的可支配收入-消费支出例子中，

$$R^2 = \hat{\beta}_1^2 \frac{\sum x_i^2}{\sum y_i^2} = \frac{0.670\,1^2 \times 7\ 425\ 000}{3\ 356\ 322} = 0.993\,4$$

说明在线性回归模型中的家庭消费支出总离差中，由家庭可支配收入的离差解释的部分占 99.34%，模型的拟合优度较高。

由(2.4.3)式知，可决系数的取值范围为 $0 \leqslant R^2 \leqslant 1$，它是一个非负的统计量，随着抽样的不同而不同，即是随抽样而变动的统计量。为此，对可决系数的统计可靠性也应进行检验，这将在第三章中讨论。

二、变量的显著性检验

变量的显著性检验，旨在对模型中被解释变量与解释变量之间的线性关系是否显著

成立作出推断，或者说考察所选择的解释变量是否对被解释变量有显著的线性影响。

从上面的拟合优度检验中可以看出，拟合优度高，则解释变量对被解释变量的解释程度就大，线性影响就强，可以推测模型线性关系成立；反之，就不成立。但这只是一个模糊的推测，不能给出一个统计上的严格的结论。因此，还必须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。**

1．假设检验

假设检验是统计推断的一个主要内容，它的基本任务是根据样本所提供的信息，对未知总体分布的某些方面的假设作出合理的判断。

假设检验的程序是：先根据实际问题的要求提出一个论断，称为统计假设，记为 H_0；然后根据样本的有关信息，对 H_0 的真伪进行判断，作出拒绝 H_0 或接受 H_0 的决策。

假设检验的基本思想是概率性质的反证法。为了检验原假设 H_0 是否正确，先假定这个假设是正确的，看由此能推出什么结果。如果导致一个不合理的结果，则表明“假设 H_0 为正确”是错误的，即原假设 H_0 不正确，因此要拒绝原假设 H_0；如果没有导致一个不合理现象的出现，则不能认为原假设 H_0 不正确，因此不能拒绝原假设 H_0。

概率性质的反证法的根据是小概率事件原理，该原理认为“小概率事件在一次试验中几乎是不可能发生的”。在原假设 H_0 下构造一个事件，这个事件在“假设 H_0 是正确”的条件下是一个小概率事件。随机抽取一组容量为 n 的样本观测值进行该事件的试验，如果该事件发生了，说明“假设 H_0 正确”是错误的，因为不应该出现的小概率事件出现了，因而应该拒绝原假设 H_0；反之，如果该小概率事件没有出现，就没有理由拒绝原假设 H_0，应该接受原假设 H_0。

2．变量的显著性检验的方法

用以进行变量显著性检验的方法主要有三种：F 检验，t 检验，z 检验。它们的区别在于构造的统计量不同。应用最为普遍的是 t 检验。几乎所有的计量经济学软件包中，都有关于 t 统计量的计算结果。我们在此只介绍 t 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

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

进一步根据数理统计学中的定义，如果真实的 σ^2 未知，而用它的无偏估计量 $\hat{\sigma}^2=\dfrac{\sum e_i^2}{n-2}$ 替代时，可构造如下统计量：

$$t=\frac{\hat{\beta}_1-\beta_1}{\sqrt{\dfrac{\hat{\sigma}^2}{\sum x_i^2}}}=\frac{\hat{\beta}_1-\beta_1}{S_{\hat{\beta}_1}} \tag{2.4.5}$$

则该统计量服从自由度为 $n-2$ 的 t 分布。因此，可用该统计量作为 β_1 显著性检验的 t 统

计量。

如果变量 X 是显著的，那么参数 β_1 应该显著地不为 0。于是，在变量显著性检验中设计的原假设与备择假设分别为

$$H_0:\beta_1=0,\qquad H_1:\beta_1\neq 0$$

给定一个显著性水平 α，比如 0.05，查 t 分布表(见附录)，得到一个临界值 $t_{\frac{\alpha}{2}}(n-2)$，则 $|t|>t_{\frac{\alpha}{2}}(n-2)$ (这里的 t 已不同于(2.4.5)式，其中 $\beta_1=0$)为原假设 H_0 下的一个小概率事件。

