第 1 章　经典单方程计量经济学模型：一元线性回归模型

经典单方程计量经济学模型是研究经济、管理和社会问题重要的模型，时至今日仍然发挥着重要作用，它是计量经济学内容体系重要的组成部分，同时也是进一步学习放宽假定条件的单方程模型和联立方程模型的基础。本章从简单的一元线性回归模型入手，介绍计量经济学模型的设定、估计及检验问题，为以后各章学习打下坚实的基础。

1.1　回归分析概述

1.1.1　回归分析基本概念

1. 回归分析的提出

回归概念最先由英国科学家弗朗西斯·高尔顿引入。高尔顿提出了相关性概念并且建立了回归分析方法，对经济计量领域贡献颇大，被誉为线性回归及其相关技术的鼻祖。1875 年，高尔顿用豌豆做实验，阐明了豌豆尺寸大小的遗传规律。选取大小不一的豌豆 7 组，每组种 10 粒。最后，将原始豌豆种子（亲本）与新豌豆种子（后代）的大小进行比较。绘制结果时，发现后代和亲本并不完全相同：体型偏小豌豆倾向得到体型更大的后代，而体型偏大豌豆倾向得到体型偏小的后代。高尔顿将这种现象称为“向均值回归”（趋向于祖先的平均类型），后来又称之为“回归平均”。有一种普遍现象，如果某个个体在某个时期出现了某种极端特征，如某项指标明显高于或低于均值，那么，在未来的某个时期，这类个体或后代普遍会减弱其最开始出现的极端特征，这种趋势线被称为“回归”效应。这种效应也得到学者的验证。正如高尔顿进一步发现的那样，较矮的父辈往往有比自己更高的后代，而较高的父辈往往有比自己更矮的后代。需要注意的是，随着父母身高的增加，后代的平均身高也会增加，如图 1.1 所示。

在研究失业率和通货膨胀率（用货币工资变化率表示）关系时，新西兰经济学家威廉·菲利普斯提出了具有较大影响的菲利普斯曲线。1958 年，菲利普斯以英国 1867—1913 年失业率与货币工资变化率的原始数据为研究样本，得出了失业率与通货膨胀率此消彼长的交替关系。由图 1.2 可以看出，通货膨胀率较高时，失业率较低；通货膨胀率较低时，失业率较高。换句话说，降低通货膨胀率要以增加失业人口数量为代价，提高就业率要以提高通货膨胀率为代价，经济学家所期盼的低通货膨胀率和低失业率共存的现象无法实现。

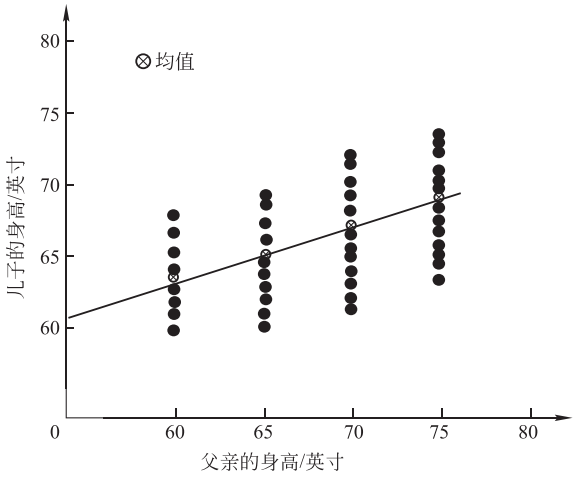

图 1.1　给定父亲身高时儿子身高的假想分布

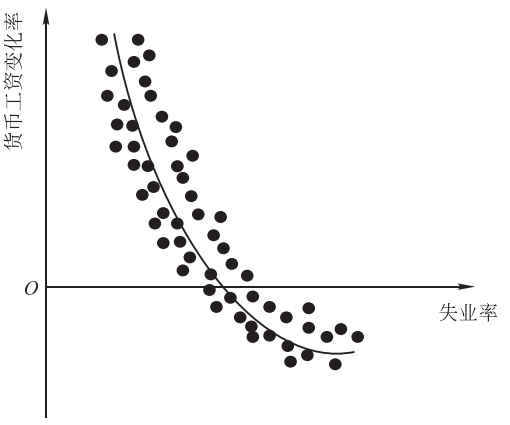

图 1.2　菲利普斯曲线

2. 变量间的函数关系与统计关系

从哲学上讲，世界上的任何事物与现象都不能孤立存在，都与周围的其他事物和现象有某种联系或关联。万事万物之间的联系可以大致分为两类，一类是确定的函数关系，另一类是不确定的统计关系。在经济变量之间，一般不存在精确的函数关系，以相关关系居多。计量经济学主要是探寻不同经济变量之间的相关关系以及相关关系的作用机理及经济规律。

确定性现象间的关系常常表现为函数关系。例如，牛顿发现的宇宙中万物相互吸引的万有引力定律，$F=(km_1m_2)/r^2$，m_1 和 m_2 表示物体的质量，r 表示距离，k 为比例常数。类似的例子有欧姆定律、波尔定律、能量守恒定律等。

不确定现象之间的关系通常以统计的形式出现。例如，研究某种农作物产量 Y 与化肥施用量 X 之间的关系，容易发现，Y 随 X 的变化呈现规律性的变化：在化肥施用量达到饱和之前，Y 随 X 的增加而增加；在化肥施用量达到饱和之后，Y 随的 X 增加而减少。但是，Y 与 X 之间不存在确定的函数关系。之所以无法明确给出两者之间的函数关系，是因为农作物产量 Y 除了受化肥施用量 X 的影响以外，还受到光照、温度、水分、土壤、技术、田间管理等因素的影响，而这几个影响因素无法进行准确的量化。虽然无法研究两者之间的函数关系，但对两者之间统计关系的研究，也可以用来指导农业生产。作为非确定性变量，农作物产量 Y 也称为随机变量。

变量之间的函数关系和相关关系也不是完全绝对的，在某些特定条件下也可以互相转化。例如，在观察决定论现象的过程中，经常会产生测量误差，此时函数关系多以相关关系的形式表示；相反，如果能够确定非决定论现象的影响因素，则所有变量纳入依赖关系中，变量之间的相关关系可以转换为函数关系。相关分析和回归分析主要用于研究非决定论现象之间的统计相关性。

3. 变量间的函数关系与统计关系

回归分析研究一个变量对另一组变量的依赖关系，但并不一定是因果关系，可能仅仅是相关关系。一个统计关系，即使相关性很强，也不能确立为因果关系。对因果关系的判断，

必须来自统计学以外，比如利用先验理论、经济学理论或管理学理论等。

4. 术语与符号

在进入正式的回归理论分析之前，先来斟酌一下有关术语与符号的问题。被解释变量和解释变量两个名词在文献中都有过种种其他描述，见表 1.1。

表 1.1　被解释变量和解释变量的其他描述

被解释变量（explained variable）	解释变量（explanatory variable）
因变量（dependent variable）	自变量（independent variable）
预测子（predictand）	预测元（predictor）
回归子（regressand）	回归元（regressor）
响应（response）变量	刺激（stimulus）变量
内生（endogenous）变量	外生（exogenous）变量
结果（outcome）变量	协变量（covariate）
被控变量（controlled variable）	控制变量（control variable）

计量经济学所使用的术语依赖于传统习惯和个人偏好，本书中使用的术语是解释变量和被解释变量，或者是更中立的回归变量。如果只研究被解释变量对解释变量的依赖性，如消费支出对实际收入的依赖性，则这类研究称为一元回归分析或双变量回归分析。但是，在研究一个被解释变量对多个解释变量的依赖性时，例如，农作物的产量依赖于气温、降雨、光照和肥料的应用，则称为多元回归分析。也就是说，在一元回归中只有一个解释变量，在多元回归中至少有两个不同的解释变量。

“random”和“stochastic”是同义词，都是随机的意思。一个随机变量的含义是：随机变量以特定概率取特定值，概率值可正可负。除非另作声明，字母 Y 一律指被解释变量，而字母 X（X_1，X_2，…，X_k）一律指解释变量。其中 X_k 代表第 k 个解释变量。下标 i 或 t 则指第 i 次或第 t 次观测，这样 X_{ki}（或 X_{kt}）就指对变量 X_k 的第 i（或 t）次观测。N（或 T）指总体中的观测总个（次）数，而 n（或 t）则指样本中的观测值总个数。作为一种惯例，观测值下标 i 将用于横截面数据（cross-sectional data）（在一个时间点上对不同对象收集的数据），而下标 t 将用于时间序列数据（timeseries data）（对同一个研究对象在不同时期收集的数据）。

5. 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来看，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

$$r_{XY}=\frac{\sigma_{XY}}{\sqrt{V(X)V(Y)}} \tag{1-1}$$

其中，σ_{XY}是变量 X 和 Y 的协方差，V（X）和 V（Y）分别是变量 X 和变量 Y 的方差。如果给出 X 和 Y 的一组样本（X_i，Y_i），$i=1$，2，…，n，则样本相关系数为

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

其中，$\overline{X}$ 与 $\overline{Y}$ 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因果关系，这时可以通过回归分析来研究它们间的具体依存关系。例如，就像经济学中所阐述的边际效应一样，消费支出与可支配收入之间不但密切相关，而且有因果关系，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。结论可得每 1 元的变化所引起的消费支出变化将会越来越小，这便是回归分析的结果。

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或因变量，后一个变量称为解释变量或自变量。

相关分析和回归分析具有密切的关联性，既有相同点又有不同点。相同点表现在，相关分析和回归分析均可以描述非决定论变量的统计依赖性，并能测量依赖的程度。不同点表现在以下几点。第一，相关分析只测量统计变量之间的关联程度，变量的地位是对称的，只考察变量之间最基本的关系。回归分析侧重于分析相关变量之间的因果关系，变量间的位置关系不能调换，如下雨和打伞之间具有明显的因果关系，下雨是原因，打伞是结果，位置不可以调换。第二，相关关系的用途是通过样本来预测总体，是解决“怎么样”，因果关系的用途是解释两个变量之间的影响机制，是解决“为什么”。第三，在经济学研究中更注重回归分析，在统计学研究中更注重相关分析。一般来说，相关关系是经济学家研究的起点，因果关系是经济学家研究的目标。本书主要研究回归分析。

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。

1.1.2 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平均值，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。

［案例 1-1］ 国内生产总值（GDP）是一个国家或地区经济核算最重要的指标之一，它能够大致反映一个国家或地区的经济水平和发展状况。一个国家的税收与财政收入密切相关。税收的增加意味着财政收入的增加，财政收入增加，意味着国家用于文化产业、科学技术、脱贫攻坚、军事装备、环境治理、稀有矿产矿床勘探、城市改善的投入就会增加，对公共服务的投入也会增加。因此，税收总额的增加对整个国家的经济发展极为重要。税收和国

内生产总值的关系是什么？表 1.2 为我国 2000—2019 年国内生产总值和税收收入。

表 1.2　我国 2000—2019 国内生产总值和税收收入　　单位：亿元

年份	税收	国内生产总值	年份	税收	国内生产总值
2000	12 581.5	100 280.1	2010	73 210.8	401 512.8
2001	15 301.4	109 655.2	2011	89 738.4	472 881.6
2002	17 636.5	120 332.7	2012	100 614.3	519 470.1
2003	20 017.3	135 822.8	2013	110 530.7	568 845.2
2004	24 165.7	159 878.3	2014	119 175.3	643 974.0
2005	28 778.5	184 937.4	2015	124 922.2	685 505.8
2006	34 804.4	216 314.4	2016	130 360.7	743 585.5
2007	45 622.0	265 810.3	2017	144 369.9	827 121.7
2008	54 223.8	314 045.4	2018	156 402.9	919 281.1
2009	59 521.6	340 506.9	2019	158 000.5	990 865.1

以表 1.2 中的数据绘出国内生产总值 X 与税收收入 Y 的散点图（如图 1.3 所示）。从图 1.3 中可以看出，虽然不同的税收收入存在差异，但整体来说，随着国内生产总值的增加，税收收入也在增加，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。