在参数估计完成后，可以很容易计算 t 的数值。如果发生了 $|t|>t_{\frac{\alpha}{2}}(n-2)$，则在 α 的显著性水平下拒绝原假设 H_0，即变量 X 是显著的，通过变量显著性检验；如果未发生 $|t|>t_{\frac{\alpha}{2}}(n-2)$，则在显著性水平 α 下不拒绝原假设 H_0，表明变量 X 是不显著的，未通过变量显著性检验。

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：

$$t=\frac{\hat{\beta}_0-\beta_0}{\sqrt{\dfrac{\hat{\sigma}^2\sum X_i^2}{n\sum x_i^2}}}=\frac{\hat{\beta}_0-\beta_0}{S_{\hat{\beta}_0}} \tag{2.4.6}$$

同样地，该统计量服从自由度为 $n-2$ 的 t 分布，检验的原假设一般仍为 $\beta_0=0$。

在例 2.1.1 及例 2.3.1 的可支配收入-消费支出例子中，首先计算 σ^2 的估计值：

$$\hat{\sigma}^2=\frac{\sum e_i^2}{n-2}=\frac{\sum y_i^2-\hat{\beta}_1^2\sum x_i^2}{n-2}$$

$$=\frac{3\ 356\ 322-0.670\,1^2\times 7\ 425\ 000}{10-2}=2\ 780.56$$

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

$$S_{\hat{\beta}_1}=\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2}}=\sqrt{\frac{2\ 780.56}{7\ 425\ 000}}=\sqrt{0.000\ 37}=0.019$$

$$S_{\hat{\beta}_0}=\sqrt{\frac{\hat{\sigma}^2\sum X_i^2}{n\sum x_i^2}}=\sqrt{\frac{2\ 780.56\times 53\ 650\ 000}{10\times 7\ 425\ 000}}=44.82$$

t 统计量的计算结果分别为

$$t_1=\frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}=\frac{0.670\,1}{0.019}=35.27$$

$$t_0=\frac{\hat{\beta}_0}{S_{\hat{\beta}_0}}=\frac{142.28}{44.82}=3.17$$

给定一个显著性水平 α=0.05，查 t 分布表中自由度为 8(在这个例中 $n-2=8$)，

α=0.05 的临界值，得到$t_{\frac{\alpha}{2}}(8)=2.306$。$|t_1|>t_{\frac{\alpha}{2}}(n-2)$，说明解释变量家庭可支配收入在5%的显著性水平下显著，即通过了变量显著性检验；同样地，$|t_0|>t_{\frac{\alpha}{2}}(n-2)$,表明在 5%的显著性水平下，拒绝截距项为零的假设。

三、参数检验的置信区间估计

假设检验可以通过一次抽样的结果检验总体参数可能值的范围（最常用的假设为总体参数值为零），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$\hat{\beta}_j$离真实的参数值β_j有多“近”($j=0,1$)，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**(random interval)$(\hat{\beta}_j-\delta,\ \hat{\beta}_j+\delta)$包含参数$\beta_j$的真值的概率为$1-\alpha$，即

$$P(\hat{\beta}_j-\delta\leqslant\beta_j\leqslant\hat{\beta}_j+\delta)=1-\alpha$$

如果存在如上述这样的一个区间，称之为**置信区间**(confidence interval)；$1-\alpha$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)。

在变量的显著性检验中已经知道

$$t=\frac{\hat{\beta}_j-\beta_j}{S_{\hat{\beta}_j}}\sim t(n-2)\qquad j=0,1$$

这就是说，如果给定置信度$1-\alpha$，从t分布表中查得自由度为$n-2$的临界值$t_{\frac{\alpha}{2}}$，那么t值处在$(-t_{\frac{\alpha}{2}},t_{\frac{\alpha}{2}})$的概率是$1-\alpha$，表示为

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

即

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

$$P(\hat{\beta}_j-t_{\frac{\alpha}{2}}\times S_{\hat{\beta}_j}<\beta_j<\hat{\beta}_j+t_{\frac{\alpha}{2}}\times S_{\hat{\beta}_j})=1-\alpha$$