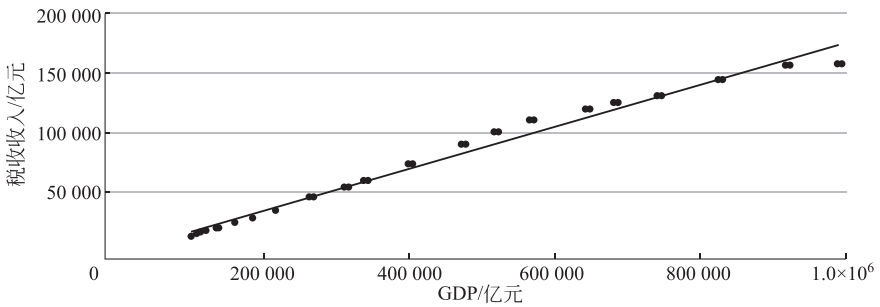

图 1.3　不同国内生产总值与税收收入分布

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，或称为总体回归曲线。相应的函数为

$$E(Y \mid X)=f(X) \tag{1-3}$$

该函数称为（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，U 形边际成本函数以二次多项式的形式出现等。将税收收入看成是国内生产总值的线性函数时，式(1-3)可进一步写成

$$E(Y \mid X)=\beta_0+\beta_1 X \tag{1-4}$$

其中，β_0，β_1是未知参数，称为回归系数。式（1-4）也称为线性总体回归函数。由于线性函数处理起来比较简单，参数估计、检验及预测方面相对容易。因此，为了好处理，模型在设定的时候经常假设为线性形式，尽管有时候可能存在偏差。在模型设定时，解释变量只能出现一次项，不能出现二次项及更高次项。当然，还存在一些函数，形式上含有二次项或更高次项，是非线性的，但可以通过变化转化为线性形式，进而继续按照线性形式进行研究。

1.1.3 随机干扰项

在上述税收收入和国内生产总值关系的例子中，总体回归函数描述了所考察总体的税收收入总体来说随国内生产总值变化的规律，但对某一个具体的年份，其税收收入不一定恰好就是国内生产总值水平下的税收收入的平均值$E(Y)$。图 1.3 显示，个别年份税收收入聚集在给定国内生产总值水平下税收收入$E(Y)$的周围。

对每个统计年份，记为

$$\mu=Y-E(Y) \tag{1-5}$$

其中，称μ为观察值Y围绕它的期望值$E(Y)$的离差，它是一个不可观测的随机变量，称为随机误差项，通常又不加区别地称为随机干扰项。

由式（1-5），个别年份税收收入为

$$Y=E(Y)+\mu \tag{1-6}$$

或者在线性假设下为

$$Y=\beta_0+\beta_1 X+\mu \tag{1-7}$$

即给定国内生产总值X，个别年份可表示为两部分之和：① 国内生产总值水平下平均税收收入$E(Y)$，称为系统性部分或确定性部分；② 其他随机部分或非系统性（nonsystematic）部分μ。

式（1-6）或式（1-7）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。

在总体回归函数中引入随机干扰项，主要有以下 6 方面的原因。

（1）表示未知的影响因素。由于对研究的总体理解不充分，未知的影响因素很多，无法导入模型。因此，只能使用随机干扰项来表示这些未知的影响因素。

（2）表示不完整的数据。虽然可以看出某个重要原因对被解释变量产生了很大的影响，但是无法得到具体的数据。例如，经济理论指出，税收收入除了受国内生产总值的影响，还会受纳税人诚信的影响，但后者实际上很难收集。因此，必须从模型中排除此变量，并将其包含在随机干扰项中。

（3）表示许多微弱的因素。在构建模型时，可以发现一些影响因素，虽然可以收集相关数据，但对被解释变量只有很小的影响。考虑到模型的简洁性和收集数据的成本，这些影响

较弱的变量可以在建模过程中省略，它们的影响可以纳入到随机干扰项中。

（4）表示数据观测误差。在观察和测量数据时，由于各种原因可能会产生测量误差，产生的误差被放入到随机误差项中。

（5）表示模型设定错误。由于经济问题的复杂性，人们无法给出模型的准确形式。因此，构建的计量经济学模型和真实的模型之间可能存在差异。这种差异被纳入到随机干扰项中。

（6）变量固有的随机性。即使模型没有设定误差，数据也没有观测误差，但有些变量本质上是随机的，这种情况也会随机影响被解释变量。这种效果只能被纳入到随机误差项中。

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，是在模型设定过程产生的，是可以避免的。

1.1.4　样本回归函数

总体回归函数显示了调查对象总体的解释变量和被解释变量的平均变化规律，但由于整体信息往往较难获得（比如经济成本过高等），因此获得总体回归函数并不容易。实际上，是通过样本来估计总体，通过样本信息来推测总体信息。

案例 1-1 给出 21 世纪以来我国的税收和国内生产总值之间的关系。样本点近似地分布在一条直线的两侧。由于样本点来自总体，因此可以用样本点的拟合线来近似代表总体拟合线。该直线称为样本回归线，其函数形式记为

$$\hat{Y}=f(X)=\hat{\beta}_0+\hat{\beta}_1 X \tag{1-8}$$

该函数称为样本回归函数（sample regression function，SRF）。

将式（1-8）看成式（1-7）的近似替代，则$\hat{Y}$就为 E（Y）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

$$Y=\hat{Y}+\hat{\mu}=\hat{\beta}_0+\hat{\beta}_1 X+e \tag{1-9}$$

其中，e 称为（样本）残差项（或剩余项），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。

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 $Y=E(Y)+\mu=\beta_0+\beta_1 X+e$，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 1.4 绘出了总体回归线与样本回归线的基本关系。

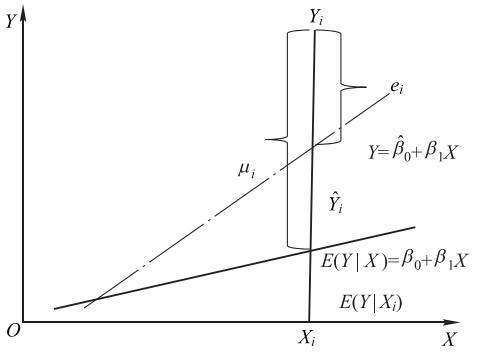

图 1.4　总体回归线与样本回归线的基本关系

1.2　一元线性回归模型的基本假设

计量经济学中单方程的模型可以分为线性模型和非线性模型两个大类。而线性回归模型是线性模型的一种，它通过回归分析的方法建立线性模型，可以分析经济现象中的因果关系。

一元线性回归模型只有一个解释变量，其一般形式为

$$Y_i=\beta_0+\beta_1 X_i Y_i=\beta_0+\beta_1 x_i+u_i \tag{1-10}$$

其中，Y_i为被解释变量，X_i为解释变量，β_0和β_1为待估参数，U_i为随机干扰项。

如同数学模型和物理模型需要一定的假设条件一样，计量经济学在建立模型时，为了保证参数估计值具有良好的特性，也需要对模型作一些基本的假设。

1.2.1　第一类假设：对模型设定的假设

假设 1：一方面模型选择的解释变量和被解释变量是正确的，另一方面模型选择的函数形式是正确的。

1.2.2　第二类假设：对解释变量的假设

假设 2：解释变量 X_i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_i的样本方差趋于一个非零的有限常数，即

$$\sum_{i=1}^{n}(X_i-\overline{X})^2/n\to Q,n\to\infty \tag{1-11}$$

回归分析的目的是揭示变量间的因果关系，被解释变量 Y 的变化往往是通过改变解释变量 X 的值来实现。因此，解释变量 X 必须足够可变。将样本方差限制为非零有限的常数，主要目的是排除无界变量作为解释变量的出现，无解变量会使现实研究和理论推导变得没有意义。

1.2.3　第三类假设：对随机干扰项的假设

假设 3：给定解释变量 X_i 的任何值，随机干扰项的 U 均值为零，即

$$E(U \mid X_i)=0 \tag{1-12}$$

这个假设意味着解释变量 Y 的值发生变化时，μ 的期望值不会发生变化，或者说，μ 的期望值始终保持为零，这也说明 μ 和 X 之间没有相关性。因此，当满足假设 3 时，可以说 X 是外部解释变量或者说来自系统之外，如果假设 3 不成立，那么 X 是内生变量或者说来自系统内部。

需要注意的是，当随机干扰项 μ 的条件零均值假设成立时，根据期望迭代法则一定有如下非条件零均值性质

$$E(\mu_i)=E(E(\mu_i \mid X))=E(0)=0 \tag{1-13}$$

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

$$\sigma_{X\mu_i}=E(X\mu_i)-E(X)E(\mu_i)=E(X\mu_i)=0 \tag{1-14}$$

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

$$\sigma_{X_i\mu_i}=E(X_i\mu_i)=0 \tag{1-15}$$

这时，也称 X 是同期外生的或称 X 与 μ 同期不相关。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 U 具有给定 X_i 任何值条件下的同方差性及不序列相关性，即

$$V(\mu_i \mid X)=\sigma^2 \quad i=1,2,\cdots,n \tag{1-16}$$

$$\sigma_{\mu_i\mu_j}=0 \quad i\neq j \tag{1-17}$$

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。

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

$$V(\mu_i)=\sigma^2 \tag{1-18}$$

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

$$V(\mu_i \mid X)=E(\mu_i^2 \mid X)-[E(\mu_i \mid X)]^2=E(\mu_i^2 \mid X)=\sigma^2 \tag{1-19}$$

或

$$V(\mu_i)=E(\mu_i^2)-[E(\mu_i)]^2=E(\mu_i^2)=\sigma^2 \tag{1-20}$$

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，式（1-17）可等价地表示为

$$\sigma_{\mu_i\mu_j}=E[(\mu_i \mid X)(\mu_j \mid X)]=0 \tag{1-21}$$

假设 5：随机干扰项服从零均值、同方差的正态分布，即

$$\mu_i \mid X \sim N(0,\sigma^2) \tag{1-22}$$

在小样本下该假设十分重要，而在大样本的情况下，正态性假设可以放松。由中心极限定理可知，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。

以上假设就是线性回归模型的经典假设（classical assumption），满足以上假设的线性回归模型被称为经典线性回归模型（classical linear regression model，CLRM）。其中前四个假设也被专门称为高斯-马尔可夫假设，这些假设会保证本章估计方法有良好的效果。

1.3 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就是指在一组样本观测值 $\{(X_i,Y_i)\mid i=1,2,\cdots,n\}$ 下，运用计量经济学的方法，得出参数的估计值和回归线。本教材涉及普通最小二乘法（OLS）、最大似然法（ML）和矩估计法（MM），其中以普通最小二乘法为主。

1.3.1 参数估计的普通最小二乘法

1829 年，德国数学家高斯提出了普通最小二乘法，该方法得到普遍的应用，并成为计量经济学发展的基石。要详细了解这个计量方法，首先了解普通最小二乘原理。

已知一组样本观测值 $\{(X_i,Y_i)\mid i=1,2,\cdots,n\}$，普通最小二乘法（ordinary least squares，OLS）需要将这些已知数据在样本回归函数中最大程度地拟合，也就是回归的拟合值尽可能地接近样本观测值，两者之间的误差越小越好。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。

$$Q=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=\sum_{i=1}^{n}[Y_i-(\hat{\beta}_0+\hat{\beta}_1X_i)]^2 \tag{1-23}$$

利用平方和表示距离的原因是，样本拟合值 $\hat{Y}_i$ 与真实观测值 Y_i 之差有正有负，如果仅仅加总求和可能正负项互相抵消。运用平方和能够较好反映拟合值和观测值之间的距离：平方和越小，说明两者之间的距离越小，拟合程度越高。

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

$$\begin{cases}\sum(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)=0\\\sum(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)X_i=0\end{cases} \tag{1-24}$$

或

$$\begin{cases}\sum Y_i=n\hat{\beta}_0+\hat{\beta}_1\sum X_i\\\sum Y_iX_i=\hat{\beta}_0\sum X_i+\hat{\beta}_1\sum X_i^2\end{cases} \tag{1-25}$$

解得

$$\begin{cases}\hat{\beta}_0=\dfrac{\sum X_i^2\sum Y_i-\sum X_i\sum Y_iX_i}{n\sum X_i^2-(\sum X_i)^2}\\ \hat{\beta}_1=\dfrac{n\sum Y_iX_i-\sum Y_i\sum X_i}{n\sum X_i^2-(\sum X_i)^2}\end{cases}\tag{1-26}$$

式（1-24）或式（1-25）称为正规方程组，记为

$$\begin{aligned}\sum X_i^2&=\sum(X_i-\overline{X})^2\\&=\sum X_i^2-\frac{1}{n}(\sum X_i)^2\end{aligned}$$

$$\begin{aligned}\sum X_iY_i&=\sum(X_i-\overline{X})(Y_i-\overline{Y})\\&=\sum X_iY_i-\frac{1}{n}\sum X_iY_i\end{aligned}$$

式（1-26）的参数估计量可以写成

$$\begin{cases}\hat{\beta}_1=\dfrac{\sum x_iy_i}{\sum x_i^2}\\ \hat{\beta}_0=\overline{Y}-\hat{\beta}_1\overline{X}\end{cases}\tag{1-27}$$