于是得到$1-\alpha$的置信度下β_j的置信区间是

$$(\hat{\beta}_j-t_{\frac{\alpha}{2}}\times S_{\hat{\beta}_j},\hat{\beta}_j+t_{\frac{\alpha}{2}}\times S_{\hat{\beta}_j})\tag{2.4.7}$$

在例 2.1.1 与例 2.3.1 中，如果给定$\alpha=0.01$，查表得

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

从假设检验中已经得到

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

于是，根据(2.4.7)式计算得到 β_1，β_0 的置信区间分别为(0.606 4, 0.733 8)和(–8.09,292.65)。显然，参数 β_1 的置信区间小于 β_0 的置信区间。

由于置信区间在一定程度上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。**如何才能缩小置信区间呢？**从(2.4.7)式不难看出：(1)**增大样本容量** n。样本容量变大，可使样本参数估计量的标准差减小；同时，在同样的显著性水平下，n 越大，t 分布表中的临界值越小。(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成正比，模型的拟合优度越高，残差平方和应越小。

模型的参数一般具有特定的经济意义。例如，在例 2.1.1 中，参数 β_1 表示边际消费倾向。当经过模型估计得到 $\hat{\beta}_1=0.670\,1$ 后，我们能否说“边际消费倾向为 0.6701”呢？不能。根据置信区间，我们只能说“边际消费倾向以 0.99 的置信水平处于以 0.6701 为中心的区间(0.606 4, 0.733 8)中”。

§2.5 一元线性回归分析的应用：预测问题

计量经济学模型的一个重要应用是经济预测。对于一元线性回归模型

$$\hat{Y}_i=\hat{\beta}_0+\hat{\beta}_1X_i$$

如果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 $\mathrm{E}(Y\mid X=X_0)$ 或个别值 Y 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。原因在于两方面：一是模型中的参数估计量是不确定的；二是随机干扰项的影响。所以，我们得到的仅是预测值的一个估计值，预测值仅以某一个置信度处于以该估计值为中心的一个区间中。预测在更大程度上说是一个区间估计问题。**

一、预测值是条件均值或个别值的一个无偏估计

在总体回归函数为 $\mathrm{E}(Y\mid X)=\beta_0+\beta_1X$ 的情况下，Y 在 $X=X_0$ 时的条件均值为

$$\mathrm{E}(Y\mid X=X_0)=\beta_0+\beta_1X_0$$

通过样本回归函数 $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X$，求得 $X=X_0$ 条件下的拟合值为

$$\begin{aligned}\hat{Y}_0&=\hat{\beta}_0+\hat{\beta}_1X_0\\ \mathrm{E}(\hat{Y}_0)&=\mathrm{E}(\hat{\beta}_0+\hat{\beta}_1X_0)\\ &=\mathrm{E}(\hat{\beta}_0)+X_0\mathrm{E}(\hat{\beta}_1)\\ &=\beta_0+\beta_1X_0\end{aligned}\qquad(2.5.1)$$

另一方面，在总体回归模型为 $Y=\beta_0+\beta_1X+\mu$ 的情况下，Y 在 $X=X_0$ 条件下的值为

$$\begin{aligned} Y_0 &= \beta_0+\beta_1X_0+\mu \\ \mathrm{E}(Y_0) &= \mathrm{E}(\beta_0+\beta_1X_0+\mu) \\ &= \beta_0+\beta_1X_0+\mathrm{E}(\mu) \\ &= \beta_0+\beta_1X_0 \end{aligned} \tag{2.5.2}$$

(2.5.1)式与(2.5.2)式说明，在 $X=X_0$ 条件下，样本估计值 $\hat{Y}_0$ 是总体均值 $\mathrm{E}(Y\mid X=X_0)$ 和个别值 Y_0 的无偏估计，因此可用 $\hat{Y}_0$ 作为 $\mathrm{E}(Y\mid X=X_0)$ 与 Y_0 的预测值。

二、总体条件均值与个别值预测值的置信区间

1．总体条件均值预测值的置信区间

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

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

则 $$\mathrm{E}(\hat{Y}_0)=\mathrm{E}(\hat{\beta}_0)+X_0\mathrm{E}(\hat{\beta}_1)=\beta_0+\beta_1X_0$$

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

可以证明（参见《计量经济学学习指南与练习（第二版）》，潘文卿，李子奈编著，高等教育出版社，2015。）

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

因此

$$\begin{aligned} \mathrm{Var}(\hat{Y}_0) &= \frac{\sigma^2\sum X_i^2}{n\sum x_i^2}-\frac{2X_0\bar{X}\sigma^2}{\sum x_i^2}+\frac{X_0^2\sigma^2}{\sum x_i^2} \\ &= \frac{\sigma^2}{\sum x_i^2}\left(\frac{\sum X_i^2-n\bar{X}^2}{n}+\bar{X}^2-2X_0\bar{X}+X_0^2\right) \\ &= \frac{\sigma^2}{\sum x_i^2}\left[\frac{\sum x_i^2}{n}+(X_0-\bar{X})^2\right] \\ &= \sigma^2\left[\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum x_i^2}\right] \end{aligned}$$

故

$$\hat{Y}_0\sim N\left\{\beta_0+\beta_1X_0,\sigma^2\left[\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum x_i^2}\right]\right\} \tag{2.5.3}$$

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

$$t=\frac{\hat{Y}_0-(\beta_0+\beta_1X_0)}{S_{\hat{Y}_0}}\sim t(n-2) \tag{2.5.4}$$

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 $1-\alpha$ 的置信度下，总体均值 $\mathrm{E}(Y\mid X_0)$ 的置信区间为

$$\hat{Y}_0-t_{\frac{\alpha}{2}}\times S_{\hat{Y}_0}<\mathrm{E}(Y\mid X_0)<\hat{Y}_0+t_{\frac{\alpha}{2}}\times S_{\hat{Y}_0} \tag{2.5.5}$$

2．总体个别值预测值的置信区间

由 $Y_0=\beta_0+\beta_1X_0+\mu$ 知

$$Y_0\sim N(\beta_0+\beta_1X_0,\sigma^2)$$

于是

$$\hat{Y}_0-Y_0\sim N\left\{0,\sigma^2\left[1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum x_i^2}\right]\right\} \tag{2.5.6}$$

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

$$t=\frac{\hat{Y}_0-Y_0}{S_{\hat{Y}_0-Y_0}}\sim t(n-2) \tag{2.5.7}$$

其中

$$S_{\hat{Y}_0-Y_0}=\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum x_i^2}\right]}$$

从而在 $1-\alpha$ 的置信度下，Y_0 的置信区间为

$$\hat{Y}_0-t_{\frac{\alpha}{2}}\times S_{\hat{Y}_0-Y_0}<Y_0<\hat{Y}_0+t_{\frac{\alpha}{2}}\times S_{\hat{Y}_0-Y_0} \tag{2.5.8}$$

在例 2.1.1 及例 2.3.1 的可支配收入-消费支出例子中，得到的样本回归函数为

$$\hat{Y}_i=142.28+0.6701X_i$$

则在 $X_0=1\,000$ 处，

$$\hat{Y}_0=142.28+0.6701\times1\,000=812.38$$

它可作为总体均值 $\mathrm{E}(Y\mid X=1\,000)$ 或 Y 的个别值在 $X=1\,000$ 处的预测的估计值。而

$$\mathrm{Var}(\hat{Y}_0)=2\,780.56\left[\frac{1}{10}+\frac{(1\,000-2\,150)^2}{7\,425\,000}\right]=773.31$$

$$S(\hat{Y}_0)=27.81$$

因此，总体均值 $\mathrm{E}(Y\mid X=1\,000)$ 的 95%的置信区间为

$$812.38-2.306\times27.81<\mathrm{E}(Y\mid X=1\,000)<812.38+2.306\times27.81$$

或为

$$(748.25,\ 876.51)$$

同样地，对于 Y 在 $X=1\,000$ 的个别值 Y_0，易知其 95%的置信区间为

$$812.38-2.306\times59.61<Y\mid_{X=1\,000}<812.38+2.306\times59.61$$

或为

$$(674.92,\ 949.84)$$

如图 2.5.1 所示，如果对每个 X 值求其总体均值 $\mathrm{E}(Y\mid X)$ 的 95%的置信区间，将区间端点连接起来，可以得到关于总体回归函数的**置信带(域)**(confidence band)。同样地，对