式（1-27）称为普通最小二乘法估计量的离差形式，其中，$x_i=X_i-\overline{X}$，$y_i=Y_i-\overline{Y}$。由于$\hat{\beta}_0$，$\hat{\beta}_1$的估计结果是从最小二乘原理得到的，所以称为普通最小二乘估计量。

如果$\hat{y}_i=\hat{Y}_i-\overline{Y}$，则有

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

可得

$$\hat{y}_i=\hat{\beta}_1x_i\tag{1-28}$$

其中，用到了正规方程组的第一个方程

$$\sum e_i=\sum[Y_i-(\hat{\beta}_0+\hat{\beta}_1X_i)]=0$$

式（1-28）也称为样本回归函数的离差形式。

估计值（estimate）和估计量（estimator）是有区别的。“估计值”又称为“点估计”，是某一个根据样本数据计算出来的具体数值。把式（1-26）看成$\hat{\beta}_0$和$\hat{\beta}_1$的一个表达式，它就成为Y_i的函数，而Y_i是随机变量，则$\hat{\beta}_0$和$\hat{\beta}_1$也是随机变量，因此就称之为“估计量”。估计量是一个样本的函数，而估计值是该函数代入样本后的一个值。

1.3.2　参数估计的最大似然法

最大似然法（maximum likelihood，ML），是一种根据最大似然原理来估计概率模型参数的方法。相较于最小二乘原理，最大似然原理更深刻地揭示了通过样本来估计总体，因此在计量经济学理论中占据一席之地，成为计量经济学理论发展的基础。而最大似然法也成为了

其他估计方法的基础，成功地估计了一些特殊的计量经济学模型。

普通最小二乘法和最大似然法虽然都是计量经济学参数的估计方法，但原理差别较大。普通最小二乘法本质是最合理的参数估计量应该使模型能最好地拟合样本数据；而最大似然法本质是最合理的参数估计量应该使从模型中抽取该样本观测值的概率最大。最大似然法的直观理解是：一个随机试验有 n 种可能的结果 A_1，A_1，…，A_n，出现的结果是 A_1，可以认为条件对 A_1 出现更有利，A_1 出现的概率最大。最大似然法就是要选择最合理的参数值，使样本出现的概率最大。

以正态分布为例，如果已经得到容量为 n 的样本观测值，什么样的总体最可能产生已知的样本呢？首先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$\{(X_i,\ Y_i)\mid i=1,\ 2,\ \cdots,\ n\}$，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

$$P(Y_i)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(Y_i-\beta_0-\beta_1X_i)^2}\quad i=1,2,\cdots,n$$

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

$$\begin{aligned}L(\beta_0,\beta_1,\sigma^2)&=P(Y_1,Y_2,\cdots,Y_n)\\&=\frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\mathrm{e}^{-\frac{1}{2\sigma^2}\sum(Y_i-\beta_0-\beta_1X_i)^2}\end{aligned}\tag{1-29}$$

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为

$$L^*=\ln L=-n\ln(\sqrt{2\pi}\sigma)-\frac{1}{2\sigma^2}\sum(Y_i-\beta_0-\beta_1X_i)^2\tag{1-30}$$

对 L^*求最大值，等价于对$\sum(Y_i-\beta_0-\hat{\beta}_1X_i)^2$求最小值。设$\hat{\beta}_0$，$\hat{\beta}_1$满足该最值条件，即

$$\begin{cases}\dfrac{\partial}{\partial\hat{\beta}_0}\sum(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)^2=0\\[2ex]\dfrac{\partial}{\partial\hat{\beta}_1}\sum(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)^2=0\end{cases}$$

解得模型的参数估计量为

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

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。

1.3.3　参数估计的矩估计法

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。式（1-24）或式（1-25）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

$$E(\mu_i)=0,\quad \sigma_{X_i\mu}=E(X_i,\mu_i)=0$$

于是，相应的样本矩条件可写成

$$\frac{1}{n}\sum(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)=0 \tag{1-31}$$

$$\frac{1}{n}\sum(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)X_i=0 \tag{1-32}$$

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即，因此得到的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。

1.3.4　最小二乘估计量的统计性质

仅仅得到模型的参数估计还是远远不够的，有必要继续评估所估计的参数是否准确，以确定它是不是可以表示所抽取总体中的实际参数值。在抽取样本过程中并不是一成不变的，必然存在一定波动并且估计方法的选择也是因人而异，这些都会导致估计值和实际值发生偏差。差异的存在就要求人们将参数估计量的统计属性视为重要标准来判断估计量的优劣，这些基本的特性包括：线性性、无偏性、有效性、一致性、渐近无偏性和渐近有效性六个方面。

线性性、无偏性、有效性被称为估计量的有限样本性质或小样本性质，将包含有类似性质的估计量叫作最佳线性无偏估计量（BLUE）。在进行估计时，具有 BLUE 性质的估计量有时候难以找到，尤其在样本容量有限时。此时就将样本容量扩大时存在的估计量渐近性质纳入思考范围。一致性、渐近无偏性、渐近有效性被称为无限样本属性或大样本渐近性质。综上，在样本数量无法进行最佳线性无偏估计的时候，就需要转变思路，通过增加样本抽取的数量来检验所估计参数值的大样本特征。

需要补充说明的是，在小样本估计时，有效性和无偏性是最核心的性质，线性性有则更好，没有也可以接受。在大样本估计时，渐近性的地位更明显。可以证明，在满足基本假定条件时，普通最小二乘参数估计是具有最小方差的线性无偏估计。

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。由式（1-27）可知

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

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

其中，$k_i=\frac{x_i}{\sum x_i^2}$。同样可得

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

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

其中，$w_i=\frac{1}{n}-\overline{X}k_i$。

2. 无偏性

无偏性，即以 X 的所有样本值为条件，估计量$\hat{\beta}_0$和$\hat{\beta}_1$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

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

易知

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

故

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

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

同样的，容易得出

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

3. 有效性（最小方差性）

同一个参数可以有多个无偏估计量，不同的估计量方差也不相同。估计值围绕真实值波动幅度越大，对应的估计量的方差也越大，同时估计值对真值代表的有效性也越弱。所有无偏估计量中方差最小的那个称为最有效。

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

$$V(\hat{\beta}_1\mid X)=V(\sum k_i Y_i\mid X)=\sum k_i^2 V[(\beta_0+\beta_1 X_1+\mu_i)\mid X]$$

$$=\sum k_i^2 V(\mu_i\mid X)=\sum\left(\frac{x_i}{\sum x_i^2}\right)^2\sigma^2=\frac{\sigma^2}{\sum x_i^2}\qquad(1-33)$$

$$V(\hat{\beta}_0\mid X)=V(\sum w_i Y_i\mid X)=\sum w_i^2 V[(\beta_0+\beta_1 X_1+\mu_i)]$$

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

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

$$=\left(\frac{1}{n}+\frac{\overline{X}^2}{\sum x_i^2}\right)\sigma^2=\frac{\sum x_i^2+n\overline{X}^2}{n\sum x_i^2}\sigma^2=\frac{\sum X_i^2}{n\sum x_i^2}\sigma^2 \tag{1-34}$$

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

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

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

$$V(\hat{\beta}_0^*)\geqslant V(\hat{\beta}_1)$$

由以上分析可以看出，在满足零均值、同方差、序列不相关等基本假设条件下，普通最小二乘估计量具有线性、无偏性和有效性，是最佳线性无偏估计量，这就是著名的高斯—马尔可夫定理。高斯—马尔可夫定理的作用为在满足基本假定的条件下，普通最小二乘是具有最小方差的无偏估计量，没有必要再寻求其他方法进行估计。即使再用其他方法估计，估计量的方差也不小于最小二乘估计量，估计的效果也不优于最小二乘估计。

上文已经证明，在小样本下，普通最小二乘法估计量具有较好的性质。同样，在大样本条件下，普通最小二乘法也具有较好的性质。例如，对$\hat{\beta}_1$的一致性来说，易知

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

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

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

其中，P 为概率，等式右边第二项分子是 X 与 u 的样本协方差的概率极限，它等于总体协方差 $\sigma_{X\mu}$，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，因此得到

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

1.3.5 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。由普通最小二乘估计量的线性性可以知道：$\hat{\beta}_1 = \sum k_i Y_i$、$\hat{\beta}_0 = \sum w_i Y_i$。因此，$\hat{\beta}_0$，$\hat{\beta}_1$ 的概率分布依赖 Y_i 的分布。假设随机干扰项 μ_i 和 Y_i 服从正态分布，那么$\hat{\beta}_0$ 和$\hat{\beta}_1$ 也服从正态分布。于是有

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

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

$$\sigma_{\hat{\beta}_0} = \sqrt{\frac{\sigma^2 \sum x_i^2}{n\sum x_i^2}} \tag{1-35}$$

$$\sigma_{\hat{\beta}_1} = \sqrt{\frac{\sigma^2}{\sum x_i^2}} \tag{1-36}$$

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（如图 1.5 所示）。

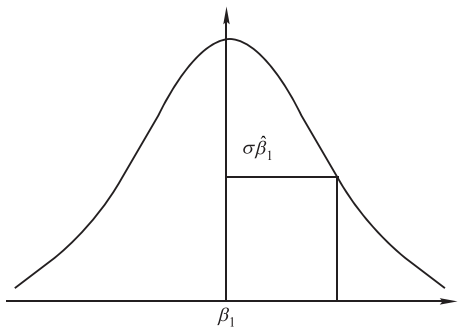

图 1.5 判断估计量的可靠性

2. 随机干扰项μ_i的方差 σ^2的估计

因为随机干扰项μ_i 是理论上的值，不能进行观测，只能借助 μ_i 的估计量——残差 e_i 来对总体方差 σ^2 实施估计。可以证明 σ^2 的最小二乘估计量为

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} \tag{1-37}$$

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的最大似然估计量为

$$\hat{\sigma}^2 = \frac{1}{n}\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \frac{\sum e_i^2}{n} \tag{1-38}$$

在矩估计法中，由于有总体矩条件 $V(\mu_i)=E(\mu_i^2)=\sigma^2$ 其对应的样本矩条件即为

$$\hat{\sigma}^2=\frac{1}{n}\sum(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)^2=\frac{\sum e_i^2}{n} \tag{1-39}$$

对照式（1-37）知，σ^2 的最大似然估计量与矩估计量都不具有无偏性，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 $\hat{\beta}_0$ 的样本方差和标准差的估计量分别为

$$S_{\beta_1}^2=\frac{\hat{\sigma}^2}{\sum x_i^2} \tag{1-40}$$

$$S_{\beta_1}^2=\frac{\hat{\sigma}}{\sqrt{\sum x_i^2}} \tag{1-41}$$

$$S_{\beta_0}^2=\frac{\hat{\sigma}^2\sum X_i^2}{n\sum x_i^2} \tag{1-42}$$

$$S_{\beta_0}=\hat{\sigma}\sqrt{\frac{\sum X_i^2}{n\sum x_i^2}} \tag{1-43}$$

1.4　一元线性回归模型的统计检验

所谓回归分析，就是指利用抽取样本所求的参数以及回归线来估计真实的参数与回归线的过程。事实上，只有在充足的重复样本选取的前提下才能更好地实现回归分析，样本量如果不充足，通过样本估计不一定能得到较好的效果。在具体某一次抽样估计时，样本的参数估计多大程度上靠近总体参数，需要进行下一步的统计检验。本章涉及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置信区间的估计。

1.4.1　拟合优度检验

拟合优度即拟合的优秀程度。如果拟合线能穿过全部样本观测点，则拟合的结果是完美的。但这种情况几乎不会出现，尤其当样本量比较大的时候。多数情况下，样本点集聚于回归线的周围，即样本残差项（剩余项）有正有负，不全是 0。人们希望样本点尽量“紧密”地分布在回归线的周围，即残差项越小越好。拟合优度就是衡量这种“紧密”程度的指标。检验的方法是：基于所抽取的样本数据，构造能够表征点与线“紧密”程度的统计量，并计算出统计量的值，然后用该统计值和判别标准进行比较，最后得出检验的结论。

1. 总离差平方和的分解

已知由一组样本观测值 $\{(X_i, Y_i) \mid i=1, 2, \cdots, n\}$ 得到如下样本回归直线（如图 1.6 所示）

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

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

$$y_i=Y_i-\overline{Y}_i=(Y_i-\hat{Y})+(\hat{Y}_i-\overline{Y})=e_i+\hat{y}_i \tag{1-44}$$