每个 X 值求 Y 的个别值的 95%的置信区间，将区间端点连接起来，可以得到关于个别值 Y_0 的**置信带(域)**。可以看出，Y 的个别值 Y_0 的置信带比其总体均值的置信带宽。

对于 Y 的总体均值 $E(Y_0)$ 与个别值 Y_0 的预测区间(置信区间)，有：(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，置信带的宽度在 X 的均值处最小，在其附近进行预测(插值预测)精度高；X 越远离其均值，置信带越宽，预测精度将下降。

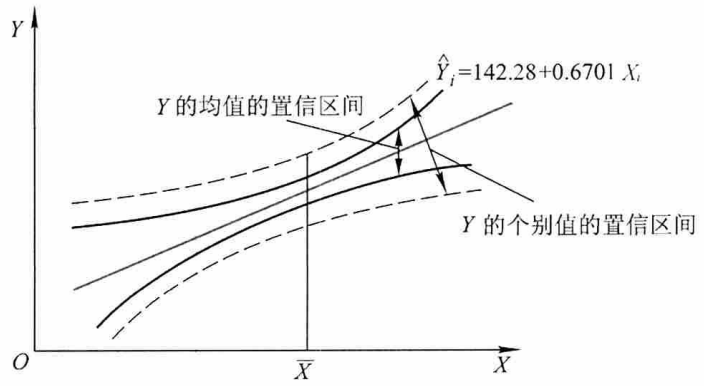

图 2.5.1 Y 均值与个别值的置信区间

§2.6 建 模 实 例

本节通过一个**截面数据**（cross-sectional data）的实例演示计量经济模型建立的一般过程。

例 2.6.1

为考察中国居民 2013 年人均**可支配收入**与人均**消费支出**的关系，表 2.6.1 给出了中国内地 31 个省、市、自治区以当年价测算的居民家庭年人均可支配收入(X)与年人均消费支出(Y)两组数据。由于表中是同一年份中不同地区居民家庭的人均可支配收入与人均消费支出，因此也称为**截面数据**。

1．建立模型

本例中我们假设拟建立如下一元回归模型：

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

表 2.6.2 给出了采用 Eviews 软件对表 2.6.1 中的数据进行回归分析的计算结果。一般可写出如下回归分析结果：

$$\hat{Y}_i = 477.12 + 0.707\,1X_i$$

$$(1.18) \qquad (34.629)$$

$$R^2 = 0.976\,4 \qquad F = 1\,199.15$$

其中括号内的数为相应参数的 t 检验值，R^2 是可决系数，F 是一个重要的检验统计量，其含义将在后面的章节中介绍。

表 2.6.1 中国内地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 单位：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	40 830.0	29 175.6	湖 北	16 472.5	11 760.8
天 津	26 359.2	20 418.7	湖 南	16 004.9	11 945.9
河 北	15 189.6	10 872.2	广 东	23 420.7	17 421.0
山 西	15 119.7	10 118.3	广 西	14 082.3	9 596.5
内蒙古	18 692.9	14 877.7	海 南	15 733.3	11 192.9
辽 宁	20 817.8	14 950.2	重 庆	16 568.7	12 600.2
吉 林	15 998.1	12 054.3	四 川	14 231.0	11 054.7
黑龙江	15 903.4	12 037.2	贵 州	11 083.1	8 288.0
上 海	42 173.6	30 399.9	云 南	12 577.9	8 823.8
江 苏	24 775.5	17 925.8	西 藏	9 746.8	6 310.6
浙 江	29 775.0	20 610.1	陕 西	14 371.5	11 217.3
安 徽	15 154.3	10 544.1	甘 肃	10 954.4	8 943.4
福 建	21 217.9	16 176.6	青 海	12 947.8	11 576.5
江 西	15 099.7	10 052.8	宁 夏	14 565.8	11 292.0
山 东	19 008.3	11 896.8	新 疆	13 669.6	11 391.8
河 南	14 203.7	10 002.5			

资料来源：《中国统计年鉴》(2014)。

表 2.6.2 中国内地城镇居民人均消费支出对人均可支配收入的回归

Dependent Variable: Y
Method: Least Squares
Date: 01/11/15 Time: 10:01
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		