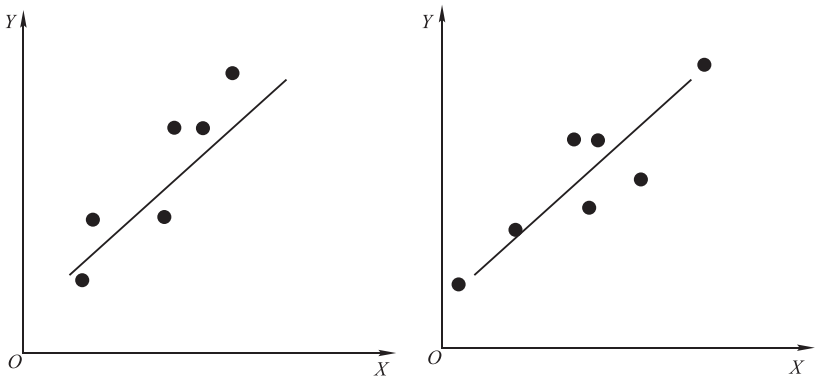

图 1.6　OLS 法样本回归直线

图 1.7 表示了这种分解，其中，$\hat{y}_i = Y_i - \overline{Y}$ 是样本回归线理论值（回归拟合值）与观测值 Y 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 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。

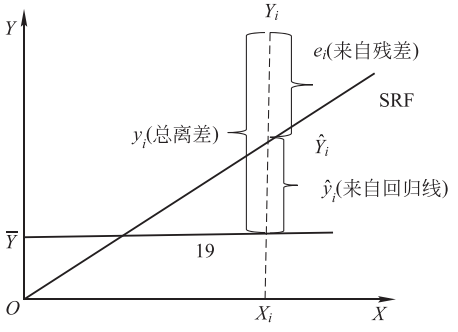

图 1.7　离差分解示意图

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

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2 \tag{1-45}$$

$\sum y_i^2 = \sum (Y_i - \overline{Y})^2 = \text{TSS}$ 称为总离差平方和，反映样本观测值总体离差的大小。

$\sum \hat{y}_i^2 = \sum (\hat{Y}_i - \overline{Y})^2 = \text{ESS}$ 称为回归平方和，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，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。

式（1-45）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原因。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。

ESS 反映样本观测值与估计值偏离的大小，但不能直接用它作为拟合优度检验的统计量。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，可以用可决系数 R^2 来检验模型的拟合优度，表示为

$$R^2=\frac{\mathrm{ESS}}{\mathrm{TSS}}=1-\frac{\mathrm{RSS}}{\mathrm{TSS}} \tag{1-46}$$

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的越好。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，即 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

$$R^2=\hat{\beta}_1^2\left(\frac{\sum x_i^2}{\sum y_i^2}\right) \tag{1-47}$$

这里用到了样本回归函数的离差形式来计算回归平方和

$$\mathrm{ESS}=\sum\hat{y}_i^2=\sum(\hat{\beta}_1x_i)^2=\hat{\beta}_1^2\sum\hat{x}_i^2$$

由式（1-46）可知，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，是随抽样而变动的统计量。因此，对可决系数的统计可靠性也应进行检验，这将在后面章节讨论。

1.4.2　变量的显著性检验

变量的显著性检验，就是为了判断模型选择的解释变量是否能显著的影响被解释变量。上文介绍的拟合优度检验更多地侧重整个模型的检验效果。但整个模型效果比较好，不代表每个变量效果都好。如一个团队优秀，并不代表每一位成员都优秀，也许有滥竽充数者。因此，在检验完团队后，还要对成员进行检验。变量的显著性检验就是基于这个基本思想展开的。

变量的显著性检验是利用一种样本结果来证实或证伪一个虚拟假设的过程。基本原理为小概率事件原理，主要是“无效假设”“检验无效假设”。选择显著性检验的目的是消除“第一类错误”“第二类错误”。从统计学角度看，在原假设为真时，决定放弃原假设，即“弃真”，称为第一类错误，出现的概率记为 α；在原假设不真时，决定不放弃原假设，即“取伪”，称为第二类错误，出现概率记为 β。通常只限定犯第一类错误的最大概率 α，不考虑犯第二类错误的概率 β。这样的假设检验又称为显著性检验，概率 α 称为显著性水平。

1. 假设检验

作为统计检验的重要内容之一，假设检验的任务是根据样本数据，对未知总体分布的某一假设进行判断。

假设检验的程序是：首先根据问题的需求提出原假设，记为 H_0。原假设是基于问题的一个统计推断，真伪在检验前无法知道，需要依据样本数据来辨别 H_0 的真伪，作出拒绝 H_0 或不拒绝 H_0 的决策。

在假设检验时，反证法的思想得到了应用。为了判断原假设 H_0 的真伪，先假定原假设 H_0 是真的，由此结论入手，观察下一步会得出什么样的结论。如果得出结论明显有悖常理，这意味着原假设 H_0 被证伪；如果得出结论是正确的，则意味着原假设 H_0 被证实。

所谓小概率事件就是说某种现象在进行单次试验的过程中极不可能发生的情况，恰好体现了反证法的原理。预先在原假设 H_0 下设计一个情形并且在原假设成立时为小概率事件。在总体中随机抽取 n 个样本进行该情形的试验，如果产生了这个几乎不可能发生的事，就说明原假设成立这个结论是不对的，那么必须拒绝原假设 H_0；如果这个接近不会发生的事件真的没有出现，就选择不拒绝原假设 H_0。

2. 变量的显著性检验方法

费希尔、内曼和皮尔逊所提出的显著性检验方法（test-of-significance approach）既可以检验统计假设，也补充了置信区间方法。显著性检验是一个虚拟假设真伪的检验程序，是利用样本结果来证实的，关键点在于一个检验统计量（test statistic）及其在虚拟假设下的抽样分布。

F 检验、t 检验、z 检验是三种常见的变量显著性的检验方法。它们主要在构造的统计量方面存在差异。由于 t 检验的应用最为普遍，几乎覆盖所有的计量经济学软件包，因此本书对 t 检验进行介绍。

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

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

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 $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$ 替代时，可构造如下统计量

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2}}} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \tag{1-48}$$

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选假设分别为

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

在统计学中，当拒绝虚拟假设时，在统计上是显著的；反之，当不拒绝虚拟假设时，在统计上是不显著的。

给定一个显著性水平 α，比如 0.05，查 t 分布表，得到一个临界值 $t_{\frac{\alpha}{2}}(n-2)$，则 $|t| >$

$t_{\frac{\alpha}{2}}(n-2)$ 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，未通过变量显著性的检验。

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

$$t=\frac{\hat{\beta}_0-\beta_0}{\sqrt{\frac{\hat{\sigma}^2\sum X_i^2}{n\sum x_i^2}}}=\frac{\hat{\beta}_0-\beta_0}{S_{\hat{\beta}_0}} \tag{1-49}$$