2．模型检验

从回归估计的结果看，模型拟合较好。可决系数 R^2 =0.976 4，表明居民人均消费支出变化的 97.64%可由人均可支配收入的变化来解释。从斜率项的 t 检验值看，大于 5%显著性水平下自由度为 n–2=29 的临界值 $t_{0.025}(29)$=2.045，且该斜率值满足 0<0.707 1<1，符合经济理论中边际消费倾向在 0 与 1 之间的绝对收入假说，表明 2013 年，中国城镇居民家庭人均可支配收入每增加 1 元，人均消费支出增加 0.707 1 元。

3．预测

假设我们需要关注 2013 年人均可支配收入在 20 000 元这一档的中国家庭的人均消费支出问题。由上述回归方程可得该类家庭人均消费支出的预测值：

$$\hat{Y}_0 = 477.12 + 0.707\,1 \times 20\,000 = 14\,619.1\,(\text{元})$$

下面给出该类家庭人均消费支出 95%置信度的预测区间。

由于人均可支配收入 X 的样本均值与样本方差为

$$\mathrm{E}(X)=18\,282.2 \qquad \mathrm{Var}(X)=58\,984\,198$$

于是，在 95%的置信度下，$\mathrm{E}(Y_0)$的预测区间为

$$14\,619.1 \pm 2.045 \times \sqrt{\frac{21\,395\,274}{31-2} \times \left[\frac{1}{31} + \frac{(20\,000-18\,282.2)^2}{(31-1)\times 58\,984\,198}\right]}$$

$$=14\,619.1 \pm 323.5$$

或　　（14 295.6,14 942.6）

如果我们想知道某地区某家庭人均可支配收入为 20 000 元时，该家庭人均消费支出的个值预测，则仍通过上述样本回归方程得到 14 619.1 元的消费支出预测值。

同样地，在 95%的置信度下，该家庭人均消费支出的预测区间为

$$14\,619.1 \pm 2.045 \times \sqrt{\frac{21\,395\,274}{31-2} \times \left[1 + \frac{1}{31} + \frac{(20\,000-18\,282.2)^2}{(31-1)\times 58\,984\,198}\right]}$$

$$=14\,619.1 \pm 1\,786.1$$

或　　（12 833.0, 16 405.2）

本章练习题

1．为什么计量经济学模型的理论方程中必须包含随机干扰项？

2．下列计量经济学方程哪些是正确的？哪些是错误的？为什么？

（1）$Y_i = \alpha + \beta X_i \qquad i=1,2,\cdots,n$

（2）$Y_i = \alpha + \beta X_i + \mu_i \qquad i=1,2,\cdots,n$

（3）$Y_i = \hat{\alpha} + \hat{\beta} X_i + \mu_i \qquad i=1,2,\cdots,n$

（4）$\hat{Y}_i=\hat{\alpha}+\hat{\beta}X_i+\mu_i \qquad i=1,2,\cdots,n$

（5）$Y_i=\hat{\alpha}+\hat{\beta}X_i \qquad i=1,2,\cdots,n$

（6）$\hat{Y}_i=\hat{\alpha}+\hat{\beta}X_i \qquad i=1,2,\cdots,n$

（7）$Y_i=\hat{\alpha}+\hat{\beta}X_i+\hat{\mu}_i \qquad i=1,2,\cdots,n$

（8）$\hat{Y}_i=\hat{\alpha}+\hat{\beta}X_i+\hat{\mu}_i \qquad i=1,2,\cdots,n$

其中，带“ ^ ”者表示“估计值”。

3．一元线性回归模型的基本假设主要有哪些？违背基本假设的计量经济学模型是否就不可以估计？

4．线性回归模型

$$Y_i=\alpha+\beta X_i+\mu_i,\qquad i=1,2,\cdots,n$$

的零均值假设是否可以表示为 $\frac{1}{n}\sum_{i=1}^{n}\mu_i=0$？为什么？

5．假设已经得到关系式 $Y=\beta_0+\beta_1X$ 的最小二乘估计，试回答：

（1）假设决定把变量 X 的单位扩大 10 倍，这样对原回归的斜率和截距会有什么样的影响？如果把变量 Y 的单位扩大 10 倍，又会怎样？

（2）假定给 X 的每个观测值都增加 2，对原回归的斜率和截距会有什么样的影响？如果给 Y 的每个观测值都增加 2，又会怎样？

6．假使在回归模型 $Y_i=\beta_0+\beta_1X_i+\mu_i$ 中，用不为零的常数 δ 去乘每一个 X 值，这会不会改变 Y 的拟合值及残差？如果对每个 X 都加上一个非零常数 δ，又会怎样？