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 $\beta_0=0$。

用显著性检验的语言说，如果一个统计的值落在临界域内，那么这个统计量就是在统计上显著的。这时选择拒绝虚拟假设。同理，如果一个检验统计的值落在接受域中，那么它就是统计上不显著的，这时选择不拒绝虚拟假设。

1.4.3　参数置信区间的估计

假设检验是利用样本的结果来估计总体参数的大致范围，参数估计的结果随着抽样样本的变化而变化，样本参数估计值不一定等于总体回归参数的真实值。虽然样本估计值不一定等于总体真实值，但可以通过样本统计量构造总体参数的大致区间，这就是置信区间。置信区间的长度可以刻画样本参数与总体真实参数之间的接近程度。置信区间的本质是误差范围。

利用样本检验的结果可以估计总体参数真实值的大致范围，但没有给出到底距离参数真实值有多远。置信区间给出了两者接近的程度。置信区间就是以样本估计值为中心，以统计量的置信上限和置信下限为上下界构成的区间。一般常用计量软件都会报告检验的置信区间。

要衡量参数估计值 $\hat{\beta}_j$ 与总体真值 β_j 的距离，可以预先设定一个概率 α（$0<\alpha<1$），这里 α 通常取 5%，也是犯第一类错误的概率。并计算出一个大于 0 的数 δ，使得区间（$\hat{\beta}_j-\delta$，$\hat{\beta}_j+\delta$）包含参数 β_j 的真值的概率为 $1-\alpha$（通常选 95%），即

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

若存在上述这样的一个区间，称之为置信区间（confidence interval）或者接受域；这个区间之外的区域称为临界域或者拒绝域。$1-\alpha$ 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）。若建立了 σ^2 的 99%置信界限，并且事先声称这些界限将包含真实的 σ^2，那么将有 99%的概率是正确的。

经过检验，如果统计量的值居于接受域之内，则称该统计量是显著的，可以拒绝原假设；如果统计量的值居于临界域之内，则称该统计量是不显著的，可以不拒绝原假设。补充说明的是，“不拒绝”和“接受”的含义不完全相同，尽管在口语中二者经常互相替代。如同法律上，经常说的宣判“无罪”，而不是说宣判“清白”。

1.5 案例分析

粮食安全问题不仅是一个经济问题，更是一个政治问题，它是维护经济发展、社会稳定和国家安全的重要基础。习近平总书记高度重视粮食生产和安全，始终把解决好十几亿人口的吃饭问题作为我们党治国理政的头等大事，强调“农业基础地位任何时候都不能忽视和削弱，手中有粮、心中不慌在任何时候都是真理。为考察我国粮食产量（万 t）与种植面积（千 hm^2）之间的关系，表 1.3 报告了 2019 年我国 31 个省、自治区、直辖市粮食作物产量和面积数。

表 1.3　2019 年 31 个省级行政区粮食产量和种植面积

省（自治区、直辖市）	产量/万 t	种植面积/千 hm^2	省（自治区、直辖市）	产量/万 t	种植面积/千 hm^2
北京	28.8	46.5	辽宁	2 430.0	3 488.7
天津	223.3	339.3	吉林	3 877.9	5 644.9
河北	3 739.2	6 469.2	黑龙江	7 503.0	14 338.1
山西	1 361.8	3 126.2	上海	95.9	117.4
内蒙古	3 052.5	6 827.5	江苏	3 706.2	5 381.5
浙江	592.1	977.4	重庆	1 075.2	1 999.3
安徽	4 054.0	7 287.0	四川	3 498.5	6 279.3
福建	493.9	822.4	贵州	1 051.2	2 709.4
江西	2 157.5	3 665.1	云南	1 870.0	4 165.8
山东	5 357.0	8 312.8	西藏	103.9	184.8
河南	6 695.4	10 734.5	陕西	1 231.1	2 998.9
湖北	2 725.0	4 608.6	甘肃	1 162.6	2 581.1
湖南	2 974.8	4 616.4	青海	105.5	280.2
广东	1 240.8	2 160.6	宁夏	373.2	677.4
广西	1 332.0	2 747.0	新疆	1 527.1	2 203.6
海南	145.0	272.6			

数据来源：2020 年中国农村统计年鉴。

1. 建立模型

建立如下回归模型：

$$Y=\beta_0+\beta_1 X_1+\mu$$

利用 Stata 软件回归结果如下：

Source	SS	df	MS	Number of obs	=	31
				F(1, 29)	=	765.16
Model	110575154	1	110575154	Prob > F	=	0.0000
Residual	4190880.87	29	144513.133	R-squared	=	0.9635
				Adj R-squared	=	0.9622
Total	114766035	30	3825534.49	Root MSE	=	380.15

chanliang	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mianji	.5711662	.0206484	27.66	0.000	.5289354	.613397
_cons	-16.35967	103.1414	-0.16	0.875	-227.3076	194.5883

一般可写成如下回归结果

$$Y_i = -16.36 + 0.57X_i$$

2. 模型检验

根据回归结果，模型拟合效果较好，拟合优度 $R^2 = 0.96$，表明粮食产量变化的 96%可以由种植面积来作出解释。粮食种植面积的系数为 0.57，表明粮食播种面积每增加 1 000 hm^2，粮食产量增加 0.57 万 t。

3. 预测

为了保护农村生态环境和践行绿水青山就是金山银山的概念，部分不适合种植粮食作物的地区要退出粮食生产，我国粮食播种面积可能要减少。根据中国国家统计局发布《2020 年国民经济和社会发展统计公报》，2020 年我国粮食种植面积为 11 677 万 hm^2，假如粮食播种面积要减少 1%，即减少 1 167.7 千 hm^2，我国粮食减少 665.6 万 t。

课后习题

1. 在构建计量经济学模型时，设置随机干扰项的原因是什么？

2. 判断下面设定的计量经济学模型是否正确，错误的解释原因。

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

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

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

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

3. 一元线性回归模型需要满足哪些基本假设条件？违背这些条件还能估计吗？

4. 线性回归模型 $Y_i = \alpha + \beta X_i + \mu_i$，$i = 1, 2, \cdots, n$ 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. 在一元线性回归模型中，如果把解释变量 X 的单位扩大十倍，被解释变量 Y 的单位保持不变，对估计参数会产生什么样的影响？如果把被解释变量 Y 的单位扩大十倍，解释变量 X 的单位保持不变，对估计参数会产生什么样的影响？