7．假设有人做了如下的回归：

$$y_i=\hat{\beta}_0+\hat{\beta}_1x_i+e_i$$

其中，y_i,x_i 分别为 Y_i,X_i 关于各自均值的离差。问 $\hat{\beta}_1$ 和 $\hat{\beta}_0$ 将分别取何值？

8．令 $\hat{\beta}_{YX}$ 和 $\hat{\beta}_{XY}$ 分别为 Y 对 X 的回归和 X 对 Y 的回归中的斜率，证明：

$$\hat{\beta}_{YX}\hat{\beta}_{XY}=r^2$$

其中，r 为 X 与 Y 之间的线性相关系数。

9．记样本回归模型为 $Y_i=\hat{\beta}_0+\hat{\beta}_1X_i+e_i$，试证明：

（1）估计的 Y 的均值等于实测的 Y 的均值：

$$\overline{\hat{Y}}=\overline{Y}$$

（2）残差和为零，从而残差的均值为零：

$$\sum e_i=0,\qquad \overline{e}=0$$

（3）残差项与 X 不相关：

$$\sum e_iX_i=0$$

（4）残差项与估计的 Y 不相关：

$$\sum e_i\hat{Y}_i=0$$

10．证明：一元线性回归总离差平方和的分解式中：

$$\sum \hat{y}_i e_i = 0$$

11．下面数据是依据 10 对 X 和 Y 的观察值得到的：

$$\sum Y_i = 1\,110\,,\qquad \sum X_i = 1\,680\,,\qquad \sum X_i Y_i = 204\,200$$

$$\sum X_i^2 = 315\,400\,,\qquad \sum Y_i^2 = 133\,300$$

假定满足所有的经典线性回归模型的假设。求：

（1）β_0，β_1 的估计值及其标准差；

（2）可决系数 R^2；

（3）对 β_0，β_1 分别建立 95%的置信区间。利用置信区间法，你可以接受零假设：$\beta_1 = 0$ 吗？

12．下表是中国内地某年各地区税收 Y 和国内生产总值 GDP 的统计资料。

单位：亿元

地区	Y	GDP	地区	Y	GDP
北　京	1 435.7	9 353.3	湖　北	434.0	9 230.7
天　津	438.4	5 050.4	湖　南	410.7	9 200.0
河　北	618.3	13 709.5	广　东	2 415.5	31 084.4
山　西	430.5	5 733.4	广　西	282.7	5 955.7
内蒙古	347.9	6 091.1	海　南	88.0	1 223.3
辽　宁	815.7	11 023.5	重　庆	294.5	4 122.5
吉　林	237.4	5 284.7	四　川	629.0	10 505.3
黑龙江	335.0	7 065.0	贵　州	211.9	2 741.9
上　海	1 975.5	12 188.9	云　南	378.6	4 741.3
江　苏	1 894.8	25 741.2	西　藏	11.7	342.2
浙　江	1 535.4	18 780.4	陕　西	355.5	5 465.8
安　徽	401.9	7 364.2	甘　肃	142.1	2 702.4
福　建	594.0	9 249.1	青　海	43.3	783.6
江　西	281.9	5 500.3	宁　夏	58.8	889.2
山　东	1 308.4	25 965.9	新　疆	220.6	3 523.2
河　南	625.0	15 012.5			

要求，以手工和运用 Eviews 软件(或其他软件)：

（1）作出散点图，建立税收随国内生产总值 GDP 变化的一元线性回归方程，并解释斜率的经济意义；

（2）对所建立的回归方程进行检验；

（3）若该年某地区国内生产总值为 8 500 亿元，求该地区税收收入的预测值及预测区间。